LREC-COLING 2024

DeTermIt! Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2024)

Workshop Proceedings

Editors
Giorgio Maria Di Nunzio, Federica Vezzani, Liana Ermakova, Hosein Azarbonyad, and Jaap Kamps

21 May, 2024
Torino, Italia

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024

ISBN 978-2-493814-15-9
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Preface from the General Chairs

Automatic Text Simplification (ATS) is the process that involves the reduction of linguistic complexity within a text to enhance its comprehensibility and readability. ATS plays a pivotal role in enhancing content and conveying clear, unambiguous information, while serving as a valuable preprocessing step, making texts more 'manageable' for various tasks like information extraction and retrieval. On a broader scale, ATS holds significant societal implications, particularly in assisting individuals with low literacy levels or those encountering challenges in reading comprehension.

One of the main barriers in text understanding is unfamiliar context and terminology. Even in developed countries, up to 30% of the population can only comprehend texts written with a basic vocabulary. Lexical simplification strives to enhance text comprehensibility for a broad audience by substituting intricate vocabulary and phrases with simpler alternatives while retaining the initial intended significance. Several initiatives emerged to help citizens with reading disabilities, e.g. the French project ALECTOR aims to leverage document accessibility for children with dyslexia. [1] EasyText.AI [2] focuses on text simplification for people with cognitive disabilities and provides simplifications of COVID-19-related texts in multiple languages. Finally, identification of difficult terms for second language learners can be helpful to optimize and personalize learning materials.

The *DeTermIt! Evaluating Text Difficulty in a Multilingual Context* workshop explores the theoretical and practical perspectives surrounding the evaluation of text difficulty in a multilingual context. In today's interconnected world, where information dissemination knows no linguistic bounds, it is mandatory to ensure that knowledge is accessible to diverse audiences, regardless of their language proficiency.

From a *theoretical* point of view, this workshop discusses the development of refined models and strategies for ATS. Additionally, the workshop promotes the study of the identification of common patterns and challenges encountered in different languages, which can lead to the creation of more effective tools and multilingual resources and promoting linguistic inclusivity. From a *practical* standpoint, the workshop considers the role of multilingual resources and their application in simplifying complex terminology. The development and utilization of language resources, such as bilingual and multilingual glossaries, translation memories, and terminology databases, are pivotal in achieving this goal. Furthermore, we analyze the effectiveness of machine translation and natural language processing techniques in aiding the simplification of text, and their implications for cross-linguistic text difficulty assessment.

The central inquiries in this workshop revolve around two key aspects: first, the theoretical elements that identify complexity within the text, and second the experimental analysis for simplifying the text to align with the reading proficiency of the target audience.

This first edition of DeTermIt! 2024 is co-located with the LREC-COLING 2024 joint conference and held in Turin, on May 21, 2024.[3]

The submitted papers went through a double-blind review process that required at least three reviews by members of the international scientific committee. We accepted 18 papers out of 29 submissions (62% acceptance rate): 12 long papers and 6 short papers.

[1] https://anr.fr/Project-ANR-16-CE28-0005

[2] https://easytext.ai/

[3] https://determit2024.dei.unipd.it/

Overall, these contributions encompass a diverse range of topics, showcasing the breadth and depth of research in the field of automatic text simplification. Papers deal with the development and refinement of text simplification systems in various languages, such as German, Finnish, French, and Arabic, reflecting a global interest in linguistic accessibility. Additionally, some studies explore innovative approaches to simplify complex scientific, legal, and governmental texts, aiming to enhance readability and comprehension. Multilingualism is a recurring theme, with papers addressing the challenges and opportunities of simplification across different linguistic contexts. Furthermore, advancements in lexical complexity prediction and the evaluation of simplification techniques through quantitative and qualitative research methodologies are examined, highlighting the interdisciplinary nature of the field.

The keynote speaker is Prof. Sara Carvalho (University of Aveiro, Portugal) with the title "Clear Communication, Better Healthcare: Leveraging Terminological Data for Automatic Text Simplification". By exploring the systematic representation and organization of terminological data, the talk is aimed at demonstrating how the double-dimensional approach to terminology has an impact on the development of ATS tools, ultimately enhancing patient-provider interactions and driving better healthcare outcomes.

Giorgio Maria Di Nunzio - Università degli Studi di Padova, Italy
Federica Vezzani - Università degli Studi di Padova, Italy
Liana Ermakova - Université de Bretagne Occidentale, France
Hosein Azarbonyad - Elsevier, The Netherlands
Jaap Kamps - University of Amsterdam, The Netherlands

Organizing Committee

General Chairs
Giorgio Maria Di Nunzio - Università degli Studi di Padova, Italy
Federica Vezzani - Università degli Studi di Padova, Italy
Liana Ermakova - Université de Bretagne Occidentale, France
Hosein Azarbonyad - Elsevier, The Netherlands
Jaap Kamps - University of Amsterdam, The Netherlands

Scientific Commitee
Hosein Azarbonyad - Elsevier, The Netherlands
Florian Boudin - Nantes University, France
Lynne Bowker - University of Ottawa, Canada
Sara Carvalho - Universidade de Aveiro, Portugal
Rute Costa - Universidade NOVA de Lisboa, Portugal
Giorgio Maria Di Nunzio - Università degli Studi di Padova, Italy
Eric Gaussier - University Grenoble Alpes, France
Natalia Grabar - CNRS, France
Jaap Kamps - University of Amsterdam, The Netherlands
Rodolfo Maslias - TermNet, Austria
Ana Ostroški Anić - Institute of Croatian Language and Linguistics, Croatia
Horacio Saggion - University Pompeu Fabra
Grigorios Tsoumakas - Aristotle University of Thessaloniki
Sara Vecchiato - University of Udine, Italy
Federica Vezzani - Università degli Studi di Padova, Italy
Cornelia Wermuth - KU Leuven, Belgium

Table of Contents

Reproduction & Benchmarking of German Text Simplification Systems

Regina Stodden
Department of Computational Linguistics
Faculty of Arts and Humanities
Heinrich Heine University Düsseldorf, Germany
regina.stodden@hhu.de

Abstract

The paper investigates the reproducibility of various approaches to automatically simplify German texts and identifies key challenges in the process. We reproduce eight sentence simplification systems including rules-based models, fine-tuned models, and prompting of autoregressive models. We highlight three main issues of reproducibility: the impossibility of reproduction due to missing details, code, or restricted access to data/models; variations in reproduction, hindering meaningful comparisons; and discrepancies in evaluation scores between reported and reproduced models. To enhance reproducibility and facilitate model comparison, we recommend the publication of model-related details, including checkpoints, code, and training methodologies. Our study also emphasizes the importance of releasing system generations, when possible, for thorough analysis and better understanding of original works. In our effort to compare reproduced models, we also create a German sentence simplification benchmark of the eleven models across seven test sets. Overall, the study underscores the significance of transparency, documentation, and diverse training data for advancing reproducibility and meaningful model comparison in automated German text simplification.

Keywords: Text Simplification, Reproduction Study, German Sentence Simplification Benchmark

1. Introduction

Text simplification (TS) is a Natural Language Processing (NLP) task that aims to enhance the accessibility and understandability of textual content for a diverse audience. This process involves the transformation of complex language structures into simpler and more straightforward forms to be better understandable for a specific target group, e.g., people with varying linguistic abilities, cognitive impairments, or those learning a new language (Alva-Manchego et al., 2020).

In recent years, German TS has also gained more attraction resulting in a few sentence simplification models, e.g., ZEST (Mallinson et al., 2020), sockeye-APA-LHA (Spring et al., 2021), mBART-DEplain-APA (Stodden et al., 2023), or custom-decoder-ats (Anschütz et al., 2023). But even if the NLP community has increased efforts in better reproducibility of (new) research by designing checklists on responsibility[1] or asking for reproducibility studies (Branco et al., 2020), some NLP models are still not easily reproducible. This has also hampered German TS because the access to resources is often restricted, not enough information are named for reproduction, models and code are unavailable, or system outputs are not made accessible to other researchers.

Therefore, in this work, we try to reproduce existing German TS models and re-generate their system outputs to facilitate analysing different German TS approaches or creating evaluation methods for German TS. Further, we discuss whether the reproduced models match or differ from the original models by analysing automatic TS metrics. To compare the reproduced models with each other, we also create a German sentence simplification benchmark on 7 test sets, including the system outputs of all 11 TS models. We make the code, the system outputs (if permitted by license), and the system evaluation reports available to increase the reproduction of this work in future German TS research. All materials are provided in https://github.com/rstodden/easse-de.

2. Related Work

The most similar works to ours are reproduction studies of English text simplification systems. Cooper and Shardlow (2020) and Arvan et al. (2022), for example, both reproduced the work on English TS by Nisioi et al. (2017): they trained a TS model using the provided code on a to-be-processed dataset and evaluate whether they can simulate the original findings. In our work, we will do the same for 7 German TS models.

Popović et al. (2022), in comparison, do not focus on the reproduction of a TS model but tried to repeat the human evaluation study proposed

[1] https://aclrollingreview.org/responsibleNLPresearch/

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 1–15

in Nisioi et al. (2017). Unfortunately, we cannot replicate this for German TS as human evaluation is rarely performed, and insufficient information would be available to repeat the process.

3. Method

For our reproduction study, we first describe the selection of models (see subsection 3.1) and then explain on which data we have trained and evaluated them (see subsection 3.2). Afterwards, we explain more about how we check the extent of the reproduction, whether the model seems totally different, rather close or identical to the original model (see subsection 3.3).

3.1. Models

Based on a literature review, we found some sentence simplification models for German, which have been proposed in recent years. We split the lines of research into

(i) rule-based models, e.g., rule-based model by Suter et al. (2016), *DISSIM* (Niklaus et al., 2019), and **hda-etr** (Siegel et al., 2019) (see subsection 4.1),

(ii) training sequence-to-sequence generation models, e.g., **sockeye-APA-LHA** (Spring et al., 2021) and other sockeye variants (Ebling et al., 2022) (see subsection 4.2),

(iii) fine-tuning sequence-to-sequence generation models, e.g., **mBART_DEplain-APA** (Stodden et al., 2023), **mBART_DEplain-APA+web** (Stodden et al., 2023), or *mT5-*MULTISIM (Ryan et al., 2023) (see subsection 4.3),

(iv) zero shot simplification, e.g., *ZEST* (Mallinson et al., 2020),

(v) prompting autoregressive language models, e.g., **BLOOM** in Ryan et al. (2023) or Ponce et al. (2023) (see subsection 4.4), or *ChatGPT* in Manning (2023) or Deilen et al. (2023) (see subsection 4.4), and

(vi) combining autoregressive language models and sequence-to-sequence models, e.g., **custom-decoder-ats** (Anschütz et al., 2023) (see subsection 4.5).

We tried to reproduce all of the listed models. Unfortunately, for some models (see models highlighted in italics), neither the code nor the prompts are available, and they require too much computing power (i.e., mT5-MultiSim) to reproduce the model. Further, no system generations are available for these models, which could have been used for comparisons. Hence, we could only reproduce 6 models and their corresponding system outputs (see models highlighted in boldface). In section 4, we will describe each of the (reproduced) models in more detail and describe how we have reproduced them.

For the German TS benchmark, we also propose three new TS systems, i.e., mT5 (Xue et al., 2021) fine-tuned on a manually aligned news corpus, i.e., DEplain-APA (Stodden et al., 2023a), and mT5 fine-tuned on an automatically aligned web corpus, i.e., Simple German Corpus (Toborek et al., 2023) (and Toborek et al. 2023 for more corpus description paper). For both models, i.e., mT5-DEplain-APA[2] and mT5-SGC[3], we use the same hyperparameters (see Appendix A), the code and the system outputs also available in the Github repository. Additionally, we train sockeye on DEplain-APA with the same parameters as those used for sockeye-APA-LHA. This model is further called sockeye-DEplain-APA.

3.2. Training & Test Data

For training, fine-tuning, or prompting the models, we used the same training and evaluation data as named in the original work (if available).

hda-etr is a rule-based system which require no training data and was not evaluated on any test data yet. The training and/or evaluation data to reproduce trimmed_mbart_sent (i.e., DEplain-APA (Stodden et al., 2023a) and DEplain-web (Stodden et al., 2023b))[4], BLOOM (i.e., TextComplexityDE (TCDE19) (Naderi et al., 2019) and GEOlino (Mallinson et al., 2020)[5]), and encoder-decoder-ats (i.e., 20Minuten (Rios et al., 2021)[6]) are available, pre-split into training, development, and test set which enhanced the reproduction process. The APA-LHA data (Spring et al., 2021) to reproduce sockeye-APA-LHA is available upon request[7], but the data is randomly split into training and test sets each time when pre-processing the data. Hence, our experiments for sockeye-APA-LHA are conducted on a different split than in the original paper.

Additionally, we evaluate the models on the Simple German Corpus (Toborek et al., 2023); a manually aligned test set of web texts corresponding to the training data for mT5-SGC.

[2] https://huggingface.co/DEplain/mt5-DEplain-APA

[3] https://huggingface.co/DEplain/mt5-simple-german-corpus

[4] https://github.com/rstodden/DEPlain

[5] https://github.com/XenonMolecule/MultiSim

[6] https://github.com/ZurichNLP/20Minuten

[7] https://zenodo.org/records/5148163

3.3. Evaluation

Strategies on how to evaluate the similarity of reproduced models to the original models are (i) similarity of system outputs, (ii) comparison of automatic metrics measured on the new system outputs and the reported scores in the original papers, or (iii) comparison of human judgements on the system outputs.

In our study, the first strategy only applies to one model (i.e., trimmed_mbart_sent), as for the other models, the system outputs of the original TS models have not been made available. Hence, only for the trimmed_mbart_sent model we can compare how similar the published system generations are to our reproduced system generations. As similarity measurement, we check for exact matches in both sets of generations and apply the BERT-Score-F1 (Zhang* et al., 2020)[8]. Further, we did not validate the reproduced models by manual evaluation, as evaluation using manual judgements is often not conducted. If conducted, for example in Mallinson et al. (2020), no system generations (also no reproduced generations) are available to be analyzed.

Our strategy for validation of the reproduced models is to compare the reported scores of TS metrics, e.g., SARI (Xu et al., 2016), BLEU (Papineni et al., 2002), BERT-Score Precision (BS_P) (Zhang* et al., 2020), or the German adaptation of Flesch Reading Ease (FRE) (Amstad, 1978) with the scores measured for the system generations of the reproduced systems.

Most of the German TS papers describe that they are evaluating their systems using the implementation of the metrics in EASSE (Alva-Manchego et al., 2019)[9], i.e., Trienes et al. (2022), Ryan et al. (2023), Stodden et al. (2023)[10], and Ponce et al. (2023).

Mallinson et al. (2020) use their own version of SARI, BLEU, and FRE-BLEU, Anschütz et al. (2023) did not use EASSE as it does not include ROUGE. Hence, they use the implementations of BLEU, SARI, and ROUGE provided in Huggingface (Wolf et al., 2020). In other papers, e.g., Spring et al. (2021) or Rios et al. (2021), it is not mentioned which implementation of SARI or BLEU has been used. We have generated the metric scores for all models using the metrics implementation described in the original paper. If no details on the implementation were provided, we have generated the scores with the EASSE-DE package (Stodden, 2024).

[8]For both metrics we have used their Huggingface implementation, i.e., https://huggingface.co/spaces/evaluate-metric/exact_match and https://huggingface.co/spaces/evaluate-metric/bertscore.

[9]EASSE is a Python package (Alva-Manchego et al., 2019) which is designed for the ease of evaluation of English sentence simplification.

[10]They are using the German version of EASSE, i.e., EASSE-DE (Stodden, 2024).

4. TS Models & Reproduction

In the following, we briefly summarize the TS systems for which we can reproduce results and argue why we couldn't or haven't reproduced the other models (see subsection 4.6). See Table 1 for an overview of all reproduced models.

4.1. Rule-based Models

Siegel et al. (2019) implement some rules of easy-to-read guidelines ("Leichte Sprache") as a rule-based simplification model. In more detail, it contains the following two rules: substitution of complex words and compound splitting. Their model, called hda-etr, focuses only on lexical simplification. Siegel et al. include their rules into LanguageTool[11], a re-writing tool that assists in giving recommendations on how to correct or improve a given input text. For hda-etr a working code is provided[12], containing also a graphical interface for highlighting infringements against easy guidelines. For better performance, we re-implemented the code without the infringements and interface. The updated code can be found at https://github.com/rstodden/easy-to-understand_language.

4.2. Training Sequence-to-sequence Models

In recent years, the same department, i.e., the computational linguistics department of the University of Zurich, has published a few research papers including very similar TS models (see (Säuberli et al., 2020; Spring et al., 2021; Ebling et al., 2022)). They trained a sequence-to-sequence model with a transformer architecture using the Sockeye framework (Domhan et al., 2020) among others on APA-LHA-OR-B1 and APA-LHA-OR-A2 (Spring et al., 2021).

(Säuberli et al., 2020) experimented with a former and smaller version of APA-LHA and Sockeye. They report results of their base Sockeye architecture as well as additional experiments with, e.g., smaller batch sizes or extensions with linguistic features. They also experimented with data augmentation strategies, i.e., adding non-parallel simplifications (NULL2TRG), adding identical pairs with the simplifications on both sides of the pair (TRG2TRG), and adding pairs including

[11]https://github.com/languagetool-org/languagetool

[12]https://github.com/hdaSprachtechnologie/easy-to-understand_language

System Name	Reference	Type	Training Data	# Simp. Pairs	URL
hda-etr	Siegel et al. (2019)	rule-based	-	-	https://github.com/hdaSprachtechnologie/easy-to-understand_language
sockeye-APA-LHA	Spring et al. (2021) & Ebling et al. (2022)	seq2seq	APA-LHA OR-A2 & APA-LHA OR-B1	8,455 & 9,268	https://github.com/ZurichNLP/RANLP2021-German-ATS
sockeye-DEplain-APA	-	seq2seq	DEplain-APA	10,660	https://huggingface.co/DEplain
mBART-DEplain-APA	Stodden et al. (2023)	fine-tuned seq2seq	DEplain-APA	10,660	https://huggingface.co/DEplain/trimmed_mbart_sents_apa
mBART-DEplain-APA+web	Stodden et al. (2023)	fine-tuned seq2seq	DEplain-APA+web	10,660 + 1,594	https://huggingface.co/DEplain/trimmed_mbart_sents_apa_web
mT5-DEplain-APA	-	fine-tuned seq2seq	DEplain-APA	10,660	https://huggingface.co/DEplain
mT5-SGC	-	fine-tuned seq2seq	SGC	4,430	https://huggingface.co/DEplain
BLOOM-zero	Ryan et al. (2023)	zero-shot AR model	-	-	https://github.com/XenonMolecule/MultiSim
BLOOM-sim-10	Ryan et al. (2023)	few-shot AR model	TCDE19 & GEOlino	200 & 959	https://github.com/XenonMolecule/MultiSim
BLOOM-random 10	Ryan et al. (2023)	few-shot AR model	TCDE19 & GEOlino	200 & 959	https://github.com/XenonMolecule/MultiSim
custom-decoder-ats	Anschütz et al. (2023)	AR model + fine-tuned seq2seq	Simplified, monolingual German data & 20Minuten	544,467 & 17,905	https://huggingface.co/josh-oo/custom-decoder-ats

Table 1: Overview of German TS models including training details (i.e., training data and size of training samples). Each line separates different model types. The models in italics are newly proposed in this work.

back-translated simplifications and original simplifications (BT2TRG). Compared to their base model, adding more data decreased their SARI and BLEU scores, except for adding TRG2TRG, their overall best-performing system. As far as we know, these (and the ones by Anschütz et al. (2023)) are the only experiments with augmented data for German TS. Unfortunately, the experiments cannot be reproduced as neither the corpus, the models, the code, nor enough details regarding building the models are available.

However, we reproduce another Sockeye variant for German TS, which was proposed by (Spring et al., 2021; Ebling et al., 2022). However, we ran into a few issues which made our results incomparable to the original results and reported scores. First, due to non-solvable conflict errors of required Python packages, we need to update the sockeye version from 2.3.8 to 3.1.14[13]. The technical differences between both implementations are listed in Appendix B. Further, a data split into training, development and test set were neither provided nor fixed through parameters in the code.

4.3. Fine-tuning Sequence-to-sequence models (transfer-learning)

Rios et al. (2021) are the first who used mBART (Liu et al., 2020) for German document simplification. The main improvements of their approach compared to the standard mBART are to maximize the input length (to 4096), reduce the vocabulary to the 20k most frequent German tokens, and add a special language tag to specify the target language level (de_A1, de_A2, or de_B1). This approach has also been adapted for document simplification of news and web texts (Stodden et al., 2023), for paragraph simplification of clinical notes (Trienes et al., 2022), and for sentence simplification of news and web texts (Stodden et al., 2023). An overview of the adaptations and the different hyperparameters can be found in Appendix H.

As this work focuses on sentence simplification, we will just include the models proposed in Stodden et al. (2023), i.e., trimmed_mbart_sents_DEplain-APA (further called mBART-DEplain-APA) and trimmed_mbart_sents_DEplain-APA+web (further called mBART-DEplain-APA+web). Compared to Rios et al. (2021), they reduce the vocabulary to 35k and use one universal language tag (de_SI). As the names suggest the models are trained on DEplain-APA (Stodden et al., 2023a) or DEplain-APA plus DEplain-web (Stodden et al., 2023b). The checkpoints of the models, instructions on how to use them and their system generations for three test sets are available on Huggingface[14] and GitHub[15]. Hence, for reproduction we could use the Huggingface's text-to-text-generation pipeline to generate the system outputs on all test sets.

[13] Sockeye 3 (Hieber et al., 2022) is a neural machine translation pipeline.

[14] https://huggingface.co/DEplain/trimmed_mbart_sents_apa and https://huggingface.co/DEplain/trimmed_mbart_sents_apa_web

[15] https://github.com/rstodden/DEPlain/tree/main/G__Automatic_Text_Simplification_Experiments/generated_outputs

4.4. Autoregressive Language Models

Ryan et al. (2023) experimented with few-shot and zero-shot learning on a multi-lingual simplification corpus (including German) (Ryan et al., 2023b) using the autoregressive language model BLOOM (with 176 billion parameters) (Workshop, 2023). As examples in the few-shot setting they used either k random sentences pairs or k pairs in which source sentences are most similar to the to-be-tested sentence. We reproduced their experiments with the provided code and data.

4.5. Autoregressive Language Models + Sequence-to-sequence Models

For custom-decoder-ats (Anschütz et al., 2023), first, Anschütz et al. have fine-tuned an autoregressive language model on simplified language and then have combined it with a fine-tuned sequence-to-sequence model.

For custom-decoder-ats[16] (Anschütz et al., 2023) the checkpoint of the model and instructions on how to use it are available on Huggingface. Hence, we could use the provided code and Huggingface's text-to-text-generation pipeline to generate the system outputs on all test sets.

4.6. No Reproduction

We have not reproduced some of the models, the reasons for that are as follows: (Mallinson et al., 2020) propose a zero-shot cross-lingual sentence simplification model called *ZEST*. Although the code is available, we could not reproduce the ZEST model and regenerate its outputs.

Ryan et al. (2023) have proposed a multilingual sentence simplification model named mT5-MULTISIM. They fine-tuned mT5 (Xue et al., 2021) on several corpora, including three German corpora, i.e., GEOlino (Ryan et al., 2023a), TCDE19 (Ryan et al., 2023c), and German News[17]. Due to limited computing power, we could not reproduce mT5-MULTISIM as it was originally trained on 3 GPUs with the size of 48 GB for each.

Schlippe and Eichinger (2023) also used a T5 model for training their German TS model, but they use the multilingual model Flan-T5 (Chung et al., 2022). Their training and evaluation data is not available. Hence, we haven't included this model in our reproduction study.

Ponce et al. (2023) also experiment with BLOOM, but with the version with 7 billion parameters[18] and on structural simplification, i.e., split and rephrase. They do not provide enough information to reproduce their approach (e.g., prompt missing, few-shot or zero-shot?) as it is only a small side project of their work.

Some researchers experiment on German TS with ChatGPT, e.g., (Deilen et al., 2023), (Manning, 2023) or Schlippe and Eichinger (2023), but we do not include this approaches as we are focusing on open, non-proprietary language models.

5. Reproduction Results

To check whether the reproduced models are identical to the models described in the original work, we compare the newly measured scores with those reported in the original papers.

5.1. hda-etr

For hda-etr, unfortunately, no automatic scores are provided in the original paper, hence, we cannot compare whether our re-implementation works as expected. However, to enable comparisons in future work, in section 6, we report results of hda-etr on a few test sets.

5.2. Sockeye-APA-LHA

As previously mentioned, our reproduction of sockeye-APA-LHA was trained on a different model version with different training data and will also be evaluated on different test sets of APA-LHA. The comparison of reported and reproduced results also reflects this (see Appendix C): the BLEU scores differ between roughly 1.0 and for SARI, even between 4.0 and 9.0 points. Hence, unfortunately, our reproduced Sockeye-APA-LHA model is not comparable to the original model, and the conclusions we can draw from the reproduction might not be the same as the original model.

5.3. BLOOM

For the three different approaches using BLOOM, i.e., zero-shot BLOOM, random 10-shot BLOOM and similarity 10-shot BLOOM, our reproduced system generations seem to be slightly different than the original system generations (see Appendix D). For all approaches on GEOlino, the SARI scores differ by less than 1.5 points. However, for TCDE19, the gap between the SARI scores is up to 2.5 points. If we also compare the baseline results, we can see that these are identical. Hence, we can exclude different data splits as a possible reason. It remains unclear why the numbers are that different, either the provided code is slightly different from the one used for the reported experiments, the evaluation method is different, or predictions of BLOOM are not fixed.

[16] https://huggingface.co/josh-oo/custom-decoder-ats

[17] Unfortunately, although it should be available on request, we do not yet have access to this corpus.

[18] https://huggingface.co/bigscience/bloom-7b1

5.4. custom-decoder-ats

To check whether custom-decoder-ats still meets the results reported in the original paper, we reproduced the results on the 20min corpus. The reported results of Anschütz et al. (2023) differ only slightly from the reproduced results (BLEU: roughly 0.3 and SARI: roughly 2.0, see Appendix E). Hence, we argue that the reproduced model is fairly comparable to the described model.

5.5. mBART-DEplain-APA & mBART-DEplain-APA+web

Even if the checkpoints of the mBART models are provided to reproduce the system generations, the scores of the reproduced models differ from the reported scores in the original paper (see Appendix F). For example, the SARI scores of mBART-DEplain-APA differ by roughly 2 points on DEplain-web or roughly 4 points on DEplain-APA when comparing reproduced and reported scores. However, the scores of the reproduced baseline are identical to the reported baseline using the EASSE-DE evaluation framework. Hence, we can argue that the same evaluation approach and the same test data have been used, and these are not the reasons for the differences.

We also compared the similarity of the reproduced system generations of mBART-DEplain-APA and mBART-DEplain-APA+web with the provided system generations of the original models by measuring their exact match and BERTScore-F1. As can be seen in Table 2, the exact matches for mBART-DEplain-APA are on each test set lower than 50% and for mBART-DEplain-APA+web varying between 49% and 74%. However, the BERTScores show that the predictions per instance are quite similar for both models, even if, again, the scores are higher for DEplain-APA+web. This confirms the previous findings; thus, the uploaded model must be slightly different from the one used to report the results in the original paper.

	exact↑	BS mean↑	BS min	BS std
DEplain-APA	42.24	0.9589	0.6587	0.0546
DEplain-web	17.98	0.9163	0.4885	0.0740
TCDE19	9.20	0.8889	0.7253	0.0739

(a) mBART-DEplain-APA

	exact↑	BS mean↑	BS min	BS std
DEplain-APA	73.68	0.9827	0.7328	0.0391
DEplain-web	56.99	0.9628	0.4623	0.0694
TCDE19	48.80	0.9593	0.7360	0.0600

(b) mBART-DEplain-APA+web

Table 2: Similarity between copied system generations and reproduced system generations by exact match (in %), and BERT-Score F1 values (mean, minimum, and standard deviation).

6. German TS Benchmark

The previous results show that most of the reproduced models are similar to the results of the original models. However, the results of the models are not comparable to each other as they are evaluated on different test sets and with different metrics implementations. To unify the evaluation reports and build a German sentence simplification benchmark, we evaluate the reproduced models and three new models on seven German sentence simplification test sets, i.e., APA-LHA-OR-A2, APA-LHA-OR-B1, DEplain-APA, DEplain-web, Simple German Corpus (SGC), TCDE19, and GEOlino. We first describe the evaluation approach, then report the models' results per domain of the test set, and finally compare the results across all test sets.

6.1. Method

All models are automatically evaluated against one reference[19] and on the same evaluation metrics, i.e., SARI (Xu et al., 2016), BLEU (Papineni et al., 2002), BS_P (Zhang* et al., 2020), and FRE (Amstad, 1978). Although the metrics have been criticized regarding their suitability for text simplification evaluation (e.g., see Sulem et al. 2018, Tanprasert and Kauchak 2021, or Alva-Manchego et al. 2021), we are reporting them due to missing alternatives. Following the recommendation of Alva-Manchego et al. (2021), we use BS_P as the main evaluation metric. If the score is high, we verify it with other metrics, such as SARI, BLEU, and FRE. In addition, as recommended by Tanprasert and Kauchak (2021) and Alva-Manchego et al. (2019), we also report linguistic features to get more insights into the system-generated simplifications, i.e., compression ratio and sentence splits.

For the measurement of the metrics and features, we are using the evaluation framework, i.e., EASSE-DE, a multi-lingual adaptation of the EASSE evaluation framework (Stodden, 2024). In comparison to EASSE, EASSE-DE includes, for example, German tokenization, German readability metrics, and a multi-lingual version of BERTScore. In Appendix G, more details are provided regarding the settings used for evaluation with EASSE-DE. We do not manually evaluate the models as this is out of the scope of this work.

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio↓	Sent. splits↑
hda_LS	3.02	14.02	0.12	37.55	1.14	1.04
sockeye-APA-LHA	**13.59**	**51.77**	**0.35**	68.65	0.64	0.99
sockeye-DEplain-APA	4.79	40.32	0.25	**70.25**	0.71	1.25
mBART-DEplain-APA	4.73	30.28	0.23	57.55	0.85	1.33
mBART-DEplain-APA+web	4.56	25.89	0.23	56.35	0.84	1.16
mT5-DEplain-APA	4.65	34.47	0.24	58.10	0.58	1.09
mT5-SGC	2.78	39.79	0.28	**70.25**	**0.48**	1.00
BLOOM-zero	2.44	26.83	0.19	51.85	0.82	1.29
BLOOM-10-random	2.64	33.05	0.24	57.95	0.64	0.98
BLOOM-10-similarity	5.10	38.05	0.29	64.60	0.59	0.98
custom-decoder-ats	0.28	37.05	0.08	52.60	3.16	**2.91**
Identity baseline	3.50	3.90	0.18	44.70	1.00	1.00
Reference baseline	100	100	1.00	69.55	0.60	0.97
Truncate baseline	2.60	17.49	0.19	54.25	0.79	1.00

Table 3: Evaluation on APA-LHA-OR-A2.

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio↓	Sent. splits↑
hda_LS	4.54	15.49	0.15	36.15	1.15	1.10
sockeye-APA-LHA	**11.00**	**44.93**	**0.32**	61.90	0.70	0.97
sockeye-DEplain-APA	3.57	39.4	0.25	**70.65**	0.68	1.26
mBART-DEplain-APA	5.32	30.94	0.26	57.65	0.86	1.37
mBART-DEplain-APA+web	5.81	26.61	0.25	56.05	0.85	1.19
mT5-DEplain-APA	4.92	35.70	0.26	57.70	0.57	1.10
mT5-SGC	2.54	39.36	0.29	70.45	**0.48**	1.00
BLOOM-zero	3.41	27.56	0.21	56.80	0.84	1.34
BLOOM-10-random	5.18	32.43	0.26	56.25	0.71	0.98
BLOOM-10-similarity	6.21	37.22	0.27	62.00	0.72	0.98
custom-decoder-ats	0.52	37.59	0.07	49.70	3.78	**3.51**
Identity baseline	5.47	4.89	0.22	43.70	1.00	1.00
Reference baseline	100	100	1.00	62.60	0.68	0.98
Truncate baseline	4.59	18.36	0.22	53.85	0.79	1.00

Table 4: Evaluation on APA-LHA-OR-B1.

6.2. News Test Sets: APA-LHA-OR-A2 & APA-LHA-OR-B1 & DEplain-APA

Although, mBART-DEplain-APA, mT5-DEplain-APA, sockeye-DEplain-APA, and sockeye-APA-LHA are trained on alignments of the same source, i.e., news of the Austrian Press Agency, sockeye-APA-LHA achieves clearly better BS_P (difference > 5), SARI (difference > 9) and BLEU scores (difference > 5) on both APA-LHA test sets (see Table 3 and Table 4). In contrast, sockeye-DEplain-APA, mBART-DEplain-APA and mT5-DEplain-APA perform much better on DEplain-APA than sockeye-APA-LHA (see Table 5) with respect to BS_P (difference > 16), SARI (difference > 4), and BLEU (difference > 8). Hence, as expected, the models are most suitable on the test set of the corpus that they have been trained on (APA-LHA vs. DEplain-APA). Besides computational reasons, this might also be due to the different alignment strategies (APA-LHA: automatically vs. DEplain-APA: manually) or the different extent of the complex-simple pairs (APA-LHA: OR to A2 or B1 vs. DEplain-APA: B2 to A2) of both corpora.

However, mBART-DEPlain-APA, mT5-DEplain-APA, and sockeye-DEplain-APA are all trained on the same training data. Hence, their differences in performance seem to be due to their system architectures. When evaluating on DEplain-APA, sockeye-DEplain-APA splits the sentences most often, whereas mT5-DEplain-APA compresses most sentences. Further, the mBART model achieves the best results concerning BS_P and BLEU, but sockeye-DEplain-APA achieves the highest SARI score and a much lower BS_P score (difference = 11). More experiments with different hyperparameters and training sets are required to confirm this finding.

Further, we can compare the mBART models with respect to a data augmentation strategy because both models are trained in an identical setting except for additional training data in mBART-DEplain-APA+web. The augmented data (automatically aligned and from different domains) seems to reduce the quality of the system generations on the news domain as on all three test sets: the BLEU, SARI and BS_P scores are lower for mBART-DEplain-APA+web than mBART-DEplain-APA.

[19]Unfortunately, no test set contains more than one reference. Therefore, the results should be considered with caution as the suitability of the evaluation metrics has been checked on (English) test sets with multiple references.

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio↓	Sent. splits↑
hda_LS	22.3	26.06	0.55	64.60	1.00	1.00
sockeye-APA-LHA	11.84	40.16	0.37	63.70	0.94	0.97
sockeye-DEplain-APA	19.58	**44.14**	0.53	71.45	0.94	1.09
mBART-DEplain-APA	**28.49**	38.72	**0.64**	65.30	0.99	1.07
mBART-DEplain-APA+web	28.03	33.81	**0.64**	65.20	0.98	1.05
mT5-DEplain-APA	22.32	39.41	0.61	63.20	0.87	1.04
mt5-SGC	8.12	37.92	0.48	**71.65**	**0.74**	1.00
BLOOM-zero	16.14	35.43	0.53	65.10	0.87	1.14
BLOOM-10-random	17.97	35.93	0.57	65.50	0.91	1.00
BLOOM-10-similarity	20.97	41.27	0.57	65.70	0.93	1.07
custom-decoder-ats	1.24	36.42	0.16	53.00	7.41	**5.07**
Identity baseline	26.89	15.25	0.63	58.75	1.00	1.00
Reference baseline	100.00	100.00	1.00	65.80	1.03	1.20
Truncate baseline	16.11	27.20	0.55	66.10	0.80	1.01

Table 5: Evaluation on DEplain-APA.

6.3. Web Test Sets: DEplain-web & Simple-German-Corpus

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio↓	Sent. splits↑
sockeye-APA-LHA	0.24	32.41	0.13	69.55	0.74	0.90
sockeye-DEplain-APA	3.44	36.24	0.24	76.7	0.76	1.32
mBART-DEplain-APA	13.50	33.11	0.40	69.65	0.90	1.30
mBART-DEplain-APA+web	**17.99**	34.07	**0.44**	69.05	0.85	1.16
mT5-DEplain-APA	6.80	**37.15**	0.36	70.90	0.63	1.10
mt5-SGC	2.50	36.56	0.37	**78.10**	**0.47**	0.93
BLOOM-zero	10.88	30.58	0.35	70.30	0.85	1.28
BLOOM-10-random	11.06	30.90	0.39	68.55	0.69	0.98
BLOOM-10-similarity	11.62	37.03	0.42	70.05	0.63	0.98
custom-decoder-ats	0.72	34.92	0.10	57.15	5.41	**3.79**
Identity baseline	20.85	11.93	0.42	62.95	1.00	1.00
Reference baseline	100.00	100.00	1.00	77.90	0.94	1.84
Truncate baseline	17.28	24.58	0.40	67.05	0.82	1.02

Table 6: Evaluation on DEplain-web.

Focusing on the web test sets, mBART-DEplain-APA+web performs best on DEplain-web (wrt. BS_P and BLEU, see Table 6) and BLOOM-10-similarity best on SGC (wrt. BS_P, SARI, and BLEU, see Table 7). Although mt5-SGC and mBART-DEplain-APA+web are both trained on complex-simple pairs of the web domain, both achieve comparable low BS_P scores on SGC. A reason for that might be the mix of topics, different alignment types (automatic vs. manual), or

a mix of language varieties (Easy German, Plain German, and others) in their training data.

customer-decoder-ats and sockeye-APA-LHA perform the worst on both datasets (wrt. BS_S). Following the compression ratio and sentence split values, customer-decoder-ats seems to hallucinate by extending the original text with many additional sentences. This might be because customer-decoder-ats is originally built to simplify longer texts. Sockeye-APA-LHA appears to underperform on test sets for other target groups or domains other than its training data.

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio ↓	Sent. splits↑
hda_LS	6.34	20.22	0.25	41.15	1.00	1.03
sockeye-APA-LHA	0.33	35.50	0.13	63.70	0.80	0.82
sockeye-DEplain-APA	1.35	37.86	0.18	**71.05**	0.79	1.01
mBART-DEplain-APA	5.70	32.77	0.31	58.15	0.97	1.00
mBART-DEplain-APA+web	6.56	29.80	0.33	44.95	1.61	1.09
mT5-DEplain-APA	2.81	35.92	0.30	51.45	0.76	0.88
mt5-SGC	3.30	43.62	0.37	58.55	**0.61**	0.85
BLOOM-zero	3.76	31.95	0.25	53.55	0.81	1.07
BLOOM-10-random	4.64	33.16	0.30	51.50	0.75	0.92
BLOOM-10-similarity	**13.32**	**44.66**	**0.38**	58.65	0.92	1.13
custom-decoder-ats	0.44	36.53	0.06	32.05	8.83	**3.68**
Identity baseline	7.46	6.51	0.29	41.15	1.00	1.00
Reference baseline	100.00	100.00	1.00	65.40	1.25	1.81
Truncate baseline	4.66	20.12	0.28	50.50	0.81	0.87

Table 7: Evaluation on SGC.

6.4. Knowledge Acquiring Test Sets: GEOlino & TCDE19

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio ↓	Sent. splits↑
hda_LS	55.22	34.20	0.76	61.50	1.00	1.00
sockeye-APA-LHA	0.69	18.94	0.15	69.45	1.05	0.92
sockeye-DEplain-APA	7.27	24.71	0.33	77.3	0.96	1.15
mBART-DEplain-APA	50.56	**44.29**	0.74	70.75	1.04	1.15
mBART-DEplain-APA+web	**55.35**	44.28	**0.79**	64.60	0.97	1.08
mT5-DEplain-APA	28.43	36.93	0.65	67.95	0.80	1.04
mt5-SGC	11.92	28.75	0.55	**78.30**	**0.70**	0.94
BLOOM-zero	28.18	32.15	0.59	67.85	0.87	1.26
custom-decoder-ats	0.77	22.05	0.08	46.55	14.61	**4.76**
Identity baseline	67.12	26.81	0.86	61.50	1.00	1.00
Reference baseline	100.00	100.00	1.00	66.00	0.95	1.32
Truncate baseline	45.39	29.78	0.75	63.80	0.83	1.00

Table 8: Evaluation on GEOlino (n=663).

We also evaluate on two test sets with simplification of knowledge-acquiring platforms, i.e. GEOlino simplification of science for children and TCDE19 with simplifications of Wikipedia texts for non-native German speakers. For both corpora, only test sets and no training sets exist. Therefore, BLOOM-10-random and BLOOM-10-similarity cannot be evaluated as no samples exist that could be added during prompting.

Further, currently, no training data for sentence simplification in the same domain or for the same target group of these test sets exists. Therefore, the presented results in Table 8 and Table 9 can be seen in the out-of-domain evaluation of the TS systems. mBART-DEplain-APA+web performs best on both test sets with respect to BLEU and BERTScore whereas mBART-DEplain-APA achieves best SARI scores.

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio ↓	Sent. splits↑
hda_LS	20.66	26.92	0.45	33.65	1.00	1.01
sockeye-APA-LHA	0.13	29.87	0.14	**69.05**	0.43	0.97
sockeye-DEplain-APA	0.68	31.79	0.19	65.0	0.51	1.42
mBART-DEplain-APA	13.69	**39.14**	0.50	51.10	0.76	1.57
mBART-DEplain-APA+web	**17.75**	37.37	**0.55**	43.65	0.74	1.29
mT5-DEplain-APA	2.84	35.09	0.40	46.60	0.40	1.14
mt5-SGC	1.05	32.98	0.38	64.40	**0.31**	0.97
BLOOM-zero	9.46	34.96	0.42	45.55	0.78	1.75
custom-decoder-ats	1.73	32.87	0.22	27.70	1.54	**4.22**
Identity baseline	27.31	14.99	0.55	28.10	1.00	1.00
Reference baseline	100.00	100.00	1.00	51.20	0.95	2.04
Truncate baseline	20.17	26.45	0.52	37.65	0.81	1.00

Table 9: Evaluation on TCDE19 (n=250).

6.5. Comparison Across Domains

In this section, we analyse the reproduced models' results across all test sets. For a better overview of the capabilities of the models across the test sets, in Appendix I, we provide the BS_P scores of all models on all test sets and in Appendix J for SARI. The tables also include the rank of the model per test set.[20]

Comparing the performance of the models across all test sets, the scores of hda_LS are always close to the scores of the identity baseline, which might be due to only minimal changes in the original sentences. Of all models, custom-decoder-ats still produces the most complex sentences with respect to FRE and compression ratio. On all test sets, the readability seems even lower for custom-decoder-ats than for the original complex texts (see identity baselines). The reason for that is hallucination in the system outputs, which could be explained by the model's design as it is trained for document simplification, in which the texts are, by nature, longer than in sentence simplification corpora. mt5-SGC has the lowest compression ratio on all test sets, possibly due to the very short sentences in its training data, which are mostly texts in Easy German.

Overall, no system ranks best across all test sets (wrt. BS_P and SARI). On average, BLOOM-10-similarity performs best (wrt. BS_P) if similar examples are available, whereas mBART-DEplain-APA+web achieves on average, the best ranks following BS_P on all seven test sets, and sockeye-DEplain-APA performs best on both settings wrt. SARI. The additional data, i.e., massive data during pre-training in BLOOM and additional web data for mBART, seems to have a positive effect on the system generations or at least the evaluation

[20] BLOOM-10-random and BLOOM-10-similarity require training samples each time when generating a simplified sentence, which is not available for all test sets (e.g., TCDE19 or GEOlino). In addition, when simplifying texts in practice, i.e., as an intra-lingual translation tool, also no simplification examples would be made available. In order to integrate this limitation, BLOOM-10-random and BLOOM-10-similarity will be penalized in our evaluation on TCDE19 and GEOlino with the highest rank equal to the worst result.

scores. In comparison, sockeye-DEplain-APA is only trained on simple-complex pairs of DEplain-APA, and, therefore, the model cannot transfer well to other domains as it only performs very well on the news test sets.

However, the transfer learning of pre-trained models appears to be more effective for BLOOM than for mBART or mT5, which might be due to its larger pre-training data size. Further, BLOOM has been prompted with only a few samples, but it still outperforms the smaller language models, even though they have been fine-tuned on many task-relevant samples. We can also confirm the findings of (Ryan et al., 2023) that BLOOM-10-similarity generates better simplifications than BLOOM-10-random and better than BLOOM-zero on all test sets with respect to BS_P and SARI. For more comparisons of mT5 and BLOOM (also including the capability of TS models across multiple languages), we refer the interested reader to Ryan et al. (2023).

For the sockeye models, we assume that the size of APA-LHA and DEplain-APA is too small to train a model from scratch. It could be a promising approach to combine similar training data with each other to increase the training size for sockeye, e.g., a combination of APA-LHA, DEplain-APA, and/or SGC because a positive effect of data combination has been revealed for mBART-DEplain-APA+web compared to mBART-DEplain-APA.

7. Conclusion & Discussion

We have reproduced different approaches on how to simplify German texts automatically. However, we have also revealed some new issues regarding models' reproduction and have confirmed previously named problems with respect to the training data and the evaluation process.

We found the following three main issues with the models, i.e.,

(i) impossibility of reproduction, e.g., due to missing details, missing code, not-available or restricted-access data, or restricted-access language models,

(ii) differences in reproduction and, therefore, less comparison, e.g., due to different data splits, and

(iii) differences in evaluation scores for reported scores and scores of reproduced models due to different system outputs or different implementations of metrics.

For better reproducibility and better comparison between ATS models, we recommend publishing as many details and materials related to the models as possible with respect to copyright and licenses, e.g., publishing (i) the checkpoints of the trained or fine-tuned models and code how to reuse them, or (ii) the code and a description of how to rebuild and re-train the model, including model versions and used prompts.

Additionally, we also recommend publishing the system generations (if not restricted by copyright) to enable further analysis of the results. In our reproduction study, in the comparison of the reported and reproduced scores, we have seen that even if the ATS models or the code are available for reproduction, the system generations seem to be different from those described in the original works. Hence, some analysis of the original work might not hold when reproducing.

We have also shown that, for example, due to limited computing resources, system generations cannot always be reproduced even if the code or the model is provided. We argue that the system generations are helpful for understanding the original work better and can also be valuable for building better evaluation metrics.

To compare the reproduced models with each other, we have built a German sentence simplification benchmark on 7 test sets. We found, as expected, that models achieve the best scores if they are evaluated and trained on the same corpus. We have also shown that some models, especially mBART-DEplain-APA+web (wrt. SARI and BERT-Score), achieve good scores on test sets on which domain or target group they were not trained. Hence, the models seem to have learned some universal simplification. Nevertheless, we want to emphasize that simplicity is subjective. Hence, for each person and each target group, a text is easier or more difficult to read. Following this, a text simplification model should also learn to simplify for a specific target group and not for many target groups at the same time (Gooding, 2022; Stajner, 2021). Therefore, we recommend not mixing training data from texts written for different target groups but evaluating the models only on texts written for the target group of interest. Due to limited resources, this is currently impractical. Hence, we have presented approaches with mixed training data and evaluated across texts of different target groups.

However, the analysis with respect to SARI or BERT-Score allows us to draw different conclusions: Following their scores, different models are ranked as best models. More work regarding the suitability and interpretability of evaluation metrics (especially regarding test sets with only one reference) is required for a more reliable interpretation of this German TS benchmark.

8. Bibliographical References

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Studenten-Schreib-Service. PhD Thesis.

Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, and Georg Groh. 2023. Language models for German text simplification: Overcoming parallel data scarcity through style-specific pre-training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1147–1158, Toronto, Canada. Association for Computational Linguistics.

Mohammad Arvan, Luís Pina, and Natalie Parde. 2022. Reproducibility of exploring neural text simplification models: A review. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 62–70, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with REPROLANG2020. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5539–5545, Marseille, France. European Language Resources Association.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. ArXiv preprint, arXiv:2210.11416.

Michael Cooper and Matthew Shardlow. 2020. CombiNMT: An exploration into neural text simplification models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5588–5594, Marseille, France. European Language Resources Association.

Silvana Deilen, Sergio Hernández Garrido, Ekaterina Lapshinova-Koltunski, and Christiane Maaß. 2023. Using ChatGPT as a CAT tool in easy language translation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 1–10, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. 2020. The sockeye 2 neural machine translation toolkit at AMTA 2020. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 110–115, Virtual. Association for Machine Translation in the Americas.

Sarah Ebling, Alessia Battisti, Marek Kostrzewa, Dominik Pfütze, Annette Rios, Andreas Säuberli, and Nicolas Spring. 2022. Automatic text simplification for german. *Frontiers in Communication*, 7.

Sian Gooding. 2022. On the ethical considerations of text simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126,

Online. Association for Computational Linguistics.

Sabine Manning. 2023. KI-Tools für Einfache Sprache: (3) Bard und GPT-4 im Vergleich. https://multisprech.org/2023/11/16/ki-tools-fuer-einfache-sprache-3-bard-und-gpt-4-im-vergleich/. Last change: 2023-11-16; Last access: 2024-01-11.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2019. DisSim: A discourse-aware syntactic text simplification framework for English and German. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 504–507, Tokyo, Japan. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

David Ponce, Thierry Etchegoyhen, Jesús Calleja Pérez, and Harritxu Gete. 2023. Split and rephrase with large language models. ArXiv preprint, arXiv:2312.11075.

Maja Popović, Sheila Castilho, Rudali Huidrom, and Anya Belz. 2022. Reproducing a manual evaluation of the simplicity of text simplification system outputs. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 80–85, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161, Online and in Dominican Republic. Association for Computational Linguistics.

Michael Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

Andreas Säuberli, Sarah Ebling, and Martin Volk. 2020. Benchmarking data-driven automatic text simplification for German. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 41–48, Marseille, France. European Language Resources Association.

Tim Schlippe and Katharina Eichinger. 2023. Multilingual text simplification and its performance on social sciences coursebooks. In *Artificial Intelligence in Education Technologies: New Development and Innovative Practices*, pages 119–136, Singapore. Springer Nature Singapore.

Melanie Siegel, Dorothee Beermann, and Lars Hellan. 2019. Aspects of linguistic complexity: A german - norwegian approach to the creation of resources for easy-to-understand language. In *11th International Conference on Quality of Multimedia Experience QoMEX 2019, Berlin, Germany, June 5-7, 2019*, pages 1–3. IEEE.

Nicolas Spring, Annette Rios, and Sarah Ebling. 2021. Exploring German multi-level text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1339–1349, Held Online. INCOMA Ltd.

Sanja Stajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.

Regina Stodden. 2024. EASSE-DE: Easier Automatic Sentence Simplification Evaluation for German. ArXiv preprint, arXiv:2404.03563.

Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Julia Suter, Sarah Ebling, and Martin Volk. 2016. Rule-based automatic text simplification for german. In *13th Conference on Natural Language Processing (KONVENS 2016)*.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics.

Vanessa Toborek, Moritz Busch, Malte Boßert, Christian Bauckhage, and Pascal Welke. 2023. A new aligned simple German corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11393–11412, Toronto, Canada. Association for Computational Linguistics.

Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus, and Christin Seifert. 2022. Patient-friendly clinical notes: Towards a new text simplification dataset. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 19–27, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

BigScience Workshop. 2023. Bloom: A 176b-parameter open-access multilingual language model. Version 4.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

9. Language Resource References

Alva-Manchego, Fernando and Martin, Louis and Scarton, Carolina and Specia, Lucia. 2019. *EASSE: Easier Automatic Sentence Simplification Evaluation*. GitHub. PID https://github.com/feralvam/easse.

Felix Hieber and Michael Denkowski and Tobias Domhan and Barbara Darques Barros and Celina Dong Ye and Xing Niu and Cuong Hoang and Ke Tran and Benjamin Hsu and Maria Nadejde and Surafel Lakew and Prashant Mathur and Anna Currey and Marcello Federico. 2022. *Sockeye 3: Fast Neural Machine Translation with PyTorch*.

Mallinson, Jonathan and Sennrich, Rico and Lapata, Mirella. 2020. *GEOlino*. GitHub. PID https://github.com/Jmallins/ZEST-data.

Babak Naderi and Salar Mohtaj and Kaspar Ensikat and Sebastian Möller. 2019. *TextComplexityDE*. GitHub. PID https://github.com/babaknaderi/TextComplexityDE.

Rios, Annette and Spring, Nicolas and Kew, Tannon and Kostrzewa, Marek and Säuberli, Andreas and Müller, Mathias and Ebling, Sarah. 2021. *20Minuten*. GitHub. PID https://github.com/ZurichNLP/20Minuten.

Ryan, Michael and Naous, Tarek and Xu, Wei. 2023a. *GEOlino-small*. GitHub. PID https://github.com/XenonMolecule/MultiSim/tree/main/data/German.

Ryan, Michael and Naous, Tarek and Xu, Wei. 2023b. *MultiSim*. GitHub. PID https://github.com/XenonMolecule/MultiSim.

Ryan, Michael and Naous, Tarek and Xu, Wei. 2023c. *TextComplexityDE-small*. GitHub. PID https://github.com/XenonMolecule/MultiSim/tree/main/data/German.

Spring, Nicolas and Rios, Annette and Ebling, Sarah. 2021. *LHA Sentence Alignments Extracted From the Austria Press Agency Corpus*. Zenodo. PID https://doi.org/10.5281/zenodo.5148163.

Regina Stodden. 2024. *EASSE-DE: Easier Automatic Sentence Simplification Evaluation for German*. GitHub. PID https://github.com/rstodden/easse-de.

Regina Stodden and Omar Momen and Laura Kallmeyer. 2023a. *DEplain-APA*. Zenodo. PID https://doi.org/10.5281/zenodo.8304430.

Stodden, Regina and Momen, Omar and Kallmeyer, Laura. 2023b. *DEplain-web*. GitHub. PID https://github.com/rstodden/DEPlain.

Toborek, Vanessa and Busch, Moritz and Boßert, Malte and Bauckhage, Christian and Welke, Pascal. 2023. *Simple German Corpus*. GitHub. PID https://github.com/buschmo/Simple-German-Corpus.

A. Hyperparamter mT5

parameter name	value
epochs	10
model	mt5-base
prefix	"simplify to plain German: "
max length	128:128
learning rate	0.001
batch size	4
metric	SARI
optimzer	adafactor

Table 10: Hyperparameter for fine-tuning mT5

B. Hyperparameter Sockeye

C. Reproduction results of sockeye-APA-LHA

D. Reproduction results of BLOOM

E. Reproduction results of customer-decoder-ats

F. Reproduction Results of mBART-DEplain-APA and mBART-DEplain-APA+web

G. EASSE-DE settings

• lowercasing: False, • tokenizer: spacy, • test set: custom, • metrics: bleu,sari,bertscore,fre • language: DE

	Sockeye-APA-LHA	Spring et al. (2021)
Sockeye version	3.1.34	< 2.3.17
num_layers	6	6
optimized_metric	'bleu'	'bleu'
max_num_checkpoint_not_improved	10	10
checkpoint_improvement_threshold	0.001	
seed	42	1
batch_type	'sentence'	word
batch_size	256	2048
optimizer	'adam'	'adam'
max_seq_len	95	95
label_smoothing	0.3	0.3
transformer_model_size	512	512
transformer_attention_heads	4	4
transfromer_feed_forward_num_hidden	2048	2048
transformer_dropout_attention	0.1	0.1
transformer_dropout_act	0	0
transformer_dropout_prepost	0.1	0.1
embed_dropout	0.3	0.3
transformer_positional_embedding_type	'fixed'	'fixed'
initial_learning_rate	0.0002	0.0002
learning_rate_reduce_factor	0.9	0.9
learning_rate_schedule_type	'plateau-reduce'	'plateau-reduce'
update_interval	1	2
vocabular size	20000	20000
init		xavier
Init-scale		3
Init-xavier-factor-type		avg
architecture		transformer

Table 11: Hyperparameters of our reproduction and the ones reported in Spring et al. (2021).

	copied		reproduced	
System	**BLEU↑**	**SARI↑**	**BLEU↑**	**SARI↑**
Sockeye-APA-LHA	12.3	40.73	11.40	45.20

(a) APA-LHA OR-B1

	copied		reproduced	
System	**BLEU↑**	**SARI↑**	**BLEU↑**	**SARI↑**
Sockeye-APA-LHA	15.20	42.04	14.15	52.17

(b) APA-LHA OR-A2

Table 12: Reproduced and copied results for sockeye on APA-LHA (Spring et al., 2021).

H. Hyperparameter mBART

I. Overview of BERT-Score Precision per model and test set

J. Overview of SARI results per model and test set

	TCDE19 (n=25)		GEOlino (n=25)	
System	**copied SARI↑**	**repro. SARI↑**	**copied SARI↑**	**reprod. SARI↑**
zero-shot	32.26	34.96	29.59	28.75
random 10-shot	38.07	35.49	35.42	36.92
similarity 10-shot	38.93	39.86	39.7	40.36
Identity Baseline	15.42	15.42	27.45	27.44
Truncate Baseline	26.81	26.81	30.7	30.74

Table 13: Reproduced and copied results for BLOOM. The identity baseline results are taken from the code, all other copied scores are taken from the original paper (Ryan et al., 2023)
.

	BLEU↑	SARI↑	ROUGE-L↑
german_gpt FT	4.8	42.74	17.93

(a) 20Minuten (copied)

	BLEU↑	SARI↑	ROUGE-L↑
german_gpt FT	4.12	41.85	17.23

(b) 20Minuten (reproduced)

Table 14: Reproduced and copied results for 20Min and custom-decoder-ats (Anschütz et al., 2023).

System	BLEU↑	SARI↑	BS_P↑	FRE↑
mBART-DEplain-APA	28.25	34.818	0.639	63.072
mBART-DEplain-APA+web	28.506	34.904	0.64	62.669
Identity baseline	26.89	15.25	0.63	58.75

(a) copied

System	BLEU↑	SARI↑	BS_P↑	FRE↑
mBART-DEplain-APA	30.01	39.12	0.48	-
mBART-DEplain-APA+web	29.62	34.44	0.47	-
Identity baseline	28.50	15.88	0.45	-

(b) reproduced & EASSE

System	BLEU↑	SARI↑	BS_P↑	FRE↑
mBART-DEplain-APA	28.49	38.72	0.64	65.3
mBART-DEplain-APA+web	28.03	33.81	0.64	65.2
Identity baseline	26.89	15.25	0.63	59.23

(c) reproduced & EASSE-DE

Table 15: Reproduced and copied results for mBART-DEplain-APA and mBART-DEplain-APA+web (Stodden et al., 2023) on DEplain-APA.

System	BLEU↑	SARI↑	BS_P↑	FRE↑
mBART-DEplain-APA	15.727	30.867	0.413	64.516
mBART-DEplain-APA+web	17.88	34.828	0.436	65.249
Identity baseline	20.85	11.931	0.423	60.825

(a) copied

System	BLEU↑	SARI↑	BS_P↑	FRE↑
mBART-DEplain-APA	14.41	33.15	0.20	-
mBART-DEplain-APA+web	18.95	34.11	0.25	-
Identity baseline	21.65	12.34	0.23	-

(b) reproduced & EASSE

System	BLEU↑	SARI↑	BS_P↑	FRE↑
mBART-DEplain-APA	13.5	33.11	0.4	69.65
mBART-DEplain-APA+web	17.99	34.07	0.44	69.05
Identity baseline	20.85	11.93	0.42	62.95

(c) reproduced & EASSE-DE

Table 16: Reproduced and copied results for mBART-DEplain-APA and mBART-DEplain-APA+web (Stodden et al., 2023) on DEplain-web.

	Rios et al. (2021)	Rios et al. (2021)	Trienes et al. (2022)	Stodden et al. (2023)	Stodden et al. (2023)
model	standard mbart	small mbart	mBART-large-cc25	mBART	long-mbart
model-url			facebook/mbart-large-cc25	facebook/mbart-large-cc25	facebook/mbart-large-cc25
max length	1024:1024	1024:4096		256:256	2048:1024
learning rate			0.00003	0.00003	0.00003
lr_schedule_type	'plateau-reduce'	'plateau-reduce'		'plateau-reduce'	'plateau-reduce'
batch size	1024:1024	4	4	16	1
optimizer			adamW	adam	adam
warm-up			10% of train + linear decay		
beam size			5	6	6
vocabulary size	250k	20k		35k	35k
attention window	x	512		512	512
attention dropout	0.1	0.1		0.1	0.1
dropout	0.3	0.3		0.3	0.3
label smoothing	0.2	0.2		0.2	0.2
early stopping	rougeL	rougeL		rougeL	rougeL
language tags	de_DE:[de_A1|de_A2|de_B1]		yes, but not specified	de_DE:de_Sl	de_DE:de_Sl

Table 17: Hyperparamters of mBART in different papers.

	APA-LHA-OR-B1		**APA-LHA-OR-A2**		**DEplain-APA**		**DEplain-web**		**SGC**		**GEOlino**		**TCDE19**		**AVG rank**	
	BS_P	**Rank**	**BS_P**	**Rank**	**BS_P**	**Rank**	**BS_P**	**Rank**	**BS_P**	**Rank**	**BS_P**	**Rank**	**BS_P**	**Rank**	**(5 sets)**	**(7 sets)**
hda_LS	0.12	10	0.15	10	0.55	6	*n/a*	11	0.25	8	0.76	2	0.45	3	9	7.14
sockeye-APA-LHA	**0.35**	**1**	**0.32**	**1**	0.37	10	0.13	9	0.13	10	0.15	8	0.14	9	6.2	6.86
sockeye-DEplain-APA	0.25	4	0.25	8	0.53	8	0.24	8	0.33	9	0.19	7	0.18	8	7.4	7.43
mBART-DEplain-APA	0.23	8	0.26	6	**0.64**	**2**	0.4	3	0.31	4	0.74	3	0.5	2	4.6	4
mBART-DEplain-APA+web	0.23	8	0.25	8	**0.64**	**2**	**0.44**	**1**	0.33	3	**0.79**	**1**	**0.55**	**1**	4.4	**3.43**
mT5-DEplain-APA	0.24	6	0.26	6	0.61	3	0.36	6	0.3	6	0.65	4	0.4	5	5.4	5.14
mT5-SGC	0.28	3	0.29	2	0.48	9	0.37	5	0.37	2	0.55	6	0.38	6	4.2	4.71
BLOOM-zero	0.19	9	0.21	9	0.53	8	0.35	7	0.25	8	0.59	5	0.42	4	8.2	7.14
BLOOM-10-random	0.24	6	0.26	6	0.57	5	0.39	4	0.3	6	*n/a*	11	*n/a*	11	5.4	7
BLOOM-10-similarity	0.29	2	0.27	3	0.57	5	0.42	2	**0.38**	**1**	*n/a*	11	*n/a*	11	**2.6**	5
custom-decoder-ats	0.08	11	0.07	11	0.16	11	0.1	10	0.06	11	0.08	9	0.22	7	10.8	10

Table 18: Overview of BERT-Score Precision values per model and test set including ranks per test set. The last two columns contain the averages across all test sets (n=7) and all test sets with available training data (n=5).

	APA-LHA-OR-B1		**APA-LHA-OR-A2**		**DEplain-APA**		**DEplain-web**		**SGC**		**GEOlino**		**TCDE19**		**AVG rank**	
	SARI	**Rank**	**SARI**	**Rank**	**SARI**	**Rank**	**SARI**	**Rank**	**SARI**	**Rank**	**SARI**	**Rank**	**SARI**	**Rank**	**(5 sets)**	**(7 sets)**
hda_LS	14.02	11	15.49	11	26.06	11	*n/a*	11	20.22	11	34.2	4	26.92	9	11	9.71
sockeye-APA-LHA	**51.77**	**1**	**44.93**	**1**	40.16	3	32.41	8	35.5	6	18.94	9	29.87	8	3.8	5.14
sockeye-DEplain-APA	40.32	2	39.4	2	**44.14**	**1**	36.24	4	24.71	3	31.79	7	37.86	7	**2.4**	**3.71**
mBART-DEplain-APA	30.28	8	30.94	8	38.72	5	33.11	7	32.77	8	**44.29**	**1**	**39.14**	**1**	7.2	5.43
mBART-DEplain-APA+web	25.89	10	26.61	10	33.81	10	34.07	6	29.8	10	44.28	2	37.37	2	9.2	7.14
mT5-DEplain-APA	34.47	6	35.7	6	39.41	4	**37.15**	**1**	35.92	5	36.93	3	35.09	3	4.4	4
mT5-SGC	39.79	3	39.36	3	37.92	6	36.56	3	43.62	2	28.75	6	32.98	5	3.4	4
BLOOM-zero	26.83	9	27.56	9	35.43	9	30.58	10	31.95	9	32.15	5	34.96	4	9.2	7.86
BLOOM-10-random	33.05	7	32.43	7	35.93	8	30.9	9	33.16	7	n/a	11	n/a	11	7.6	8.57
BLOOM-10-similarity	38.05	4	37.22	5	41.27	2	37.03	2	**44.66**	**1**	n/a	11	n/a	11	2.8	5.14
custom-decoder-ats	37.05	5	37.59	4	36.42	7	34.92	5	36.53	4	22.05	8	32.87	6	5	5.57

Table 19: Overview of SARI scores per model and test set including ranks per test set. The last two columns contain the averages across all test sets (n=7) and all test sets with available training data (n=5).

Complexity-Aware Scientific Literature Search
Searching for Relevant and Accessible Scientific Text

Liana Ermakova[†], **Jaap Kamps**[‡]
[†]Université de Bretagne Occidentale, HCTI, France, liana.ermakova@univ-brest.fr
[‡]University of Amsterdam, The Netherlands, kamps@uva.nl

Abstract

We conduct a series of experiments on ranking scientific abstracts in response to popular science queries issued by laypersons. We show that standard IR ranking models optimized on topical relevance are indeed ignoring the individual user's context and background knowledge. We also demonstrate the viability of complexity-aware retrieval models that retrieve more accessible relevant documents or ensure these are ranked prior to more complex documents on the topic. More generally, our results help remove some of the barriers to consulting scientific literature by laypersons and hold the potential to promote science literacy in the general public.

Lay Summary: *In a world of misinformation and disinformation, access to objective evidence-based scientific information is crucial. The general public ignores scientific information due to its perceived complexity, resorting to shallow information on the web or in social media. We analyze the complexity of scientific texts retrieved for a layperson's topic, and find a great variation in text complexity. A proof of concept complexity-aware search engine is able to retrieve both relevant and accessible scientific information for a layperson's information need.*

Keywords: Complexity-Aware Information Retrieval, Text Complexity and Readability, Lay Access to Scientific Text.

1. Introduction

The internet and social media drastically altered both the process of generating information and the way we consume it. The internet gives us far easier access to objective scientific information, which is a natural antidote against the pervasive misinformation and disinformation on the Web. In reality, only a small number of non-specialists refer to scientific sources, opting instead for superficial information disseminated on the internet and social media. One of the primary motives for avoiding the scientific literature is its perceived complexity. Even in developed countries, up to 30% of the population can only comprehend texts written with a basic vocabulary (Štajner et al., 2022).

Traditionally Information Retrieval (IR) systems are evaluated according to their efficiency in retrieving documents topically related to a query but this paradigm ignores the widely varying backgrounds and expertise levels of individual users, who may strictly prefer more accessible information on the topic over highly advanced documents. Specialized scholarly search engines, such as Google Scholar, DBLP, or PubMed, are designed to assist experts in scientific literature review (Gusenbauer and Haddaway, 2020) and thus do not target the accessibility of retrieved documents to laypersons. However, retrieved scientific documents might be too difficult for a user who might not understand these documents. As a result, these documents might be completely useless for a user even if they are relevant to the query.

We assume an information retrieval or retrieval augmented generation setting with a closed collection. Despite promising results of LLMs for multiple NLP tasks, including the application of ChatGPT for biomedical QA (Jahan et al., 2023; Ateia and Kruschwitz, 2023), these models still suffer from problems such as hallucinations (Ji et al., 2023; Ateia and Kruschwitz, 2023; Ermakova et al., 2023a) or non-determinism and its potential cascading effect (Ateia and Kruschwitz, 2023). For example, ChatGPT provides correct or partially correct answers in half of the cases but the provided references only exist in a small fraction of the answers (Zuccon et al., 2023). This model's instability and hallucinations reduce the reliability of the provided answers for a scientific request. Arguably, these generative models even increase the need for grounded scientific evidence to validate generated responses.

In this paper, our main aim is to investigate the viability of complexity-aware retrieval models aiming to retrieve scientific information for non-expert users. Specifically, we aim to answer the following research questions:

- How difficult are scientific abstracts?
- Are current retrieval models sensitive to text complexity?
- How effective are complexity-aware retrieval models?

To answer these research questions, we conducted a series of experiments on ranking scientific abstracts in response to popular science queries. As traditional ad-hoc retrieval benchmarks, such as TREC collections, are not aimed

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 16–26

at evaluating the complexity of the retrieved documents, we conducted our experiments on a specialized scientific retrieval corpus for a broad audience. The CLEF SimpleText track (Ermakova et al., 2021, 2022, 2023b) was the first to investigate 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. The CLEF SimpleText track studies both the initial ranking of scientific abstracts in response to a popular science query and the use of emerging text simplification (e.g., Wu and Huang, 2022; Laban et al., 2021) approaches to rewrite complex text in order to make them accessible. This paper investigates whether the initial ranking stage can already be made aware of the text complexity of retrieved abstracts, and attempts to rank more accessible literature first.

The rest of this paper is structured in the following way. In Section 2, we discuss related work on ranking scientific text and related work on quantifying text complexity. In Section 3, we analyze the difficulty of scientific abstracts. In Section 4, we discuss traditional lexical and neural ranking models and analyze both their retrieval effectiveness as well as the text complexity of retrieved results. In Section 5, we introduce two complexity-aware ranking approaches and analyze the trade-offs between retrieval effectiveness and complexity of the retrieved results. We end in Section 6 with discussion and conclusions.

2. Related Work

This section discusses related work. First, we discuss prior work on retrieving scientific text with particular emphasis on the data used in the experiments of this paper. Second, we discuss prior work on quantifying text complexity, with particular emphasis on the common readability measures used in our analysis.

2.1. Scientific Text Retrieval

The origins of the field of IR and its Cranfield/TREC evaluation paradigm are based on searching academic literature (Cleverdon, 1962, 1967). The constantly growing number of scientific publications makes the use of automatic tools necessary, including information retrieval or summarization (Guo et al., 2021). Although specialized scientific documents have long been considered by IR systems (Jones and Van Rijsbergen, 1976), they are not sensitive to the complexity of the text. Moreover, academic search systems, including Google Scholar, PubMed, and Web of Science, are traditionally designed for scientific domain experts to assist them in doing systematic reviews, meta-analyses (Gusenbauer and Haddaway, 2020). Knowledge extraction from published scholarly literature for business and research applications is another popular area of research but it also targets specialists in a particular domain rather than laypersons (Thakur and Kumar, 2022).

Given the escalating worries about public misinformation in various countries and the rise of disinformation campaigns orchestrated by organizations, addressing how to effectively educate a wide audience about the progress in technology and science is a major concern (Scheufele and Krause, 2019).

The CLEF SimpleText track shifted the focus to laypersons searching scientific literature (Ermakova et al., 2021, 2022, 2023b). The track covers a wider range of topics on automatic scientific text simplification, from language simplification to terminology extraction and explanation. For the analysis in this paper, we use the data of the CLEF SimpleText Track's Task 1 retrieving scientific abstracts in response to a popular science query:

Corpus The Corpus consists of 4.9 million bibliographic records, including 4.2 million academic abstracts with corresponding detailed information about authors, affiliations, and citations from the Citation Network Dataset (12th version released in 2020)[1] (Tang et al., 2008).

Context There are 40 popular science articles, with 20 from *The Guardian*[2] and 20 from *Tech Xplore*.[3] These journalistic articles were used to construct search requests on popular science topics.

Requests There are 114 queries with 1-4 queries per context article, 47 queries are based on *The Guardian* and 67 on *Tech Xplore*.

Train Data The SimpleText organizers provide relevance judgments for 29 queries (corresponding to 15 Guardian articles, G01–G15), with 23 queries having more than 10 relevant abstracts. The approaches of this paper haven't been trained on this data, but it can serve as an additional evaluation for unsupervised approaches.

Assessments For the evaluation, we used the relevance assessments released for the SimpleText test data for 34 queries associated with the 5 articles from The Guardian (G16–G20, 17 queries) and 5 articles from Tech Xplore (T01–T05, 17 queries).

[1] https://www.aminer.cn/citation
[2] https://www.theguardian.com/science
[3] 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	*Elementary*						*Jr. High*		*High School*				*Undergrad.*				*Grad.*	*PhD*		
	Primary						*Secondary*						*University*					*PhD*		
	Compulsory												*Higher Edu.*							
Age	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

Table 2: Flesch-Kincaid Grade Level of CLEF SimpleText data

Data	**Size**	**Length**		**FKGL**	
		Mean	**Median**	**Mean**	**Median**
Corpus (scientific abstracts)	4,894,063	901	913	13.87	13.90
News (popular science)	40	5,504	5,540	12.69	12.80

For details of the exact task setup and results, we refer the reader to the detailed overview of the track in (Ermakova et al., 2023b).

2.2. Text Complexity

This paper performs an initial analysis of the complexity of the scientific abstracts retrieved for a popular science query. The most used way to quantify text complexity is by using readability measures (Zamanian and Heydari, 2012). To quantify the complexity, we use the popular Flesch-Kincaid Grade Level (FKGL) measure based on lexical and grammatical complexity (Flesch, 1948). 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 reasonable approximation when averaging over larger sets of data. Readability measures have been criticized ever since their invention (e.g., Štajner et al., 2012), but are the most used initial indicators of text complexity in NLP and IR.

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. Corpus Analysis

In this section, we will investigate our first research question: *How difficult are scientific abstracts?* Specifically, we apply readability measures to analyze the text complexity of the scientific data used in our experiments.

Table 2 shows an analysis of the text complexity of the corpus and of popular science context. As shown in Table 2, the average (median) length of the abstracts is 901 (913) tokens, and the average (median) complexity of the abstracts is 13.87 (13.9) 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 translating to students halfway 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 12.69 (12.8) FKGL, confirming that a FKGL around 12, translating to the readability level of a high school diploma, is appropriate for laypersons.

Is every single abstract too complex for an average citizen? We down-sampled the corpus by taking every 500th article, resulting in an arbitrary sample of 8,513 non-empty abstracts. Figure 1 (top) 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 (bottom) 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.

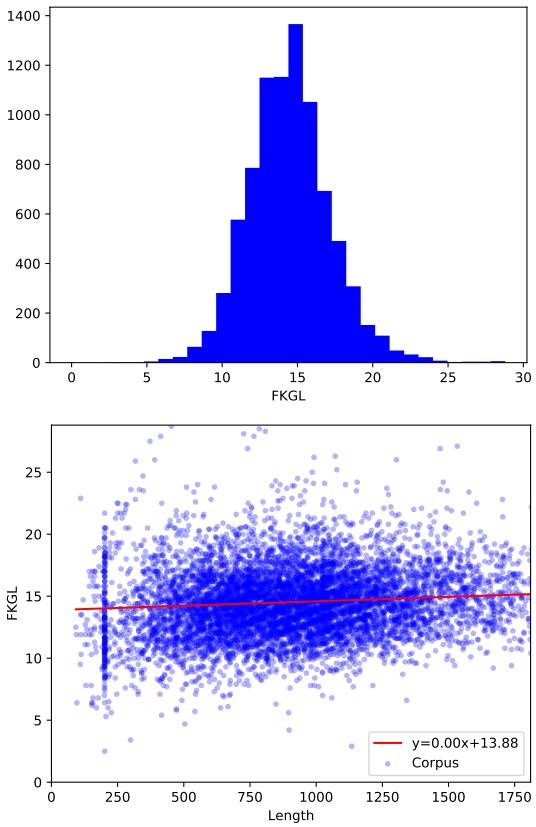


Figure 1: Distribution of text complexity in Flesch-Kincaid Grade Levels (top) and by length (bottom).

Our corpus analysis confirms the common assumption 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 significant fraction of abstracts is within the readability levels of most adult citizens. In the rest of this paper, we will investigate how information retrieval approaches can be made aware of the text complexity and prioritize the retrieval of relevant and accessible abstracts for the request at hand.

4. Effectiveness and Text Complexity

In this section, we will study our second research question: *Are current retrieval models sensitive to text complexity?* Specifically, we will use traditional and neural rankers for scientific text. First, we will evaluate the results in terms of retrieval effectiveness. Second, we will analyze the retrieved results in terms of their text complexity.

4.1. Lexical and Neural Ranking Models

We first conduct a standard IR evaluation of scientific text retrieval, using the corpus of scientific abstracts and popular science requests from the CLEF SimpleText track (Ermakova et al., 2023b).

First, we use a representative traditional ranker BM25 which is based on TF-IDF and normalized document length (Robertson et al., 2009). BM25 is commonly used in traditional search engines, including ElasticSearch,[4] Apache Solr,[5] and Terrier.[6] We used the ElasticSearch implementation of BM25 to retrieve 1,000 results for each keyword query which serves as a first-stage retrieval for the neural re-ranking models. Second, we use a representative neural cross-encoder re-ranker which is a re-implementation of BERT for query-based passage re-ranking (Nogueira and Cho, 2019). This model has shown effective retrieval performance even when applied in zero-shot to new data. Specifically, we apply an MSMARCO-trained model available from Hugging Face.[7] We use this neural cross-encoder re-ranker in a zero-shot way to re-rank either the top 100 or the top 1k retrieved abstracts by the BM25 run.

4.2. Retrieval Effectiveness

We first look at the retrieval effectiveness in the same way as in any other IR evaluation based on topical relevance judgments. Table 3 shows the performance of the three retrieval models on the train and test data, and we make a number of observations. We use standard IR evaluation measures:

- MRR (Mean Reciprocal Rank), which shows a harmonic mean of the ranks;
- Precision@k aiming to compute the share of relevant documents in the top-k retrieved results;
- NDCG (Normalized Discounted Cumulative Gain) considering both the relevance of the items and their position in the list;
- Bpref, preference-based metric that considers whether relevant documents are ranked above irrelevant ones;
- MAP (Mean Average Precision), the mean of the average precision scores for each query.

Comparing the BM25 and the neural re-rankers on the test data, we see that the cross-encoders lead to considerable improvement in retrieval effectiveness, on all evaluation measures. In particular, NDCG@10 increases from 0.3911 up to 0.4782 for

[4] https://www.elastic.co/
[5] https://opensolr.com/
[6] http://terrier.org/
[7] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2

Table 3: Retrieval effectiveness on CLEF SimpleText train (top) and test (bottom)

Run	MRR	Precision			NDCG			Bpref	MAP
		5	10	20	5	10	20		
BM25 1k	0.5605	0.4345	0.3655	0.3161	0.3606	0.3627	0.4385	0.4226	0.4072
CE 100	0.5252	0.3241	0.3034	0.2448	0.2701	0.2947	0.3472	0.4012	0.3033
CE 1k	0.4608	0.2759	0.2379	0.1701	0.2312	0.2307	0.2582	0.3335	0.2001
BM25 1k	0.6424	0.4353	0.4059	0.2990	0.4165	0.3911	0.3315	0.2502	0.1895
CE 100	0.7050	0.5118	0.4912	0.3657	0.5004	0.4782	0.4007	0.2616	0.2011
CE 1k	0.6329	0.4765	0.4735	0.3578	0.4502	0.4448	0.3816	0.2797	0.2051

Table 4: Analysis of output (over all 114 queries)

Run	Queries	Top	Year		Length		FKGL	
			Avg	Med	Avg	Med	Avg	Med
BM25 1k	114	10	2012.0	2014	1000.0	995.5	14.0	13.9
CE 100	114	10	2011.7	2013	1102.3	1041.5	14.2	14.1
CE 1k	114	10	2011.8	2014	1142.3	1047.0	14.2	14.1

the CE 100 run, suggesting that the relevant documents have higher ranks, especially in the top positions. The results on the train data are less impressive, but inspection reveals very high fractions of unjudged documents at the top of the neural runs, as no neural IR system contributed to the pools of the train data. Hence, the test data reflects the quality of these runs.

4.3. Text Complexity

We saw that modern IR models perform well in terms of retrieval effectiveness, but how complex are the retrieved abstracts? Table 4 shows an analysis of text complexity of the top 10 results of the lexical and neural models.

We see that the top 10 of the traditional BM25 model retrieves texts of a similar complexity level as the corpus (shown in Table 2 above) with an FKGL of around 14 (with a mean of 14.0, and a median of 13.9). The neural re-rankers also retrieve abstracts with this complexity level, with a slightly higher mean of 14.2 and median of 14.1. To remind, FKGL level 14 corresponds to university-level education, higher than can be taken for granted by a layperson user. Our results indicate that both traditional lexical rankers and modern neural re-rankers focus indeed solely on the topical relevance of abstracts—is the abstract on the topic of the request—and ignore other aspects such as the text complexity.

In this section, we saw that lexical and in particular neural rankers are highly effective in retrieving scientific text. This observation is consistent with the retrieval effectiveness of these models in other domains, and it's reassuring that their effectiveness extends to the domain of scientific text ranking. Their increased effectiveness is already making important potential contributions to the findability of scientific literature, and hence the UNESCO SGDs, at least for expert searchers who have sufficient expertise and language proficiency levels.

5. Complexity-Aware Search

In this section, we explore our third research question: *How effective are complexity-aware retrieval models?* We are interested in making the IR approach aware of the complexity of the text, with the intent to retrieve relevant and accessible texts to our layperson user. We first analyze the distribution of complexity in the retrieved set of abstracts. We then propose straightforward approaches to combine evidence for relevance and readability into the ranking and evaluate these approaches in terms of retrieval effectiveness and in terms of the resulting text complexity. Can we trade-off between these two requirements in ways more suitable for laypersons searching scientific text?

5.1. Analysis of Complexity

What subset of abstracts is selected by a general query based on the popular science newspaper articles? We use the default ElasticSearch engine, 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 aver-

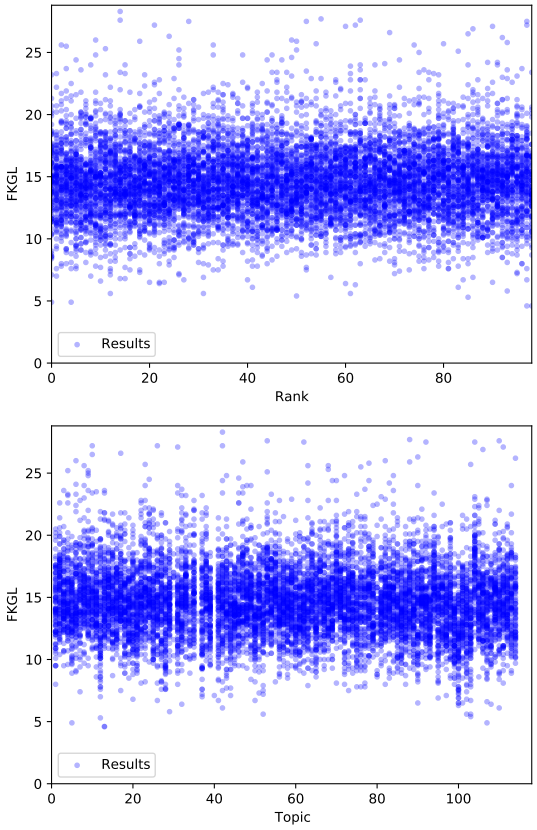


Figure 2: Distribution of text complexity: Top 100 results BM25 over 114 queries by rank (top) and topic (bottom).

age (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 the distribution of FKGL (not shown) is very similar.

Figure 2 shows the distribution of FKGL readability levels over the rank of retrieval (top) and over each individual query (bottom). In both cases, we see that the standard retrieval engine is completely blind to the text complexity and exclusively focuses 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).

5.2. Complexity-Aware Retrieval Models

Based on the observations above, we explore the viability of complexity-aware retrieval (CAR) models that combine both the relevance and text complexity of a given abstract.

Complexity-Aware Retrieval Filter Our first approach is based on a straightforward global filter, that will only allow the retrieval of abstracts with a favorable readability level. In reality, we use a fudge factor to ensure all selected abstracts receive a higher relevance score than those filtered out.[8] In pseudo-code for FILTER:

if (fkgl <= median_fkgl)
  then combined_score = relevance_score + 10
  else combined_score = relevance_score

We use a global median FKGL of 14 to create interpretable experimental conditions where we prioritize the more accessible half of the corpus and actively demote the less accessible half.

Complexity-Aware Retrieval Combine The neural cross-encoder provides a well-behaved score distribution with a small fraction of documents per topic receiving a positive relevance score. We invert the FKGL level so that lower FKGL levels are more desirable, in a way that the median FKGL level becomes a zero score. In pseudo-code for COMBINE:

if (relevance_score > 0)
  then combined_score = relevance_score
    + (median_fkgl - fkgl)
  else combined_score = relevance_score

Unlike in the rigorous filter, here a high relevance score can still overturn a less desirable FKGL, and a very desirable FKGL can overturn a low relevance score.

We opt for simple and straightforward approaches where we are in full control of the experimental parameters and obtain clear and interpretable outcomes. For the experiments in the rest of this section, we focus on the cross-encoder reranking model.

5.3. Effectiveness and Text Complexity

How will promoting readability fare? Will this be sufficient to retrieve accessible abstracts? And at what cost in performance, as we are trading off against standard retrieval effectiveness?

5.3.1. Text Complexity

Let us first look at whether our complexity-aware retrieval approaches are indeed factoring in the text complexity of the retrieved abstracts. Table 5 shows the text complexity of the top 10 results for all of the 114 queries.

[8] This is following William S. Cooper, ACM SIGIR Salton winner in 1994, who promoted both strict mathematical rigor but also the use of simple experimental stimuli to test controllable and interpretable outcomes. We choose a boost factor of 10 based on the distributional analysis before, which ensures a cohort ranking in which our filter pushes below median FKGL abstracts to the top of the ranking while preserving the internal ranking of each cohort.

Table 5: Analysis of complexity-aware retrieval results (over all 114 queries)

Run	Queries	Top	Year		Length		FKGL	
			Avg	Med	Avg	Med	Avg	Med
CE 1k	114	10	2011.8	2014	1142.3	1047.0	14.2	14.1
CE 1k CAR combine	114	10	2011.6	2014	992.9	909.0	11.2	11.2
CE 1k CAR filter	114	10	2011.5	2014	1056.8	982.0	12.2	12.4

Table 6: Complexity-Aware Retrieval effectiveness on train(top) and test (bottom)

Run	MRR	Precision			NDCG			Bpref	MAP
		5	10	20	5	10	20		
CE 100	0.5252	0.3241	0.3034	0.2448	0.2701	0.2947	0.3472	0.4012	0.3033
CE 100 CAR combine	0.4371	0.3172	0.3069	0.2466	0.2190	0.2489	0.2795	0.3998	0.2838
CE 100 CAR filter	0.5946	0.3517	0.3138	0.2655	0.3008	0.3041	0.3241	0.3906	0.3009
CE 1k	0.4608	0.2759	0.2379	0.1701	0.2312	0.2307	0.2582	0.3335	0.2001
CE 1k CAR combine	0.3182	0.2000	0.1966	0.1655	0.1423	0.1633	0.2240	0.3211	0.1714
CE 1k CAR filter	0.4952	0.2759	0.2414	0.1563	0.2390	0.2431	0.2531	0.3249	0.1934
CE 100	0.7050	0.5118	0.4912	0.3657	0.5004	0.4782	0.4007	0.2616	0.2011
CE 100 CAR combine	0.6779	0.4529	0.3971	0.3456	0.4415	0.4016	0.3642	0.2658	0.1792
CE 100 CAR filter	0.7349	0.5294	0.4353	0.3309	0.5252	0.4511	0.3716	0.2597	0.1790
CE 1k	0.6329	0.4765	0.4735	0.3578	0.4502	0.4448	0.3816	0.2797	0.2051
CE 1k CAR combine	0.5880	0.4412	0.4147	0.3098	0.3854	0.3706	0.3250	0.2700	0.1865
CE 1k CAR filter	0.6403	0.5000	0.4765	0.2941	0.4754	0.4533	0.3334	0.2727	0.1936

We observe that our complexity-aware rankers are indeed returning more accessible scientific abstracts to our lay users. The CAR Filter approach retrieves abstracts of FKGL around 12 (mean 12.2, median 12.4) and the CAR Combine approach FKGL around 11 (mean and median 11.2). To put these text complexity levels in context, an FKGL of 11-12 corresponds to the final years of compulsory education and even lower than the journalistic text used as context for the search requests.

That is, the complexity-aware retrieval approaches are indeed effective in retrieving more accessible scientific abstracts corresponding to the reading level of the targeted lay user.

5.3.2. Retrieval Effectiveness

Let us now look at the performance in terms of retrieval effectiveness. Recall that our baselines are highly effective cross-encoder rankers exhibiting competitive zero-shot performance on many collections and domains. Our CAR approaches try to avoid retrieving complex, but potentially relevant abstracts, so we may observe a trade-off in terms of retrieval effectiveness. Table 6 shows the results. First, we observe that the CAR Combine approach leads to a loss of performance, with NDCG@10 on the train data dropping 16% to 28%. Recall this may still be a reasonable trade-off approach: CAR Combine reduces the FKGL considerably to 11 and strictly focuses on retrieving only accessible content, and still obtains an effectiveness that can exceed the BM25 model. It is reasonable to assume our lay user would prefer to see more accessible abstracts first. Second, the CAR Filter approach fares even better. We would expect some trade-off between retrieval effectiveness and text complexity, and see indeed some small drop at higher recall levels. However, we see a gain in performance on early precision. On the main measure NDCG@10 however, we even observe small gains in retrieval effectiveness up to +5% on the train data and up to +2% on the test data.

In this section, we investigated the viability of complexity-aware rankers aiming to retrieve relevant and accessible abstracts for lay users. First, in line with our analysis of the distribution of text complexity per topic, We observed that we can factor text complexity into the ranking models, and created different types of rankers that promote relevant and accessible text to the front of the ranking. Second, we expected some trade-off in effectiveness between pure-relevance rankers and complexity-aware rankers. However, our experiments demonstrate that the cost can be quite small: it can even lead to minor gains in retrieval effectiveness.

Third, more generally, perhaps most important is the potential positive effect on the user experience of these models by retrieving abstracts fitting the background and education level of our users. This, in turn, holds great promise to increase science literacy and broaden the audience of objective scientific information to the general public.

6. Discussion and Conclusions

The main aim of this paper was to investigate the viability of complexity-aware retrieval models aiming to retrieve scientific information for non-expert users. Scientific literacy is crucial for all citizens, yet traditional IR systems and specialized scholarly search engines seem to cater to expert users.

We conducted an extensive analysis of both relevance and complexity and made a number of observations. Our first research question was: *How difficult are scientific abstracts?* We found that scientific abstracts had high complexity levels on average, confirming the common assumption that scientific literature is complex, but also a remarkable spread of complexity levels. Our second research question was: *Are current retrieval models sensitive to text complexity?* We found that current lexical and neural retrieval models focus exclusively on topical relevance and retrieve scientific abstracts with a complexity similar to the overall corpus. Our third research question was: *How effective are complexity-aware retrieval models?* We found that complexity-aware retrieval models combining relevance and text complexity are effective in reducing the text complexity of retrieved results. One of the more effective strategies is a straightforward filter that demotes those abstracts with undesirable text complexity in the ranking. We expected to have to trade off the retrieval effectiveness with the accessibility of scientific abstract, however, we observed no loss of retrieval effectiveness.

More generally our experiments demonstrate the viability of building complexity-aware rankers sensitive to the background expertise and language proficiency levels of our searchers. This has the potential to greatly improve the user experience of lay users searching scientific literature. Complexity-aware retrieval is a step to make IR more inclusive and sustainable by making scientific knowledge and health-related information more accessible to a wider audience including people with a lower level of education or learning disabilities and thus reducing inequality.

Our conclusions prompt the need for further study of complexity-aware IR. In the future, we plan to investigate in-depth more advanced techniques to evaluate the complexity of texts as well as the accessibility of scientific texts from the perspective of users with different backgrounds.

7. Ethics and Limitations

Complexity-aware ranking is an important step forward to more quality education by making scientific research really open, accessible, and understandable for everyone. Difficult scientific texts are less accessible for non-native speakers (Siddharthan, 2002), young readers, people with reading disabilities (Gala et al., 2020; Chen et al., 2016), needed for reading assistance (e.g. congenitally deaf people) (Inui et al., 2003) or lower level of education. Thus, complexity-aware models could help to reduce inequality and contribute to the inclusiveness and sustainability of natural language processing and information retrieval. Complexity-aware retrieval models can help to make science results accessible for anyone, promoting equal access to education, and health-related information, and ultimately more equal employment opportunities.

The popularization of science is one of UNESCO's oldest programs (UNESCO, b). Education is at the core of UNESCO programs to reach its sustainable development goals (UNESCO, a). This paper investigates how IR can promote science literacy, making significant direct contributions to SGD 4 (quality education), and SGDs 5 and 10 (reduced inequalities), and SDG 3 (increasing well-being), see Figure 3. Moreover, through education it has an indirect impact on all the 17 sustainable development goals (SDGs) of UNESCO.

Figure 3: UNESCO Sustainable Development Goals, with particular contributions to SDG 4, as well as SDG 3, SDG 5, and SGD 10. Based on https://en.unesco.org/sustainabledevelopmentgoals.

The current paper presents a proof of concept of the viability of complexity-aware search. For this reason, we opted for technically simple and interpretable manipulation of very standard classical and modern neural retrieval rankings. This ensures that our results hold for entire classes of systems, but presents no final claims on what would constitute an optimal approach.

Similarly, we equate perceived text complexity with the very crude approximations provided by traditional readability measures. These readability measures have been widely studied and widely used in the literature, ensuring that our results can be directly compared. An additional advantage of these readability measures is that they are clearly interpretable in terms of grammatical and lexical

complexity, strengthening the general conceptual results of the paper.

However, the perceived complexity of scientific text, and the real-world barriers to accessing scientific documents, as well as the key science literacy we may need to provide to lay users, is far more complex. This would need to address missing background knowledge and vernacular, including terminological explanations aiming for the laypersons. For example, explaining a medical condition as *angina pectoris* in precise medical terms may be less helpful than its imprecise relation to heart attacks. Similarly, a technical definition of an advanced term like *differential privacy* may be less helpful than explaining that this is a soft precondition for protecting a lay user's privacy. Such lay explanations seem more general and categorical (this is a type of cancer, privacy protection, ...).

We hope and expect that our general results showing that search engines can be made sensitive to text complexity, will inspire a novel research line in NLP and IR, developing different search technology that can avoid overly complex search results, and appropriate NLP technology that can help laypersons understand the retrieved scientific information. Such future technology should empower lay users, and let them interactively explore scientific information rather than become another single gatekeeper to information. This involves attention to learning aspects, and improving their science literacy, in ways that positive reinforcement of laypersons interest and use of objective science. This can be a natural antidote against shallow information on the web and in social media, often published for their monetary or political value and not their information value or lay user's interests.

8. Acknowledgments

We want to thank in particular the colleagues and the students who participated in data construction, evaluation and reviewing. Liana Ermakova is supported in part by the MaDICS (https://www.madics.fr/ateliers/simpletext/) research group and the French National Research Agency (project *ANR-22-CE23-0019-01*). Jaap Kamps is funded in part by the Netherlands Organization for Scientific Research (NWO CI # CISC.CC.016, NWO NWA # 1518.22.105), and the University of Amsterdam (AI4FinTech program). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

9. Bibliographical References

Samy Ateia and Udo Kruschwitz. 2023. Is ChatGPT a Biomedical Expert? – Exploring the Zero-Shot Performance of Current GPT Models in Biomedical Tasks. ArXiv:2306.16108 [cs].

Ping Chen, John Rochford, David N. Kennedy, Soussan Djamasbi, Peter Fay, and Will Scott. 2016. Automatic Text Simplification for People with Intellectual Disabilities. In *Artificial Intelligence Science and Technology*, pages 725–731. WORLD SCIENTIFIC.

C. W. Cleverdon. 1962. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, College of Aeronautics, Cranfield UK.

C. W. Cleverdon. 1967. The Cranfield tests on index language devices. *Aslib*, 19:173–192.

Liana Ermakova, Patrice Bellot, Pavel Braslavski, Jaap Kamps, Josiane Mothe, Diana Nurbakova, Irina Ovchinnikova, and Eric SanJuan. 2021. Overview of simpletext 2021 - CLEF workshop on text simplification for scientific information access. In *CLEF'21: Proceedings of the Twelfth International Conference of the CLEF Association*, volume 12880 of *Lecture Notes in Computer Science*, pages 432–449. Springer.

Liana Ermakova, Sarah Bertin, Helen McCombie, and Jaap Kamps. 2023a. Overview of the CLEF 2023 SimpleText Task 3: Scientific text simplification. In *Working Notes of CLEF 2023: Conference and Labs of the Evaluation Forum*, volume 3497 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Liana Ermakova, Eric SanJuan, Stéphane Huet, Hosein Azarbonyad, Olivier Augereau, and Jaap Kamps. 2023b. Overview of the CLEF 2023 SimpleText Lab: Automatic simplification of scientific texts. In *CLEF'23: Proceedings of the Fourteenth International Conference of the CLEF Association*, volume 14163 of *Lecture Notes in Computer Science*, pages 482–506. Springer.

Liana Ermakova, Eric SanJuan, Jaap Kamps, Stéphane Huet, Irina Ovchinnikova, Diana Nurbakova, Sílvia Araújo, Radia Hannachi, Élise Mathurin, and Patrice Bellot. 2022. Overview of the CLEF 2022 SimpleText Lab: Automatic simplification of scientific texts. In *CLEF'22: Proceedings of the Thirteenth International Conference of the CLEF Association*, volume 13390 of *Lecture Notes in Computer Science*, pages 470–494. Springer.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):p221 – 233.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C Ziegler. 2020. Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In *Language Resources and Evaluation for Language Technologies (LREC)*.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

Michael Gusenbauer and Neal R. Haddaway. 2020. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, 11(2):181–217. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.1378.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proc. of the second international workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, pages 9–16, USA. ACL.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 326–336, Toronto, Canada. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

K Sparck Jones and Cornelis Joost Van Rijsbergen. 1976. Information retrieval test collections. *Journal of documentation*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. 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*, pages 6365–6378. Association for Computational Linguistics.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Dietram A Scheufele and Nicole M Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16):7662–7669.

Advaith Siddharthan. 2002. An architecture for a text simplification system.

Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility*, pages 14–22. Citeseer.

Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical simplification benchmarks for English, Portuguese, and Spanish. *Frontiers in Artificial Intelligence*, 5.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: Extraction and mining of academic social networks. In *KDD'08*, pages 990–998.

Khusbu Thakur and Vinit Kumar. 2022. Application of Text Mining Techniques on Scholarly Research Articles: Methods and Tools. *New Review of Academic Librarianship*, 28(3):279–302. Publisher: Routledge _eprint: https://doi.org/10.1080/13614533.2021.1918190.

UNESCO. 1950b. Impact of science on society.

UNESCO. 2017a. *Education for Sustainable Development Goals: learning objectives*. Unesco.

Shih-Hung Wu and Hong-Yi Huang. 2022. CYUT Team2 SimpleText Shared Task Report in CLEF-2022. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, CEUR Workshop Proceedings, Bologna, Italy. CEUR-WS.org.

Mostafa Zamanian and Pooneh Heydari. 2012. Readability of texts: State of the art. *Theory & Practice in Language Studies*, 2(1).

Guido Zuccon, Bevan Koopman, and Razia Shaik. 2023. ChatGPT Hallucinates when Attributing Answers. ArXiv:2309.09401 [cs].

10. Language Resource References

Our experiments are based on the corpus (Liana Ermakova and Eric SanJuan and Stéphane Huet and Jaap Kamps, 2023a), the lay search requests (Liana Ermakova and Eric SanJuan and Stéphane Huet and Jaap Kamps, 2023b), and the relevance judgments (Liana Ermakova and Eric SanJuan and Stéphane Huet and Jaap Kamps, 2023c,d).

Liana Ermakova and Eric SanJuan and Stéphane Huet and Jaap Kamps. 2023a. *CLEF 2023 SimpleText Corpus*. CLEF/ANR SimpleText Project. SimpleText Project, CLEF SimpleText Task 1, 1.0. PID https://simpletext-project.com/.

Liana Ermakova and Eric SanJuan and Stéphane Huet and Jaap Kamps. 2023b. *CLEF 2023 SimpleText Popular Science Queries*. CLEF/ANR SimpleText Project. SimpleText Project, CLEF SimpleText Task 1, 1.0. PID https://simpletext-project.com/.

Liana Ermakova and Eric SanJuan and Stéphane Huet and Jaap Kamps. 2023c. *CLEF 2023 SimpleText Relevance Judgments (Test)*. CLEF/ANR SimpleText Project. SimpleText Project, CLEF SimpleText Task 1, 1.0. PID https://simpletext-project.com/.

Liana Ermakova and Eric SanJuan and Stéphane Huet and Jaap Kamps. 2023d. *CLEF 2023 SimpleText Relevance Judgments (Train)*. CLEF/ANR SimpleText Project. SimpleText Project, CLEF SimpleText Task 1, 1.0. PID https://simpletext-project.com/.

Beyond Sentence-level Text Simplification
Reproducibility Study of Context-Aware Document Simplification

Jan Bakker, Jaap Kamps
University of Amsterdam
Amsterdam, The Netherlands
jan.bakker@student.uva.nl, kamps@uva.nl

Abstract

Previous research on automatic text simplification has focused on almost exclusively on sentence-level inputs. However, the simplification of full documents cannot be tackled by naively simplifying each sentence in isolation, as this approach fails to preserve the discourse structure of the document. Recent Context-Aware Document Simplification approaches explore various models whose input goes beyond the sentence-level. These model achieve state-of-the-art performance on the Newsela-auto dataset, which requires a difficult to obtain license to use. We replicate these experiments on an open-source dataset, namely Wiki-auto, and share all training details to make future reproductions easy. Our results validate the claim that models guided by a document-level plan outperform their standard counterparts. However, they do not support the claim that simplification models perform better when they have access to a local document context. We also find that planning models do not generalize well to out-of-domain settings.

Lay Summary: *We have access to unprecedented amounts of information, yet the most authoritative sources may exceed a user's language proficiency level. Text simplification technology can change the writing style while preserving the main content. Recent paragraph-level and document-level text simplification approaches outcompete traditional sentence-level approaches, and increase the understandability of complex texts.*

Keywords: Generative Text Simplification, Machine Learning for Natural Language Processing, Reproducibility Study.

1. Introduction

To date, most research on automatic text simplification has focused on sentence-level inputs. However, the simplification of full documents cannot be tackled by naively simplifying each sentence in isolation, as this approach fails to preserve the discourse structure of the document. Cripwell et al. (2023b) proposed to guide the simplification of each sentence by a document-level plan specifying how it should be simplified—should it be copied, deleted, split or rewritten? Their planning model leverages both the context of each sentence and its internal structure in order to predict a simplification operation. Although this approach was able to outperform the baseline end-to-end systems, it is still limited in that the simplification model has no direct access to the context of each sentence.

In their follow-up paper, Cripwell et al. (2023a) explored various systems that use a local document context within the simplification process itself, either by working at the paragraph level or attending over an additional input representation. In doing so, they achieved state-of-the-art performance on the Newsela-auto dataset, even when not relying on plan-guidance. Figure 1 shows a Wiki-auto example input and the output of one of the sentence-level and paragraph-level text simplification models.

In this paper, we aim to replicate their experiments on another dataset, namely Wiki-auto, in order to assess the generalizability of their findings. Furthermore, we investigate the ability of the models trained on Newsela-auto to adapt to new domains by evaluating them on Wiki-auto. The rest of this paper is structured in the following way. Section 2 discusses the exact scope of our reproducibilty study. Section 3 details the experimental data, models, and setup. Section 4 presents the planning and simplication results on Wiki-auto, both under within-domain and out-of-domain conditions. We end the paper with discussion and conclusions in Section 5. An appendix provides additional evaluation measures and further examples of output of the various models.

2. Scope of Reproducibility

This section discusses the exact scope of our reproducibilty study.

We identify two main claims made by Cripwell et al. (2023a) about document-level simplification which we aim to verify:

1. Considering all metrics, text-only models that take as input either a sentence ($\text{BART}_{\text{sent}}$) or a whole document ($\text{BART}_{\text{doc}}$, LED_{doc}) underperform compared to models that have access to a local document context ($\text{BART}_{\text{para}}$, LED_{para}, ConBART).

2. Plan-guided models outperform their standard counterpart on all metrics.

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 27–38

Complex document

Silvano "Nano" Campeggi (] ; January 23, 1923 – August 29, 2018) was an Italian artist who designed and produced the artwork for the posters of many classic Hollywood films. His iconic images are associated with the golden era of Hollywood and Campeggi is now generally regarded as the most important graphic artist and poster designer in the history of American cinema.

In the following decades, Campeggi designed and produced the poster and advertising graphics for over 3000 films, working not only under contract with the MGM studios, but also with Warner Brothers, Paramount, Universal, Columbia Pictures, United Artists, RKO, Twentieth-Century Fox and several other movie studios. Sixty- four of the films he illustrated won Oscars, including "Casablanca", "Ben-Hur", "Singin' in the Rain", "An American in Paris", "West Side Story", "Exodus", "Breakfast at Tiffany's", and "Gigi".

Campeggi died on 29 August 2018, at the age of 95.

Simplification plan-guided sentence-level BART model

Silvano "Nano" Campeggi (]; January 23, 1923 – August 29, 2018) was an Italian artist. He designed and produced the artwork for the posters of many classic Hollywood movies. His iconic images are associated with the golden era of Hollywood and Campeggi is generally regarded as the most important graphic artist and poster designer in the history of American cinema.

Sixty-four of the films he illustrated won Oscars, including "Casablanca", "Ben-Hur", "Singin' in the Rain", "An American in Paris", "West Side Story", "Exodus", "Breakfast at Tiffany's", and "Gigi".

Campeggi died on 29 August 2018 in Milan at the age of 95.

Simplification plan-guided paragraph-level BART model

Silvano "Nano" Campeggi (]; January 23, 1923 – August 29, 2018) was an Italian artist. He designed and produced the artwork for the posters of many classic Hollywood movies. His iconic images are associated with the golden era of Hollywood.

Campeggi illustrated over 3000 movies, including "Casablanca", "Ben-Hur", "Singin' in the Rain", "An American in Paris", "West Side Story", "Exodus", "Breakfast at Tiffany's", and "Gigi".

Campeggi died on 29 August 2018, at the age of 95.

Figure 1: Wiki-auto example of plan-guided text simplification at the sentence- and paragraph-level.

These claims are made in the Results and Discussion section of the original paper, under the subsections *Context Awareness Matters* and *The Utility of Planning*.

While the authors of the original paper only performed their simplification experiments on Newsela-auto, we replicate their experiments on Wiki-auto.[1] In a sense, our paper adds a missing table to Cripwell et al. (2023a), as the earlier Cripwell et al. (2023b) evaluated their planning models on both datasets. They found the accuracy on Newsela-auto to be significantly higher, which they attributed to Wiki-auto being an inferior simplification corpus. Indeed, the pairs of complex-simple documents in Wiki-auto were automatically collated and aligned, while the Newsela dataset contains news articles that were manually rewritten at different levels of simplification (Xu et al., 2015). However, the Newsela dataset requires a license to use, making it difficult to fully reproduce the results obtained by the original authors. Furthermore, replicating their experiments on another dataset allows us to assess whether the aforementioned claims generalize to new domains. Lastly, by evaluating their pretrained models on Wiki-auto, we are able to gain insight into the out-of-domain performance of these models.

[1] Replication according to the ACM definition: different team, different experimental setup. Also replication according to the NeurIPS definition: same code and analysis, but different data.

3. Methodology

This section details our methodology: first, the experimental data; second, the experimenal models; third, the experimental setup; and fourth, the computational requirements of our experiments.

The authors of the original paper made their code, open-source datasets and several pretrained models available on GitHub.[2] Because their code is of high quality, running it allows us to use the exact same model architectures, training and evaluation scripts for our replication study. We describe the data, models and our experimental setup in the following subsections.

3.1. Data

WikiLarge (Zhang and Lapata, 2017) is a dataset of complex-simple document pairs that were automatically collated from English Wikipedia and Simple English Wikipedia. Wiki-auto (Jiang et al., 2020) was derived from WikiLarge by aligning the simple output document with the complex input document at both the sentence and paragraph level. For all experiments, we utilize the preprocessed version of Wiki-auto from Cripwell et al. (2023b). In this version, each complex document consists of only the aligned paragraphs, and each simple document consists of only the aligned sentences within the aligned paragraphs. Moreover, each complex sentence is annotated with a simplification operation - delete, copy, rewrite or split - based on the simple sentences to which it is aligned. For example, if a complex sentence is aligned to multiple simple sentences, it is assigned the split operation. Documents with lots of deletion are removed from dataset; we refer to the original paper for more details on the preprocessing procedure.

Since the authors made their Wiki-auto datasets publicly available, we did not have to preprocess the data ourselves. However, as these datasets were only used for training and evaluating the planning models, they do not contain information on which sentences belong to the same paragraph. Meanwhile, fine-tuning certain simplification models also requires paragraph pairs. Therefore, we constructed a preprocessed paragraph-level

[2] https://github.com/liamcripwell/plan_simp

Data	Copy	Rephrase	Split	Delete
Wiki-auto	20.64	39.01	11.18	29.17
Newsela-auto	26.06	35.49	21.75	16.69

Table 1: Operation class distributions of Wiki-auto and Newsela-auto in percentages.

dataset by combining the information from the original Wiki-auto data with the datasets shared by the authors.

To illustrate the difference between the preprocessed Wiki-auto and Newsela-auto datasets, we highlight some characteristics also reported by the original authors. First, the number of document pairs is significantly higher for Wiki-auto (85,123) than for Newsela-auto (18,319). Second, the average number of sentences per complex document is much smaller for Wiki-auto (5.4) than for Newsela-auto (38.6). Third, percentage-wise, the Wiki-auto dataset contains more rephrase and delete operations, and less copy and split operations than the Newsela-auto dataset. The exact percentages are shown in Table 1.

3.2. Planning models

Cripwell et al. (2023b) experimented with several planning models, whose task is to predict a simplification operation - delete, copy, rewrite or split - for each sentence in a complex document. For example, their RoBERTa-based *classifier* simply takes a tokenized sentence as input and outputs a prediction score for each operation class. Their *contextual classifier* additionally attends over a high-level representation of the document context. This is a sequence of vector encodings for the sentences surrounding the input sentence, combined with custom positional embeddings indicating their relative distance to it.

On both Wiki-auto and Newsela-auto, the contextual classifier achieved the highest accuracy. Specifically, the best-performing variants used dynamic context, weight initialization and a context window radius of 13 sentences. During inference, using dynamic context means that the left context consists of previously simplified sentences, rather than complex ones. During training, the ground truth simplifications are used. Weight initialization means that the RoBERTa layers of the contextual classifier are initialised with weights from the context-independent classifier. For Newsela-auto, the most accurate variant also included document positional embeddings into the context, indicating the document quintile (1-5) that a given sentence falls into. This variant was used for plan-guidance by Cripwell et al. (2023a). Similarly, in this work, we fine-tune both planners - with and without document positional embeddings - on Wiki-auto, and utilize the variant with the highest accuracy to guide our simplification models.

3.3. Simplification models

We train all document simplification models from the original paper on Wiki-auto. That is, we fine-tune them on pairs of complex inputs and simple outputs. The original authors distinguished three model categories, each of which we briefly describe here.

3.3.1. Text-only

Text-only models take only a text sequence as input. They are trained by fine-tuning BART and a Longformer encoder-decoder to perform simplification on documents (BART_{doc}, LED_{doc}), paragraphs ($\text{BART}_{\text{para}}$, LED_{para}), and sentences ($\text{BART}_{\text{sent}}$). The sentence- and paragraph-level models are iteratively applied over a document in order to simplify it.

3.3.2. Context-aware

ConBART is a modification of the BART architecture, that takes both a sentence and a high-level representation of its document context as input. This context representation is constructed using the same strategy as for the planning models, with a context window radius of 13 sentences and a dynamic context mechanism. ConBART is iteratively applied over the sentences in a document in order to simplify it.

3.3.3. Plan-Guided

Each of the proposed models can be modified to take a simplification operation as control-token at the beginning of each text input. During training, the ground-truth operations are used as control-tokens. At inference time, the operations are generated by a planning model. The resulting systems are referred to as $\hat{O} \rightarrow h$, where h is the simplification model. If the ground-truth operations are used during inference, the resulting systems are referred to as $O \rightarrow h$. Furthermore, to align with the original paper, we rename $\hat{O} \rightarrow \text{BART}_{\text{sent}}$ to PG_{Dyn} and $O \rightarrow \text{BART}_{\text{sent}}$ to $\text{PG}_{\text{Oracle}}$.

3.4. Experimental setup

We use the code provided by the original authors for our experiments. It is complete, readable and runs without errors. Furthermore, it is well-documented, including instructions on how to leverage the pretrained models. The exact arguments used to train each planning and simplification model are not documented. Still, we are largely able to recover them

from careful inspection of the code and the training details outlined in the original paper. We use these arguments to train our models on Wiki-auto, and share them on GitHub[3] to make reproduction easy. We also provide our code for constructing the preprocessed paragraph-level dataset.

3.4.1. Training details

Despite being able to recover most arguments, we have to make a few assumptions about the training procedure. First of all, the authors mention training their simplification models until convergence, without defining convergence. We implement early stopping and train until the first epoch at which the validation loss does not improve. Then we select the model checkpoint from the epoch before. The authors also do not specify when to stop training the planning models. We decide to train them for 10 epochs, and select the checkpoint with the lowest validation macro F1-score. Moreover, there are some inconsistencies between the training details reported by Cripwell et al. (2023b) and Cripwell et al. (2023a). Both papers report different learning rates for their simplification models, and whereas the first paper mentions enforcing a minimum output length for BART_{doc}, the second does not. However, both papers report the same results for those models that they have in common. We use the training details specified in the second paper, since this is the one that we aim to replicate.

3.4.2. Inference

Following the original authors, we perform inference using beam search with a beam size of 5 and a maximum length of 1024 tokens. Furthermore, for our out-of-domain experiments, we utilize all models that were pretrained on Newsela-auto and made available by the authors. These include one planning model, which is the best variant of the contextual classifier, and four simplification models, namely LED_{para} and the plan-guided modifications of $\text{BART}_{\text{sent}}$, ConBART and LED_{para}. Because Wiki-auto does not have multiple simplification levels, we manually specify a target reading level of 3 (the second simplest) for our experiments.

3.4.3. Evaluation metrics

We evaluate each model using the same evaluation scripts and metrics as the original authors. Thus, we evaluate the planning models using the F1-score for each operation class, as well as the micro and macro averages. To evaluate the simplification models, we leverage BARTScore (Yuan et al., 2021) and SMART (Amplayo et al., 2022) as analogs for meaning preservation and fluency. Furthermore, we assess readability using the Flesch-Kincaid grade level (FKGL) (Kincaid et al., 1975), and simplicity using SARI (Xu et al., 2016).

[3] https://github.com/JanB100/doc_simp

Planning model	Training time
Classifier	62
Dyn. context	97
+ docpos	102

Table 2: Training time per planning model in minutes. **Dyn. Context** is the contextual classifier with r = 13, dynamic context and weights initialised using the classifier weights.

Simplification model	Training time
BART_{doc}	72
$\text{BART}_{\text{sent}}$	111
$\text{BART}_{\text{para}}$	54
LED_{doc}	146
LED_{para}	136
ConBART	109

Table 3: Training time per simplification model in minutes.

3.5. Computational requirements

We run all training and inference processes on two NVIDIA A100 GPUs with 40 GB memory. In line with the original paper, we use a batch size of 32 to train the planning models on Wiki-auto. The time needed to train each planning model for 10 epochs on 2 GPUs is shown in Table 2. Note that because of weight initalization, one can only train the contextual classifier after the context-independent classifier has been trained.

The original authors used a batch size of 16 to train their simplification models on Newsela-auto. However, using the same batch size to train on Wiki-auto results in memory issues. Therefore, we leverage a batch size of 8 and accumulate the gradients over 2 batches. The time needed to train each simplification model without plan-guidance on 2 GPUs is shown in Table 3. The training times with plan-guidance are approximately equal. We refer to the original paper for statistics on inference times and parameter counts.

4. Results and Discussion

This section presents in results of our experiments on Wiki-auto. First, the planning results. Second, the text simplication results. Third, the effectiveness under out-of-domain conditions.

Model	Copy	Rephrase	Split	Delete	Micro	**Macro**
Classifier	40.0 (42.1)	53.0 (52.9)	42.3 (42.6)	48.9 (49.0)	48.2 (48.4)	46.0 (46.7)
Dyn. context	**45.7** (44.8)	56.0 (57.9)	**42.9** (42.4)	**57.1** (54.8)	**52.8** (52.8)	**50.5** (50.0)
+ docpos	44.2 (43.7)	**58.6** (55.4)	39.8 (43.6)	52.1 (56.7)	52.4 (52.3)	48.7 (49.9)

Table 4: Reproduced (and original) Planning Accuracy (class and average F1-scores) on Wiki-auto. **Dyn. Context** is the contextual classifier with r = 13, dynamic context and weights initialised using the classifier weights.

System	**BARTScore ↑**			**SMART ↑**			**FKGL ↓**	**SARI ↑**	**Length**	
	P ($r \to h$)	R ($h \to r$)	F1	P	R	F1			Tok.	Sent.
Input	-2.48	-1.65	-2.06	55.9	64.1	59.3	9.64	16.7	155.3	5.5
Reference	-0.61	-0.61	-0.61	100	100	100	6.59	97.2	97.1	4.5
BART_{doc}	-2.04	-2.09	-2.07	**62.9**	53.9	57.2	9.66	45.2	96.6	2.3
$\text{BART}_{\text{sent}}$	-2.11	-1.91	-2.01	58.1	62.8	59.7	6.95	43.1	111.5	5.2
$\text{BART}_{\text{para}}$	-2.01	-1.90	-1.96	62.0	62.6	61.6	7.69	43.7	107.6	4.5
LED_{doc}	-2.21	-1.61	**-1.91**	60.7	**68.3**	**63.7**	8.42	34.3	145.7	5.5
LED_{para}	-2.26	**-1.60**	-1.93	60.1	68.0	63.3	8.73	31.1	151.0	5.6
ConBART	-2.19	-1.81	-2.00	58.5	64.9	60.9	7.54	39.4	128.6	5.4
PG_{Dyn}	**-1.85**	-2.05	-1.95	61.3	59.9	59.9	6.46	**48.6**	90.2	4.4
$\hat{\text{O}} \to$ ConBART	-1.86	-2.03	-1.95	61.5	60.1	60.1	6.54	48.4	92.5	4.4
$\hat{\text{O}} \to \text{BART}_{\text{para}}$	-1.86	-2.04	-1.95	60.7	59.8	59.6	6.40	48.4	93.3	4.5
$\hat{\text{O}} \to \text{LED}_{\text{para}}$	-1.87	-1.94	**-1.91**	62.5	61.7	61.4	7.11	47.2	102.6	4.5
$\text{PG}_{\text{Oracle}}$	**-1.57**	-1.72	-1.65	67.5	67.7	67.5	6.39	**56.4**	89.6	4.5
$\text{O} \to$ ConBART	-1.59	-1.70	-1.65	67.7	67.8	67.7	6.48	56.1	91.9	4.5
$\text{O} \to \text{BART}_{\text{para}}$	-1.58	-1.73	-1.66	67.0	67.1	67.0	6.28	56.1	91.1	4.5
$\text{O} \to \text{LED}_{\text{para}}$	-1.62	**-1.63**	**-1.62**	**69.0**	**69.1**	**69.0**	7.04	55.0	100.9	4.5

Table 5: **Results of document simplification systems on Wiki-auto.** For BARTScore, *h* is the hypothesis and *r* is the reference.

4.1. Planning results

Table 4 summarizes the results of training and evaluating our planning models on Wiki-auto. The planning accuracies of our models are close to those originally reported in Cripwell et al. (2023b, Table 2), indicating a succesful reproduction. In particular, the improvement of the contextual classifiers over the context-free classifier is the biggest for the delete operation, and the smallest for the split operation. This confirms the intuition of the original authors that deletion is mostly context dependent, while splitting is mostly context independent. However, all F1-scores are relatively low. As indicated by the authors, this is likely a result of Wiki-auto being an inferior simplification corpus. In line with the original results, we find the macro F1-score of the contextual classifier on Wiki-auto to be optimal when not using document positional embeddings. We hypothesize that the small document lengths (as shown in Section 3.1) make these embeddings redundant, and utilize the contextual classifier without document positional embeddings for our plan-guided simplification systems.

4.2. Simplification results

Table 5 shows the results of training and evaluating our document simplification systems on Wiki-auto. It corresponds to the Newsela-auto results in Cripwell et al. (2023a, Table 3). We leverage these results to assess the main claims made by the original authors:

1. Considering all metrics, text-only models that take as input either a sentence ($\text{BART}_{\text{sent}}$) or a whole document ($\text{BART}_{\text{doc}}$, LED_{doc}) underperform compared to models that have access to a local document context ($\text{BART}_{\text{para}}$, LED_{para}, ConBART).

2. Plan-guided models outperform their standard counterpart on all metrics.

The first claim is concerned with all models that are not guided by a simplification plan. Considering only those models, we find that $\text{BART}_{\text{sent}}$

Complex document

Silvano "Nano" Campeggi (] ; January 23, 1923 – August 29, 2018) was an Italian artist who designed and produced the artwork for the posters of many classic Hollywood films. His iconic images are associated with the golden era of Hollywood and Campeggi is now generally regarded as the most important graphic artist and poster designer in the history of American cinema.

In the following decades, Campeggi designed and produced the poster and advertising graphics for over 3000 films, working not only under contract with the MGM studios, but also with Warner Brothers, Paramount, Universal, Columbia Pictures, United Artists, RKO, Twentieth-Century Fox and several other movie studios. Sixty- four of the films he illustrated won Oscars, including "Casablanca", "Ben-Hur", "Singin' in the Rain", "An American in Paris", "West Side Story", "Exodus", "Breakfast at Tiffany's", and "Gigi".

Campeggi died on 29 August 2018, at the age of 95.

Simple document

Silvano "Nano" Campeggi (]; January 23, 1923 – August 29, 2018) was an Italian artist. He designed and produced the artwork for the posters of many classic Hollywood movies.

Campeggi was known for his poster designs for "Casablanca", "Singin' in the Rain", and "Breakfast at Tiffany's".

Campeggi died on August 29, 2018 in Florence from respiratory failure at the age of 95.

Simplification by pretrained $\hat{O} \rightarrow$ LED$_{para}$

Silvano "Nano" Campeggi (; born January 23, 1923 – August 29, 2018) was an Italian artist. He designed and produced the artwork for the posters of many Hollywood movies. Campeggi is now known as the most important picture artist and poster designer in the history of American movies.

Over the following many years, Campeggi designed and produced the poster and advertising pictures for many movies. Sixty-four of the films he illustrated won awards. They included "Casablanca," "Ben-Hur", "Singin' in the Rain," "An American in Paris" and "Gigi."

Campeggi died on 29 August 2018, when he was 95.

Simplification by BART$_{para}$

Silvano ""Nano"" Campeggi (]; January 23, 1923 – August 29, 2018) was an Italian artist. He designed and produced the artwork for the posters of many classic Hollywood movies.

Campeggi designed and produced the poster and advertising graphics for over 3000 movies, working not only under contract with the MGM studios, but also with Warner Brothers, Paramount, Universal, Columbia Pictures, United Artists, RKO, Twentieth Century Fox and several other movie studios. Sixty-four of the movies he illustrated won Oscars, including "Casablanca", "Ben-Hur", "Singin' in the Rain", "An American in Paris", "West Side Story", "Exodus", "Breakfast at Tiffany's", and "Gigi".

Campeggi died on 29 August 2018, at the age of 95

Simplification by PG$_{Dyn}$

Silvano "Nano" Campeggi (]; January 23, 1923 – August 29, 2018) was an Italian artist. He designed and produced the artwork for the posters of many classic Hollywood movies. His iconic images are associated with the golden era of Hollywood and Campeggi is generally regarded as the most important graphic artist and poster designer in the history of American cinema.

Sixty-four of the films he illustrated won Oscars, including "Casablanca", "Ben-Hur", "Singin' in the Rain", "An American in Paris", "West Side Story", "Exodus", "Breakfast at Tiffany's", and "Gigi".

Campeggi died on 29 August 2018 in Milan at the age of 95.

Simplification by $\hat{O} \rightarrow$ BART$_{para}$

Silvano "Nano" Campeggi (]; January 23, 1923 – August 29, 2018) was an Italian artist. He designed and produced the artwork for the posters of many classic Hollywood movies. His iconic images are associated with the golden era of Hollywood.

Campeggi illustrated over 3000 movies, including "Casablanca", "Ben-Hur", "Singin' in the Rain", "An American in Paris", "West Side Story", "Exodus", "Breakfast at Tiffany's", and "Gigi".

Campeggi died on 29 August 2018, at the age of 95.

Figure 2: A complex-simple document pair from Wiki-auto, along with the corresponding outputs of three document simplification systems trained on Wiki-auto and one system pretrained on Newsela-auto.

and BART$_{para}$ perform best overall. While LED$_{doc}$ achieves the highest BARTScore and SMART F1-scores, its outputs are much longer than the references. Furthermore, whereas BART$_{doc}$ obtains the highest SARI scores, its outputs are not more readable than the inputs according to FKGL. This is largely a result of the sentences being relatively long, which SARI does not account for since it is a token-based metric. Thus, BART$_{sent}$ and BART$_{para}$ perform best overall and therefore the claim does not hold; BART$_{sent}$ even outperforms its contextual modification (ConBART) in terms of SARI. This suggests that having access to a local document context is more advantegeous for models performing simplification on Newsela-auto than for models performing simplification on Wiki-auto.

Regarding the second claim, we find that plan-guided models significantly outperform their standard counterparts in terms of SARI and FKGL. Although this is not necessarily true for SMART and BARTScore, the differences in F1-scores are small. Thus, we find that the claim largely holds. The underlying intuition is that document simplification is a highly complex task, and therefore decomposing it into two easier tasks, namely planning and generation, makes the full task simpler. Our results demonstrate that this is true even when the accuracy on the planning subtask is relatively low, and that using an oracle plan further increases performance across every metric.

Furthermore, we observe that the outputs of the text-only LED models are approximately as long as the inputs, and therewith much longer than the references and the outputs of all other models. We also find that this problem can be overcome by using a planning model in combination with the simplification model. However, our $\hat{O} \rightarrow$ LED$_{para}$ system does not outperform $\hat{O} \rightarrow$ BART$_{para}$, as was the case in the original paper. This is because the Longformer architecture was designed to process long text sequences, and the input paragraphs and documents in Newsela-auto are substantially longer than those in Wiki-auto.

In any case, it is important to realize that automatic evaluation metrics have their limitations. Specifically, when considering all metrics, we found that sentence-level models do not underperform compared to models that have access to a local document context (Claim 1). Nevertheless, it is conceivable that the latter class of models performs

Model	Copy	Rephrase	Split	Delete	Micro	**Macro**
Dyn. context + docpos	21.3	45.6	25.1	23.8	33.5	29.0

Table 6: Planning Accuracy (class and average F1-scores) on Wiki-auto for a model trained on Newsela-auto.

System	**BARTScore ↑**			**SMART ↑**			**FKGL ↓**	**SARI ↑**	**Length**	
	P ($r \to h$)	R ($h \to r$)	F1	P	R	F1			Tok.	Sent.
Input	-2.48	-1.65	-2.06	55.9	64.1	59.3	9.64	16.7	155.3	5.5
Reference	-0.61	-0.61	-0.61	100	100	100	6.59	97.2	97.1	4.5
LED_{para}	-2.68	-2.60	-2.64	39.4	45.5	41.4	4.55	35.5	91.4	5.7
PG_{Dyn}	-2.84	-2.81	-2.82	38.4	43.6	40.1	4.69	35.6	96.3	6.2
$\hat{\text{O}} \to$ ConBART	-2.89	-2.86	-2.88	37.7	42.6	39.3	4.55	35.6	93.6	6.2
$\hat{\text{O}} \to \text{LED}_{\text{para}}$	**-2.52**	**-2.52**	**-2.52**	**41.8**	**47.3**	**43.7**	4.87	**36.6**	98.4	5.9
$\text{PG}_{\text{Oracle}}$	-2.21	-2.47	-2.34	50.8	51.5	**51.1**	5.47	**44.9**	79.7	4.5
$\text{O} \to$ ConBART	-2.27	-2.52	-2.40	49.7	50.3	49.9	5.29	44.6	77.8	4.6
$\text{O} \to \text{LED}_{\text{para}}$	**-2.03**	**-2.30**	**-2.17**	**50.9**	**51.9**	**51.1**	5.32	43.7	82.2	4.7

Table 7: **Results on Wiki-auto for document simplification systems trained on Newsela-auto.** For BARTScore, *h* is the hypothesis and *r* is the reference.

better according to human judgements, because intuitively they should be better able to preserve the discourse structure of the document.

Figure 2 shows an example of a complex document from Wiki-auto, along with the simple document to which it is aligned and the corresponding outputs of four simplification systems. First of all, note that the simple document is no direct simplification of the complex document, as the last paragraph contains additional information. This is a result of the complex-simple document pairs in Wiki-auto being automatically collated. Second, note that the last sentence of the simplification created by PG_{Dyn} contains a factual error. This demonstrates that these systems are prone to hallucination, and therefore they should only be used in practice when their outputs are checked by humans. Most importantly, the right part of Figure 2 illustrates the effects of plan-guidance and access to a local document context onto the output. For example, we observe that $\text{BART}_{\text{para}}$ and $\hat{\text{O}} \to \text{BART}_{\text{para}}$ leave out different sentences, which shows that leveraging a document-level plan can make a difference even when the simplification model already operates at the paragraph-level. Conversely, we also observe that $\hat{\text{O}} \to \text{BART}_{\text{para}}$ merges multiple sentences in the second paragraph, while PG_{Dyn} is unable to do so. This reveals the ability of the simplification model to take advantage of operating at the paragraph-level, even when it is guided by a document-level plan.

4.3. Out-of-domain results

Table 6 shows the accuracy of the planning model, which was pretrained on Newsela-auto, when it is evaluated on Wiki-auto. The macro F1-score is close to that of a random classifier (25.0), indicating a poor out-of-domain performance. In particular, what the planner has learned about when to copy, split or delete a sentence does not at all generalize to Wiki-auto. Only for the rephrase operation does the acquired knowledge partially generalize, and 39.01% of the sentences in Wiki-auto fall into this class (Table 1). However, the class F1-score is still significantly lower than that of the same model trained on in-domain data (Table 4).

Table 7 displays the results of the full document simplification systems, which were pretrained on Newsela-auto, when they are evaluated on Wiki-auto. In terms of SARI, we find that LED_{para} performs better than its standard counterpart trained on in-domain data (Table 5). We interpret this as a certain capacity of generalization. Furthermore, we notice that the plan-guided models do not obtain significantly better results than LED_{para}. This is unsurprising given the poor out-of-domain performance of the planning model. However, we also find that leveraging the planner does not harm performance. Using oracle plans significantly increases performance, which demonstrates that plan-guidance can still be helpful when using simplfication models in an out-of-domain setting.

Compared to the simplification models trained on Wiki-auto, the models trained on Newsela-auto achieve significantly lower FKGL scores, indicating

that their outputs are easier to read. BARTScore, SMART and SARI compare these outputs to the references. As the references come from Wiki-auto, it is rather predictable that the best models trained on Wiki-auto achieve significantly better scores than the models trained on Newsela-auto. Even so, these results demonstrate that the models trained on Newsela-auto and Wiki-auto perform different types of transformations.

The difference between the in-domain and out-of-domain results can best be illustrated using an example. The lower left part of Figure 2 shows the output of the $\hat{\text{O}} \rightarrow \text{LED}_{\text{para}}$ system pretrained on Newsela-auto, given an input from Wiki-auto. In contrast to the other systems, $\hat{\text{O}} \rightarrow \text{LED}_{\text{para}}$ simplifies "graphic" to "picture", and "at the age of" to "when he was". Similar observations can be made upon inspection of more examples. This is because the system was essentially pretrained to rewrite news articles to a lower grade level, and this is not the same as rewriting English Wikipedia articles to Simple English Wikipedia articles. Yet, despite being less similar to the references, the outputs of the pretrained systems on Wiki-auto are in general fluent and easy to understand.

5. Conclusion

This section summarizes the main conclusions from our replication study of the paper Context-Aware Document Simplification (Cripwell et al., 2023a). The original paper evaluates a variety of document simplification systems on the Newsela-auto dataset, which requires a license to use. We leverage the code of the original authors to replicate their experiments on an open-source dataset, namely Wiki-auto, and share the exact arguments that we use to make reproduction easy. The accuracies of our planning models are close to those originally reported by the authors. Furthermore, we verify the claim that models guided by a document-level plan outperform their standard counterparts. We cannot verify the claim that models with access to a local document context perform better than those operating at the sentence- or document-level. Lastly, we evaluate the pretrained models shared by the original authors on Wiki-auto, and find that the planning model does not generalize well, while the simplification models partially generalize.

6. Ethics and Limitations

The motivation of this paper is the unavailability of the Newsela dataset used in (Cripwell et al., 2023a). The used Wiki-auto data (Zhang and Lapata, 2017) is freely available, hence offers an easy starting point for investigating document-level text simplification models and approaches. However, the alignment is of less quality than the unavailable Newsela data, and there is a need for a new open-access data set based on direct document-level text simplifications.

Our experiments are restricted to English and Encyclopedic data and we welcome research on text simplification in other languages and document genres.

7. Acknowledgements

Experiments in this paper were carried out on the National Supercomputer Snellius, supported by SURF and the HPC Board of the University of Amsterdam. Jan Bakker is partly supported by a conference grant of the master AI program at the University of Amsterdam. Jaap Kamps is funded in part by the Netherlands Organization for Scientific Research (NWO CI # CISC.CC.016, NWO NWA # 1518.22.105), and the University of Amsterdam (AI4FinTech program). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

8. Bibliographical References

Reinald Kim Amplayo, Peter J. Liu, Yao Zhao, and Shashi Narayan. 2022. Smart: Sentences as basic units for text evaluation.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023a. Context-aware document simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13190–13206, Toronto, Canada. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023b. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. Association for Computational Linguistics.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

9. Language Resource References

Jan Bakker. 2024a. *Document Simplification on Wiki-Auto (Data and Code)*. University of Amsterdam. Document Simplification Project, Document Simplification, 1.0. PID https://github.com/JanB100/doc_simp.

Jan Bakker. 2024b. *Document Simplification on Wiki-Auto (Models)*. University of Amsterdam. Document Simplification Project, Document Simplification, 1.0. PID https://huggingface.co/janbakker.

A. Data, code, and trained models

We share all our data, code, and pretrained models on GitHub (Jan Bakker, 2024a, https://github.com/JanB100/doc_simp) and HuggingFace (Jan Bakker, 2024b, https://huggingface.co/janbakker) building on the earlier code-base[4] and Wiki-auto corpus.[5] As Wiki-Auto is freely available, this offers an easy starting point for any researcher wanting to explore paragraph-level or document-level text simplification.

B. Additional Evaluation Results

Table 8 shows extra evaluation results for the document simplification systems trained and evaluated on Wiki-auto (complementing Table 5).

Table 9 shows extra evaluation results on Wiki-auto for the document simplification systems trained on Newsela-auto (complementing Table 7).

C. Example Simplifications

In addition to Figure 2, Table 10 and Table 11 show the outputs of four document simplification systems on two more examples from Wiki-auto.

[4] https://github.com/liamcripwell/plan_simp
[5] https://github.com/chaojiang06/wiki-auto

System	BARTScore Faith. ($s \rightarrow h$)	BLEU ↑	ROUGE-L ↑	SARI ↑	add	keep	delete
Input	-0.60	34.4	59.3	16.7	0.0	50.2	0.0
Reference	-1.65	100	100	97.2	96.1	97.2	98.5
$BART_{doc}$	-1.05	36.8	61.2	45.2	16.6	55.8	63.2
$BART_{sent}$	-0.92	39.9	63.8	43.1	17.8	56.1	55.5
$BART_{para}$	-0.90	41.2	64.9	43.7	17.5	57.8	55.7
LED_{doc}	-0.78	**42.7**	64.5	34.3	17.1	57.1	28.6
LED_{para}	-0.74	41.6	63.7	31.1	14.8	56.0	22.6
ConBART	-0.84	39.7	63.4	39.4	16.5	55.3	46.3
PG_{Dyn}	-1.02	39.9	64.1	**48.6**	**19.2**	58.8	**67.8**
$\hat{O} \rightarrow$ ConBART	-1.02	39.9	64.7	48.4	19.0	58.8	67.5
$\hat{O} \rightarrow BART_{para}$	-0.96	41.5	64.7	47.2	19.1	**59.5**	62.9
$\hat{O} \rightarrow LED_{para}$	-0.96	41.5	**65.3**	47.2	19.1	**59.5**	62.9
PG_{Oracle}	-1.02	51.3	**73.7**	**56.4**	23.2	**68.7**	**77.2**
$O \rightarrow$ ConBART	-1.02	51.3	73.6	56.1	22.9	68.5	76.9
$O \rightarrow BART_{para}$	-1.07	50.8	73.4	56.1	**23.5**	68.4	76.4
$O \rightarrow LED_{para}$	-0.96	**52.4**	**73.7**	55.0	23.4	**68.7**	72.9

Table 8: **Extra results of document simplification systems on Wiki-auto.** For BARTScore, *s* is the source and *h* is the hypothesis.

System	BARTScore Faith. ($s \rightarrow h$)	BLEU ↑	ROUGE-L ↑	SARI ↑	add	keep	delete
Input	-0.60	34.4	59.3	16.7	0.0	50.2	0.0
Reference	-1.65	100	100	97.2	96.1	97.2	98.5
LED_{para}	-1.62	22.4	49.5	35.5	5.1	42.3	59.1
PG_{Dyn}	-1.78	20.1	48.6	35.6	4.5	40.7	61.7
$\hat{O} \rightarrow$ ConBART	-1.89	19.4	48.0	35.6	4.2	40.1	**62.6**
$\hat{O} \rightarrow LED_{para}$	-1.44	**23.5**	**52.0**	**36.6**	**5.2**	**44.4**	60.3
PG_{Oracle}	-1.65	**31.0**	60.1	**44.9**	6.5	**54.6**	73.7
$O \rightarrow$ ConBART	-1.76	29.9	59.1	44.6	6.1	53.6	**74.1**
$O \rightarrow LED_{para}$	-1.30	**31.0**	**60.4**	43.7	**6.9**	54.1	70.2

Table 9: **Extra results on Wiki-auto for document simplification systems trained on Newsela-auto.** For BARTScore, *s* is the source and *h* is the hypothesis.

System	Output
Complex	*Ralph Steven Greco (May 25, 1942 – March 31, 2019) was the Johnson and Johnson Distinguished Professor, Emeritus of Surgery at Stanford University School of Medicine. He was a leader of the resident Well Being in surgery movement and surgical training program leader.* *Greco married to Irene L. Wapnir, M.D., professor of surgery at Stanford. Together they had 3 children. He died on March 31, 2019 at the age of 76.* *SARI = 22.6, FKGL = 2.51*
Simple	*Ralph Steven Greco (May 25, 1942 – March 31, 2019) was an American surgeon and sculptor. He was the Johnson and Johnson Distinguished Professor, Emeritus of Surgery at Stanford University School of Medicine. He was well-known for his advocacy for the well-being of surgery residents.* *He was married to Irene L. Wapnir. The couple had three children. Greco died from prostate cancer on March 31, 2019 in Stanford, California.* *SARI = 100.0, FKGL = 3.81*
$\text{BART}_{\text{para}}$	*Ralph Steven Greco (May 25, 1942 – March 31, 2019) was an American surgeon. Greco was the Johnson and Johnson Distinguished Professor, Emeritus of Surgery at Stanford University School of Medicine. He was a leader of the resident Well Being in surgery movement and surgical training program leader.* *Greco died on March 31, 2019 in Stanford, California at the age of 76.* *SARI = 55.4, FKGL = 7.21*
PG_{Dyn}	*Ralph Steven Greco (May 25, 1942 – March 31, 2019) was an American surgeon. He was the Johnson and Johnson Distinguished Professor, Emeritus of Surgery at Stanford University School of Medicine. He was a leader of the resident Well Being in surgery movement and surgical training program leader.* *Greco married Irene L. Wapnir, M.D., professor of surgery at Stanford University. Together they had 3 children. He died on March 31, 2019 at the age of 76.* *SARI = 32.6, FKGL = 2.91*
$\hat{\text{O}} \rightarrow \text{BART}_{\text{para}}$	*Ralph Steven "Ralph" Greco (May 25, 1942 – March 31, 2019) was an American surgeon. He was the Johnson and Johnson Distinguished Professor, Emeritus of Surgery at Stanford University School of Medicine. Greco died in Palo Alto, California from complications of a stroke on March 31 at the age of 90. He is a leader of the resident Well Being in surgery movement and surgical training program leader.* *Greco married to Irene L. Wapnir, M.D., professor of surgery at Stanford. Together they had 3 children. Greco died on March 31, 2019 at the age of 76.* *SARI = 34.2 , FKGL = 3.11*
Pretrained	*Ralph Steven Greco was a top doctor in California. Greco worked at the Stanford University School of Medicine from 1942 to 1954. He was also a leader in the movement and training program.* *Greco married to Irene L. Wapnir, a professor of surgery at Stanford. Together they had 3 children. He died on March 31, 2019, at the age of 76.* *SARI = 35.7, FKGL = 2.82*

Table 10: A complex-simple document pair from Wiki-auto, along with the outputs of three simplification systems trained on Wiki-auto and the $\hat{\text{O}} \rightarrow \text{LED}_{\text{para}}$ system pretrained on Newsela-auto.

System	Output
Complex	*In organic chemistry and biochemistry, a side chain is a chemical group that is attached to a core part of the molecule called the "main chain" or backbone. The side chain is a hydrocarbon branching element of a molecule that is attached to a larger hydrocarbon backbone. It is one factor in determining a molecule's properties and reactivity. A side chain is also known as a pendant chain, but a pendant group (side group) has a different definition.* *The placeholder R is often used as a generic placeholder for alkyl (saturated hydrocarbon) group side chains in chemical structure diagrams. To indicate other non-carbon groups in structure diagrams, X, Y, or Z are often used.* *SARI = 14.9, FKGL = 10.40*
Simple	*A side chain in organic chemistry is a chemical group. It is attached to the molecule's "main chain" or backbone. A side chain is also known as a pendant chain.* *Alkyl (saturated hydrocarbon) group side chains are represented by the placeholder R. Other non-carbon groups can be represented by X, Y, or Z.* *SARI = 100.0, FKGL = 4.52*
$\text{BART}_{\text{para}}$	*In organic chemistry and biochemistry, a side chain is a chemical group that is attached to a core part of the molecule called the "main chain" or backbone. The side chain, also known as a pendant chain, is one factor in determining a molecule's properties and reactivity. A pendant group (side group) has a different definition.* *The placeholder R is often used as a generic placeholder for alkyl group side chains in chemical structure diagrams. To indicate other non-carbon groups in structure diagrams, X, Y, or Z are often used.* *SARI = 27.7, FKGL = 9.34*
PG_{Dyn}	*In organic chemistry and biochemistry, a side chain is a chemical group that is attached to a core part of the molecule. The side chain is a hydrocarbon branching element of a molecule that is attached to a larger hydrocarbon backbone.* *The placeholder R is often used as a generic placeholder for alkyl group side chains in chemical structure diagrams.* *SARI = 28.5, FKGL = 11.9*
$\hat{\text{O}} \rightarrow \text{BART}_{\text{para}}$	*In organic chemistry and biochemistry, a side chain is a chemical group that is attached to a core part of the molecule called the "main chain" or backbone. The "side chain" is a hydrocarbon branching element of a molecule that is connected to a larger hydrocarbon backbone. A pendant group (side group) has a different definition.* *The placeholder R is often used as a generic placeholder for alkyl group side chains in chemical structure diagrams.* *SARI = 29.1, FKGL = 10.2*
Pretrained	*In organic chemistry, a side chain is a chemical group. It is attached to a part of the molecule called the "main chain" or backbone. The side chain is a group of atoms that are attached to a larger part of the molecule. It is one of the things that makes a molecule different. A side chain is also known as a pendant chain. But a pendant group (side group) has a different definition.* *SARI = 43.9, FKGL = 4.86*

Table 11: A complex-simple document pair from Wiki-auto, along with the outputs of three simplification systems trained on Wiki-auto and the $\hat{\text{O}} \rightarrow \text{LED}_{\text{para}}$ system pretrained on Newsela-auto.

Towards Automatic Finnish Text Simplification

Anna Dmitrieva, Jörg Tiedemann
University of Helsinki
{name.surname}@helsinki.fi

Abstract

Automatic text simplification (ATS/TS) models typically require substantial parallel training data. This paper describes our work on expanding the Finnish-Easy Finnish parallel corpus and making baseline simplification models. We discuss different approaches to document and sentence alignment. After finding the optimal alignment methodologies, we increase the amount of document-aligned data 6.5 times and add a sentence-aligned version of the dataset consisting of more than twelve thousand sentence pairs. Using sentence-aligned data, we fine-tune two models for text simplification. The first is mBART, a sequence-to-sequence denoising auto-encoder proven to show good results for monolingual translation tasks. The second is the Finnish GPT model, for which we utilize instruction fine-tuning. This work is the first attempt to create simplification models for Finnish using monolingual parallel data in this language. The data has been deposited in the Finnish Language Bank (Kielipankki) and is available for non-commercial use, and the models are accessible through Huggingface.

Keywords: automatic text simplification, parallel dataset, Finnish

1. Introduction

In recent years, the number of non-English text simplification corpora has grown significantly. For example, there exist a number of simplification datasets for other European languages such as French (see, for example, Alector (Gala et al., 2020), CLEAR (Grabar and Cardon, 2018)), German (see: Klexicon (Aumiller and Gertz, 2022), Patient-friendly Clinical Notes (Trienes et al., 2022)), Italian (see: AdminIT (Miliani et al., 2022)), and others (more examples can be found in Ryan et al., 2023). It is worth noting that the past decade saw a growing movement toward media accessibility in European countries, including legal action such as implementing the Directive EU 2016/2102[1] on the accessibility of the websites and mobile applications of public sector bodies. This is one of the reasons why the interest in accessible communication studies for European languages other than English has increased.

In Finland, Easy Language is well-established in practice (Leskelä, 2021), and Easy Language content such as news, books, and websites is produced regularly. Nevertheless, the first parallel Finnish-Easy Finnish dataset (Dmitrieva et al., 2022) has been introduced only very recently (Dmitrieva and Konovalova, 2023). This dataset, however, is rather small, with only 1919 entries, and aligned only on the document level. In this work, we increase the size of this dataset by adding more aligned document pairs and producing a sentence-aligned version. Information on the base dataset can be found in Section 3, and our work on document and sentence alignment is described in Section 4. We then train different sentence simplification models to provide a baseline for automatic Finnish text simplification. Modeling is described in Section 5.

[1] http://data.europa.eu/eli/dir/2016/2102/oj

2. Related work

Using news as a data source is a popular approach to building simplification corpora for languages that have simplified news sources. We will name just a few examples. For instance, Ebling et al. (2022) describe a dataset consisting of articles from the Austria Press Agency (Austria Presse Agentur, APA). At this press agency, four to six news items covering the topics of politics, economy, culture, and sports are manually simplified into two language levels, B1 and A2, each day (Ebling et al., 2022). Rios et al. (2021) describe another parallel German simplification dataset based on news articles from the Swiss news magazine "20 Minuten" that consists of full articles paired with shortened, simplified summaries that serve as a quick "tl;dr" for the reader. Goto et al. (2015) describe a data set consisting of Japanese news sentences and their corresponding simplified Japanese news sentences sourced from a web resource called NEWS WEB EASY (Tanaka et al., 2013) offered by the NHK [Japan Broadcasting Corporation]. Finally, Newsela (Xu et al., 2015), a well-known simplification dataset with simplifications for four different grade levels, available in English and Spanish, is also news-based.

Since simplification can be viewed as a monolingual translation problem, researchers sometimes use tools intended for multilingual alignment of machine translation corpora to align monolingual simplification data. For example, Spring et al. (2023) use Vecalign (Thompson and Koehn, 2019) among other sentence aligners to analyze alignment qual-

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 39–50

ity for automatic simplification of German texts, and Stodden et al. (2023) experiment with Vecalign and Bertalign (Liu and Zhu, 2022) to develop a new parallel dataset for German simplification. Vecalign also includes a tool that can be used for document alignment (Thompson and Koehn, 2020). Most of the alignment strategies require pre-trained embeddings, which can also be utilized on their own for parallel text detection (Spring et al., 2023; Stodden et al., 2023; Aumiller and Gertz, 2022). Specialized tools for monolingual alignment, such as MASSAlign (Paetzold et al., 2017) and CATS (Customized Alignment for Text Simplification) (Štajner et al., 2018), are also used for alignment, often in conjunction with other methods (see, for instance, Ebling et al., 2022).

In this work, we use two different architectures to create simplification models. BART (Lewis et al., 2020) models, including multilingual BART/mBART (Liu et al., 2020), are widely used for automatic text simplification and have shown good results for English (Martin et al., 2022), German (Trienes et al., 2022; Stodden et al., 2023), Spanish (Alarcón et al., 2023), and other languages. GPT models are used for simplification less often but still have shown good results, for example, for English (Maddela et al., 2023) and Russian (Shatilov and Rey, 2021). We use a GPT model trained on multiple Finnish resources (Luukkonen et al., 2023).

3. Data

We use three datasets as sources for our research: the Parallel Corpus of Finnish and Easy-to-read Finnish (Dmitrieva et al., 2022), the Yle Finnish News Archive 2011-2018 (Yleisradio, 2017) which we call the "general" archive because it consists of all news that appeared on yle.fi during these years, and Yle News Archive Easy-to-read Finnish 2011-2018 (Yleisradio, 2019). All of these datasets are available in the Language Bank of Finland [Kielipankki] under the CLARIN ACA-NC license (Academic - Non-Commercial Use, Attribution, No Redistribution, Other). The first parallel dataset is based on Yle articles from 2019 to 2020, so we are using articles from earlier times to increase the amount of parallel data.

YLE news in Easy Finnish comes on air every day in the form of short (around 5 minutes) radio and TV broadcasts relaying the most important recent events. The radio broadcast then appears on YLE's website in the form of an article, where each paragraph details its own piece of news. The target audience of Easy Finnish news is very broad, with the main target groups being immigrants, older adults, and people with intellectual disabilities (Kulkki-Nieminen, 2010).

The editors at YLE choose the material to simplify for Easy Finnish news themselves. There is no time frame for how recent the "regular" Finnish article should be, but the editors mostly select articles that came out in the 24 hours before the Easy Finnish broadcast airs (Dmitrieva and Konovalova, 2023). Therefore, for document alignment, we enforce the same limitation as Dmitrieva and Konovalova (2023) did in the original dataset and only align Standard Finnish and Easy Finnish documents from the same date. Unfortunately, we could not match articles prior to September 2014. Easy Finnish articles from before this date are mixed into the general news archive without any clear identifiers. Therefore, in this paper, we are working with articles from September 2014 to December 2020. We leave the identification of earlier Easy Finnish news in the general archive for future work.

4. Dataset augmentation

In this work, we first align more Standard and Easy Finnish articles and then produce a sentence-aligned version of the entire dataset. For both tasks, we use embedding models to produce document and sentence vectors. Here is the complete list of the embeddings that we use:

1. LASER[2] (we used the laserembeddings library[3]),
2. LaBSE (Feng et al., 2022; we used the version from sentence-transformers[4]),
3. MPNet (Song et al., 2020; we used the version from sentence-transformers[5]),
4. DistilUSE (multilingual knowledge distilled version of multilingual Universal Sentence Encoder (Yang et al., 2020)), also from sentence-transformers[6].

Three of these four models are multilingual sentence-BERT networks (Reimers and Gurevych, 2019). We have selected DistilUSE as a benchmark since it has been used in the making of the original dataset (Dmitrieva and Konovalova, 2023). We also chose MPNet because it has shown

[2] https://github.com/facebookresearch/LASER

[3] https://github.com/yannvgn/laserembeddings

[4] https://huggingface.co/sentence-transformers/LaBSE

[5] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

[6] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

the best average performance among multilingual models (see https://www.sbert.net/docs/pretrained_models.html), and LaBSE because it has shown good performance on the task of aligning simplified and regular sentences (Stodden et al., 2023). The last model that we've selected is LASER (Artetxe and Schwenk, 2019), which is the default model for Vecalign. We are using the original LASER model because it has Finnish embeddings.

We use the NLTK (Bird et al., 2009) Punkt sentence tokenizer (Kiss and Strunk, 2006) for sentence segmentation in this work. Segmentation does not appear to be an issue for our data in most cases since it has been professionally proofread before publishing and then crawled from the original website as is.

4.1. Document alignment

We use two approaches to document alignment.

The first approach is, following the previous work (Dmitrieva and Konovalova, 2023), a simple comparison of document vectors made by averaging the embeddings of all sentences in a document. We use cosine similarity to find the closest vectors.

The second method that we use is a technique proposed in Thompson and Koehn (2020). First, we use the provided script[7] for obtaining document embeddings for candidate generation. This method can be used with different sentence embeddings, so we try it with all four types of embeddings mentioned above. We set the K nearest neighbors to 5 and keep all other parameters default (such as $J = 16$ and $\gamma = 20$). We also experiment with dimensionality reduction for all embeddings to see how different the results can be. Following the original paper (Thompson and Koehn, 2020), we set the new dimensionality to 128. For sentence-transformers, we use the dimensionality reduction technique proposed within the library[8]. For LASER, we use the PCA (principle component analysis) module from scikit-learn (Pedregosa et al., 2011).

Lastly, we use a simplified version of the candidate re-scoring method from Thompson and Koehn (2020) to re-score the output of the models that performed best during candidate generation. We only do this for the documents aligned with the Vecalign method and, following the original paper, use Vecalign with LASER sentence embeddings. Our formula for re-scoring is simply

$$S(E,F) = \frac{1}{len(E)} \sum_{e,f \in a(E,F)} sim(e,f) \quad (1)$$

where E and F are the source and target documents respectively, $a(E,F)$ is the alignment between these documents, and sim is the cosine similarity between sentences. Unlike in the original paper, we do not divide by the total number of alignments, because the mismatch in sizes of source and target documents is so high that it does not make sense to penalize for unaligned sentences. Instead, we divide by the number of sentences in the Easy Finnish document, because that would be the maximum possible number of alignments. We also do not take into account the probability that both documents are in the correct language because our task is monolingual.

It should be noted that we treat document and sentence alignments as exclusive. So, if document 1 aligns with document 2, no other document can align with documents 1 or 2. In all document alignment methods, we employ a simple strategy to find the best match for each document after obtaining K best candidates. For all Easy Finnish documents, we find five possible Standard Finnish matches, obtaining a matrix of distances or similarities. Then, we find the maximum (for similarities) or minimum (for distances, which is what the Vecalign method returns) value in the matrix. We lock that document pair, eliminate it from the matrix, and look for the next highest or lowest value.

4.1.1. Evaluation

We use the Parallel Corpus of Finnish and Easy-to-read Finnish (Dmitrieva et al., 2022) to compare document alignment methods. This dataset has document pairs with "positive", "neutral", and "negative" labels. The "positive" label means that the human annotator working on the dataset was positive that the Easy Finnish document is the simplified version of the source document, the "negative" label means the opposite (the documents in the pair talk about different things), and the "neutral" label was given when the annotator was not sure. There are 1257 "positive", 470 "negative", and 192 "neutral" article pairs in the dataset (Dmitrieva and Konovalova, 2023). The labels were given after automatic pre-alignment had been performed, i.e. the annotator did not look for the pairs herself. During alignment evaluation, we only compare the document pairs that are present in both the predicted sample and the annotated dataset, so the support is different in every case. When counting the "strict" scores, we consider "neutral" documents to be positive, and when counting "lax" (relaxed) scores, we consider the "neutral" documents to be negative.

[7] https://github.com/thompsonb/vecalign/blob/master/standalone_document_embedding_demo.py

[8] https://github.com/UKPLab/sentence-transformers/blob/master/examples/training/distillation/dimensionality_reduction.py

We experimented with different thresholds for cosine similarity and distance scores. In our case, the distance is the cosine distance computed within the scikit-learn's nearest neighbors algorithm and defined as 1.0 minus the cosine similarity. For both metrics, there are 9 possible thresholds from 0.1 to 0.9. We have reached the conclusion that in the majority of cases, good F1 scores can be obtained with the highest (for distance) or lowest (for similarity) possible thresholds, which also let us obtain the highest number of pairs, i.e., have the best possible recall while still having high precision. Table 1 contains the evaluation results for the document alignment algorithm from Thompson and Koehn (2020) [the second approach], and Table 2 contains the results of document comparison with just cosine similarity between averaged sentence vectors [the first approach].

It appears that LaBSE and LASER embeddings are giving the best results in all cases. That is why we decided only to try the candidate re-scoring method (Thompson and Koehn, 2020) on the results obtained with these embedding models. However, in our case, candidate re-scoring proved not to be particularly helpful. Not only did the precision decrease, but we also got comparatively low support scores, which means that the set of document pairs that this algorithm retrieved matches the document pairs in the "true" data set rather vaguely. It can be seen that just the candidate generation algorithm from Vecalign worked best in our case. Using full-size embeddings as opposed to truncated embeddings gave only a slight improvement to the performance (same as in the original paper (Thompson and Koehn, 2020)), which means that in a more data-dense setting, truncated embeddings can be used.

4.2. Sentence alignment

For sentence alignment, we wanted the aligners to adhere to as many of the following criteria as possible:

- One-to-one, one-to-many, many-to-one, many-to-many sentence alignments are all possible.
- Crossing alignments/crossing links are allowed. Between document 1 with sentences A, B, C (here and in all examples below sentences are given in the exact order) and document 2 with sentences a, b, c, d, we can have alignments such as BC -> a and A -> d.
- Sentences within an alignment are consecutive. Between document 1 with sentences A, B, C and document 2 with sentences a, b, c, d, we cannot have alignments such as AC -> bd. We also cannot have alignments such as A -> ba; only A -> ab is possible.
- Alignments are exclusive. Between document 1 with sentences A, B, C and document 2 with sentences a, b, c, d, we cannot have both alignments A -> a and B -> a; only one of them can be chosen.
- If the method uses embeddings, it should be possible to change the embedding model.

We were unable to find a method that would satisfy all the criteria, so we opted for those that came close. We also designed a simple cosine similarity-based method to use as a baseline, satisfying all the above criteria. As another baseline, we use MASSAlign with TF-IDF-based text comparison, i.e. without any embeddings.

The first method that we use is **Vecalign** for sentence alignment (Thompson and Koehn, 2019). It is based on the similarity of sentence embeddings and a dynamic programming approximation, which is fast even for long documents. Vecalign is language-agnostic because it can work with any embeddings. It does not provide crossing alignments but satisfies all other requirements.

Our second aligner is **Bertalign** (Liu and Zhu, 2022), which works in two steps. The first step finds the optimal paths for 1-to-1 alignments based on the top-k most semantically similar target sentences for each source sentence using the bidirectional encoder representations from transformer-based cross-lingual word embeddings. The second step relies on search paths found in the previous step to recover all valid alignments with more than one sentence on each side of the bilingual text (ibid.). Bertalign outperforms Vecalign on the English-Chinese bilingual alignment (Liu and Zhu, 2022) and also on German-Easy German monolingual alignment (Stodden et al., 2023). This method also does not provide crossing alignments but satisfies all other requirements.

Both Vecalign and Bertalign let the user set the maximum number of consecutive sentences that can be aligned at once (maximum overlap size). We set this number to 3 in all experiments. We chose this threshold because in the manually aligned golden test set for sentence alignment evaluation described in paragraph 4.2.1, this is the maximum number of consecutive sentences appearing in one alignment, and 3:n and n:3 alignments are seen very rarely, so we did not see a reason to go over that limit.

We employ two baselines. The first is **MASSAlign** (Paetzold et al., 2017), which does not utilize embeddings at all. It uses a vicinity-driven approach in which it first creates a similarity matrix between the paragraphs/sentences of aligned documents/paragraphs, using a standard bag-of-words TF-IDF model, then finds a starting point to begin the search for an alignment path (ibid.). MASSAlign

		Strict			Lax				
Embeddings	**Dist↓**	**p**	**r**	**f1**	**p**	**r**	**f1**	**sup-1**	**sup-2**
Truncated embeddings									
LaBSE-128	0,9	0,723	**1,000**	0,840	0,820	**1,000**	**0,901**	1439	1439
MPNet-128	0,9	0,718	**1,000**	0,836	0,814	**1,000**	0,898	1453	1453
DistilUSE-128	0,9	0,712	**1,000**	0,832	0,808	**1,000**	0,894	1473	1473
LASER-128	0,9	**0,730**	0,993	**0,841**	**0,823**	0,993	0,900	1319	1329
Full-size embeddings									
LaBSE	0,9	0,728	**1,000**	0,842	0,824	**1,000**	0,903	1424	1424
MPNet	0,9	0,717	**1,000**	0,835	0,814	**1,000**	0,897	1473	1473
DistilUSE	0,9	0,711	**1,000**	0,831	0,807	**1,000**	0,893	1504	1504
LASER	0,9	**0,729**	**1,000**	**0,843**	**0,826**	**1,000**	**0,905**	1188	1188
After candidate rescoring									
LaBSE rescored	n/a	0,701	**1,000**	0,824	**0,805**	**1,000**	**0,892**	743	743
LASER rescored	n/a	**0,706**	**1,000**	**0,828**	0,803	**1,000**	0,891	595	595

Table 1: Document alignment with Vecalign document embeddings (Thompson and Koehn, 2020). "Sup-1" is support-1, the number of pairs deemed "positive" (true pairs) under the current threshold. "Sup-2" is support-2, the number of document pairs in the predicted sample that match the document pairs in the true dataset.

		Strict			Lax				
Embeddings	**Cos. sim.↑**	**p**	**r**	**f1**	**p**	**r**	**f1**	**sup-1**	**sup-2**
LaBSE	0,68	0,717	**1,000**	0,835	**0,812**	**1,000**	**0,896**	1613	1613
MPNet	0,55	0,701	**1,000**	0,825	0,797	**1,000**	0,887	1628	1628
DistilUSE	0,47	0,689	**1,000**	0,816	0,783	**1,000**	0,878	1710	1710
LASER	0,80	**0,719**	**1,000**	**0,836**	0,810	**1,000**	0,895	1574	1575

Table 2: Document alignment by comparing averaged sentence embeddings.

	Strict			Lax		
Embeddings	**p**	**r**	**f1**	**p**	**r**	**f1**
Vecalign						
LaBSE	0,786	0,305	0,439	0,847	0,7	0,766
MPNet	0,788	0,3	0,435	**0,852**	0,704	**0,771**
DistilUSE	0,789	0,314	0,449	0,841	0,65	0,733
LASER	**0,801**	**0,426**	**0,556**	0,839	0,668	0,744
Bertalign						
LaBSE	0,745	0,179	0,289	0,813	0,596	0,688
MPNet	0,77	0,269	0,399	0,822	0,601	0,694
DistilUSE	0,738	0,166	0,271	0,802	0,561	0,66
LASER	0,694	0,081	0,145	0,749	0,408	0,528
Cos. sim. matrix						
LaBSE	0,34	0,368	0,353	0,585	**0,726**	0,648
MPNet	0,304	0,305	0,304	0,607	0,691	0,646
DistilUSE	0,301	0,336	0,318	0,514	0,632	0,567
LASER	0,311	0,269	0,288	0,601	0,614	0,608
MASSAlign						
n\a	0,57	0,238	0,335	0,774	0,318	0,451

Table 3: Sentence alignment by different methods. "P" stands for "precision", "r" for recall, and "f1" for f1-score.

does not allow crossing alignments and sometimes returns non-exclusive alignments, but it has shown competitive results on the monolingual alignment task (Stodden et al., 2023; Spring et al., 2023). We use it with default values as in the example script[9], since we found out empirically that it is possible to obtain sensible alignments with these values. As a stop-words list, we use the stop-words list for Finnish from NLTK. The other baseline that we use is a simple algorithm similar to the one described in Section 4.1 for choosing the best documents out of K best. We embed all sentences and concatenations of consecutive sentences (of length 1 <= len <= 3) and obtain a **cosine similarity matrix**. Then, we look for the greatest value in this matrix, lock that alignment, eliminate all the sentences that go into that alignment (if we align sentences AB to sentence b, we must also eliminate rows A, B, ABC, BC, ab, abc, bc, bcd), and look for the next highest value. This method satisfies all our criteria.

4.2.1. Evaluation

We use the script provided in the Vecalign repository[10] to score our alignments. In order to obtain a gold test set, we manually aligned 50 randomly chosen "positive" document pairs from the Parallel Corpus of Finnish and Easy-to-read Finnish (Dmitrieva et al., 2022). There are 1638 singular sentences in Standard Finnish documents and 291 sentences in Easy Finnish documents. Between these documents, there are 223 non-zero alignments in the golden test set, of which 160 are one-to-one, 47 are one-to-many or many-to-one, and 16 are many-to-many ("many" was never higher than 3). The results can be viewed in Table 3.

It can be seen that Vecalign with LASER embeddings outperforms all other methods. Bertalign seems to work way worse on our data than, for example, on German monolingual data (Stodden et al., 2023). We have come to the conclusion that the performance of different alignment methods depends greatly on the nature of the data since even different monolingual corpora on the same language align differently: compare, for example, the results in Spring et al. (2023) and Stodden et al. (2023) that both deal with German-Easy German alignment. However, in Spring et al. (2023), Vecalign also demonstrated good performance. Unfortunately, we were unable to obtain good results with MASSAlign or Bertalign like Stodden et al. (2023) did. However, it should be noted that while annotating the golden test set, we concluded that a big part of our data may be difficult to align even for humans. The bigger the length difference between the Easy Finnish and Standard Finnish documents was, the harder it was to find true matches between the sentences.

Vecalign provides a score for all non-zero alignments, which reflects the cost of the alignment. The smaller the number is, the better the alignment. Zero scores are given to zero alignments (when the sentence is not aligned to any other sentence). We evaluated score thresholds from 0.1 to 0.9 on the golden test set and then empirically. To us, it appears that alignments with the score <= 0.65 can be confidently chosen for further use.

	2019-20	2014-18	Total
Documents			
Pairs	1257	7004	8261
$\text{Words}_{\text{reg}}$	471565	1700469	2172034
$\text{Words}_{\text{easy}}$	69179	402274	471453
Sentences			
Pairs	2994	8950	11944
$\text{Words}_{\text{reg}}$	41056	116684	157740
$\text{Words}_{\text{easy}}$	26699	80926	107625

Table 4: Dataset statistics. "reg" stands for Standard Finnish, or regular, texts, "easy" stands for Easy Finnish. We only consider "positive" document pairs and sentence pairs with a score equal to or below 0.65.

4.3. Dataset statistics

The statistics of our new dataset can be seen in Table 4. We have increased the amount of documents 6.5 times and added a sentence-aligned version of 11944 sentence pairs. We only considered pairs with the score <= 0.65. If the score limit is lifted, the total number of non-zero pairs in the entire dataset would be 56088.

5. Modeling

In addition to increasing the amount of Finnish simplification data, we also present the first baseline models for automatic Finnish sentence simplification. As mentioned before, we worked with two different architectures:

- mBART (Liu et al., 2020): a multilingual version of BART, a denoising autoencoder for pre-training sequence-to-sequence models, particularly effective when fine-tuned for text generation (Lewis et al., 2020). We use mBART cc25, a model with 12 encoder and decoder layers trained on 25 languages' monolingual corpus[11].

[9] https://ghpaetzold.github.io/massalign_docs/examples.html

[10] https://github.com/thompsonb/vecalign/blob/master/score.py

[11] https://github.com/facebookresearch/fairseq/tree/main/examples/mbart

	Highest SARI	Epoch
mBART	37.612	10
Finnish GPT	44.63	10

Table 5: Model evaluation results for sentence simplification.

Feature	mBART	FinnGPT	Target
Compression	0.710	0.680	0.743
Sentence splits	0.828	0.831	0.875
Levenshtein	0.782	0.610	0.559
Exact copies	0.181	0.036	0.020
Additions	0.057	0.297	0.403
Deletions	0.339	0.559	0.618

Table 6:
Quality estimation reports from EASSE.

- Finnish GPT: a Generative Pretrained Transformer with 1.5B parameters for Finnish. We use the XL version[12] and fine-tune it according to the authors' instructions[13].

We fine-tune mBART with default parameters as in the original instruction referenced above. For Finnish GPT, we employ instruction fine-tuning (Ouyang et al., 2022) and use the instruction "Mukauta selkosuomeksi" [translate to Easy Finnish]. We trained both models for 10 epochs.

For evaluation, we use the SARI metric, which uses an arithmetic average of n-gram precisions and recalls of editing operations: addition, keeping, and deletions between the source, output, and references (Xu et al., 2016). SARI is widely used for evaluating text simplification. It has some drawbacks, such as not being able to consider grammaticality or coherence, but it does have a good correlation with human judgments of simplicity (ibid.). Due to SARI's popularity, our results can be compared easily to any past works on simplification for other languages and future works on Finnish simplification. We use the code from the EASSE library (Alva-Manchego et al., 2019). The evaluation results can be seen in Table 5.

We also provide quality estimation features available in EASSE: the compression ratio of the simplification with respect to its source sentence, the Levenshtein similarity between source and simplification (calculated as Levenstein ratio in characters), the average number of sentence splits performed by the system, the proportion of exact matches (i.e. original sentences left untouched), the average proportion of added words and deleted words (Alva-Manchego et al., 2019). We do not report the lexical complexity score because, to the best of our knowledge, it is not language-agnostic in the current implementation. For comparison, we provide the quality estimation values between the source and target documents. The values can be seen in Table 6.

As can be seen, none of the systems has achieved the level of compression between the actual target sentences and source sentences. However, both mBART and Finnish GPT are close to the correct amount of sentence splitting. The higher Levenstein similarity, the number of exact copies, and the lesser amount of additions and deletions lead us to believe that mBART is a more conservative model, which can explain lower SARI scores.

Some examples of simplifications produced by models can be found in Table 7. It can be seen that mBART indeed makes fewer changes to the original sentence, sometimes leaving the smaller sentences unchanged. However, it should be noted that sometimes the target sentence also does not change the source much, so it is not necessarily an undesirable behavior. Finnish GPT seems to produce shorter and easier sentences but does not really change word order, which would be beneficial in the third example. The shortening can probably be explained by the fact that a lot of manual simplifications in our dataset also shorten the original sentences quite a lot. Sometimes, none of the models get to the simplification degree that the target sentence shows: for instance, in the 1st example, although the models performed some simplification, such as using a number to spell percentages and changing some words, it is still not as simple as the target. Nevertheless, the outputs are coherent and provide at least some degree of simplification.

The finetuned Finnish GPT can be found in the following repository: https://huggingface.co/Helsinki-NLP/simple-finnish-gpt3-xl. The finetuned mBART model can be accessed here: https://huggingface.co/Helsinki-NLP/fin-simple-mBART, or through the following PID: http://urn.fi/urn:nbn:fi:lb-2024011801.

6. Conclusions and future work

We have described the process of expanding a news-based parallel Finnish-Easy Finnish dataset, adding more document-aligned data and a sentence-aligned version. The augmented dataset spans almost 6.5 years of news broadcasts and has more than eight thousand document pairs and twelve thousand sentence pairs. We also made a small, manually aligned golden test set for sen-

[12] https://huggingface.co/TurkuNLP/gpt3-finnish-xl

[13] https://github.com/spyysalo/instruction-finetune

Source	Target	mBART	Finnish GPT
Suomen Apteekkariliiton mukaan tilastot osoittavat, että mielialalääkkeiden euromääräinen myynti on kasvanut Suomessa noin yhdeksän prosenttia viime vuodesta.	Mielialalääkkeiden myynti on kasvanut Suomessa tänä vuonna.	Suomen Apteekkariliitto kertoo, että mielialalääkkeiden myynti on kasvanut Suomessa noin 9 prosenttia viime vuodesta.	Suomen Apteekkariliitto sanoo, että mielialalääkkeiden myynti on kasvanut apteekeissa noin 9 prosenttia viime vuodesta.
According to the Finnish Association of Pharmacists, statistics show that the sales of mood drugs in euros in Finland have increased by around nine percent since last year.	Sales of mood medications have increased in Finland this year.	The Finnish Association of Pharmacists says that sales of mood medications have grown in Finland by about 9 percent since last year.	The Finnish Association of Pharmacists says that sales of mood medications have increased in pharmacies by about 9 percent since last year.
Onnettomuus tapahtui vähän ennen puolta päivää paikallista aikaa Coraton ja Andrian välisellä rataosuudella Etelä-Italiassa.	Onnettomuus sattui lähellä Andrian kaupunkia.	Coraton ja Andrian välisellä rataosuudella Etelä-Italiassa on tapahtunut onnettomuus.	Onnettomuus tapahtui vähän ennen puolta päivää.
The accident happened shortly before half a day at the local time on the line between Corato and Andria in southern Italy.	The accident happened near the city of Andria.	An accident has occurred on the track section between Corato and Andria in southern Italy.	The accident happened shortly before noon.
Sankarikoiria Suomessa on nimetty jo vuosia.	Suomessa on valittu sankarikoiria jo monta vuotta.	Sankarikoiria Suomessa on nimetty jo vuosia.	Sankarikoiria on Suomessa.
Hero dogs in Finland have been named for years.	Hero dogs have been chosen in Finland for many years.	Hero dogs in Finland have been named for years.	There are heroic dogs in Finland.

Table 7: Example simplifications. Finnish texts are from news articles (copyright: Yleisradio), and English texts are translations of the sentences above.

tence alignment. Currently, all these datasets can be found on Kielipankki (Dmitrieva and Yleisradio (2024a); Dmitrieva and Yleisradio (2024b)). We have obtained robust results on document alignment; however, despite trying multiple aligners that have been proven to work well for monolingual alignment, the predictive values for sentence alignment were not as high. Having worked on manual sentence alignment, we can conclude that it proves to be a genuinely difficult task to perform on our dataset. We leave a possible improvement of sentence alignment for future work. Nevertheless, sentence simplification models perform fine on our sentence-aligned data in comparison to SARI scores obtained on other languages (see, for example, the fine-tuning experiment results in Ryan et al., 2023). We hope that our results can be used as a baseline for future works on Finnish sentence simplification. Another prospective task that we see is document-level simplification for Finnish. Having a good-quality document-aligned dataset will allow for experimenting with full document simplification and/or document-level planning for simplification (Cripwell et al., 2023).

7. Ethical considerations and limitations

The data described in this research is available on Kielipankki for non-commercial use. Datasets based on texts from the Yle archives cannot be deposited elsewhere for copyright reasons. Only people with login credentials from certain academic organizations or those who have obtained permission from Kielipankki will be able to download this data.

We cannot guarantee that all automatically aligned sentence or document pairs are correctly aligned. As mentioned above, due to the difficult nature of sentence alignment across our data, some erroneous sentence alignments can be expected even when the score threshold is in place. We kept the cost scores provided by Vecalign in the published data for transparency.

We acknowledge that text simplification models' output cannot be thoroughly evaluated with just automatic metrics because they do not assess grammaticality or coherence. However, we hope that increasing the amount of available simplification data will help the development of more sophisticated data-driven simplification evaluation approaches, such as LENS (Maddela et al., 2023), for languages other than English.

Most computations that required GPU, which are embedding operations and model training, were performed with a single GPU node, the GPU being a Nvidia Tesla V100 with an Xeon Gold 6230 processor. Running Finnish GPT fine-tuning with LoRA (Hu et al., 2021) required two nodes with 48 gigabytes of memory allocated per node, although we are unsure if this is the minimum memory requirement (i.e., the minimum requirement might be smaller).

8. Bibliographical References

Rodrigo Alarcón, Paloma Martínez, and Lourdes Moreno. 2023. Tuning BART models to simplify Spanish health-related content. *Procesamiento del Lenguaje Natural*, 70:111–122.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A German dataset for joint summarization and simplification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2693–2701, Marseille, France. European Language Resources Association.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. Association for Computational Linguistics.

Anna Dmitrieva and Aleksandra Konovalova. 2023. Creating a parallel Finnish-Easy Finnish dataset from news articles. In *Proceedings of the 1st Workshop on Open Community-Driven Machine Translation*, pages 21–26, Tampere, Finland. European Association for Machine Translation.

Sarah Ebling, Alessia Battisti, Marek Kostrzewa, Dominik Pfütze, Annette Rios, Andreas Säuberli, and Nicolas Spring. 2022. Automatic text simplification for German. *Frontiers in Communication*, 7:706718.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C. Ziegler. 2020. Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1353–1361, Marseille, France. European Language Resources Association.

Isao Goto, Hideki Tanaka, and Tadashi Kumano. 2015. Japanese news simplification: tak design, data set construction, and analysis of simplified text. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, USA.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Auli Kulkki-Nieminen. 2010. *Selkoistettu uutinen. Lingvistinen analyysi selkotekstin erityispiirteistä [Plain Language news: a linguistic analysis of the sfecial features of simplified text]*. Ph.D. thesis, Tampereen Yliopisto.

Leealaura Leskelä. 2021. Easy language in Finland. In Ulla Vanhatalo Camilla Lindholm, editor, *Handbook of Easy Languages in Europe*, 1 edition, volume 8 of *Easy – Plain – Accessible*, pages 149–190. Frank & Timme.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Lei Liu and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. 2023. FinGPT: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726. Association for Computational Linguistics.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Martina Miliani, Serena Auriemma, Fernando Alva-Manchego, and Alessandro Lenci. 2022. Neural readability pairwise ranking for sentences in Italian administrative language. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 849–866, Online only. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Gustavo Paetzold, Fernando Alva-Manchego, and Lucia Specia. 2017. MASSAlign: Alignment and annotation of comparable documents. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 1–4, Tapei, Taiwan. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A New Dataset and Efficient Baselines for Document-level Text Simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161, Online and in Dominican Republic. Association for Computational Linguistics.

Michael Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

A.A. Shatilov and A.I. Rey. 2021. Sentence simplification with ruGPT3. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"*, pages 1–13.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Nicolas Spring, Marek Kostrzewa, David Fröhlich, Annette Rios, Dominik Pfütze, Alessia Battisti, and Sarah Ebling. 2023. Analyzing sentence alignment for automatic simplification of German texts. In Silvana Deilen, Silvia Hansen-Schirra, Sergio Hernández Garrido, Christiane Maaß, and Anke Tardel, editors, *Emerging Fields in Easy Language and Accessible Communication Research*, pages 339–369. Frank & Timme GmbH, Berlin.

Sanja Štajner, Marc Franco-Salvador, Paolo Rosso, and Simone Paolo Ponzetto. 2018. CATS: A tool for customized alignment of text simplification corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics.

Hideki Tanaka, Hideya Mino, Tadashi Kumano, Shinji Ochi, and Motoya Shibata. 2013. News service in simplified Japanese and its production support systems. *The Best of IET and IBC*, 5:44–48.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2020. Exploiting sentence order in document alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5997–6007, Online. Association for Computational Linguistics.

Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus, and Christin Seifert. 2022. Patient-friendly clinical notes: Towards a new text simplification dataset. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 19–27, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

9. Language Resource References

Anna Dmitrieva and Aleksandra Konovalova and Yleisradio. 2022. *Parallel Corpus of Finnish and Easy-to-read Finnish from the Yle News Archive 2019-2020, source*. Kielipankki. PID http://urn.fi/urn:nbn:fi:lb-2022111625.

Anna Dmitrieva and Yleisradio. 2024a. *Parallel Corpus of Finnish and Easy-to-read Finnish from the Yle News Archive 2014-2018, source*. Kielipankki. PID http://urn.fi/urn:nbn:fi:lb-2024011701.

Anna Dmitrieva and Yleisradio. 2024b. *Parallel Sentence Aligned Corpus of Finnish and Easy-to-read Finnish from the Yle News Archive 2014-2020, source*. Kielipankki. PID http://urn.fi/urn:nbn:fi:lb-2024011703.

Yleisradio. 2017. *Yle Finnish News Archive 2011-2018*. Kielipankki. PID http://urn.fi/urn:nbn:fi:lb-2017070501.

Yleisradio. 2019. *Yle News Archive Easy-to-read Finnish 2011-2018, source*. Kielipankki. PID http://urn.fi/urn:nbn:fi:lb-2019050901.

Multilingual Resources for Lexical Complexity Prediction: A Review

Matthew Shardlow[1], Kai North[2], Marcos Zampieri[2]
[1]Manchester Metropolitan University, UK
[2]George Mason University, USA
m.shardlow@mmu.ac.uk

Abstract

Lexical complexity prediction is the NLP task aimed at using machine learning to predict the difficulty of a target word in context for a given user or user group to understand. Multiple datasets exist for lexical complexity prediction, many of which have been published recently in diverse languages. In this survey, we discuss nine recent datasets (2018-2024) all of which provide lexical complexity prediction annotations. Particularly, we identified eight languages (French, Spanish, Chinese, German, Russian, Japanese, Turkish and Portuguese) with at least one lexical complexity dataset. We do not consider the English datasets, which have already received significant treatment elsewhere in the literature. To survey these datasets, we use the recommendations of the Complex 2.0 Framework (Shardlow et al., 2022), identifying how the datasets differ along the following dimensions: annotation scale, context, multiple token instances, multiple token annotations, diverse annotators. We conclude with future research challenges arising from our survey of existing lexical complexity prediction datasets.

Keywords: Text Difficulty, Multilinguality, Lexical Complexity Prediction

1. Introduction

Estimating the complexity of words or multi-word expressions (MWE) to a reader is an important first step in automatic lexical simplification pipelines (North et al., 2023a). Lexical complexity is modeled using either Lexical Complexity Prediction (LCP) or Complex Word Identification (CWI). LCP is the task of assigning a value to a word which indicates how difficult that word will be for a reader (North et al., 2023b). This contrasts to CWI, which is the binary setting of identifying if a word requires simplification or not (Shardlow, 2013a; Paetzold and Specia, 2016; Zampieri et al., 2017). Two recent efforts to curate LCP resources were undertaken in recent shared tasks (Shardlow et al., 2021; Ermakova et al., 2022), resulting in the Complex 2.0 dataset (Shardlow et al., 2022) and the SimpleText 2022 Task 2 data. Whilst these resources focused solely on English, there have been significant efforts throughout the community to develop parallel resources in other languages (Pirali et al., 2022).

There is a great wealth of shared information between these resources and gathering them together into a single resource could benefit future multilingual complexity prediction applications. However, to unite these resources, we must understand the purposes of each resource and identify the parameters of their construction. The Complex 2.0 Framework provides seven recommendations for features of future LCP datasets. We have reproduced these with our interpretation below:

1. **Annotation Scale:** Whereas previous resources for identifying complex words had typically focussed on the binary case (complex or not), the Complex 2.0 Framework recommended the use of continuous annotations such as those resulting from aggregating over a Likert scale.
2. **Context:** The words to be assigned difficulty rankings were presented in context. Clearly context affects word sense, which affects difficulty, but also the surrounding context of a word may give some explanation or interpretation of that word.
3. **Multiple token instances:** The same word presented in different contexts gives rise to the opportunity to analyse the difficulty of a word across many occurrences.
4. **Multiple annotations per token:** Complexity is subjective and aggregating judgements from multiple diverse annotators will alleviate local subjective deviations.
5. **Diverse annotators:** Similarly, having a diverse group of annotators will help to give a representative sample of LCP annotations. The annotator pool may be targeted at a specific group (e.g., language learners, students, deaf people, et.) according to the intended application.
6. **Multiple genres:** Collecting texts from multiple genres allows for more diverse text types represented in the dataset.
7. **Multi-word expressions:** The inclusion of complexity predictions for MWEs as well as single word instances helps to give a more

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 51–59

representative sample of the target language in the resulting LCP dataset.

In this paper, we survey the currently available LCP resources and analyse these through the lens of the recommendations given in the Complex 2.0 Framework. We note that (1) We identified nine suitable resources (listed in Table 1, CWI-18 appears three times, but is counted as a single resource) which we have focused this survey on. We only consider published available datasets in languages other than English, not work building on existing datasets. (2) the existing resources represent eight languages other than English. We do not include the English components of the three previous CWI/LCP shared task datasets (Paetzold and Specia, 2016; Shardlow et al., 2021; Yimam et al., 2018) or the English SimpleText 2022 Task 2 data (Ermakova et al., 2022) in our analysis. However for multilingual completeness, we do mention the French, Spanish and German components of the CWI-2018 Shared Task dataset.

The inclusion criteria for our resources were as follows:

- The resource was published since 2018.
- The resource provides complexity values of words at the level of single semantic units (i.e., not sentence or document level).
- The complexity values arise from annotation, as opposed to prediction or correlation to frequency.
- The language of the dataset was not English. We briefly discuss the existing English datasets, which have already been surveyed extensively below.

We provide an overview of the datasets we survey in Section 2, before progressing to a feature-based survey in line with the recommendations of the Complex 2.0 framework in Section 3.

2. Datasets Overview

2.1. CAS

Focused on technical terms in medical documents in French, Koptient and Grabar (2022) categorised terms from syntactic groups into 'understood', 'unsure' or 'not understood'. The authors gather lexical and syntactic features and train supervised learning algorithms to predict the reported difficulties of syntactic groups.

2.2. CWI18

Developed for a shared task at the BEA 2018 workshop, this dataset was developed in English, Spanish, French and German using Mechanical Turk to ask annotators to identify any words in a text that were complex. Each text was presented to multiple annotators, including native and non-native speakers, with complexity judgements applied to single words and spans. The final data was returned as both binary (did any annotator find the word complex) and continuous (how many annotators found the word complex).

2.3. VYTEDU-CW

A sample of Ecuadorian University students were asked to annotate texts from the VYTEDU corpus to indicate which words were difficult to understand. VYTEDU contains transcripts of educational videos in Spanish, which are suitable for university students. The authors provide some analysis of the complex words identified in the VYTEDU corpus, noting technical terms, sophisticated vocabulary, abbreviations, metaphor, unusual terms, verb-nominalisation and compound words as sources of complexity.

2.4. CLexIS^2

Students studying either Computer Systems or Software Engineering in Ecuador were asked to identify difficult words in transcripts of recorded lectures in Spanish from their courses using a custom annotation application. Complex words are later detected using an unsupervised and supervised approach.

2.5. LLCL

Lee and Yeung (2018) provide a study on the prediction of vocabulary knowledge for foreign language learners of Chinese. As a part of this study, they describe the annotation process of a dataset of Chinese words taken from the Lexical Lists for Chinese Learning in Hong Kong. Therein, they select 5 training sets and one test set which are labelled by language learners on a 5-point scale. These annotations focus on the word itself, without context presented.

2.6. RUBible

Texts from the Russian Synodal bible are annotated in a study closely replicating the work of (Shardlow et al., 2020). 931 words are presented across 3,364 contexts, which are then annotated on a 1–5 Likert Scale. The authors compare their results to the corresponding lexical complexity prediction data for English and also provide a linear regression demonstrating the ability to predict lexical complexity in Russian based on text features.

ID	Language	Reference
CAS	French	(Koptient and Grabar, 2022)
CWI18-FR	French	(Yimam et al., 2018)
CWI18-ES	Spanish	(Yimam et al., 2018)
VYTEDU-CW	Spanish	(Ortiz Zambrano et al., 2019)
CLexIS^2	Spanish	(Ortiz Zambrano and Montejo-Ráez, 2021)
LLCL	Chinese	(Lee and Yeung, 2018)
CWI18-DE	German	(Yimam et al., 2018)
RUBible	Russian	(Abramov and Ivanov, 2022)
JaLeCon	Japanese	(Ide et al., 2023)
CWITR	Turkish	(Ilgen and Biemann, 2023)
MultiLS-PT	Portuguese	(North et al., 2024)

Table 1: The datasets we consider for our survey. We have used the name given in the associated paper as the identifier, or the abbreviated name of the corpus that the LCP annotations are based on. In the case of the Russian dataset we have used the identifier RUBible as the texts are based on the Russian Synodal Bible.

2.7. JaLeCon

News and Government texts are provided to Native Japanese speakers as well as Chinese/Korean and other learners of Japanese for annotation on a 1-4 scale. Short word units and long word units are annotated with complexity values after word segmentation, which is necessary as Japanese does not support word boundaries. Baseline experiments show that a BERT-based system is effective for LCP in Japanese.

2.8. CWITR

Turkish language texts are annotated to identify complex words for readers using the binary setting. Annotations are collected for both complex words and phrases. Paragraph level texts are presented covering Wikipedia news, Wikipedia articles, news, novel summaries, and periodicals. All annotations were collected from native speakers of Turkish. In total 25 annotators provided complexity judgements over 13,837 instances.

2.9. MultiLS-PT

The MultiLS framework promotes a unified process for the tasks of lexical complexity prediction, substitution generation and binary comparative LCP. Brazilian Portuguese data has been collected for all tasks, but here we focus solely on the lexical complexity prediction data. This data is deliberately tied to the Complex 2.0 data, presenting 5,165 annotations across Bible, News and Biomedical texts.

2.10. English Datasets

Although not the main focus of this survey, there are English datasets available for complex word identification and lexical complexity prediction. The CW Corpus (Shardlow, 2013b) provided 731 instances of complex words mined from Simple Wikipedia edit histories. Later, related shared tasks Paetzold and Specia (2016) (Yimam et al., 2018) provided data for complex word identification. The Complex 2.0 (Shardlow et al., 2021) and SimpleText (Ermakova et al., 2022) corpora both provide English data for complexity prediction in Scientific texts (SimpleText) as well as religious and news (Complex2.0). Additionally, the work of Maddela and Xu (2018) provides word complexity data for 15000 words without contexts.

3. Literature Survey

3.1. Annotation Scale

The creators of LCP resources have used varied approaches to gather annotations. In all cases the resources that we have surveyed take the approach of identifying a target group and asking them a question about the difficulty of words in a text. The annotators are required to make a decision about the words, which may be a binary decision (is this word difficult or not difficult) (Ortiz Zambrano et al., 2019; Ortiz Zambrano and Montejo-Ráez, 2021; Ilgen and Biemann, 2023), or a graded decision on a Likert-scale (Koptient and Grabar, 2022; Lee and Yeung, 2018; Abramov and Ivanov, 2022; Ide et al., 2023; North et al., 2024). There is a subtle difference in the way that binary annotations or Likert-scale annotations are applied. In the binary setting, users are presented with an entire text and asked to mark any terms that they consider to be complex, with non-complex terms left unannotated. In the Likert-scale setting, annotators are presented with one or more tokens extracted from the text and asked to assign a rating based on a scale indicating difficulty. The annotator may choose to mark the word as an easy (low end of the scale) or difficult (high end of the scale) word. Binary annotations allow for a

much quicker annotation throughput as an annotator can return several annotations per sentence by simply highlighting all words they consider complex. Likert-scale annotations offer a more subtly graded degree of complexity. For instance, binary annotations ask 'Is the given word difficult to understand?', whereas Likert-scale annotations ask 'How difficult to understand is the given word?', returning an exact complexity value.

Binary annotations can be aggregated in two ways. Firstly, a researcher may choose to identify any word in a sentence as complex if at least one annotator considered it to be complex (Ortiz Zambrano et al., 2019; Ortiz Zambrano and Montejo-Ráez, 2021; Ilgen and Biemann, 2023). This returns a broad set of complex words without making a distinction between words that are considered complex by many or few annotators. To address this, probabilistic annotations (Yimam et al., 2018) aggregate the number of annotators that selected a word as complex in a binary setting. For example, in the CWI18 data 20 annotators identified complex words in each sentence. Each complex word has a probabilistic value derived as the number of annotators out of 20 that found the word to be complex.

Likert-scale data annotations are also collected from multiple annotators per instance and aggregated using 3 (Koptient and Grabar, 2022), 4 (Ide et al., 2023) or 5 (Lee and Yeung, 2018; Abramov and Ivanov, 2022; North et al., 2024) categories. Most examples of Likert-scale based datasets that we identified use simple mean averaging over the returned annotations to deliver a final complexity value following the Complex 2.0 framework (Abramov and Ivanov, 2022; Ide et al., 2023; North et al., 2024). A notable exception to this is CAS, which takes the most common annotation from their schema ('not understood','not sure' or unnanotated) as the overall label (Koptient and Grabar, 2022).

The LLCL dataset also reports a different construction technique which spans Likert-scale and binary protocols. In this dataset, the authors present a 5-point Likert-scale which is used for annotation by the target group (foreign language learners of Chinese). Annotators select a difficulty rating for each instance from 1 (Never seen the word before) to 5 (Absolutely know the word's meaning). The final dataset is then aggregated by considering any instances with an annotation of 5 as 'non-complex' and all others as 'complex' (Lee and Yeung, 2018).

3.2. Context

The mode of presentation of context at annotation time is an important decision to make in the construction of a LCP dataset. In the binary setting, the resources that we surveyed contain examples of tokens presented within a sentence (Ortiz Zambrano et al., 2019), paragraph (Ortiz Zambrano and Montejo-Ráez, 2021) and full document (Ilgen and Biemann, 2023). Allowing a reader to observe a full context allows them to explore the complexity of the word in context, taking into account both the specific word sense used as well as contextual factors such as clue words that may help to explain the difficult word. In the Likert-scale setting, we also observed examples of words presented within an entire document (Koptient and Grabar, 2022) as well as within a full sentence (Abramov and Ivanov, 2022; Ide et al., 2023; North et al., 2024).

The LLCL corpus (Lee and Yeung, 2018) only presents the word to annotators without context as the underlying corpus consists of a word list for foreign language learners of Chinese which are not presented within context. Datasets of words with lexical complexity annotations also exist for English (Maddela and Xu, 2018) and French CEFR levels (Pintard and François, 2020).

3.3. Multiple Instances of Each Token

This recommendation from the Complex 2.0 framework indicated that datasets for lexical complexity prediction should have several instances of the same token presented in-context. The perceived complexity of a word varies greatly depending on the presentation of the word in a sentence. Take, for example, the occurrence of the rare English word 'agog' in the following 3 examples from White (2017):

(1) They were **agog**.

(2) When the boy saw the sweets he was **agog** with anticipation.

(3) His talent [as a painter] is so enormous that you look at his surfaces with your mouth **agog** at the near-impossibility of it all.

In Example 1, it is very difficult to infer the meaning of the term. We can interpret that 'agog' is an emotion or sensation which can be held by a group of people but not much more. It is not clear from such a short context if this is negative or positive, abstract or concrete. Example 2 gives more context and a reader would correctly be able to interpret that 'agog' is related to the context term of anticipation and that it is the type of feeling a child may possess when seeing sweets. Even if the reader has never seen the term previously, they can infer the meaning from these contextual clues. Finally, in Example 3, a difficult word may appear within a context where the reader is led to incorrectly infer the meaning. In this case, a reader may be led to interpret 'agog' as a synonym of 'open', whereas in this case 'agog' is used to indicate eagerness or excitement.

In the datasets that we reviewed we found that all datasets which presented a context around the word also presented multiple instances of the same token. One particular variant to this approach is CAS, which uses syntactic groups to gather syntactically related terms for annotation (Koptient and Grabar, 2022).

3.4. Multiple Token Annotations

Lexical complexity is subjective (Shardlow, 2022). Two readers given the same text may identify different words as being complex. Moreover, two readers given the same word in the same context may assign a different complexity value on a Likert-scale. One factor that affects lexical complexity is L1 vs. L2 (Gooding et al., 2021; North and Zampieri, 2023), but this does not explain the full variation and more subtle factors such as education level, specialism and environmental factors are also likely to influence perceived complexity.

All the datasets we surveyed used multiple annotators to represent a variety of subjective opinions within the datasets. The degree of repeated annotations for the same instance varies widely across datasets with the CWI18 datasets reporting as few as 2 annotations per instance (Yimam et al., 2018) ranging up to 5-7 (Koptient and Grabar, 2022; Ortiz Zambrano and Montejo-Ráez, 2021; Ilgen and Biemann, 2023; Lee and Yeung, 2018) or even more than 10 (Abramov and Ivanov, 2022; Ide et al., 2023). More annotations per instance allows for a diverse range of subjective opinions to be represented and for the aggregation of these opinions to represent some normative value that can be useful for all annotators.

One strategy for collecting multiple annotations is to use crowdsourcing (Yimam et al., 2018; Ilgen and Biemann, 2023; North et al., 2024). Many resources that we surveyed do not report whether the annotators were paid or unpaid (Koptient and Grabar, 2022; Ortiz Zambrano et al., 2019; Ortiz Zambrano and Montejo-Ráez, 2021; Lee and Yeung, 2018; Ide et al., 2023). In these cases we assume that annotators were selected from populations that did not require remuneration (such as colleagues or students). Several authors report using Mechanical Turk, but do not report the amount paid per instance (Yimam et al., 2018; Abramov and Ivanov, 2022). 2 of the resources that we surveyed do report the degree of pay for the annotators, with RUBible paying 10 cents for a batch of 10 instances and MultiLS-PT reporting payment of 2 cents per instance.

3.5. Diverse Annotators

Annotators vary between multilingual datasets. Annotators have been either hand-selected or crowd-sourced and are representative of differing target demographics. Several datasets were developed to create LCP systems for second-language (L2) learners (Lee and Yeung, 2018) and have subsequently been annotated by individuals not native to the dataset's target language. Other datasets are developed solely for identifying complex words for first-language (L1) speakers (Ortiz Zambrano et al., 2019; Abramov and Ivanov, 2022). These datasets are annotated by individuals native to the predominant language of the dataset. However, other annotator variables are often controlled, including age, level of education, or reading disability. The following paragraphs discuss the merits and flaws of datasets that have (a) employed hand-selected versus crowd-sourced annotators, alongside (b) controlled influential annotator variables.

Several multilingual datasets hand-selected their annotators making them ideal for the creation of personalised LCP systems. CAS (Koptient and Grabar, 2022) hand-selected 9 French speaking annotators to rate the complexity of medical jargon for non-expert patients. By hand-selecting their annotators, (Koptient and Grabar, 2022) were able to control the level of prior familiarity annotators had with medical terminology improving the validity of their gold complexity labels. They only selected annotators with no self-reported medical knowledge, and asked annotators to not refer to online material, including dictionaries, for assessing word difficulty. VYTEDU-CW (Ortiz Zambrano et al., 2019) and CLexIS (Ortiz Zambrano and Montejo-Ráez, 2021) hand-selected university students in Ecuador to identify complex words spoken in Spanish. They likewise controlled annotator familiarity by presenting annotators with transcripts of recorded lectures that were on a subject-matter known but not overly familiar to the annotators. LLCL (Lee and Yeung, 2018) and JaLeCon (Ide et al., 2023) hand-selected 7 and 15 L2 learners of Chinese and Japanese respectively. Both datasets make reference to L2 proficiency frameworks, with Ide et al. (2023) having only recruited annotators with at least an intermediate level of L2 proficiency.

CWI18 (Yimam et al., 2018), CWITR (Ilgen and Biemann, 2023), and MultiLS-PT (North et al., 2024) crowd-sourced annotators using Amazon Mechanical Turk (MTurk), whereas RUBible (Abramov and Ivanov, 2022) crowd-sourced their annotators from Toloka. As such, each dataset was able to obtain a substantially greater number of annotators compared to those datasets that adopted hand-selection. The CWI18-FR and CWI18-ES datasets (Yimam et al., 2018) were annotated by 22 and 54 respectively, and were recruited from a variety of countries. CWITR (Ilgen and Biemann, 2023) hired 25 annotators located in Turkey, MultiLS-PT (North et al., 2024) selected 25 an-

notators from Brazil, and RUBible (Abramov and Ivanov, 2022) gathered 10 separate annotators from Russia, Ukraine, Belarus and Kazakhstan. However, only several of these datasets attempted to control language proficiency. The CWI datasets make a distinction between native and non-native speakers yet do not explain how this distinction has been made. CWITR (Ilgen and Biemann, 2023) enforced a language proficiency exam to record Turkish language proficiency. The remaining datasets were unable to collect information regarding mother tongue, number of languages known, or L2 proficiency. Past studies have shown that discrepancies in these variables between annotators results in differing perceptions of word difficulty (Maddela and Xu, 2018; North and Zampieri, 2023). Failure to control these variables is an obvious drawback which reduces the validity of crowd-sourced datasets. This is only compensated by their larger pool of annotators and overall generalisability.

3.6. Multiple Genres

Multilingual datasets differ in genre. Several datasets contain texts pertaining to a single genre (Koptient and Grabar, 2022; Yimam et al., 2018; Ortiz Zambrano et al., 2019; Abramov and Ivanov, 2022). Other datasets consist of multiple genres (Ide et al., 2023; Ilgen and Biemann, 2023; North et al., 2024). These genres include medical-related articles, educational materials, the Bible to news and Wikipedia extracts. These genres are typically believed to be of great importance. They relate to such topics as health literacy, education, or political awareness motivating their simplification for improved accessibility (North et al., 2023b). The following paragraphs detail the types of texts provided by the single and multi-genre datasets shown within Table 1 and summarise their uses.

Single genre datasets include CAS (Koptient and Grabar, 2022), the CWI18 datasets (Yimam et al., 2018), VYTEDU-CW (Ortiz Zambrano et al., 2019), CLexIS (Ortiz Zambrano and Montejo-Ráez, 2021), and RUBible (Abramov and Ivanov, 2022). CAS provides a corpus of 100 clinical reports annotated with complex words. These clinical reports summarise a patient's medical history, diagnosis and outcome. CWI18-FR, CWI18-ES, and CWI18-DE provide 2,251, 14,280, and 7,403 complex words in context taken from Wikipedia articles (Yimam et al., 2018). Wikipedia articles are a common source of texts for LCP researchers. Public edits to pre-existing articles were previously used to gather gold complex and simplified labels (Shardlow, 2013b). Later datasets, such as the CWI18 datasets, improved their validity by incorporating human annotation. VYTEDU-CW (Ortiz Zambrano et al., 2019) and CLexIS (Ortiz Zambrano and Montejo-Ráez, 2021) gathered educational material in the form of transcripts from university lectures. These datasets are unique as they provide instances that contain elements of spoken language. The Bible is another popular text for LCP researchers. RUBible contains 3,364 extracts parallel to those found within an English sister dataset, CompLex 2.0 (Shardlow et al., 2020). RUBible is therefore a perfect dataset for the investigation of cross-lingual transfer learning in regards to LCP.

Single genre datasets allow for model specialisation, whereas multi-genre datasets are used to report model performances across multiple domains. Models trained on several single genre datasets or one multi-genre dataset can be used to investigate the performances of unique training strategies, such as transfer learning between genres and in some instances, cross-lingual transfer learning.

3.7. Multi-word Expressions

Lexical complexity prediction can be applied both to single words and to multi-word expressions (defined as a contiguous set of tokens separated by white space, with a single well-known meaning). English datasets for complex word identification and lexical complexity have taken multi-word expressions into account (Yimam et al., 2018), (Shardlow et al., 2022). In this context, we treat multi-word expressions as single lexical units, which behave as words. We assume that a complexity judgement can be made regarding a multi-word expression in the same way that it can be made for a single word. Non-compositional multi-word expressions hold some semantic value that cannot be derived from the meaning of constituent words. E.g., a hot dog is not a type of dog and may not even be hot. Similarly, the complexity of a non-compositional multi-word expression may not be easily derivable from the complexities of its constituent words.

In our multilingual resources, we observed 3 instances of datasets which report solely on single-word lexical complexity (Ortiz Zambrano and Montejo-Ráez, 2021; Lee and Yeung, 2018; North et al., 2024). All other resources took MWEs into account. The idea of MWEs comes from the English language and the idea of single- vs multi-word units may not transfer easily to other languages. For example in a language such as German, there is a heavy degree of noun compounding, where spaces between words are omitted. These behave as multi-word expressions, but appear as single words. This is particularly apparent for Japanese, which mixes syllabic and logographic characters without word boundaries. The JaLeCon dataset provides annotations over Short Unit Words (SUWs) which correspond to one or two small lexical units. Multi-word expressions are identified as Long Unit Words (LUWs), which are also annotated for complexity. (Koptient and Grabar, 2022) use syntactic groups

to form token sequences that are then annotated for complexity. These may be single words, but are often several contiguous words under a single syntactic head.

4. Discussion

The most stark difference in the resources that we have surveyed is the question that is presented to the judges of lexical complexity. In the binary setting annotators are asked to identify any complex words (and often also phrases) in a text, whereas in the Likert-scale setting annotators are asked to return a judgement on a multi-point scale for a given word (usually) in a context. This gives rise to two very different forms of lexical complexity datasets. The former refers to words or phrases which have been identified as problematic by some user. the latter refers to words or phrases which have been assigned some value judgement according to their complexity. Researchers working with both types of data should bear in mind the difference between these protocols. A 0 (non-complex) label in the binary setting implies no user found this word to be complex, whereas a 0 label in the Likert-scale setting implies that every user indicated this word to be the least complex. Similarly a 1 (complex) label in the binary setting implies that at least 1 user (depending on the aggregation protocol used) found this instance to be complex, whereas a label of 1 in the Likert-scale setting implies that every user rated the word as the most difficult complexity level.

Additionally, it is worth considering that in the binary setting a user may be asked to identify any complex words or phrases in a text, whereas in the Likert-scale setting pre-identified words are presented. Both these processes may lead to biases in datasets (reflecting tokens selected by the annotators, or tokens selected by the researchers), which should be considered when making decisions about what is desired from the resulting dataset. For example, a researcher may want only examples of complex language in an LCP dataset, in which case they may select specific tokens according to some pre-identification protocol. Alternatively, a researcher may wish to have both low-complexity and high-complexity elements in a dataset in which case they may select tokens at random.

The datasets that we have identified cover 8 languages. Including 5 Indo-European languages (French, Spanish, German, Russian and Portuguese), 6 alphabetic languages (French, Spanish, German, Russian, Turkish and Portuguese), 2 Logographic languages (Chinese and Japanese), with Japanese also exhibiting Syllabary elements (Kana). Notable exceptions include south asian langauges (e.g., Hindi, Urdu, Sinhala, Bengali) and African languages as well as other low-resource languages.

The resources that we have surveyed present a variety of languages, but also text genres incorporating encyclopaedia text, medical texts, educational texts and religious texts. Systems trained for one language or genre may be more easily adaptable to future related languages and genres. This allows for the creation of generalisable models that are able to perform well on varying types of texts.

There is some variability in the protocols used for annotation. For example, the number of annotators per instance varies from 2 to 25. It is important for dataset providers to report on these statistics and to release appropriate metadata alongside the annotations to allow future users of lexical complexity prediction datasets to fully understand the meaning of the annotations. One particular source of variability is the use of native speakers, non-native speakers or language learners as annotators. It is likely that each group will have different complexity needs and will return different subjective lexical complexity judgements. Ongoing work on personalised lexical complexity (Gooding and Tragut, 2022) could benefit from varied datasets, if the appropriate metadata for target groups is maintained.

5. Future Research Challenges

The Complex 2.0 framework and MultiLS framework describe a pattern for future dataset creation for lexical complexity prediction resources and beyond. Future resources can follow the recommendations found in these works to deliver future datasets in diverse languages conforming to robust protocols followed by previous datasets. The MLSP shared task[1] is currently seeking to create a new dataset following the MultiLS framework for both lexical complexity prediction and lexical simplification. Future work to extend these datasets with additional languages, additional annotations in existing languages and additional text types will be beneficial to the community in generating new and interesting types of data for lexical complexity prediction in diverse lingual settings. We would particularly like the community to prioritise: (a) the development of LCP resources for widely spoken languages such as Mandarin Chinese, Hindi, Arabic, Bengali and beyond. (b) the inclusion of diverse language families beyond the heavy tendency to develop resources for Indo-European languages. (c) LCP resources for low-resourced languages.

[1] https://sites.google.com/view/mlsp-sharedtask-2024/home

References

Aleksei V. Abramov and Vladimir V. Ivanov. 2022. Collection and evaluation of lexical complexity data for russian language using crowdsourcing. *Russian Journal of Linguistics*, 26(2):409–425.

Liana Ermakova, Eric Sanjuan, Jaap Kamps, Stéphane Huet, Irina Ovchinnikova, Diana Nurbakova, Sílvia Araújo, Radia Hannachi, Elise Mathurin, and Patrice Bellot. 2022. Overview of the clef 2022 simpletext lab: Automatic simplification of scientific texts. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 470–494. Springer.

Sian Gooding, Ekaterina Kochmar, Seid Muhie Yimam, and Chris Biemann. 2021. Word complexity is in the eye of the beholder. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449, Online. Association for Computational Linguistics.

Sian Gooding and Manuel Tragut. 2022. One size does not fit all: The case for personalised word complexity models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 353–365, Seattle, United States. Association for Computational Linguistics.

Yusuke Ide, Masato Mita, Adam Nohejl, Hiroki Ouchi, and Taro Watanabe. 2023. Japanese lexical complexity for non-native readers: A new dataset. In *Proceedings of the Eighteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Bahar Ilgen and Chris Biemann. 2023. Cwitr: A corpus for automatic complex word identification in turkish texts. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '22, page 157–163, New York, NY, USA. Association for Computing Machinery.

Anaïs Koptient and Natalia Grabar. 2022. Automatic detection of difficulty of French medical sequences in context. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 55–66, Marseille, France. European Language Resources Association.

John Lee and Chak Yan Yeung. 2018. Automatic prediction of vocabulary knowledge for learners of chinese as a foreign language. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–4. IEEE.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023a. Deep learning approaches to lexical simplification: A survey. *arXiv preprint arXiv:2305.12000*.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2024. MultiLS: A Multi-task Lexical Simplification Framework. *arXiv preprint arXiv:2402.14972*.

Kai North and Marcos Zampieri. 2023. Features of Lexical Complexity: Insights from L1 and L2 Speakers. *Frontiers in Artificial Intelligence*, 6(1).

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023b. Lexical complexity prediction: An overview. *ACM Comput. Surv.*, 55(9).

Jenny Ortiz Zambrano, Arturo MontejoRáez, Katty Nancy Lino Castillo, Otto Rodrigo Gonzalez Mendoza, and Belkis Chiquinquirá Cañizales Perdomo. 2019. Vytedu-cw: Difficult words as a barrier in the reading comprehension of university students. In *The International Conference on Advances in Emerging Trends and Technologies*, pages 167–176. Springer.

Jenny A. Ortiz Zambrano and Arturo Montejo-Ráez. 2021. CLexIS2: A new corpus for complex word identification research in computing studies. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1075–1083, Held Online. INCOMA Ltd.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Alice Pintard and Thomas François. 2020. Combining expert knowledge with frequency information to infer CEFR levels for words. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 85–92, Marseille, France. European Language Resources Association.

Camille Pirali, Thomas François, and Núria Gala. 2022. PADDLe: a platform to identify complex

words for learners of French as a foreign language (FFL). In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 46–53, Marseille, France. European Language Resources Association.

Matthew Shardlow. 2013a. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Shardlow. 2013b. The CW corpus: A new resource for evaluating the identification of complex words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Shardlow. 2022. Agree to disagree: Exploring subjectivity in lexical complexity. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 9–16, Marseille, France. European Language Resources Association.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2022. Predicting lexical complexity in english texts: the complex 2.0 dataset. *Language Resources and Evaluation*, 56(4):1153–1194.

Murray White. 2017. Alex Janvier and the fine art of defiance. Toronto Star.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex word identification: Challenges in data annotation and system performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 59–63, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Plain Language Summarization of Clinical Trials

Polydoros Giannouris, Theodoros Myridis, Tatiana Passali, Grigorios Tsoumakas
School of Computer Science, Aristotle University of Thessaloniki
{polydoros,tmyridis,scpassali,greg}@csd.auth.gr

Abstract

Plain language summarization, or lay summarization, is an emerging natural language processing task, aiming to make scientific articles accessible to an audience of non-scientific backgrounds. The healthcare domain can greatly benefit from applications of automatic plain language summarization, as results that concern a large portion of the population are reported in large documents with complex terminology. However, existing corpora for this task are limited in scope, usually regarding conference or journal article abstracts. In this paper, we introduce the task of automated generation of plain language summaries for clinical trials, and construct CARES (Clinical Abstractive Result Extraction and Simplification), the first corresponding dataset. CARES consists of publicly available, human-written summaries of clinical trials conducted by Pfizer. Source text is identified from documents released throughout the life-cycle of the trial, and steps are taken to remove noise and select the appropriate sections. Experiments show that state-of-the-art models achieve satisfactory results in most evaluation metrics.

Keywords: plain language summarization, lay summarization, clinical trials, simplification, summarization

1. Introduction

Lay summarization, also known as plain language summarization, is the process of distilling intricate information into clear, concise and easily digestible summaries (Vinzelberg et al., 2023). In an era of abundant specialized knowledge and technical jargon, lay summarization plays a vital role in making complex ideas, scientific findings, or technical concepts accessible and comprehensible to individuals who may lack expertise in a particular field.

Lay summarization is particularly important in communicating scientific articles to the general public, especially in the field of medicine. This became apparent during the COVID-19 pandemic when millions of scientific articles with medical content were published (Islam et al., 2020). These articles were comprehensible to only a few, resulting in misinterpretations to the extent that it led to misinformation and fake news (Brennen et al., 2020).

The state-of-the-art in lay summarization is constantly evolving, driven by cutting-edge NLP approaches, as well as by the development of appropriate datasets for supervised learning (Chandrasekaran et al., 2020; Guo et al., 2021, 2024; Goldsack et al., 2022; Attal et al., 2023). Existing datasets concern scientific publications. Another important type of scientific information is clinical trials, which comprise several long documents, including a study protocol, a statistical analysis plan and a report synopsis, as well as related scientific publications. Lay summarization of clinical trials is not only important for the general public, but also a requirement for pharmaceutical companies by regulations such as EU Regulation No 536/2014 and US Public Health Service Act 2007. However, to the best of our knowledge, this task has not been considered by the NLP community before and no relevant public datasets exist.

This work takes some first steps to fill this gap, by introducing lay summarization of clinical trials as a task, and constructing a corresponding dataset, CARES (Clinical Abstractive Result Extraction and Simplification), which pairs publicly available plain language summaries (PLSs) with relevant pieces of text from the associated documents. Table 1 shows a sample of such a pair.

Source: Programmed death ligand 1 (PD-L1, also called B7-H1 or CD274) has a known role in the suppression of T-cell responses. The PD-1 receptor is expressed on activated CD4+ and CD8+ T cells. By interaction with its ligands, PD-L1 and PD-L2, PD-1 delivers a series of strong inhibitory signals to inhibit T-cell functions. Avelumab*(MSB0010718C), a fully human antibody of the immunoglobulin G1 (IgG1) isotype, specifically targets and blocks PD-L1, the ligand for PD-1 receptor. In preclinical studies, the combination of avelumab with chemotherapy (gemcitabine, oxaliplatin, 5FU) showed improved anti-tumor activity over single-agent chemotherapy ...
Summary: Avelumab is a medicine that may work by targeting a protein called programmed death-ligand 1 (pd-l1) found on the cancer cell. Pd-l1 is involved in the bodys immune system response to cancer. When this study was started, avelumab was being tested for use in women with advanced ovarian cancer. Although avelumab is approved in other types of cancer, it is not approved for use ...

Table 1: Sample of a source and summary pair from the CARES dataset.

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 60–67

The rest of the paper is structured as follows: Section 2 presents related work in this field. Section 3 discusses the developed dataset. Section 4 covers the experiments conducted on this dataset and finally, Sections 5 and 6 introduce the conclusions and limitations of the dataset, respectively.

2. Related Work

In this section, we delve into the existing research and methodologies regarding lay summarization, particularly focusing on datasets available for the task and methods employed for generating simplified summaries from scholarly documents.

2.1. Datasets

One of the first resources in the field of lay summarization was a corpus of 572 full-text papers accompanied by lay summaries, in a variety of domains, including archaeology, hematology, and engineering, which was made available by Elsevier in the context of the 1st Workshop on Scholarly Document Processing (Chandrasekaran et al., 2020).

In biomedicine, (Guo et al., 2021) developed a dataset pairing 7,805 systematic reviews from the Cochrane database with plain language abstracts written by domain experts. (Goldsack et al., 2022) introduced two datasets: the Public Library of Science (PLOS) and eLife, each containing biomedical articles along with PLSs written by experts, the first having over 27k examples. Recently, (Attal et al., 2023) presented PLABA a dataset containing 750 abstracts from PubMed from 75 different health-related topics and expert-created adaptations at the sentence level. Lastly, (Guo et al., 2024) describe CELLS, the largest dataset of over 62k examples of parallel scientific abstracts and the corresponding expert-authored lay language summaries.

The dataset developed for our study differs from prior efforts in that CARES is the first dataset tailored specifically to *clinical trials*, instead of scientific publications.

2.2. Methods

There have been several efforts to develop models and methods for lay summarization. Specifically, (Chaturvedi et al., 2020), in their attempt to tackle CL-LaySumm20, which requested the development of non-technical summaries from scholarly documents, introduced a two-step divide-and-conquer technique. This approach involves extracting sentences from plain sections of the inputs using an unsupervised network and then performing abstractive summarization and merging them.

Furthermore, during CL-LaySumm 2020 in SDP workshop at EMNLP 2020, (Kim, 2020) achieved the top performance on the task of generating simplified summaries for scientific papers. They employed the PEGASUS (Zhang et al., 2019) model for producing the initial lay summaries, which were improved by appending important sentences to the summary of which the number of words was under a certain threshold, using a Presumm (Liu and Lapata, 2019), a BERT-based (Devlin et al., 2018) extractive summarization model.

Lastly, (Shaib et al., 2023) utilized GPT-3 in the zero-shot setting to summarize and simplify articles describing trials. They also applied this approach to the summarization of meta-analyses involving multiple documents. Despite also working on the lay summarization of clinical trials, our approach differs in that we aim to reproduce a particular document and not provide a general lay summary.

Although there is existing literature on lay summarization tasks for scientific publications and articles, we are the first to apply the generation of simplified summaries to whole clinical trials.

3. CARES Dataset

Motivated from the crucial role of high-quality parallel corpora in developing biomedical simplification models (Ondov et al., 2022), we introduce CARES, the 1st dataset for plain language summarization of clinical trials[1]. Although summaries (referred to as targets or golden summaries hereafter) are readily available, there exists no single respective technical text (source) for the entire summary. In this section, we outline our methodology for creating the dataset, as well as the process of identifying the suitable document and subsection for each component of the PLS.

3.1. Target Extraction

We start the construction of CARES from the *Plain Language Study Results Summaries* repository of Pfizer[2]. We collected the PDF files of the 176 summaries that existed in this repository, up to March 3rd, 2023. Next, we extracted their text, making sure artifacts are not introduced in the form of page numbers or identifiers present in the margins.

We found that their length often exceeds 1,200 words, which surpasses the capacity of most state-of-the-art models such as BART and PEGASUS. To address this issue, we exploited their discourse structure, inspired by the divide-and-conquer paradigm in (Gidiotis and Tsoumakas, 2020). Authors follow a question-answer structure, aimed at addressing different aspects of the clinical

[1] https://github.com/PolydorosG/CARES

[2] https://www.pfizer.com/science/clinical-trials/plain-language-study-results-summaries

trial, from its conception to its results. The respective titles of these sections are as follows:

- Q1: “Why was this study done?"
- Q2: “What happened during $\{i\}$ study?", $i \in \{$the, this$\}$
- Q3: “What were the results of the study?"
- Q4: “What $\{i\}$ did $\{j\}$ have during the study?", $i \in \{$medical problems, side effects$\}$ $j \in \{$participants, patients, children, boys, volunteers, infants$\}$
- Q5: “Were there any serious medical problems?" $\vee$ "Did $\{i\}$ have any serious $\{j\}$?", $i \in \{$(study) participants, study infants$\}$, $j \in \{$side effects, medical problems$\}$

Table 2 shows the number of examples per section, along with their average word counts. It is evident that Q1 and Q5 appear consistently in each of the initial 176 summaries, in contrast to Q2-Q4. Missing data are attributed to two factors: a) the existence of studies with different objectives, leading to certain sections being deemed irrelevant to the specific analysis being conducted and therefore not being included by the summary authors, and b) the introduction of noise during the text extraction process, despite our measures to prevent this. In such cases, portions of the text become corrupted, rendering certain sections irrecoverable.

Section	Summaries	Average # of words
$Q1$	176	325
$Q2$	175	555
$Q3$	166	287
$Q4$	171	267
$Q5$	176	130
Total	864	-

Table 2: Section headers identified in the PLSs.

3.2. Source Selection

Every clinical trial has a set of documents that describe each of its parts, from its design to the analysis of the results. Clinical studies begin with a *study protocol*. This is a detailed description of the plan that explains the objective of the clinical trial, as well as how it will be conducted. The protocol is usually accompanied by a *statistical analysis plan*. At the same time, scientific journal articles may be published for certain studies, mainly of new drugs. Finally, after the end of the trial, a *clinical study report synopsis* is created, which analyzes the results as well as the events that occurred during the study. Many of these documents exceed 100 pages of text, rendering summarization impossible for models without selecting a small portion of each document. Next, we will describe the document selected for each section, with the exception of Q3 for which no appropriate section was found.

The content of Q1 is the most general of all. It usually includes research-independent elements, such as general information about the disease and results of previous studies. Concerning the trial itself, the questions that will be answered, as well as the motivation of the researchers are described. As these are determined in advance, the most appropriate document is the study protocol. When available, we keep the study protocol’s summary, often referred to as *synopsis*. Otherwise, we use its *introduction* section, as it was found to contain most of the necessary information.

Q2 concerns the design of the clinical study. It analyzes data on the population and the separation of patients into groups. Afterward, the strategy followed regarding the administration of the substance is mentioned, as well as the type of the study, such as whether the groups are randomly selected (randomization), whether a control group is included, or whether it is single-blind or double-blind. This information is located in the *study design* part of the study protocol. Since section titles are not consistent across study protocols, we use regular expressions to isolate the particular segment.

Finally, Q4 and Q5 both refer to the side effects and medical problems experienced by the trial participants. Their difference lies in the severity, as they are analyzed separately in Q5 if they were life-threatening, required medical attention, or caused permanent damage. Due to the thematic similarity of the two questions, the source of both is found in the safety results section of the clinical study report synopsis.

Despite an initial choice of both document and section within the selected document, we find that source lengths remain prohibitively large for neural models. A major reason for this was the introduction of noise during text extraction. Despite preprocessing steps, including cropping margins and automatically identifying and removing text from tabular data, errors in these steps may persist, introducing a large volume of artifacts. For this reason, we proceed to evaluate each sentence of the large initial source section with regard to its similarity to the target.

Let $\mathcal{Y} = \{Y_1, Y_2, ..., Y_n\}$ be a set of golden summaries, and $\mathcal{X} = \{X_1, X_2, ..., X_n\}$ be a set of the respective candidate, uncleaned sources. We tokenize each initial source into a sequence of m sentences, $X_i = [x_1, x_2, ..., x_m]$. We then quantify the similarity of each sentence with the summary using the ROUGE-L (Lin, 2004) recall score R_{LCS}:

$$R_{LCS}(x_i, Y_i) = \frac{LCS(x_i, Y_i)}{l}, \quad (1)$$

where l is the length of sentence x_i and $LCS(x_i, Y_i)$

is the length of the longest common subsequence between x_i and Y_i.

Although ROUGE was used in previous work to match sentences of the summary with parts of a document and automatically create source-target pairs for training (Gidiotis and Tsoumakas, 2020), no special care has been taken for entities. Given the simplified vocabulary of the summaries, we propose the addition of entity matches between candidate sources and targets as an anchor between complex and simplified text. In the context of factual consistency of summarization, Nan et al. (2021) proposed the use of named entity recall and precision. We adapt this concept in order to measure entity-level recall R_E. Specifically, we extend the proposed method to also include numbers and percentages. Since decimals are not included in PLSs, we identify such digits in the source. These numbers are then rounded to the nearest integer to best align the source and target formats. The final similarity score is thus defined as:

$$S(x_i, Y_i) = \alpha * R_{LCS}(x_i, Y_i) + (1 - \alpha) * R_E(r(NE_{x_i}), NE_{Y_i}) \quad (2)$$

where α is a hyper parameter, NE_{x_i}, NE_{Y_i} the named entities detected in source sentence x_i and PLS Y_i respectively, and $r(\cdot)$ the simplification function applied to source entities.

Finally, after evaluating each sentence's similarity to the respective target, we select sentences in descending order until scores reach a threshold or the maximum length is exceeded. Sentences are reordered before forming the dataset to best replicate the document's structure. In case of no appropriate source sentences (i.e. no sentences pass the fixed similarity threshold), the examples are removed completely. An example of the source extraction pipeline is provided in Figure A.1.

The final dataset consists of 478 source-target pairs. The final word counts, along with the number of examples in each split are presented in Table 3. Note that to best reproduce real-world settings, we make sure each summary's subsections are not present in different splits. Lastly, to aid future research, we publish both the selected sources as well as the entire source documents and targets.

	Length		Summaries		
Section	**Source**	**Target**	**Train**	**Val.**	**Test**
Q1	713	330	87	13	9
Q2	665	570	79	12	8
Q4	406	279	102	18	13
Q5	405	129	104	19	14

Table 3: Source and summary length after our similarity-based filtering method, along with number of examples in train, test and validation splits.

4. Experiments

To facilitate future work, we benchmark our dataset using state-of-the-art summarization models BART and PEGASUS. We run all experiments on an Nvidia Tesla T4 with 16 GB of memory, using the open-source Hugging-Face implementations (Wolf et al., 2019) for a maximum of 10 epochs. We monitor the models' performance on the validation set for each epoch and select the best model according to ROUGE-L score. All BART models were initialized from the "facebook/bart-large" model, and PEGASUS from "google/pegasus-large". Finally, entity recognition was performed using spaCy (Honnibal and Montani, 2017).

Since Q4 and Q5 of the same PLS may have the same source, we train models under two settings. The first consists of training separate models for each section, referred to as BART and PEGASUS, treating them as a distinct summarization task. For the second approach, in order to utilize the whole training set in a single model we prepend the section's title to each source employing special tokens to tag it, before feeding the document to the model (Passali and Tsoumakas, 2022). These models are referred to as BART$_{TAG}$ and PEGASUS$_{TAG}$.

We evaluate generated summaries using both ROUGE and named entity recall and precision, to evaluate entity-level factual consistency. The experimental results reported in Table 4 are highly promising, with the BART models outperforming PEGASUS on most sections. However, determining what constitutes a *good* ROUGE score can vary depending on the domain and the specific task at hand. In our investigation, we observed that the ROUGE scores of models trained on our dataset align with those reported in similar studies on analogous datasets. It is worth noting that while individual models exhibit superior performance compared to the tagging method, this enhanced performance is achieved at the expense of requiring four times as many models.

Regarding the somewhat subpar performance in Q2, we attribute it to the open-endedness of the question rather than our regular expressions. To further investigate this we calculate the ROUGE scores between the selected sources and golden summaries. As can be seen in Table 5, contrary to the model performance, our retrieval approach appears to be most successful in Q2. Therefore we ascribe the relatively bad performance, to Q2 being a harder section to simplify and summarize.

Finally, we notice that models trained on the entire dataset (BART$_{TAG}$ and PEGASUS$_{TAG}$), despite generally showcasing lower ROUGE scores, are able to more accurately generate entities. This observation is consistent with previous claims that ROUGE alone is inadequate to quantify factual con-

	Model	ROUGE – 1	ROUGE – 2	ROUGE – L	$\text{Precision}_{\text{NE}}$	$\text{Recall}_{\text{NE}}$
Q1	BART	**51.8**	**25.3**	**31.3**	46.79	35.34
	BART_{TAG}	50.7	24.0	30.6	**47.36**	**44.60**
	PEGASUS	34.0	11.7	22.7	35.51	42.34
	PEGASUS_{TAG}	35.0	10.9	22.4	43.46	41.73
Q2	BART	**47.1**	19.3	25.7	69.98	**44.13**
	BART_{TAG}	46.5	**19.5**	**26.8**	**70.05**	38.77
	PEGASUS	34.0	13.0	25.4	58.32	39.94
	PEGASUS_{TAG}	35.0	12.4	24.2	69.50	26.85
Q4	BART	**73.9**	**62.4**	67.0	53.90	48.74
	BART_{TAG}	70.7	58.6	63.8	**58.71**	**57.09**
	PEGASUS	66.2	57.0	60.8	54.86	46.85
	PEGASUS_{TAG}	69.7	61.1	**68.7**	60.61	48.64
Q5	BART	**62.3**	**47.4**	**53.7**	30.52	56.22
	BART_{TAG}	61.0	46.6	53.1	35.42	**58.60**
	PEGASUS	50.4	39.4	46.0	**39.05**	48.97
	PEGASUS_{TAG}	56.3	44.8	51.2	31.59	37.24

Table 4: ROUGE F1 and named entity results of BART and PEGASUS models on our dataset. We mark the best performances with bold. BART_{TAG} and PEGASUS_{TAG} are trained on the entire dataset.

Section	R-1	R-2	R-L
$Q1$	28.00	6.55	13.48
$Q2$	32.91	8.34	14.42
$Q4$	27.34	5.83	13.84
$Q5$	18.47	4.78	10.89

Table 5: ROUGE scores between selected source segments and golden summaries.

sistency (Kryściński et al., 2019b).

Following previous work on lay summarization (Guo et al., 2022), we report the average Coleman-Liau readability score (Coleman and Liau, 1975) for the source, gold summary and model-generated summary for BART_{TAG} in Table 6. This score evaluates the simplicity of a passage, by providing an estimate of the years of education required to understand it. A lower score suggests a simpler writing style. We confirm that PLSs offer greater readability than the respective source segments. We also find that the $BART_{TAG}$ consistently exhibits readability levels are consistently closer to the desired target, reflecting its effectiveness in producing simplified versions of the source.

Section	Source	Summary	Model Summary
$Q1$	14.5	11.6	11.1
$Q2$	12.0	10.2	11.0
$Q4$	13.5	11.7	12.1
$Q5$	13.2	12.0	12.6
Average	13.3	11.4	11.7

Table 6: Coleman-Liau readability scores for source, golden and BART_{TAG} summaries.

Despite impressive ROUGE scores, we note the factual inconsistency of generated summaries, which has previously been reported by several authors as a problem in abstractive summarization (Kryściński et al., 2019b; Cao et al., 2018; Kryściński et al., 2019a). Qualitative analysis shows that this problem can be largely attributed to three reasons: i) Missing information, where identified sources do not contain all necessary information to accurately produce summary entities, ii) Typos, where entities are "mistyped", due to the model's dictionary (e.g. letters missing from a substance's name), iii) Hallucinations, where entities are made up due to biases present in the training set (e.g. stating that a study was performed in the US rather than the UK). We present representative examples for some identified causes of factual inconsistencies in Table 7 of Appendix A.2.

5. Conclusion

This work introduced the task of automatic generation of lay summaries for *clinical trials* and constructed the first related dataset to support training and evaluation. To enable the use of transformer models for this task, we proposed the division of each golden summary into thematic subsections with appropriate length. Additionally, we located the source of each section from an array of documents and proposed similarity measures as a means of improving source quality. To facilitate future research, we benchmarked our dataset with popular summarization models using several metrics and found that BART performs well on all thematic sections. Finally, we noted challenges in the form of factual inconsistency of generated summaries, attributable to both model biases and source imperfections.

6. Limitations

Although CARES utilizes all publicly available PLSs by Pfizer, it remains smaller than datasets available for other summarization tasks. This is largely attributed to plain summaries being made mandatory in recent years. Another limitation of CARES is the inclusion of summaries by a single sponsor. Although the general format is similar between trials of different sponsors, we cannot guarantee models trained on CARES will generalize well across different sponsors.

7. Bibliographical References

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

J Scott Brennen, Felix M Simon, Philip N Howard, and Rasmus Kleis Nielsen. 2020. *Types, sources, and claims of COVID-19 misinformation*. Ph.D. thesis, University of Oxford.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Philipp Mayr, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview of the first workshop on scholarly document processing (SDP). In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 1–6, Online. Association for Computational Linguistics.

Rochana Chaturvedi, Saachi ., Jaspreet Singh Dhani, Anurag Joshi, Ankush Khanna, Neha Tomar, Swagata Duari, Alka Khurana, and Vasudha Bhatnagar. 2020. Divide and conquer: From complexity to simplicity for lay summarization. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 344–355, Online. Association for Computational Linguistics.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2022. Cells: A parallel corpus for biomedical lay language generation. *arXiv preprint arXiv:2211.03818*.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2024. Retrieval augmentation of large language models for lay language generation. *Journal of Biomedical Informatics*, 149:104580.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):160–168.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, S. M. Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, Abrar Ahmad Chughtai, and Holly Seale. 2020. Covid-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103(4):1621 – 1629.

Seungwon Kim. 2020. Using pre-trained transformer for better lay summarization. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 328–335, Online. Association for Computational Linguistics.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019a. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019b. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. *arXiv preprint arXiv:2102.09130*.

Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11):1976–1988.

Tatiana Passali and Grigorios Tsoumakas. 2022. Topic-aware evaluation and transformer methods for topic-controllable summarization.

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain Marshall, Junyi Jessy Li, and Byron Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1387–1407, Toronto, Canada. Association for Computational Linguistics.

Oliver Vinzelberg, Mark David Jenkins, Gordon Morison, David McMinn, and Zoe Tieges. 2023. Lay text summarisation using natural language processing: A narrative literature review.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.

A. Appendix

A.1. Source extraction

Figure A.1 presents an example of the source selection pipeline. Initially, target sections are identified based on their titles. Subsequently, the most relevant document and subsection within it are extracted. Given that this text may contain irrelevant sentences or noise artifacts, the proposed similarity score, as defined in Equation 2, is utilized to assess the alignment between the candidate source and the target. Finally, we obtain the clean source segment by filtering out sentences that fail to reach a similarity threshold.

A.2. Hallucination examples

Table 7 contains examples for each of the identified types of model hallucinations. In the first example, the model incorrectly calculated that 5 out of 17 equates to 17%, which is inaccurate. The second example highlights a typographical error where the drug "palbociclib" was mistakenly spelled as "palbocciclib". Finally, in the third case, the model erroneously stated that a vaccine had been approved both in the United States and the European Union when, in reality, it was only approved in Europe. These errors demonstrate the importance of carefully assessing the outputs of NLP models, as they can sometimes produce inaccuracies or hallucinate information that differs from the factual reality.

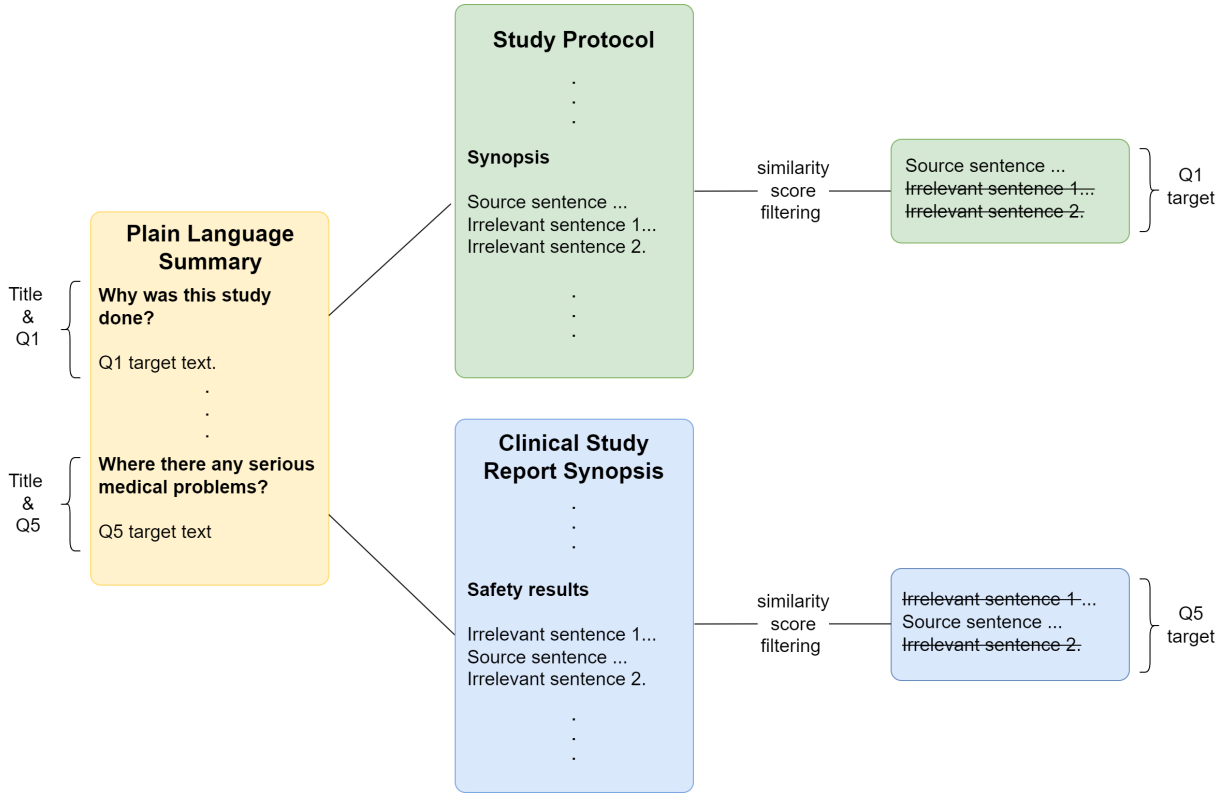


Figure 1: An example of source extraction.

Type	Example
Numerical Error	In this study, 5 out of 17 (17%) participants who received pregabalin 5 mg/kg/day had at least 1 medical problem ...
Typo	This study compared 2 groups of patients to find out if patients taking palbociclib in combination with letrozole had their cancer get better compared to patients taking a placebo ... The patients and researchers did not know who took palbocciclib...
Hallucinations	**Target summary** However, invasive meningococcal disease may be prevented with a vaccine. A vaccine is a type of medicine that helps people fight off germs. Meningococcal disease is caused by the meningococcus germ. There are different types of this germ. For example, meningococcal type a disease is caused by the meningococcus a germ. Menacwy-tt (nimenrix) is a vaccine approved in Europe for the prevention of meningococcal disease. **Model Generated Summary** However, invasive disease may be prevented with a vaccine. A vaccine is a type of medicine that helps people fight off germs. Menacwy-tt (nimenrix) is a vaccine approved in the United States, the US, and the European Union for the prevention of invasive disease.

Table 7: Error examples from BART$_{TAG}$ generated summaries. Model mistakes and hallucinations are marked in red, while the corresponding correct information is highlighted in blue.

Enhancing Lexical Complexity Prediction Through Few-Shot Learning with GPT-3

Jenny Ortiz-Zambrano[1],César Espín-Riofrío[2], Arturo Montejo-Ráez [3]
[1,2]Guayaquil University, [3]Jaen University
[1,2]Av. Delta S/N y Av. Kennedy - Guayaquil - Ecuador,
[3]Campus Las Lagunillas s/n. 23071 - Jaén - Spain
{jenny.ortizz, cesar.espinr}@edu.ec, amontejo@ujaen.es

Abstract

This paper describes an experiment to evaluate the ability of the GPT-3 language model to classify terms regarding their lexical complexity. This was achieved through the creation and evaluation of different versions of the model: text-Davinci-002 y text-Davinci-003 and prompts for few-shot learning to determine the complexity of the words. The results obtained on the CompLex dataset achieve a minimum average error of 0.0856. Although this is not better than the state of the art (which is 0.0609), it is a performing and promising approach to lexical complexity prediction without the need for model fine-tuning.

Keywords: GPT-3, Few-shot Learning, Lexical Complexity Prediction

1. Introduction

Reading involves a complex process that goes beyond coming across words or sections that are difficult for the reader to understand. Therefore, it is essential to properly understand the content of the texts to build coherent mental representations and fully understand their meaning (van den Broek, 2010).

Advancements in information technologies enable individuals to access a wealth of information across diverse domains, including education, information, social, health, government, and even scientific literature. Nonetheless, a considerable portion of the population faces obstacles in accessing this information due to significant reading challenges. These hurdles include lengthy sentences, technical jargon, and hard linguistic constructions that impede their comprehension of the text. Among those particularly impacted are individuals with intellectual disabilities and those with limited education. Surprisingly, even university students, with their advanced education and specialized knowledge in various subjects, can be part of groups struggling with reading disabilities (Alarcón García, 2022).

Lexical simplification (LS) is an automated process that substitutes words considered challenging for a particular target audience with easier alternatives while maintaining the original sentence's meaning intact. LS has an important role in Text Simplification (TS) and aims to enhance text accessibility for diverse groups of individuals (North et al., 2023a). Deep learning and, more recently, large language models (LLM) and prompt learning, have transformed our approach to various natural language processing (NLP) tasks including lexical simplification (LS) (North et al., 2023b).

The main objective of this article is to demonstrate how the Transformers GPT-3 based language model can classify text in terms of lexical complexity. This was achieved through the creation and evaluation of different versions of the model and prompts for few-shot learning to determine the complexity of the words.

This paper is organized as follows: in Section 2, a brief overview of the state-of-art is provided in complex word identification is provided. Section 3 explains GPT-3 for solving NLP tasks. Section 4 presents the experimental settings. Section 5 our solution and the results obtained with different variations on prompting are detailed. Section 6 presents a discussion about the results obtained compared to those proposed in SemEVal 2021, allowing us to present an analysis of our findings and highlight its importance and the contributions of the model in the field of predicting lexical complexity. Finally, in Section 7, conclusions and some insights on planned work are provided.

2. Previous work

Previous innovative forms of lexical simplification involved complicated systems with multiple components, each requiring extensive technical mastery and fine-tuned interaction to achieve maximum performance (Aumiller and Gertz, 2023). Recent advances in deep learning, particularly with the advent of large language models (LLMs) can be fine-tuned quickly. The high performance of these models sparked renewed interest in LS (North et al., 2023b). More advanced deep learning models,

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 68–76

such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT-3 (Brown et al., 2020) and others, are capable of automatically generating, selecting, and classifying candidate substitutions with performance superior to traditional approaches (North et al., 2023b).

With a capacity of 175 billion parameters, GPT-3 stands out for its deep knowledge of the language, its processing power, and its ability to learn from large volumes of online text data. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks (Brown et al., 2020). Due to these qualities, GPT-3 can perform a wide variety of natural language-related tasks with never-before-seen ease, including text generation and classification (Kublik and Saboo, 2022). The immense magnitude of the model allows it to produce results of high quality, precision, and diversity in the generated content. This development has raised a great deal of interest and concern in various fields, including Natural Language Processing (NLP), the machine learning industry, the media, the AI ethics communities, and society at large (Chan, 2023).

Despite being a generative model, GPT-3 can take different approaches to classify text, including zero-shot classification (where no examples are provided to the model), as well as one or few-shot classification (where some examples are presented to the model). In zero-shot learning, no prior training or adjustment to the labeled data is required. Currently, GPT-3 produces results for invisible data, but to perform zero-shot classification with GPT-3, we must provide you with a compatible prompt (Kublik and Saboo, 2022). In the few-shot learning, some examples of the task to be solved are provided. GPT-3's exceptional ability to learn in just a few tries, which is unprecedented in Natural Language Processing (NLP) models, is a prominent and notable feature (Chan, 2023).

SimpleText@CLEF-2022 Task[1] 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 (Ermakova et al., 2022). Mostert et al. (2022) 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. The findings indicate that taking text complexity into account is crucial for enhancing the accessibility of scientific information for non-experts.

Aumiller and Gertz (2023) in TSAR-2022 Shared Task on Multilingual Lexical Simplification presented two systems (Saggion et al., 2023). The initial system involved a zero-shot prompted GPT-3, where a prompt was used to request simplified synonyms based on a specific context, and the resulting simplifications were ranked. The second system was an ensemble comprising six distinct GPT-3 prompts/configurations, using average rank aggregation. Remarkably, the second system achieved the highest score for English across all metrics.

Traditional approaches are outperformed by the most advanced state-of-the-art deep learning models, such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT-3 (Brown et al., 2020) and others. GPT-3 known as Generative Pre-trained Transformer 3, is a next-generation language model based on large-scale transformers, created by OpenAI[2].

(Ortiz-Zambrano et al., 2023) participated in the CLEF 2023 Simple@Text[3] track's Task 2.1 and 2.2. In their approach, they explore zero-shot and few-shot learning strategies over the auto-regressive GPT-3 model. Several prompts to achieve those strategies were tested. The results were ranked among the top submitted runs and demonstrated a solid performance for the task of lexical complexity prediction.

(Wei et al., 2022) investigated how generating a thought sequence, composed of a series of intermediate reasoning steps, significantly improves the ability of large language models to perform complex reasoning. Specifically, we demonstrated how these reasoning abilities develop naturally in sufficiently broad language models through a simple approach called "chain-of-thought prompting" where some chain-of-thought demonstrations are provided as examples of provocation. Experiments with three large language models reveal that chain-of-thought elicitations improve performance on a variety of reasoning tasks, including arithmetic problems, common sense questions, and symbol manipulation.

According to (Zhang et al., 2022), the superior performance of the Manual-CoT approach is based on manually building proofs. For this reason, they proposed Auto-CoT as an alternative to eliminate this manual task and automatically generate demos. The Auto-CoT method presents a wide variety of questions and creates chains of reasoning to construct corresponding proofs. Results from experiments on ten publicly available reasoning datasets show that using GPT-3, Auto-CoT consistently matches or exceeds the performance of the conventional CoT approach, which requires manual proof construction.

[1] https://simpletext-project.com/2022/clef/en/task2

[2] https://openai.com/

[3] SimpleText@CLEF-2023. Available in https://simpletext-project.com/2023/clef/

3. Experimental Settings

3.1. Dataset

We used the *CompLex* corpus proposed by (Shardlow et al., 2020). *CompLex*[4] is the first English multiple domain dataset, where words are scored in a context concerning their complexity using a five-point Likert scale to label complex words in texts of three sources/domains: the Bible, Europarl and Biomedical texts. The corpus is split into single-word and multiple-word annotations. The corpus contains a total of 9,476 sentences, each annotated by approximately 7 annotators, see Table 1.

	All	Single words	Multiple words
Europarl	3,496	2,896	600
Biomed	2,960	2,480	480
Bible	3,020	2,600	420
All	9,476	7,974	1,500

Table 1: Volumetric information for the single and multiple words in each subcollection of the CompLex dataset.

Each entry contains the name of the source corpus, the sentence representing the context of the word, the targeted word to classify, and a score (in training data) that is an average of the scores given by different annotators, taking into account that each class has a predefined weight between 0 (very easy) and 1 (very difficult) in a linear distribution over possible classes, see Table 2. Sample entries are shown in Table 3. This resource was used as a benchmark collection for SemEval 2021 Task 1: Prediction of Lexical Complexity of the 15th International Workshop on Semantic Evaluation[5] (Shardlow et al., 2021).

4. Methodology

We based our work on the one presented by Mostert et al. (2022), applying GPT-3 in this text classification task, specifically, a task of identification of complex words in texts and the respective categorization of them. The experiments were carried out applying a *few-shot classification* where the purpose was to carry out an analysis of the content of the texts coming from diverse sources such as the Bible, Biomedical, and Europarl to determine that GPT-3 was able to predict the complexity degree of the word. Various experiments were carried out applying different contents in the prompt, this is because GPT-3 has no concrete way of verifying the truth, logic, or meaning of any of the millions of lines of text it generates daily. The setting for the model parameters for GPT-3 is given in Table 4.

The steps that were carried out during the development of the experimentation are the following:

- The respective model configuration was specified.
- Different prompts were built and executed with the data set in its training and evaluation phases, to ask the model to return a "soft" response in the Likert scale.
- The probabilities were obtained to know what is the priority with which the model determines its result.
- The respective evaluation metrics were calculated: MAE, MSE, and RMSE, to determine the accuracy of the results.
- The respective comparison of the results generated by the model versus the data set was performed.

4.1. Few shot prompting for Lexical Simplification

We applied the *few shot prompting* strategy by providing the model with a few samples of what we wanted it to do. In Table 5 the prompt provided to the model is shown. The prompt contains several examples that are complemented with the word to be evaluated from the text, and it also indicates which resource the text corresponds to. After prompt specification, the values are replaced by the CompLex corpus dataset with the sentence and word to be evaluated.

4.2. Construction of the different prompt

To facilitate comparison, analysis, and interpretation of the results of the model runs against the test data set assign a name for each execution as is in the Prompt Variants column, as can be seen in See Table 8. The differentiation in the construction of the various prompts intended for the GPT-3 model was based on their size, determined by the number of examples integrated for construction and training during their learning phase. This distinction was indicated by the first letter of the respective name: S to indicate small applications (Small) with an average of 2 to 4 examples, M for medium-sized ones (Medium) with an average of 5 to 6 examples, and L for those of large magnitude

[4]CompLex: Is available at https://github.com/MMU-TDMLab/CompLex

[5]Semeval 2021 - LCP SHARED TASK 2021 - https://sites.google.com/view/lcpsharedtask2021

Scale	Description	Complexity
Very Easy	Words which were very familiar to an annotator.	0
Easy	Words with which an annotator was aware of the meaning.	0.01 - 0.25
Neutral	A word which was neither difficult nor easy.	0.26 - 0.50
Difficult	Words in which an annotator was unclear of the meaning, but may have been able to infer the meaning from the sentence.	0.51 - 0.75
Very Difficult	Words that an annotator had never seen before, or were very unclear.	0.76 - 1.00

Table 2: Categories on the Likert scale proposed by (Shardlow et al., 2020)

.

Corpus	Sentence	Token	LCP score
Bible	*He sees the place of **stones**.*	stones	0.3421
Bible	*But I will stay at Ephesus until **Pentecost**,*	Pentecost	0.6250
Bible	*These are the **families** of the Levites.*	families	0.2205
Bible	*The seeds rot under their **clods**.*	clods	0.6250
Biomed	*p150CAF-1 knockdown in **ES** cells was quantified.*	ES	0.6944
Biomed	*The 2P unique region (Region I) contains an **hg***	hg	0.7500
Biomed	*on behalf of the PPE **Group**.*	Group	0.1527
Europarl	*We have taken note of your **comment**, Mr Helmer.*	comment	0.0499
Europarl	*Country **Strategy** Papers - Malaysia, Brazil*	Strategy	0.2894
Europarl	***Documents** received: see Minutes*	Documents	0.2000
Europarl	***Situation** in Darfur (vote)*	Situation	0.2115

Table 3: Examples form the CompLex dataset where the complex word is highlighted in bold.

Parameter	Values
model	text-davinci-003
prompt	orden
temperature	0
maximum tokens	5
top_p	1
presence_penalty	0
logprobs	5

Table 4: GPT-3 Model Configuration

(Large) with an average of examples between 9 and 12 examples included. This process aimed to generate multiple prompts that enable the model to offer more precise results during its evaluation. This process aimed to generate multiple prompts that enable the model to offer more precise results during its evaluation.

Next, SO comes from Source, that is, whether or not the source from which the text to be evaluated came was included in the prompt specification. We also include the *nor* operator (the result of the negation of the OR operator) and *neither* to indicate the denial of the alternatives presented, translating to “NOR” and “neither” which would mean “none” or “none”, and we have used them to express that the application does not consider any of the two previous options mentioned. See Table 9 in the *Prompt Variants* column.

4.3. Methods used to calculate the complexity level of words

To calculate the level of complexity of complex words generated by GPT-3 as a value within the range [0, 1], we explored three ways based on the categories of the complex words, as detailed below. In this way, the linguistic responses of the model are transformed into numerical values. In Table 2 the range of values that correspond to the complexity of each category was presented.

1. **Method #1 - *Middle of the range***
   Half between the lower limit and the upper limit of each range of complexity values. Scores are fixed on a per category basis. For example:

   Neutral = (0.26 + 0.50) / 2
   Neutral = 0.375

   The calculated values for each category are:

   Very Easy = 0
   Easy = 0.125
   Neutral = 0.375
   Difficult = 0.625
   Very Difficult = 0,875

   Table 6 presents in the column *Method #1* the results of an execution carried out with a total of 30 records where it can be seen that there is a large number of coincidences with the categories that correspond to the complex

```
I'm reading fragments from some sources such as the Bible, Biomed, and Europarl,
and some words are not easy to understand.  I'm classifying these words into "very
easy", "easy", "neutral", "difficult" and "very difficult".  The sentence is "neu-
tral" when it is neither "very easy", nor "easy", nor "difficult", nor "very dif-
ficult".  Several examples are:  " However, no defects in axon pathfinding along
the monosynaptic reflex arc or in muscle spindle differentiation have been noted in
PV KO mice, which develop normally and show no apparent changes in their behavior
or physical activity (Schwaller et al.  1999).  ".  I find that word "spindle" is
neutral
###
The following fragment comes from the "bible" and after reading the fragment " I
will sprinkle clean water on you, and you shall be clean:  from all your filthiness,
and from all your idols, will I cleanse you.  ".  I find that the word "filthiness"
is easy
###
The following fragment comes from the "biomed" and after reading the fragment "
Moreover, acute dosing does not recapitulate the marked learning deficits produced
in rodents [15,16] by chronic exposure to dopamine D2R antagonists [6,7] ".  I find
that the word "antagonists" is difficult
###
The following fragment comes from the "biomed" and after reading the fragment "
Thrombus formation on fissured atherosclerotic plaques is the precipitating event
in the transition from a stable or subclinical atherosclerotic disease and leads to
acute myocardial infarction, ischemic stroke or peripheral arterial occlusion.  ".
I find that word "Thrombus" is very difficult
###
The following fragment comes from the "bible" and after reading the fragment " Mount
Sinai, all it, smoked, because Yahweh descended on it in fire; and its smoke as-
cended like the smoke of a furnace, and the whole mountain quaked greatly.  ".  I
find that the word "fire" is very easy
###
The following fragment comes from the @recurso and after reading the fragment @ora-
cion I find that word @aEvaluar is
```

Table 5: Prompt example

word of the CompLex corpus. The highlighted values correspond to matches. After the execution with the test data, it was obtained a MAE=0.1293, MSE=0.0258, RMSE=0.1608.

2. **Method #2 - *Average by category***
   The average of complexity values of a category is used as the complexity degree for that category. Therefore, the scores are fixed on a per category basis. Again, scores are fixed on a per category basis. The table 6 presents in the *Method #2* column the results of the execution carried out with the test records, a total of 30, where you can see the value calculated for the level of complexity of the categories generated by the model for the complex words of the texts. After the execution with the test data, it was obtained a MAE=0.086, MSE=0.016, RMSE=0.125.

   The corresponding calculated values for each category would be the following:

   Very Easy = 0
   Easy = 0.189
   Neutral = 0.351
   Difficult = 0.588
   Very Difficult = 0.811

3. **Method #3 - *The Confidence of GPT-3***
   The Confidence Level of the model corresponds to the high percentage of precision and coherence with which the model has made use of its attention mechanism and the context to select the category to which the complex word in the text corresponds.

   We consider for the assignment of the category generated by the model the one whose confidence level is the highest. For example: If the confidence level is 90% for the Easy category, the complexity closest to the left limit of the category range is taken. If the confidence level is 80% for the Difficult category, the complexity level that is furthest to the right of the category range is assigned. In the case of Very Difficult, the same procedure as the previous ones is considered. The table 6 presents in the Method #3 column the results of the execution carried out with the test records, a total of 30, where it can be seen, the value calculated

for the level of complexity of the categories generated by the model for the words complexities of the texts according to the level of confidence of the model. After the execution with the test data, a MAE=0.191, MSE=0.047, RMSE=0.216 was obtained.

It is important to note that in the table 6, in the "Category" column of the GPT-3 section, the complexity category generated by the GPT-3 model was selected based on the highest level. high confidence in percentage terms of the probabilities associated with the predictions generated by the model that corresponds to the category of the complex word. This is observed in the execution of the model on the corpus, using a test data set composed of 30 records, as detailed in the table 7.

5. Results

Our goal is to advance research on the use of the GPT-3 model to predict word complexity in the English language by adopting a few-shot examples learning approach. We have carried out multiple iterations with the objective that GPT-3 generates more precise and coherent answers with quality and relevance, we have formulated 19 several prompts pretending to optimize the performance of the model. Through this approach, we aspire to achieve greater precision in our predictions, approaching the results obtained by the winners of the lexical complexity prediction task proposed in the framework of SemEval 2021[6].

We experimented with different prompts issued to OpenAI's largest available model: text-davinici-002 and davinici-003 as evidenced by Table 9. Our first approach uses a singular prompt template in a few-shot setting to obtain the category of word complexity: *easy - very easy - neutral - difficult - very difficult*; we further improve upon these results by combining predictions from different prompt templates as can be seen in the Table 8, the application of different runs performed with the evaluation data set. The results derived from our approach toward single word prediction yielded the following values: MAE = 0.0875, MSE = 0.0131, and R2 = 0.1930.

A test was carried out by taking a sample of 30 records to train the model applying the few-shot learning technique. The data in the column *GPT-3 Confidence Level* represents the level of lexical complexity generated by the GPT-3 model for each token in the corpus. In the table 9, we can see that the "Score Type" column in the last three rows shows runs where this strategy was applied to assign complexity levels to complex words in the corpus. This is complemented by the results presented in the table 10 in the "GPT-3 Complexity" column, which refers to the complexity generated by the model. It is from these values that the strategy for calculating lexical complexity was derived, called *GPT-3 Confidence Level*. Additionally, the table **??** shows matches where the model's complexity prediction for a token matches the complexity assigned to the token in the CompLex corpus.

[6] https://sites.google.com/view/lcpsharedtask2021

6. Discussion

In this article, we present a system proposal to resolve the task of lexical complexity prediction. Table 11 shows the results achieved by the first five classified in the evaluation carried out by the organizing entity (Shardlow et al., 2021). It is important to note that the competition involved a large number of participants, specifically 54 teams. In contrast to the performance of the first-place winner, who achieved an MAE of 0.0609, we see relatively little difference in our results in terms of the linguistic categories considered. This fact gives a dose of confidence to our approach, which, despite its simplicity, proved to be competitive compared to the proposals of several teams that opted for more complex approaches. Among these more complex approaches, the use of deep neural networks such as the BERT and ROBERTa models stands out, evidenced in teams such as JUST BLUE, RG PA, Andi, CS-UM6P, OCHADAI-KYOTO, to mention just a few examples. It is worth mentioning that only one team used a GPT model (GPT-2) in their approach.

The results generated with GPT-3 would have reached an MAE = 0.0882 as presented in Table 9. It should be noted that when running the GPT-3 model, the approach *few-shot learning* used 4 to 6 examples in various experiments so that the model can learn and then generate its response.

The best result is achieved by using the combination M-SO-05, which corresponds to 5 examples sent to the model. This practice is highly beneficial to the model, as it allows it to generate more accurate predictions. To evaluate performance, the *Means* type score was used for the *davinci-003* model which yielded the following results: MAE=0.0882, MSE=0.0136, RMSE=0.1165, and Pearson=0.5776. These indicators highlight the effectiveness of the strategy used and the model's ability to provide high-quality results. The results achieved are very encouraging since they show that the model can understand the requests made by humans in a considerable way and without much effort as when applying other models.

7. Conclusions and future Works

Using GPT-3 to classify complex words involves finding a balance between your capacity and ability

Comparing the results applying GPT-3 with Few-Shot learning in corpus CompLex							
#	Corpus CompLex			GPT-3			
	Complex word	Category	Range of Values	Category	Complexity Level		
					Method #1	*Method #2*	*Method #3*
10	voice	neutral	0.01 - 0.25	easy	0.125	0.189	87.07% - 0.032
11	darkness	**easy**	0.01 - 0.25	**easy**	0.125	0.189	76.91% - 0.058
12	behold	easy	0.26 - 0.50	neutral	0.375	0.351	29.40% - 0.381
13	camp	**easy**	0.01 - 0.25	**easy**	0.125	0.189	81.11% - 0.045
14	bonds	**easy**	0.01 - 0.25	**easy**	0.125	0.189	54.29% - 0.115
15	statutes	neutral	0.01 - 0.25	easy	0.125	0.189	51.43% - 0.127
16	snares	**easy**	0.01 - 0.25	**easy**	0.125	0.189	54.95% - 0.112
17	exhortation	**difficult**	0.51 - 0.75	**difficult**	0.189	0.588	61.30% - 0.665
18	River	**easy**	0.01 - 0.25	**easy**	0.125	0.189	86.88% - 0.033
19	generation	**easy**	0.01 - 0.25	**easy**	0.125	0.189	85.27% - 0.037
20	dainties	**difficult**	0.51 - 0.75	**difficult**	0.189	0.588	58.36% - 0.657

Table 6: Methods applied to calculate the level of lexical complexity.

Probabilities associated with the complexity category predictions generated by the GPT-3 model The results applying the *davinci-002* model and *few shots learning* approach						
#	Token	GPT-3 complexity	CompLex complexity	GPT-3 confidence level		
				Option #1	Option #2	Option #3
20	dainties	difficult	difficult	difficult 59.39%	neutral: 25.25%	easy: 10.23%
21	subjection	difficult	neutral	difficult: 92.44%	neutral: 5.0%	easy: 1.16%
22	perverseness	difficult	neutral	difficult: 92.35%	neutral: 5.26%	very: 1.3%
23	grasshoppers	easy	easy	easy: 72.01%	neutral: 13.88%	very: 9.34%
24	signet	difficult	neutral	difficult: 74.86%	neutral: 20.39%	very: 2.74%
25	snare	easy	neutral	easy: 65.44%	neutral: 18.25%	difficult: 13.4%
26	Asher	easy	neutral	easy: 76.68%	very: 9.64%	difficult: 7.44%
27	demons	difficult	easy	difficult: 59.93%	easy: 30.62%	neutral: 7.4%
28	prophet	easy	easy	easy: 88.42%	neutral: 5.55%	difficult: 3.21%
29	lion	easy	neutral	easy: 90.23%	very: 4.46%	neutral: 4.32%
30	Lion	easy	easy	easy: 84.1%	very: 12.29%	difficult: 2.07%

Table 7: Probabilities associated with the predictions generated by the GPT-3 model that correspond to the category of the complex word.

Standard applied for the construction of the *Prompt Variants*				
Size	**Source**	**Connector Logical**	**# exp**	**Emphasis**
L-M-S	**SO**	**NOR**	**05**	**Em**
M	SO		05	
M	SO	NOR	05	Em
M	SO		06	
S		NOR	04	
S		NOR	04	Em
S	SO	NOR	05	
L	SO	NOR	09	

Table 8: Standard applied for the construction of the prompt variants.

to obtain more accurate results. The result opens new perspectives in the investigation of lexical complexity. Several experiments were carried out running various prompts, a few-shot with various models of the GPT-3 Family. We have applied three strategies to calculate the level of complexity of the complex words applied in the SemEval 2021 data set. Furthermore, we found some responses where learning from a few GPT-3 examples still presents difficulties, the responses generated by the model did not match the data sets in the work proposed by (Brown et al., 2020).

The best result was generated by the text-Davinci-003 model with an MAE of 0.0882. The model has been able to interpret and generate its responses based on a few examples and complex instructions, demonstrating that the text-Davinci-003 version provides better results than text-Davinci-002.

Nowadays, GPT-3 has been intensively used and tested on many different tasks using zero-shot and few-shot learning (Huang et al., 2023). Some of them found that this model is not that good. As new models are appearing, we plan to explore how these new models Claude 2 (Wu et al., 2023), GPT-4, or LLaMA 2 (Fan et al., 2023) perform on lexical complexity prediction.

Besides, an interesting research topic is to study

Final results generated with GPT-3 The results applying the *davinci-002* and *davinci-003* models and *few shots learning* approach							
#	**Prompt Variants**	**Score Type**	**Model Version**	**Metrics**			
				MAE	**MSE**	**RMSE**	**Pearson**
1	**M-SO-05**	**Means**	**davinci-003**	**0.0882**	**0.0136**	**0.1165**	**0.5776**
2	M-SONOR-05-Em	Means	davinci-003	0.0956	0.0153	0.1238	0.5103
3	M-SONOR-05-Em	Means	davinci-002	0.1011	0.0170	0.1305	0.4661
4	S-SONOR-05	Means	davinci-003	0.1057	0.0190	0.1378	0.5016
5	M-SO-06	Means	davinci-003	0.1074	0.0199	0.1412	0.4924
6	S-SONOR-05	Means	davinci-002	0.1098	0.0208	0.1442	0.5086
7	S-NOR-04	Means	davinci-002	0,1143	0.0229	0.1512	0.4919
8	S-NOR-04	Means	davinci-003	0.1725	0.0440	0.2099	0.3826
9	S-NOR-04-Em	Means	davinci-002	0.1793	0.0512	0.2262	0.3524
10	S-NOR-04-Em	Means	davinci-003	0.1875	0.0503	0.2242	0.4477
11	M-SO-05	Half of the range	davinci-003	0.1212	0.0219	0.1480	0.5730
12	M-SONOR-05-Em	Half of the range	davinci-003	0.1292	0.0239	0.1546	0.5099
13	S-SONOR-05	Half of the range	davinci-003	0.1475	0.0310	0.1761	0.5136
14	M-SO-06	Half of the range	davinci-003	0.1555	0.0345	0.1859	0.4944
15	S-NOR-04-Em	Half of the range	davinci-003	0.2164	0.0603	0.2456	0.4650
16	S-NOR-04	Half of the range	davinci-003	0.2106	0.0580	0.2409	0.3806
17	S-SONOR-05	GPT-3 Confidence level	davinci-003	0.2333	0.0655	0.2559	0.5600
18	M-SO-06	GPT-3 Confidence level	davinci-003	0.2658	0.0816	0.2857	0.5247
19	L-SONOR-09	GPT-3 Confidence level	davinci-003	0.2431	0.0708	0.2662	0.5241

Table 9: The results applying the davinci-002 and davinci-003 models and few shots learning approach.

Probabilities associated with the level of complexity predictions generated by the GPT-3 model The results applying the davinci-002 model and few shots learning approach							
#	**Token**	**GPT3 category**	**GPT3 range**	**GPT3 complexity**	**CompLex complexity**	**CompLex range**	**Match**
20	dainties	**difficult**	0.51 - 0.75	0.5880	0.5625	**difficult**	**Yes**
21	subjection	difficult	0.51 - 0.75	0.5880	0.4375	neutral	No
22	perverseness	difficult	0.51 - 0.75	0.5880	0.4166	neutral	No
23	grasshoppers	easy	0.01 - 0.25	0.1896	0.25	easy	Yes
24	signet	difficult	0.51 - 0.75	0.5880	0.4687	neutral	No
25	snare	easy	0.01 - 0.25	0.1896	0.3194	neutral	No
26	Asher	easy	0.01 - 0.25	0.1896	0.4285	neutral	No
27	demons	difficult	0.51 - 0.75	0.5880	0.125	easy	No
28	prophet	easy	0.01 - 0.25	0.1896	0.2222	easy	Yes
29	lion	easy	0.01 - 0.25	0.1896	0.2812	neutral	No
30	Lion	easy	0.01 - 0.25	0.1896	0.1710	easy	yes

Table 10: The results of the probabilities associated with the level of complexity predictions generated by the GPT-3 model applying the davinci-002 and few shots learning approach.

#	Team Name	MAE	MSE	R^2
1	JUST_Blue	0.0609	0.0062	0.6172
2	DeepBlueAI	0.0610	0.0061	0.6210
3	OCHADAI-KYOTO	0.0617	0.0065	0.6015
4	ia pucp	0.0618	0.0066	0.5929
5	Alejandro M.	0.0619	0.0064	0.6062
	FSL with GPT-3	0.0882	0.0136	0.1613

Table 11: Results achieved by the first five classified in the SemEval 2021 International workshop.

how large language models learn about "metalinguistic" knowledge, such as lexical complexity. Is it inferred from the enormous collection of texts due to explicit references to complexity? Is it, instead, a knowledge that "emerges" from the comprehension of language itself? These are captivating questions that, in the era of large language models, could be considered central for current research in lexical complexity prediction.

8. Bibliographical References

Rodrigo Alarcón García. 2022. Lexical simplification for the systematic support of cog-

nitive accessibility guidelines. *https://doi.org/10.1145/3471391.3471400*.

Dennis Aumiller and Michael Gertz. 2023. Unihd at tsar-2022 shared task: Is compute all we need for lexical simplification. *arXiv preprint arXiv:2301.01764*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Anastasia Chan. 2023. Gpt-3 and instructgpt: technological dystopianism, utopianism, and "contextual" perspectives in ai ethics and industry. *AI and Ethics*, 3(1):53–64.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Liana Ermakova, Irina Ovchinnikova, Jaap Kamps, Diana Nurbakova, Sílvia Araújo, and Radia Hannachi. 2022. Overview of the clef 2022 simpletext task 2: complexity spotting in scientific abstracts. In *Proceedings of the Working Notes of CLEF 2022-Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th-to-8th, 2022*.

Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046*.

Cynthia Huang, Yuqing Xie, Zhiying Jiang, Jimmy Lin, and Ming Li. 2023. Approximating human-like few-shot learning with gpt-based compression.

Sandra Kublik and Shubham Saboo. 2022. Gpt-3: Building innovative nlp products using large language models. *O'Reilly Media*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Femke Mostert, Ashmita Sampatsing, Mink Spronk, and J Kamps. 2022. University of amsterdam at the clef 2022 simpletext track. *Proceedings of the Working Notes of CLEF*.

Kai North, Alphaeus Dmonte, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023a. Alexsis+: Improving substitute generation and selection for lexical simplification with information retrieval. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 404–413.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023b. Deep learning approaches to lexical simplification: A survey. *arXiv preprint arXiv:2305.12000*.

Jenny Ortiz-Zambrano, César Espin-Riofrio, and Arturo Montejo-Ráez. 2023. Sinai participation in simpletext task 2 at clef 2023: Gpt-3 in lexical complexity prediction for general audience.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2023. Findings of the tsar-2022 shared task on multilingual lexical simplification.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. Complex: A new corpus for lexical complexity predicition from likert scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. *arXiv preprint arXiv:2106.00473*.

Paul van den Broek. 2010. Using texts in science education: Cognitive processes and knowledge representation. *Science (New York, N.Y.)*, 328:453–6.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Sean Wu, Michael Koo, Lesley Blum, Andy Black, Liyo Kao, Fabien Scalzo, and Ira Kurtz. 2023. A comparative study of open-source large language models, gpt-4 and claude 2: Multiple-choice test taking in nephrology. *arXiv preprint arXiv:2308.04709*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models.

An Approach Towards Unsupervised Text Simplification on Paragraph-Level for German Texts

Leon Fruth, Robin Jegan, Andreas Henrich
University of Bamberg
An der Weberei 5, 96047 Bamberg, Germany
{leon.fruth, robin.jegan, andreas.henrich}@uni-bamberg.de

Abstract

Text simplification as a research field has received attention in recent years for English and other languages, however, German text simplification techniques are lacking thus far. We present an unsupervised simplification approach for German texts using reinforcement learning (self-critical sequence training). Our main contributions are the adaption of an existing method for English, the selection and creation of German corpora for this task and the customization of rewards for particular aspects of the German language. In our paper, we describe our system and an evaluation, including still present issues and problems due to the complexity of the German language, as well as directions for future research.

1. Introduction

Automatic text simplification (ATS) is a research field in computational linguistics. The objective of text simplification is the modification of texts in a way to make them simpler to read and understand for the target audience. Thus, helping people with low literacy levels, mentally impaired people and children (Al-Thanyyan and Azmi, 2022; Evans et al., 2014; Watanabe et al., 2009). It is closely related to other natural language processing (NLP) tasks such as text summarization.

With the advancements in deep learning, recent research addresses ATS as a mono-lingual machine translation problem (Mallinson et al., 2020): Translating a text with complex linguistic properties into a text with simple linguistic properties in the same language. For this, large-scale simplification datasets are needed. Such parallel datasets are not widely available for most languages, including German.

This work uses the approach from Laban et al. (2021) (referred to as *KiS* in this work) and adapts it to the German language. The approach bypasses the need for parallel datasets by using training based on reinforcement learning (RL) (Sutton and Barto, 2018) and rewards regarding the criteria simplicity, meaning preservation and fluency, that are jointly optimized. Since German text simplification data is limited, the dependency on a large parallel simplification dataset is circumvented using this training method. This work presents the first unsupervised ATS approach for German and one of the first, to the authors' knowledge, that simplifies on a paragraph-level. Source code, model, datasets and evaluation data are available under https://github.com/LFruth/unsupervised-german-ts.

2. Background

Linguistic complexity, a key objective in ATS, consists of lexical simplicity, replacing difficult words with simpler expressions (Carroll et al., 1998; Laban et al., 2021; Keskisärkkä, 2012), and syntactic simplicity, rewriting texts into simpler and more understandable sentences (Saggion, 2017; Alva-Manchego et al., 2019).

ATS can also be addressed through the lens of machine translation (MT), where a complex text is translated into a text of the same language with simpler linguistic properties (Coster and Kauchak, 2011; Specia, 2010).

With the introduction of transformer-based models and large-scale parallel simplification corpora such as WikiLarge (Zhang and Lapata, 2017) and Newsela (Xu et al., 2015) new approaches like ACCESS (Martin et al., 2019) have been proposed. For instance, ACCESS (Martin et al., 2019) presents a sentence simplification methodology wherein the authors introduced a parametrization mechanism to control the compression rate, the paraphrase amount, and the strength of lexical and syntactic simplification. While there also exist some larger datasets for Spanish text simplification (Agrawal and Carpuat, 2019), other languages, including German, only have very limited parallel datasets that are mostly insufficient to train a simplification model in a MT fashion (Naderi et al., 2019; Battisti et al., 2020; Rios et al., 2021; Säuberli et al., 2020; Spring et al., 2021).

An early approach for German used rule-based simplification (Suter et al., 2016), whereas another method chose a zero-shot cross-lingual technique, that was implemented to handle the lack in datasets (Mallinson et al., 2020). Reinforcement learning is applied in unsupervised models, e.g. in Zhang

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 77–89

and Lapata (2017), using the framework REINFORCE (Williams, 1992), also deployed in Nakamachi et al. (2020) with an LSTM encoder-decoder model. Newer approaches such as Anschütz et al. (2023) using style-specific pre-training also work on the lack in parallel data.

3. Method

In the following section, we describe our model architecture GUTS, short for **G**erman **U**nsupervised **T**ext **S**implification. We followed the work of KiS from Laban et al. (2021) and adapted it to simplify German paragraphs. We used the same training method k-SCST, an extension of self-critical sequence training (SCST) (Rennie et al., 2017), which is based on the REINFORCE algorithm (Williams, 1992).

3.1. Architecture

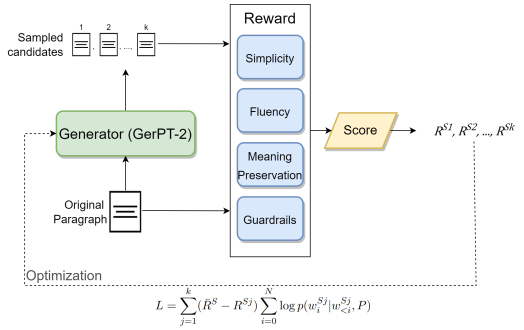


Figure 1: GUTS Learning architecture with k-SCST

Figure 1 displays how the generator learns: First k simplification candidates are sampled from the generator model, conditioned on an original paragraph. These candidates are then scored according to the reward. From the resulting rewards $R^{S1}, R^{S2}, ..., R^{Sk}$ the mean reward $\bar{R}^S$ is calculated as a baseline for the loss. The loss is computed as the difference between individual candidate rewards and the baseline, with candidates having rewards above the baseline contributing more to optimization. The probability of generating a word $p(w_i^{Sj}|...)$ is conditioned on the input paragraph P and previously generated words $w_{<i}^{Sj} = w_1^{Sj}, ...w_{i-1}^{Sj}$.

3.2. Rewards

The approach used in this work – adopted from (Laban et al., 2021) – can be described as a non-differentiable reward maximization problem. For each original paragraph P and its corresponding generated simplification S, scores in the range of $[0, 1]$ are obtained for simplicity, meaning preservation, fluency, and some guardrails. These individual reward scores are then combined into a single reward using a scoring function.

$$R = \sum_{i=0}^{N} W_i \log(s_i) \tag{1}$$

Here R denotes the total reward for a simplification. N is the number of individual scores of the reward, and s_i describes an individual score with its assigned weight W_i. This way, not every score has the same impact on the overall reward. A drawback of this scoring function is that the guardrail scores cannot zero the score since $log(0)$ is undefined. To work around this, these scores are either set to 0.0001 or 0.9999 instead.

We used the reward scores from Laban et al. (2021) and adapted them to the German language. In the following, only the scores that function differently are explained. Small changes and adaptations of the other scores are outlined in A.2.

3.2.1. Meaning Preservation

To measure how well the meaning is preserved in the generated simplification, a novel approach is presented. First, each sentence from the simplification S is aligned to the most similar sentence from the original paragraph P using sentence transformer representations (Reimers and Gurevych, 2019). By aligning these sentences, operations like sentence splitting are considered. The aligned sentences are then compared and scored with BERTScore (Zhang et al., 2020), to make use of contextual similarity between them and consider synonymity. For every sentence of P, the F1 BERTScore is computed for the aligned sentences of S. $s_{meaning}$ is calculated as follows:

$$\frac{\sum_{(se^P, se^S) \in aligned} F_{BERT}(se^P, se^S)}{|aligned| + |unaligned|} \tag{2}$$

where $aligned$ denotes the set of aligned sentence pairs from the original paragraph and the system's simplification, with the original se^P and simplified sentence se^S. The sum of the F1 BERTScores of each sentence-pair is divided by the number of aligned sentences $|aligned|$ and unaligned sentences $|unaligned|$. The set of unaligned sentences contains original and simplified sentences that were not semantically related to another sentence. Sentences from P that had no matching simplification sentence are penalized because it is assumed that information was lost during simplification. Unaligned simplified sentences that were not semantically related to any original sentence are also penalized since they are assumed to contain unnecessary or hallucinated content.

3.2.2. Hallucination Detection

A common problem for text generation tasks like ATS or summarization are factual inconsistencies. An important requirement for these tasks is that the facts from the generated text match the source text (Fischer, 2021), also referred to as faithfulness (Cao et al., 2018). In this work, we only focus on detecting the addition of named entities. First, all named entities from the generated simplification are extracted. Second, the BERTScore library (Zhang et al., 2020) is used to obtain the words from P with the highest similarity to each extracted entity. Next, the similarity value from the most related word in the original paragraph is selected for each detected entity. This value is then compared against a threshold. If the BERTScore similarity falls below this threshold, a hallucination is detected and the score $s_{hallucination}$ returns 0. Otherwise, it returns 1. Figure 3 in the appendix shows an example of the described score.

3.2.3. Article Repetition Penalty

To counter the cheating of the language model fluency score described in section A.2.3, another guardrail score was introduced that detects and penalizes the repetition of German articles like “der”, “die”, “das”. This score was introduced for this approach since the generator was abusing the repetition of high probable articles to artifically increase $s_{fluency}$. The score is set to 0 if three or more articles appear in a sequence, else it returns 1.

4. Experiments

For the generator, a German version of the medium GPT-2 model GerPT-2 (Minixhofer, 2020) was used for the experiments. More details about the training process are presented in the appendix.

4.1. Data

To test and tune the parameters of the reward scores two datasets have been used as a reference. We used the *TextComplexityDE dataset* (Naderi et al., 2019), which contains a total of 1019 sentences with simplifications, and a manually collected dataset of parallel articles from the website “Gemeinnützige Werkstätten und Wohnstätten” (Gemeinnützige Werkstätten und Wohnstätten - GWW, 2023). The latter is referred to as the *GWW dataset* and was created for this work. The GWW dataset was manually created by the authors for this work by aligning original articles with their simplified versions from the website. The dataset consists of 52 parallel articles, mainly texts for disabled people containing information and help about topics like work or living.

For the training of the generator, a dataset of short paragraphs extracted from Wikipedia articles has been generated. The raw Wikipedia articles were extracted from German Wikipedia dumps.

For the evaluation we used a dataset based on TextComplexityDE that was manually assembled, where the authors combined individual sentences to create 52 paragraphs. Besides the TextComplexityDE dataset, 300 paragraphs from the training dataset of Wikipedia articles have been randomly selected. This subset contains articles that are linguistically more diverse and difficult than those from the TextComplexityDE dataset, but have no reference simplification.

4.2. Evaluation

Since there were no comparable German models available that can simplify on a paragraph-level, a *Pivot model* is introduced for evaluation, consisting of two machine translation models (Tiedemann and Thottingal, 2020) and one simplification model (Laban et al., 2021). This Pivot model is inspired by a similar model introduced by (Mallinson et al., 2020), which the authors used as a comparison in their evaluation. First, the paragraph is translated from German to English (de-en). The KiS model can then simplify the English paragraph, before it is translated back to German by the second translation model (en-de).

Because there is no single agreed-upon measurement for simplicity (Alva-Manchego et al., 2021), a combination of reference-based and reference-less metrics has been used. SARI was integrated as a reference-based simplification metric. SARI showed the best correlation with human judgements on simplicity gain compared to other automatic metrics (Alva-Manchego et al., 2021). To measure the syntactic simplicity, the *Flesch Reading Ease* (FRE) for German has been used (Amstad, 1978). The mean FRE of the models' outputs *FRE(S)* and the average difference between the FRE value of the original text and the simplification, referred to as *FRE diff*, are calculated. For measuring the lexical simplicity improvement *Zipf diff*, the difference of the average Zipf values of all non-stop words between the original paragraph P and the simplification S are calculated. The score $s_{meaning}$ is used to capture the meaning adequacy of the simplifications. With this score, the models are rated on how well the contents from the original paragraph are preserved. Lastly, the compression rate (*Comp.*) is measured.

Table 1 displays the automatic results on the adapted TextComplexityDE dataset and on the Wikipedia paragraphs. On the TextComplexityDE dataset GUTS is slightly outperformed by the Pivot model on SARI. Both models improve on FRE and achieve, arguably, therefore syntactic simplification.

TextComplexityDE						
Model	SARI	FRE(S)	FRE diff	Zipf diff	Meaning	Comp.
manual reference	-	46.847	21.194	0.274	0.896	0.933
GUTS	0.348	37.448	11.795	0.059	**0.875**	0.789
Pivot	**0.370**	**38.712**	**13.059**	**0.206**	0.727	0.863
Wikipedia Paragraphs						
GUTS	-	**53.130**	**9.376**	-0.001	**0.819**	0.731
Pivot	-	50.187	6.402	**0.243**	0.549	0.766

Table 1: Automatic results of TextComplexityDE and Wikipedia

The Pivot model outperforms GUTS on both metrics. Both models performed reasonably well on meaning preservation. GUTS even comes close to the reference baseline, since it was directly trained on this score. The Pivot model lags behind in this area, indicating that its simplifications did not capture as much information from the original paragraph, according to $s_{meaning}$. All models tend to shorten the texts during simplification, shown by the compression values.

The evaluation on the Wikipedia paragraphs is performed with only reference-less metrics. GUTS achieves better FRE values than the pivot model for this dataset, but has worse results on the Zipf scores, showing no gain for this metric. GUTS achieves the best meaning preservation scores on this dataset.

To further evaluate the performance of GUTS, a limited manual evaluation has been conducted outlined in the following section.

4.3. Observations

In the following, the simplifications produced by the models are manually evaluated. Note that these are observations by the authors, focusing on simplification phenomena and common problems with GUTS. This is done to guide future work to improve the system.

4.3.1. Simplification Phenomena

With GUTS, some lexical simplifications in the form of substitutions with synonyms could be observed, but most of the examples were not necessarily simpler. Sometimes words that do not exist in German were used as substitutes. Many lexical changes in simplifications were not synonyms but involved shortening of words. A part of a composed word was deleted during simplification and the rest was kept. This sometimes resulted in arguably simpler words without changing the content of the text. For instance, GUTS replaced the word “Schlossräume” (English: “palace rooms”), with a Zipf value of 1.08, with “Räume” (English: “rooms”) with a value of 4.4, indicating a lexical simplification that did not significantly change the meaning of the sentence. Most of these word shortenings removed important information from the sentence and result in a misleading simplification. For example: The word “Präsidentenflugzeug” (English: “presidential plane”) was reduced to only “Präsident” (English: “president”).

For structural changes of the paragraphs rarely any sentence splittings were observed with GUTS or the Pivot model. Both models tend to delete parts of the text to make shorter sentences rather than splitting them. In many observations the arguably most important statement of the sentence is preserved. Deletions can help the reader understand texts better by removing non-essential information that may be confusing to a low literacy reader.

4.3.2. Problems

Guaranteeing the fluency and readability of a text is one of the most critical aspects of natural language generation tasks such as text simplification. One big limitation of GUTS were non-fluent text and grammatical mistakes in the generated simplifications that occurred in most of the evaluated outputs. Many of these were minor errors, like confusing German articles, e.g. using “das” instead of “der” or making mistakes with the tense of a word, for example, using the present instead of past tense. GUTS regularly produced some of the previously mentioned grammatical issues but rarely had completely incoherent outputs. The Pivot model showed the least amount of grammatical mistakes.

Another common issue were problems with faithfulness. Factual inconsistencies between source and generated texts were frequently observed with the simplifications of GUTS. One of the most common inconsistencies were numeric values, such as dates or measurements. These inconsistencies with numbers were not considered by any scoring method for the reward. For future work, the score for hallucination detection $s_{hallucination}$ could be extended to take numbers and dates into account, like Laban et al. (2021) did in their approach. The results of the Pivot model rarely contained the faithfulness issues from above. However, it rather introduced new sentences or phrases to the simplification, that were hallucinated.

5. Discussion

To bypass the data scarcity for German text simplification datasets, this work showed the first unsupervised text simplification approach for the German language. Furthermore the system is able to simplify on a paragraph-level. While many simplification phenomena happen on a paragraph-level (Alva-Manchego et al., 2019), most of the previous research on ATS has been performed on a sentence-level.

Another contribution in this work has been the novel hallucination detection method. This method is arguably implemented more dynamically than the implementation in KiS (Laban et al., 2021, §3.4.2), which directly matches the named entities in the source text and the generated text. However, their score also identifies false and hallucinated numeric values that our scoring function $s_{hallucination}$ could not do.

The meaning preservation score in this work is also a novel contribution. The score in this work presents a combination of sentence alignment and similarity measuring using BERTScore (Zhang et al., 2020), in order to rate how well the content of the original paragraph is preserved in the generated simplification.

Different problems and limitations were detected during the analysis of the rewards, the conducted experiments, and the evaluation of GUTS. For further exploration of the approach presented in this work, different parameters and settings need to be explored. Also, the individual reward scores should be investigated and improved further. GUTS lacked lexical and syntactic simplification phenomena, e.g. simpler vocabulary or sentence splitting.

Grammatical mistakes and non-fluent samples during the experiments were also an issue in this work. This is one of the most important criteria and needs to be reliable for an ATS system. Unfortunately, there is no research for measuring fluency of German texts to the authors' knowledge.

Non-factual content in the produced simplifications was another dominant issue with GUTS. This limitation is an ongoing research field for text generation tasks, such as summarization (Fischer, 2021; Cao et al., 2018; Falke, 2019). The GUTS model regularly generated simplifications with incorrect numbers and dates. Furthermore, the models sometimes even introduced hallucinations to the simplifications, which led to disinformation.

We hope that future research addresses the problems and challenges identified in this work by building upon this contribution.

6. Bibliographical References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1549–1564. Association for Computational Linguistics.

Suha Al-Thanyyan and Aqil M. Azmi. 2022. Automated text simplification: A survey. *ACM Comput. Surv.*, 54(2):43:1–43:36.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2019. Cross-sentence transformations in text simplification. In *Proceedings of the 2019 Workshop on Widening NLP@ACL 2019, Florence, Italy, July 28, 2019*, pages 181–184. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Comput. Linguistics*, 47(4):861–889.

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Abhandlung: Philosophische Fakultät I. Zürich. 1977. Studenten-Schreib-Service.

Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, and Georg Groh. 2023. Language models for german text simplification: Overcoming parallel data scarcity through style-specific pre-training.

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. A corpus for automatic readability assessment and text simplification of german. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 3302–3311. European Language Resources Association.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Hunter M Breland. 1996. Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science*, 7(2):96–99.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4784–4791. AAAI Press.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. *Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Richard Evans, Constantin Orasan, and Iustin Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations, PITR@EACL 2014, Gothenburg, Sweden, April 27, 2014*, pages 131–140. Association for Computational Linguistics.

Tobias Falke. 2019. *Automatic Structured Text Summarization with Concept Maps*. Ph.D. thesis, Darmstadt University of Technology, Germany.

Tim Fischer. 2021. *Finding Factual Inconsistencies in Abstractive Summaries*. Ph.D. thesis, Universität Hamburg.

Gemeinnützige Werkstätten und Wohnstätten - GWW. 2023. Gemeinnützige Werkstätten und Wohnstätten - GWW. https://www.gww-netz.de/de/. Last visited on May 22, 2023.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Robin Keskisärkkä. 2012. Automatic text simplification via synonym replacement. Master's thesis, Linköping UniversityLinköping University, Department of Computer and Information Science, Faculty of Arts and Sciences.

Philippe Laban, Andrew Hsi, John F. Canny, and Marti A. Hearst. 2020. The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5135–5150. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6365–6378. Association for Computational Linguistics.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5109–5126. Association for Computational Linguistics.

Louis Martin, Benoît Sagot, Éric de la Clergerie, and Antoine Bordes. 2019. Controllable sentence simplification. *CoRR*, abs/1910.02677.

Benjamin Minixhofer. 2020. GerPT2: German large and small versions of GPT2.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective assessment of text complexity: A dataset for german language. *CoRR*, abs/1904.07733.

Akifumi Nakamachi, Tomoyuki Kajiwara, and Yuki Arase. 2020. Text simplification with reinforcement learning using supervised rewards on grammaticality, meaning preservation, and simplicity. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 153–159, Suzhou, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese

BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195. IEEE Computer Society.

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in german. In *Third Workshop on New Frontiers in Summarization*, pages 152–161. ACL Anthology.

Horacio Saggion. 2017. *Automatic Text Simplification*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Andreas Säuberli, Sarah Ebling, Martin Volk, Nuria Gala, and Rodrigo Wilkens. 2020. Benchmarking data-driven automatic text simplification for german.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Computational Processing of the Portuguese Language, 9th International Conference, PROPOR 2010, Porto Alegre, RS, Brazil, April 27-30, 2010. Proceedings*, volume 6001 of *Lecture Notes in Computer Science*, pages 30–39. Springer.

Nicolas Spring, Annette Rios, and Sarah Ebling. 2021. Exploring German multi-level text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1339–1349, Held Online. INCOMA Ltd.

Julia Suter, Sarah Ebling, and Martin Volk. 2016. Rule-based automatic text simplification for german. In *Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016, Bochum, Germany, September 19-21, 2016*, volume 16 of *Bochumer Linguistische Arbeitsberichte*.

Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*, second edition. The MIT Press.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT - building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*, pages 479–480. European Association for Machine Translation.

Willian Massami Watanabe, Arnaldo Candido Junior, Vinícius Rodriguez Uzêda, Renata Pontin de Mattos Fortes, Thiago Alexandre Salgueiro Pardo, and Sandra Maria Aluísio. 2009. Facilita: Reading assistance for low-literacy readers. SIGDOC '09, page 29–36, New York, NY, USA. Association for Computing Machinery.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8:229–256.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Trans. Assoc. Comput. Linguistics*, 3:283–297.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 584–594. Association for Computational Linguistics.

A. Appendix

A.1. Datasets and Preprocessing

The German Wikipedia dump[1] from the 21st of January 2022 was downloaded and processed as follows:

1. The dump is preprocessed into articles using WikiExtractor (Attardi, 2015).
2. Empty articles are removed.
3. Articles are split into individual paragraphs at each new line ("\n"), resulting in 10.8 million paragraphs.
4. The paragraphs are further cut down into a length between 80 and 175 tokens.

The resulting dataset consists of 1,080,000 paragraphs with an average number of 4.6 sentences and 93.8 words. The data is available in the Github repository. Even though the transformer models used for this approach can handle a sequence length of at least 512 tokens, the paragraphs are cut down to a maximum of 175 tokens. This has been done to speed up each training step and limit the GPU memory consumption by the models.

The TextComplexityDE contains 23 articles split into sentences with their corresponding simplification. Since this approach aims to simplify on a paragraph-level, the individual sentences from the same Wikipedia articles were manually combined to form paragraphs. Notably, in some occasions the sentences in the composed articles were not logically sequential. While the GWW dataset contains simpler simplifications than TextComplexityDE, the information contained in the complex article and its simplification differs more. For tuning the individual reward scores the TextComplexityDE dataset and the GWW dataset have been used as a reference (see table 2 for more details).

A.2. Reward Scores

A.2.1. Lexical Simplicity

For determining lexical simplicity, the approach from Laban et al. (2021) has been used. The score relies on the observation that word frequency and difficulty are correlated (Breland, 1996). First, we strip all stop-words from the texts, as they should not be considered. Next, all remaining words are lemmatized, to have a more accurate comparison between morphologically different words with the same base form. Two sets of words are created: One set contains all words that have been removed, and one set with words that have been added during simplification. The most complex or to be precise least frequent 15% of words from both of these sets are kept, all other words are filtered out. Then the average Zipf value for each set is computed: $\overline{Zipf}_{add}$ for the added words and $\overline{Zipf}_{rem}$ for the removed words. With these values, the lexical shift $shift_{lexical}$ between the simplification S and the original paragraph P can be calculated. The score is clipped between 0 and 1 and has a ramp shape, where the score $s_{lexical}$ falls off when achieving a $shift_{lexical}$ above the target value of 0.8. An example is given in Figure 2.

A.2.2. Syntactic Simplicity

To measure the readability of their generator's output Laban et al. (2021, §3.1.1) used the readability metric FKGL. Since there was no German adaption for this metric, the adaption FRE was chosen for this score instead (Amstad, 1978). Short sentences with short words are scored well with these metrics. The objective is to reward the model for generating shorter sentences. For the syntactic score $s_{syntactic}$, the approach from KiS has been adapted. Laban et al. (2021, §3.1.1) argue that an already syntactically simple paragraph should not require any further simplification and define the target FKGL conditioned on the original paragraph's FKGL score. To calculate the score we use the same scoring function as for $s_{lexical}$.

A.2.3. Language-Model Fluency

Again, we follow the work of KiS which is based on Lau et al. (2017) showing that grammaticality of a text can be measured by observing a language models probability. The score was constructed by taking the likelihood of the original and simplified paragraph:

$$s_{fluency} = \left[\frac{\lambda + LM(P) - LM(S)}{\lambda}\right]^{+} \quad (3)$$

where $LM(P)$ and $LM(S)$ stand for the likelihood of the original and simplified paragraph that are obtained by a masked language model. If the loss of a generated simplification $LM(S)$ is higher than $LM(P)$ by λ or more, $s_{fluency}$ is set to 0. The score is clipped between 0 and 1; if $LM(S)$ is above or equal to $LM(P)$, the score is 1 otherwise the score is a linear interpolation between 0 and 1. (Laban et al., 2021, §3.2.1) For more details on the model and training used for this score see section A.3.2.

Unfortunately, adapting the LM-Fluency score $s_{fluency}$ to the German language came with new problems: The reward seemed to encourage shorter and more probable words, especially articles like "der", "die", "das" (English: "The"). This

[1] https://dumps.wikimedia.org/dewiki/

Dataset	parallel articles	avg number of sentences		avg number of words	
		original	simplification	original	simplification
TextComplexityDE	23	11.00	23.43	286.48	282.52
GWW	52	5.52	8.98	82.31	67.29

Table 2: Statistics of reference datasets TextComplexityDE and GWW

might be because articles are relatively frequent words and therefore overall very probable in German, which results in a smaller loss. It was found that just adding repeating articles to a text often decreases the overall loss of a text, therefore scoring it as more fluent. To mitigate this problem the Article Repetition Penalty was employed for this, see section 3.2.3.

A.2.4. Discriminator Fluency

The Language-Model Fluency score can be limiting as it is static and deterministic (Laban et al., 2021, §3.2.2). Therefore it can be exploited by the generator. To counter this we incorporate a score s_{discr} based on a dynamic discriminator which they used in KiS. In this case, the generator simplifies the examples and the discriminator tries to predict if a given paragraph is a generated simplification or an original paragraph written by a human. During the generator's training process, both the simplification outputs and the original paragraphs are added to the discriminators training buffer. The original paragraphs are assigned a label of 1, and the generator outputs a label of 0. When the buffer reaches n samples, the discriminator is trained and the buffer is emptied again. More details are available in section A.3.2.

A.2.5. Brevity

The brevity guardrail is a score that ensures that the length of a generated simplification falls into the range of the original paragraph. The brevity score was configured to return 0.9999 if $0.6 \leq C \leq 1.3$, otherwise it returns 0.0001.

A.3. Training Details

A.3.1. Generator

A German version of the medium GPT-2 model GerPT-2 (Minixhofer, 2020) with 345M parameters was used for the generator. The training was performed on a workstation with 64 GB of RAM, an I9-9900K processor, and two RTX 2080 Ti GPUs with 11GB memory. All training tasks performed in this work used Automatic Mixed Precision (AMP) to save memory during training and increase the speed. For optimization, AdamW was used (Loshchilov and Hutter, 2017). For experiment tracking and visualization, Weights & Biases has been utilized (Biewald, 2020).

First, the model was pre-trained on the *copy task*. Using this task, the generator learns to output an exact or close copy of the input. This is a good baseline to start the simplification process. When the generator was trained for too long on the copy task, the sampled simplification candidates during simplification training were often too similar or even an exact copy of the original text. This low diversity resulted in very similar rewards, which limited the training signal for the generator. For the *copy task*, the training script from the Summary Loop Github repository has been used (Laban et al., 2020). The generator was fine-tuned with a learning rate of $2 \cdot 10^{-5}$, with a batch size of eight examples. The model was trained on this task for about 1800 training steps (25 minutes).

For the simplification training with k-SCST a learning rate of $4 \cdot 10^{-5}$ was chosen. A batch-size of one example was applied, meaning after sampling and scoring $k = 8$ simplification candidates conditioned on one original paragraph, the generator is then optimized. The simplifications were sampled using nucleus sampling with $p = 0.95$, combined with a top-K value of $K = 5$. Additionally a setting suppressing the repetition of 5-grams in a sequence was employed during sampling to avoid repeating phrases. The p value was chosen based on the research of Holtzman et al. (2020). They argue that values between 0.9 and 1 are the most reliable, and lower values tend to generate repetitions. The value $K = 5$ was selected relatively low, as it produced the most reliable results considering the meaning preservation, hallucination and brevity scores in the beginning. In retrospect, the top-K value may have been chosen too low, limiting the diversity of the candidates and restricting the nucleus sampling capabilities.

Our main model GUTS was trained for over 110,000 steps (roughly five days). Table 3 shows how the reward scores during training were weighted.

	GUTS
$s_{lexical}$	0.5
$s_{syntactic}$	3.0
$s_{meaning}$	4.0
$s_{fluency}$	0.5
s_{discr}	0.5
$s_{brevity}$	1.0
$s_{hallucination}$	1.0
s_{ngram}	1.0
$s_{articles}$	1.0

Table 3: Score weights used for training

A.3.2. Fluency Models

The model used for $s_{fluency}$ is a German BERT base model[2], with 110M parameters. It was fine-tuned on Wikipedia articles to better capture the linguistic properties of the domain. The model was trained for roughly 20,000 steps using AdamW (Loshchilov and Hutter, 2017) as an optimizer with a learning rate of 10^{-5} and a batch size of eight examples.

The Discriminator for the score s_{discr} was trained on a buffer consisting of original paragraphs and generated simplifications, collected during the training process. When the buffer reaches 4000 samples, the discriminator is trained with the data. Afterwards the buffer gets emptied again. The same German BERT base model mentioned in the previous paragraph is used here again. It is trained using 90% of the training buffer for the discriminator. The discriminator is trained for five epochs. The end of each epoch is used as a checkpoint, where the discriminator model is saved along with the F1 performance tested on the last 10% of the training buffer. The best model of the five checkpoints is kept as the new discriminator until the training buffer reaches 4000 samples again. The model is trained using AdamW (Loshchilov and Hutter, 2017) as an optimizer with a learning rate of 10^{-5}, a batch size of 6 and AMP.

A.3.3. Further models

For the meaning preservation score $s_{meaning}$ a fine-tuned distilbert model[3], with 66M parameters, has been used for the sentence alignment. For calculating the BERTScore for $s_{meaning}$ and $s_{hallucination}$ the same German BERT base model described in section A.2.3) was utilized. For the NER tagging in $s_{hallucination}$ another finetuned distilbert model[4] was applied.

[2] https://huggingface.co/dbmdz/bert-base-german-cased

[3] https://huggingface.co/sentence-transformers/msmarco-distilbert-multilingual-en-de-v2-tmp-lng-aligned

[4] https://huggingface.co/Davlan/distilbert-base-multilingual-cased-ner-hrl

A.4. Simplification examples

A.4.1. Simplification example 1

Original Halvarsson gewann bei der Junioreneuropameisterschaft 1979 in Achenkirch die Bronzemedaille im Slalom. Die ersten Punkte in einem Weltcuprennen holte er am 10. März 1980 als Zehnter des Slaloms von Cortina d'Ampezzo. Im selben Jahr gewann er den schwedischen Riesenslalom-Meistertitel. In den zwei folgenden Wintern fuhr er im Weltcup je zwei Mal unter die besten 15, während er beim Slalomrennen der Weltmeisterschaft 1982 ausschied. In der Saison 1982/83 gelangen ihm sechs Platzierungen unter den besten 15, darunter ein sechster Platz in Kranjska Gora.

GUTS Hvar gewann bei der Junioren Europameisterschaft 1979 in Achenkirchen die Bronzemed Medaille im Slalom. Der ersten Punkte in einem Wettkampf holte er am 10." Bis zum Zehnten des Slaloms" von Cortina d'Ampezzo. Beim selben Jahr gewann er die schwedischen Riesenslalom. In den zwei Jahren fuhr er im Weltcup jeweils zwei Mal unter die guten 15, während er bei beim Slalom gegeneinander ausschied. In den Saison 1982/83 gelang ihm sechs Platzierungen unter dem besten 15, darunter einen sechster Platz in Ku.

Pivot Halvarsson gewann die Bronzemedaille bei der Junioren-Europameisterschaft 1979. Er war einer der ersten Spieler, der die Slalombarriere durchbrach, und er gewann die ersten beiden Punkte in einem Rennen am 10. März 1980. Zwei Jahre zuvor hatte er den schwedischen Riesen-Slalom gewonnen, aber dieses Mal war er am Rande des Sieges. In den folgenden zwei Wintern gewann er die schwedische Meisterschaft, einschließlich des Slaloms, und zog sich in der Saison 1982/83.

A.4.2. Simplification example 2

Original Mizuo Peck wuchs in New York auf und stand bereits in jungen Jahren auf der Theaterbühne. Sie besuchte die Theaterabteilung der State University of New York und machte ihr Filmdebüt im Jahr 2000 in dem deutschen Fernsehfilm "Nicht heulen, Husky" neben Heiner Lauterbach und Barbara Rudnik. Bekannt wurde sie aber vor allem durch ihre Rolle als Sacajawea in der dreiteiligen Filmreihe "Nachts im Museum" zwischen 2006 und 2014.

GUTS Mizuo Pecks wuchs in New York und stand bereits in den jungen Jahren auf der Bühne. Sie besuchte die Schauspielabteilung der State University von New York und machte sich ihr Filmdebüt. Bekannt wurde sie doch vor allem durch ihre Rollen als Sacajawe in der dreiteiligsten Filmreihe "Nacht" zwischen 2006 und 2015.

Pivot Mizuo Peck wurde in New York geboren und hatte bereits in jungen Jahren eine Karriere als Theaterdarstellerin. Sie besuchte die Theaterabteilung der State University of New York und debütierte Ende 2000 in einer deutschen TV-Serie mit dem Titel "Nicht heulen, Husky". Doch bald wurde sie in ihrem eigenen Film Sacajawea zur Vollzeitdarstellerin,

A.4.3. Simplification example 3

Original König Dom Manuel I. nahm ab dem Ende des 15. Jahrhunderts weitreichende Neugestaltungen in der Stadt Sintra vor. Ab dem 16. Jahrhundert nahm die Beliebtheit Sintras bei den Oberschichten zu, und adlige Familien errichteten im Kreis Sintra einige Herrenhäuser. Das Erdbeben von Lissabon 1755 rief auch in Sintra erhebliche Zerstörungen hervor, denen in der Folge umfangreiche Arbeiten des Wiederaufbaus folgten. Ende 18. Jahrhundert entstand mit der Textildruckerei und -färberei "Fábrica de Estamparia de Rio de Mouro" die erste industrielle Einrichtung im Kreis.

GUTS König Dom Manuel I nahm ab dem Ende der 15. Jahrhundert weitreichende Neugestaltungen. In der Stadt Sintra ab dem 16. Jahrhundert gab die Beliebtheit Sintras. Ab dem 16 Jahr nahm die Beliebtheit Sintra bei den Oberschichten bei, und adlige Familie errichteten im Kreis Sint. Das Erdbeben von Liss 1755 rief auch noch in Sintra erhebliche Schäden hervor, denen in den Folge umfangreiche Arbeiten des Aufbaus folgten. Ende 19. Jahrhundert entstand mit dem Textilddruckerei und -firberei "Fébrica de Estaparia de Rio de" die erste industrielle Organisation im Kreis.

Pivot Seit dem Ende des 15. Jahrhunderts wurde Sintra umfassend renoviert. Ab dem 16. Jahrhundert war die Stadt für ihre hohe Lebensqualität bekannt geworden. Von dort aus begannen Adelsfamilien, im Kreis Sintra Villen zu bauen, die eine große Anzahl von Geschäften und Restaurants umfassten. Das Erdbeben in Lissabon im Jahr 1755 verursachte auch erhebliche Schäden, was zu umfangreichen Wiederaufbauarbeiten führte. Ende des 18. Jahrhunderts wurde die erste Industrieanlage in der Gegend

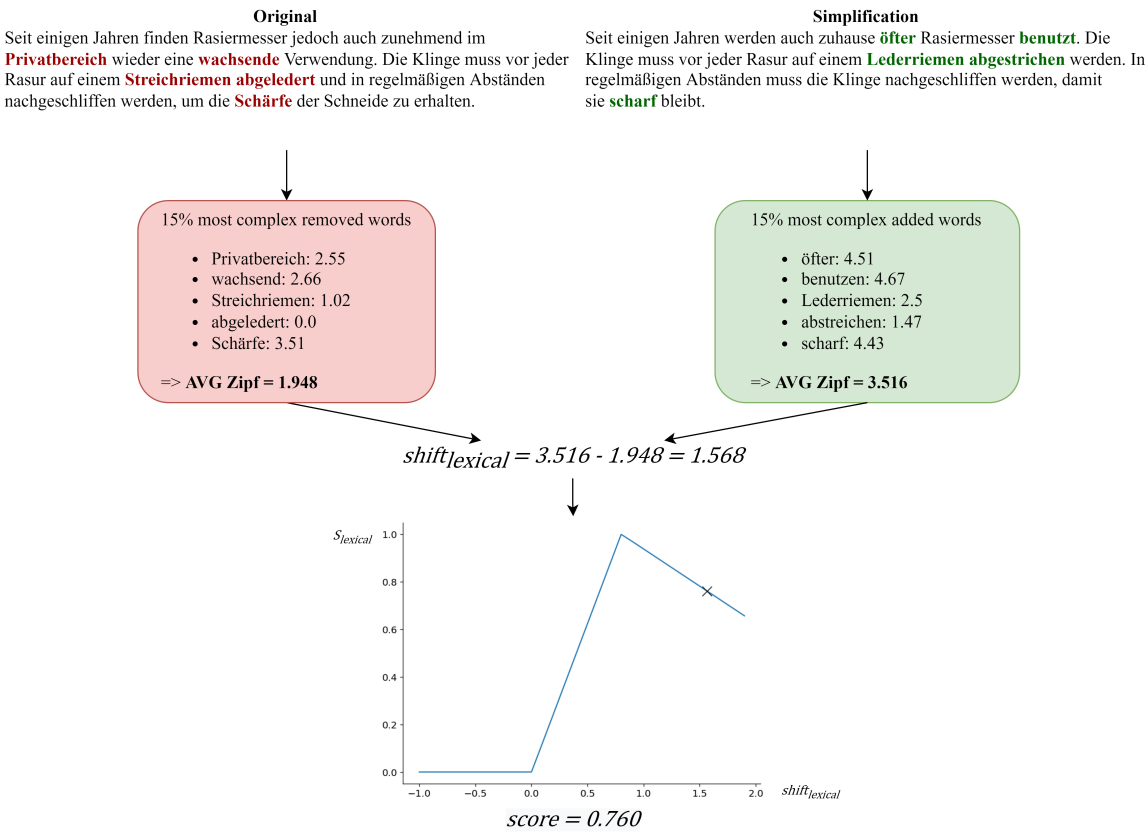


Figure 2: Example for the calculation of $s_{lexical}$

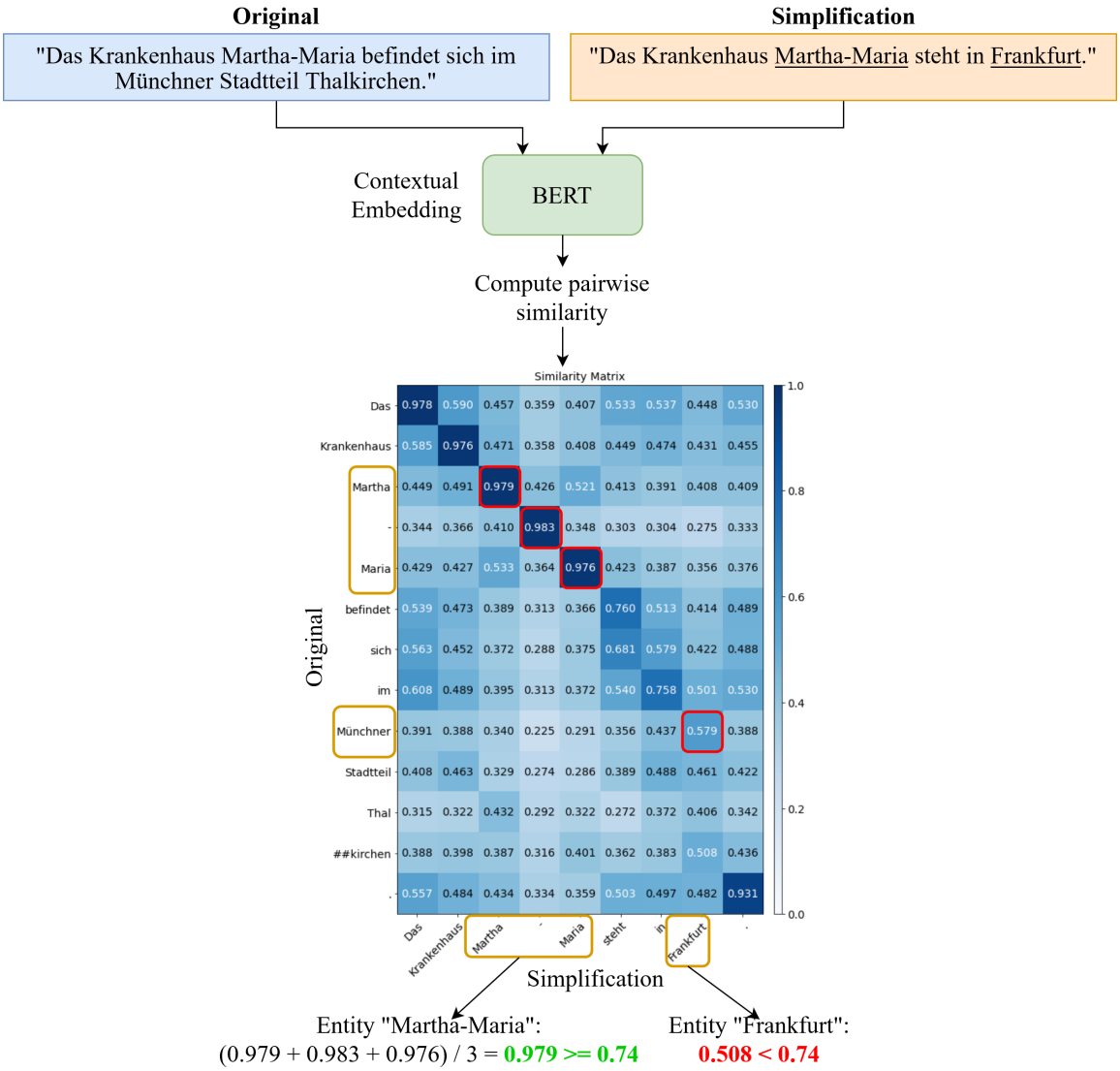


Figure 3: Hallucination detection algorithm. The confusion matrix is calculated with BERTScore (Zhang et al., 2020) using the original text and the simplification. Then, the entities in the simplification are detected: In this case “Martha-Maria” and “Frankfurt”. For each of the entities the highest similarity value in the matrix is selected. If the value is below the threshold of 0.74, it is assumed that a hallucination is present.

Simplification Strategies in French Spontaneous Speech

Lucía Ormaechea[1,2], Nikos Tsourakis[1], Didier Schwab[2], Pierrette Bouillon[1] and Benjamin Lecouteux[2]

[1] TIM/FTI, University of Geneva, 40 Boulevard du Pont-d'Arve – Geneva, Switzerland

{firstName.lastName}@unige.ch

[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG – Grenoble, France

{firstName.lastName}@univ-grenoble-alpes.fr

Abstract

Automatic Text Simplification (ATS) aims at rewriting texts into simpler variants while preserving their original meaning, so they can be more easily understood by different audiences. While ATS has been widely used for written texts, its application to spoken language remains unexplored, even if it is not exempt from difficulty. This study aims to characterize the edit operations performed in order to simplify French transcripts for non-native speakers. To do so, we relied on a data sample randomly extracted from the ORFÉO-CEFC French spontaneous speech dataset. In the absence of guidelines to direct this process, we adopted an intuitive simplification approach, so as to investigate the crafted simplifications based on expert linguists' criteria, and to compare them with those produced by a generative AI (namely, ChatGPT). The results, analyzed quantitatively and qualitatively, reveal that the most common edits are deletions, and affect oral production aspects, like restarts or hesitations. Consequently, candidate simplifications are typically register-standardized sentences that solely include the propositional content of the input. The study also examines the alignment between human- and machine-based simplifications, revealing a moderate level of agreement, and highlighting the subjective nature of the task. The findings contribute to understanding the intricacies of simplifying spontaneous spoken language. In addition, the provision of a small-scale parallel dataset derived from such expert simplifications, PROPICTO-ORFÉO-SIMPLE, can facilitate the evaluation of speech simplification solutions.

Keywords: simplification, spontaneous speech, French language, expert annotation, ChatGPT

1. Introduction

Automatic Text Simplification (ATS) aims at rewriting texts into simpler variants, by reducing their linguistic complexity, albeit preserving their original meaning (Candido et al., 2009; Horn et al., 2014). ATS has received increased attention in the past few years, in view of its significance from both societal and computational perspectives: it can assist in creating adapted texts for diverse target audiences (De Belder and Moens, 2010; Rello et al., 2013) or serve as a pre-processing step for other NLP tasks such as MT (Stajner and Popovic, 2016).

Providing a simplified version of a given text has typically been applied for newswire content (Xu et al., 2015; Saggion, 2017), healthcare-related documents (Shardlow and Nawaz, 2019; Van den Bercken et al., 2019) and Wiki-based articles (Hwang et al., 2015; Zhang and Lapata, 2017; Ormaechea and Tsourakis, 2023). Hence, ATS demonstrates its predominant application in written-based texts, while its implementation over a spoken modality remains unexplored. Yet, spoken-based texts are not exempt from difficulty.

Traditionally, features associated with complexity are strongly linked to the typical attributes found in formal written-based texts, like high lexical density or the propensity towards long subordination (Brunato et al., 2022). However, complexity also exists in spoken language, but is reflected differently from its written counterpart, mainly because the information structure is also dissimilar. Written text is the *result* of a planned language production, whereas speech, especially the spontaneous kind, is a real-time *process* (Carter and McCarthy, 2017), thus retaining traces of its on-the-fly construction like revisions, false starts, reformulations or self-corrections. Due to these phenomena, spoken language is typically disfluent, which makes speech transcripts particularly challenging to understand. Decomplexifying speech may be of particular interest when transcriptions are further used for:

- Accessibility purposes. A simplified transcript can help clarify the conveyed message and reduce its ambiguity, making it more accessible to several target audiences (*e.g.*, individuals with cognitive disabilities, foreign language learners, non-native speakers, etc).
- Ancillary purposes. Raw transcripts are often difficult to process by NLP pipelines. Providing a meaning-preserving simpler transcript may be helpful as an intermediate representation for other NLP tasks like subtitle translation (Mehta et al., 2020) or speech-to-pictograph cross-modal conversion (Ormaechea et al., 2023a).

With this article, we aim to investigate the strategies followed by experts as to simplify spontaneous French transcripts for a non-native speaking audience, and to compare the resulting simplification

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 90–102

operations with those produced by a generative AI, namely ChatGPT. More precisely, we aim to address the following research questions:

- What are the edit operations performed to obtain a simplified version of a French spontaneous speech transcript?
- How do human simplification strategies align with those adopted by ChatGPT and how suitable are they for a non-native audience?

In this way, we intend to provide an *a posteriori* characterization of the simplification strategies operated on the basis of a spoken spontaneous input. To the best of our knowledge, no such study has been conducted to date. In the absence of guidelines to direct this process, we decided to adopt an *intuitive* simplification approach (Allen, 2009). In this way, we investigate the simplifications produced based on expert linguists' criteria, and then compare them with those generated by ChatGPT.

The structure of this paper is as follows: Section 2 delves into the notion of spontaneity and describes the existing tasks that closely resemble spontaneous speech simplification. In Section 3, we discuss the input data sample, along with the survey design and ChatGPT prompts employed to collect simplifications. The analysis of these outputs, both quantitative and qualitative, is detailed in Section 4. Lastly, Section 5 provides concluding remarks, addresses limitations, and outlines potential pathways for future research.

2. Background and Related Work

2.1. Spontaneity in Speech

Spoken language exhibits differences with respect to written language which go beyond the mode of transmission used by each modality[1], and affect morphology, syntax and vocabulary (Caines et al., 2017). Among other aspects, a defining morpho-syntactic trait of speech is the lack of sentence boundaries, which are conventionally delimited in writing. From a grammatical perspective, spoken language is often characterized by the presence of disfluencies, which emerge as a result of the speaker's real-time processing and notably impact spontaneous speech. Due precisely to this *online* process, the *information packaging* (Halliday, 1985) also differs with respect to the *offline* (namely, written) one. This leads to the selection of different grammatical forms and changes in word order, which, in the case of French, are evidenced by the presence of cleft constructions (*i.e.*, *c'est lui qui a fermé la porte*) or the use of dislocated subjects (*i.e.*, *les enfants, ils arrivent*).

[1] That is, phonetics and prosody of speech versus graphemics and orthography of written language.

Unspontaneous texts constitute a revised and finalized version of a language production. Spontaneous speech, on the other side, is by nature an unfinished product. Due to the absence of prior planning, discourse unfolds in real time, consequently shedding all the traces of its elaboration, such as hesitations, reformulations, repetitions, and false starts (Blanche-Benveniste, 1997). These, unlike their written equivalent, are indelible in an oral modality, and can only lead to an elongation of the utterance (Bazillon et al., 2008). Consequently, the presence of such performance phenomena can produce concatenations of elements having a paradigmatic relation along the syntagmatic axis (Luzzati, 1998). This is evident in the spontaneous utterance illustrated in Figure 1, where disfluent features potentially hinder the correct understanding of the transcript. A simple *despontaneification* operation (see Figure 2) would result in an utterance holding an identical propositional content, and would clarify the conveyed message by eliminating paradigmatic supplements (*i.e.*, *[on va juste] euh [je vais juste]*) that stem from hesitations during the act of speaking.

2.2. Simplification and Compression in Speech

From an automated perspective, implementing simplification operations over speech appears to be an unexplored area. The existing task bearing the closest resemblance is *sentence compression* from speech transcripts (Angerbauer et al., 2019; Buet and Yvon, 2021). The aim of this process is to automatically reduce its length, generally in response to technical imperatives (Daelemans et al., 2004). This explains its relevance for subtitle generation (Luotolahti and Ginter, 2015), where technical restrictions drive the need of shrinking the text displayed on the screen. This is also triggered by the significantly faster pace of speech compared to reading, often motivating the suppression of phatic and deictic elements, as well as the condensation of information (Becquemont, 1996). Yet, the notion of *compression* must be distinguished from that of *simplification*. While the former aims at content reduction and merely preserves the most salient information, the latter seeks to generate a simpler variant without compromising the meaning.

In addition, an analogous task to speech simplification is Easy-to-Understand subtitling, in which an intralingual adaptation of subtitles is crafted to make them more accessible for viewers (Matamala, 2022). Guidelines have been proposed for this goal. While they include grammar- and style-based recommendations for simplification (Bernabé and Cavallo, 2021), they are primarily driven by the inherent spatial and temporal constraints of subtitling.

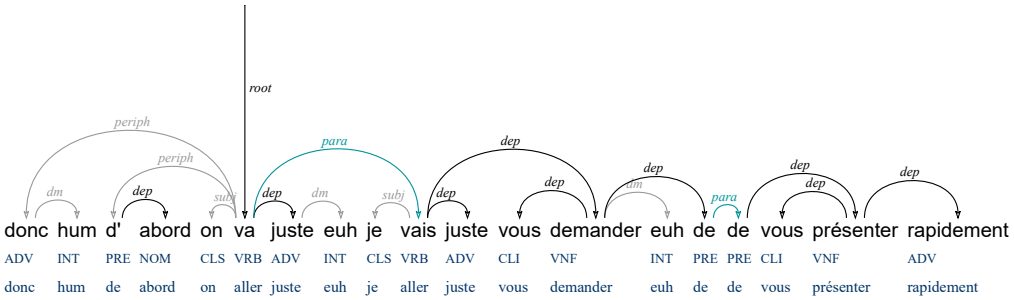


Figure 1: A spontaneous utterance extracted from the French corpus CFPP (Benzitoun et al., 2016). It displays the transcript along with the corresponding lemmas, part-of-speech tags and dependency tree.

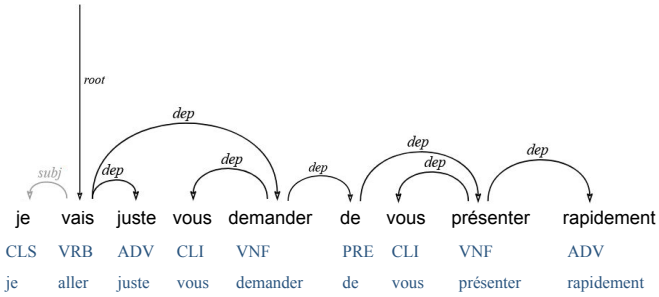


Figure 2: The despontaneified version of the sentence shown in Figure 1.

3. Methodology

As previously noted, we opted to follow an *intuitive* strategy so as to produce simplified versions of spontaneous utterances (Allen, 2009). The absence of preexisting guidelines or syllabi for this task precluded a *structural* approach. Therefore, we decided to rely on the intuition of expert linguists to obtain manual simplifications targeted to non-native speakers, enabling us to empirically investigate the mechanisms involved in simplifying spontaneous utterances in French. We also compared such outputs with those generated through ChatGPT, as we will see below.

Specifically, we decided to focus on French language given: *i*) its rich body of literature describing spontaneity phenomena (Blanche-Benveniste (1997); Bazillon et al. (2008); Luzzati (2013); Evain et al. (2022) to name just a few), and *ii*) the existence of French written-oriented simplification guidelines (Gala et al., 2020). With this work, we intend to address the still unexplored connection between the two areas.

Moreover, we decided to purposely target simplification of spontaneous utterances for a non-native speaking audience, that is, individuals that speak a given language (in this case, French), but have acquired a different first language. Although the scope of this group may be broad, due to the variety of possible cultural backgrounds, language proficiency levels or underlying mother tongues, we specifically focused on non-native speakers given the potential interest that the creation of simpler equivalents may have for this audience. Spontaneous utterances often contain slang expressions and informal register traits, which, along with a dissimilar information structure, can seem unfamiliar to a foreign speaker.

3.1. Source Dataset

In order to analyze the simplification strategies of spontaneous French speech, we resorted to ORFÉO, a well-known platform designed for the study of European French, both in its written and spoken forms (Benzitoun et al., 2016). For our study, we decided to use the latter, known as ORFÉO-CEFC (*Corpus d'Étude du Français Contemporain*), which comprises 12 existing corpora of spoken French. It features segments aligned with audio files at the sentence level, and is enriched with morphosyntactic annotations (such as POS tags or parse trees).

The subcorpora constituting ORFÉO-CEFC cover a wide range of communicative situations (*i.e.*, interviews, tales, phone calls, etc), environments (friendly, academic or familiar) and degrees of spontaneity (Blanche-Benveniste, 1997), ranging from professional situations (like those in `Reunions`) to everyday life ones (as those portrayed in `Cfpp`).

3.2. Sampling

Creating such a resource may be of interest for its eventual reuse as a test set to evaluate spontaneous speech simplification systems. For this reason, we relied on the ORFÉO-CEFC partitioning created by Pupier et al. (2022) and use their evaluation set as the population for our study (ORFÉO-TEST), which amounts to 21,459 segments.

Since the process of manual simplification is a time-consuming task, we opted to extract a subset of the previous distribution. Determining the sample size was key, insofar as we intended to: *i*) analyze a sufficiently representative subset of the original dataset examined, and *ii*) maintain a reasonable workload for annotators, so as to not compromise the stability and consistency of the task.

To ensure the reliability and knowledge extrapolation from the data sample, we performed stratified sampling, dividing the population into 12 distinct strata, each representing a subcorpus of ORFÉO-TEST (see Table 1). Sampling was proportionate, based on the number of segments in each subcorpus, and then randomly collected.

subcorpus	# utt.	%	# sampl.
Cfpb	362	1.69	2
Cfpp	3,232	15.06	15
Clapi	967	4.51	5
Coralrom	1,376	6.41	6
Crfp	2,259	10.53	10
Fleuron	217	1.01	1
Oral-Narr.	1,050	4.89	5
Ofrom	1,476	6.88	7
Reunions	1,245	5.80	6
Tcof	1,997	9.31	9
Tufs	4,525	21.09	21
Valibel	2,753	12.83	13
Total	**21,459**	**100**	**100**

Table 1: Proportional size of each stratum conforming the population. Calculation of the corresponding number of examples for a sample size of 100.

3.3. Human-Based Simplification: Survey Design

As we indicated at the beginning of Section 3, to the best of our knowledge, there are currently no guidelines on how to simplify spontaneous speech transcriptions. This renders a *structural* approach, and thus guideline-adherent, impossible as a simplification strategy. Instead, we opted for an *intuitive approach*, through which to capture the insights of professionals, and on this basis identify the sentence transformations needed to simplify spoken-based French data.

To do this, we decided to set up a manual simplification task based on the stratified sample obtained from ORFÉO-TEST. For this purpose, we enlisted 2 experts, both of them with a solid background in linguistics and a current dedication to research in this field. They both had French (in its European variety) as their native language. The task was hosted on the LimeSurvey platform, and was made accessible from February 1st until February 12th.

As for the survey structure, the sentences derived from our previous sampling were displayed on the LimeSurvey online platform successively. For each spontaneous input sentence, we asked respondents to propose a simplified version (as shown in Appendix A). As part of the instructions, we specified that the goal was to provide simpler equivalents for a French non-native speaking audience. We also asked them to list and explain their chain of thought to transform the input sentence into a simplified equivalent, thus enhancing the explicability of their decision-making. Both fields were mandatory, and we allowed back-and-forth navigation for respondents to revisit their answers.

These instructions, paired with more detailed information, were provided at the beginning of the survey. We kept them visible throughout the execution of the survey, so as to ensure the clarity of the task at hand. To combat potential fatigue during the simplification process, we provided participants with the option to interrupt the task and resume it at their convenience, but always before the due date.

On another note, after a long internal discussion, we opted to merely provide the spontaneous transcript without its corresponding audio file. This decision could have potentially facilitated the task by aiding in disambiguating certain utterances through the inclusion of paralinguistic information (*i.e.*, rhythm, tone, prosody, etc). However, we deliberately chose to challenge participants to simplify based solely on linguistic information. This decision underscores a well-known paradox: spoken language can only be studied on the basis of its written representation (Blanche-Benveniste, 1997).

3.4. Machine-Based Simplification: ChatGPT Prompting

ChatGPT has emerged as an attractive alternative for annotation and typical NLP tasks. Due to its ability to process and generate natural language text, it can assist in various tasks, such as part-of-speech tagging, identifying named entities, or even providing detailed annotations on complex datasets (Gilardi et al., 2023).

We leveraged the OpenAI API and its latest model (`gpt-4-0125-preview`) to simplify spontaneous sentences automatically[2]. The model received the original sentences as input, and produced simpler, more accessible versions of the

[2] Training data: up to December 2023.

same text. Specifically, we ensured that the model was prompted with separate messages to avoid any influence from the dialogue history. Moreover, we used the `temperature=0` setting in every API call to ensure consistency in the model's responses. The prompt included the necessary instructions and was deliberately chosen to be the same as the one given to the human experts (see Figure 7 in Appendix A). We deemed it safer to use the same prompt compared to using distinct ones. This ensured consistency in the information presented to both humans and ChatGPT, enabling a more accurate comparability between responses. The total cost for generating the simplifications of the 100 sentences was approximately 1 USD[3].

4. Results

4.1. Quantitative Evaluation

4.1.1. Taxonomization and Analysis

After the human-based completion of the survey and the machine-based generation of simplified outputs, we taxonomized the different transformations performed to convert spontaneous utterances into the proposed candidate simplifications.

To create the taxonomy, we first analyzed the chain of thought provided by the 3 respondents. On that basis, we derived a macro-categorization using the main edit-based operations: *deletion*, *replacement*, *addition*, *restructuration*, and *copy* (when no alteration to the input was made), that we later subdivided according to the observed linguistic transformations (as shown in Table 2). We then annotated the simplified sentences based on such taxonomy and computed the frequencies for each phenomenon. It should be noted that this stage proved to be more challenging than anticipated: the identification of each operation may not be easily distinguishable, as edits often ensue from jointly applying various transformations (Saggion, 2017). As a result, the computation of occurrences for each phenomenon may have been affected.

As can be seen in Figure 3, it is evident that deletions are the most prevalent among all edit operations. This is hardly surprising in the context of spoken language simplification where hesitations and errors happening during spontaneous speech delivery cannot be undone. While these aspects might be interesting from a pragmatic perspective, they do not provide any propositional content nor relevant semantic information to the sentence, and are thus erased in a simplification context.

Taking a closer look at the distribution of the different suppressed linguistic units (as seen in Plot (a) within Figure 4), it is important to note the dropping of redundant elements such as repetitions or restarts, as well as the suppression of elements related to the enunciation, such as affirmative and negative adverbs (*non, voilà, ouais*), statement verbs (*tu sais*, *je tiens à dire*) or discourse markers (*en fait*). Deletion operations also affect adjectives and adverbs that add little information to the input sentence (*toutes nos traditions* → *nos traditions*).

As for the coherence between the candidate simplifications, the three participants seem to use a similar reasoning to transform the provided inputs, prioritizing deletion operations to achieve simplification. We note, however, that ChatGPT makes more conservative decisions when generating outputs and performs fewer deletions than both humans (see Example I in Table 4 in Appendix B). Between the two linguists, there is an overall symmetry in the number and type of triggered phenomena, although Expert 1 tends to drop more items than Expert 2, especially in terms of restarts and reformulations (see again Plot (a) in Figure 4).

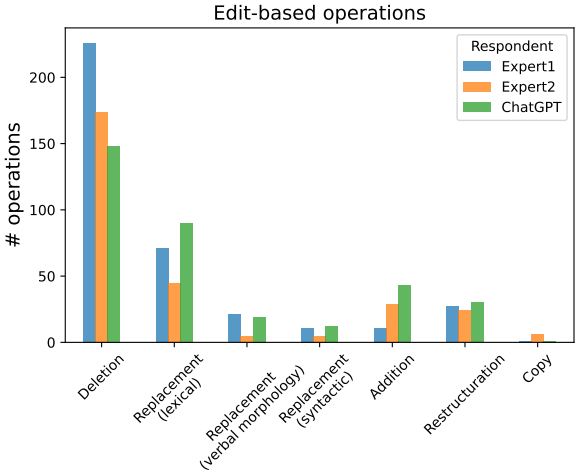


Figure 3: Overview of the distribution of edit-based operations in the analyzed data sample.

As shown in Plot (b), restructuration appeared as a much less frequent edit. All three respondents seldom performed any sentence splitting or merging modifications, very prototypical in written-based simplification. This can plausibly be explained by the typically shorter length of spoken sentences[4]. The prevailing change observed is the reordering

[3] Note that for the used model, the cost for input is 10 USD per 1 million tokens, while for output, it is 30 USD per 1 million tokens (as of April 2024).

[4] It should be noted that the notion of sentence in speech is not straightforward. The absence of sentence boundaries, which are conventionally delimited in writing, complicates the task of distinguishing each segment. For our study, we have relied on the sentential presegmentation provided by ORFÉO-CEFC.

Edit	Level	ID	Linguistic unit(s) affected, operation
Deletion		1	Repetitions
		2	Affirmation and negation words
		3	Interjections
		4	Conjunctions
		5	Discourse markers
		6	Restarts and reformulations
		7	Adverbs and adjectives
		8	Incomplete words
		9	Statement verbs
		10	Pronouns
		11	Verbs with little semantic value
Replacement	Lexical	12	Simpler synonyms for content words
		13	Compression of nominal phrases
		14	More standard equivalents for content words
		15	Smoothing of swear words
	Verbal morphology	16	Intransitive to transitive verbs
		17	Pronominal to non-pronominal verbs
		18	Change of verbal tense
		19	Compression of verbal locutions
	Syntactic	20	Passive to active voice
		21	Cleft to canonical constructions
		22	Neutralization of dislocated subjects
		23	Pronoun transformations
Restructuration		24	Reorder
		25	Sentence splitting
		26	Sentence merging
Addition		27	Explicitation or disambiguation of a word
		28	Completion of truncated sentences
		29	Clarification of uncommon terms
Copy		30	Input sentence is left unchanged

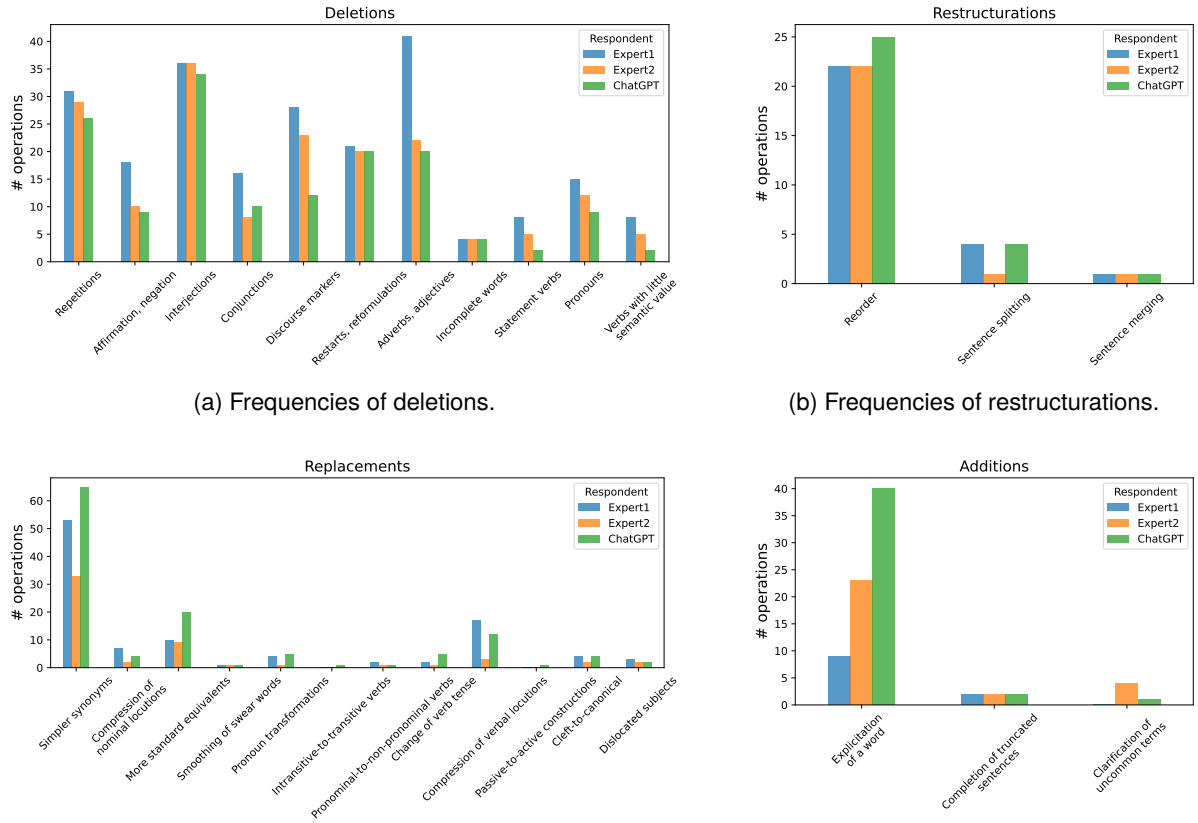
Table 2: Taxonomy of edit operations observed in the data sample, reflecting the simplification process from spoken-based transcripts.

of elements in the utterance, often driven by the search for a canonical subject-verb-object order (as seen in Example II in Table 4). As for the additions, displayed on Plot (d), the most notable category is the explicitation of a word. In this regard, the generative model seemed more inclined than humans to add extra information, with the aim of resolving eventual ambiguities from the source sentence.

Replacement operations, shown in Plot (c), were probably the most interesting edit type. We distinguished 3 linguistic levels of modification: *lexical*, *morphological*, and *syntactic*. Upon closer examination, we uncovered a preference for lexical-based edits (as shown in Figure 3), which were the second most common after deletions. Among these, the most occurring subcategory is the substitution of content words (nouns, adjectives, adverbs, and verbs), in favor of more common alternatives (*i.e.*, *confrérie* → *association* in Example III in Table 4). It is relevant to note in this regard that ChatGPT was the most prone to make changes of this type. This may have been triggered by the provided prompt, where we mentioned lexical substitution of complex terms as an example operation (see Instruction 2 shown in Figure 7).

Besides, we have noticed that the replacement of lexical units does not always stem from complex terms, but rather from slang ones. In these cases, the 3 respondents tended to use more standard equivalents, probably under the hypothesis that colloquialisms may be less familiar terms for foreign speakers. Some examples include: *gosses* → *enfants*, *monde* → *personnes* or *bouquins* → *livres*. The tendency to adopt a more formal register in the crafted simplification is also evidenced in the smoothing of profanity (as illustrated in Example IV in Table 4).

In addition, we observed a propensity to compress the constituents of phrases that do not convey much semantic content, probably on the assumption that a shorter sentence is also often perceived as simpler. This phenomenon can be observed in the shortening of nominal groups (*i.e.*, *monde du travail* → *travail*). That same principle seems to



(a) Frequencies of deletions.

(b) Frequencies of restructurations.

(c) Frequencies of replacements.

(d) Frequencies of additions.

Figure 4: A closer look to the distribution of the edit operations performed on the data sample extracted from ORFÉO-TEST.

apply in the morphology dimension, where verbal locutions or periphrases are often compressed into shorter forms (*i.e.*, *faire la demande* → *demander*), and compound or less frequent tenses tend to be converted into simple or more frequent ones (*elle jouait déjà de la guitare* → *elle joue la guitare*), if meaning is not altered. In addition, even to a lesser extent, we find instances in which pronominal verb forms are replaced by non-pronominal alternatives, and intransitive ones are substituted by transitive variants.

As for changes in syntax, these occur far less frequently than lexical transformations. This may be due to the fact that syntactic edits typically affect a larger span within the utterance than lexical ones, making them intrinsically less numerous. In any case, it is interesting to note that syntactic operations have mainly been applied to constructions that exhibit a marked information structure. For this reason, cleft clauses and dislocated subjects, common in spoken French, are reverted to their canonical non-marked forms (see Example V in Table 4). Finally, it is worth noting that the conversion of passive constructions into active voice is anecdotal. Although diathesis change is a well-established operation in the field of ATS, the use of passive voice is inherently rare in French, and is even less common in a spoken modality.

Overall, the results show that the most common edit operations in spontaneous speech simplification are deletions. The proposed simplifications are often sentential equivalents stripped of any oral marks such as enunciation elements (discourse markers, interjections), hesitations (inherent to the live construction of a message), or the use of slang and profanity (infrequent in a written form). As a result, the proposed simplified outputs are often *writified*, register-standardized versions of the inputs that strictly include their propositional content.

4.1.2. Respondents' Agreement

In the next evaluation step, we seek to understand the level of agreement among participants. We opted for the Jaccard Index because of its adeptness in quantifying the similarity between different answers. This choice was made since participants' responses are not limited to single, mutually exclusive categories but can include multiple selections. This metric calculates the ratio of

the intersection to the union of the sets of choices, providing a clear, normalized value ranging from 0 to 1. Our approach involves comparing the selections of each respondent with every other respondent for the same sentence pair to assess how similar their choices are. The results of this pairwise comparison are: $J(Exp_1, Exp_2) = 0.54$, $J(Exp_1, GPT) = 0.52$, $J(Exp_2, GPT) = 0.51$. Overall, the values suggest a moderate level of agreement among the respondents, with none of the pairs showing a particularly high or low level of consensus. This indicates a generally consistent understanding or interpretation of the operations for making simplifications, but it also highlights the subjective nature of the task (Dmitrieva et al., 2021; Ormaechea et al., 2023b), where individual differences in judgment can lead to variations in the chosen operations.

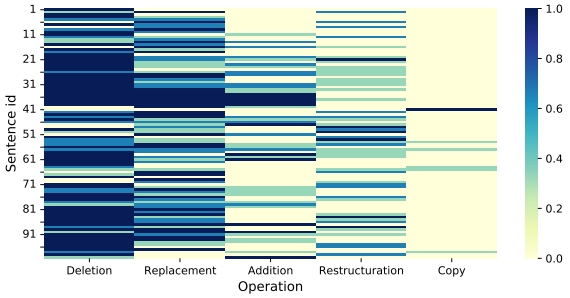


Figure 5: Agreement heatmap across sentences based on the 5 macro-categories.

Generally, there is a trade-off between the level of detail in the taxonomy definition and the desired level of agreement, as finer granularity often leads to more diverse operations and, consequently, reduced consensus (Heineman et al., 2023). To obtain a better insight into the chosen operations, we resorted to the analysis shown in the heatmap of Figure 5, where each cell represents the consensus among the respondents for one specific macro-category and input sentence. The darkest color indicates that all participants performed the same macro-operation on a given sentence. Based on the heatmap analysis, we observed that the agreement among respondents varied depending on the executed operation: in 69.7% of cases, deletion was performed on the same sentence by the three participants, which signifies a consensus on its utility for simplification. For the other cases, we observed a lesser consensus: 41.7% for replacement, 25% for restructuration, 12.2% for addition, and 16.7% for copy.

4.2. Qualitative Evaluation

To assess the suitability of the produced human- and machine-based simplification sentences for a foreign-speaking audience, we conducted a qualitative intrinsic evaluation with three master-level non-native French students. Specifically, they were asked to score the given simplification on a five-point Likert scale (see Table 3), on the basis of two criteria: *i*) *simplicity gain* (S_G): how much simpler is the candidate simplification compared to the original sentence?; and *ii*) *meaning preservation* (M_P): how much of the meaning in the original sentence is preserved in the candidate simplification?

Simplicity gain	Meaning preservation
5 – Much simpler	5 – Fully preserved
4 – Somewhat simpler	4 – Mostly preserved
3 – Same difficult	3 – Partially preserved
2 – More difficult	2 – Completely different
1 – Unintelligible	1 – Unintelligible

Table 3: Labels assigned to each score. Inspired on the taxonomy by Yamaguchi et al. (2023).

Judges were shown the original sentences along with the simplified versions proposed by one of the three respondents in a random order (see Figure 8 in Appendix C). Based on their assessment, we observed that three judges, each with slightly different but closely aligned evaluations, agreed that Expert 1 was the most proficient at providing simpler sentences (see Figure 6). Whereas Expert 1 achieves high S_G scores, Expert 2 makes more conservative decisions, leading to a lower gain, yet obtaining a higher average than Expert 1 in the M_P dimension. ChatGPT receives an intermediate mean score for both criteria, and seems to find a trade-off between these two seemingly inverse tendencies.

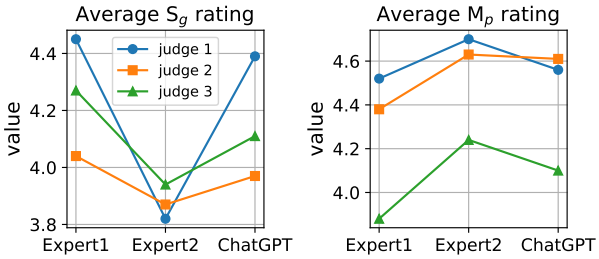


Figure 6: Average rating on S_G and M_P dimensions from the judges.

5. Conclusions and Further Work

In this paper, we have presented a taxonomy of the simplification strategies applied on the basis of French spontaneous transcripts for a non-native audience. To date, research on simplification has been primarily based on written sources, but seldom on spoken-based ones. Due to the lack of guidelines allowing us to steer this process, we

adopted an *intuitive* approach to characterize the strategies employed to simplify this kind of data. By means of a survey-based study, we collected a set of simplifications from 2 native French-speaking linguists. More precisely, we asked them to provide an explainable simplified version of 100 spoken utterances randomly selected from ORFÉO-TEST. Additionally, we have compared human-crafted speech simplifications, with machine-generated ones. Based on the quantitative evaluation, ChatGPT tends to suppress fewer elements when generating simplified outputs compared to human experts. As for the qualitative evaluation, it suggests an inverse correlation between S_G and M_P criteria. Results show that Expert 1 achieves higher S_G than ChatGPT, but the latter strikes a more balanced compromise between the two dimensions.

With this work, we provide a multi-reference set that allows to map the existing ORFÉO-TEST audio-transcript pairs with simpler counterparts. Assuming that the intuitions provided by experts serve as ground truth simplified sentences, this resource can be further used to assess automated solutions for generating spontaneous speech simplifications. For these reasons, we have released on a GitHub repository the resulting set mapping the original transcripts to their corresponding expert simplifications, named PROPICTO-ORFÉO-SIMPLE[5].

Furthermore, by annotating edit operations, we enable a finer-grained evaluation and a better understanding of the patterns that a model would have applied. This can promote greater explainability compared to conventional scores used to assess model performance (*i.e.*, BLEU or SARI). These overall metrics often provide little information about the simplification operations that the system has learned. Additionally, this in-depth examination can further serve as the groundwork for defining guidelines on speech-based simplification.

As for the limitations of the study, the lack of context in the manual sentence-level simplification was pointed out by the experts as a difficulty for its completion. Of course, providing context would have facilitated the task, especially within spontaneous speech, which is by nature interactive and conversational. However, we chose random proportionate sampling with the aim of favoring a better representativeness of the extracted sample. Consequently, the resulting data being analyzed lacked context as the sentences comprising it originated from various strata and were not linked to a single conversation.

[5] PROPICTO-ORFÉO-SIMPLE is made available on the following link: https://www.ortolang.fr/market/corpora/propicto.

Acknowledgements

This work is part of the PROPICTO (French acronym standing for *PRojection du langage Oral vers des unités PICTOgraphiques*) project, funded by the Swiss National Science Foundation (N°197864) and the French National Research Agency (ANR-20-CE93-0005).

We would also like to express our sincere gratitude to the linguists who took the time to perform the manual simplification task, and to the participants who completed the qualitative intrinsic evaluation.

6. Bibliographical References

David Allen. 2009. A study of the role of relative clauses in the simplification of news texts for learners of English. *System*, 37:585–599.

Katrin Angerbauer, Heike Adel, and Ngoc Thang Vu. 2019. Automatic Compression of Subtitles with Neural Networks and its Effect on User Experience. In *Interspeech 2019*, pages 594–598. ISCA.

Thierry Bazillon, Vincent Jousse, Frédéric Béchet, Yannick Estève, Georges Linarès, and Daniel Luzzati. 2008. La parole spontanée : transcription et traitement. In *Traitement Automatique des Langues, Volume 49, Numéro 3 : Recherches actuelles en phonologie et en phonétique : interfaces avec le traitement automatique des langues*, pages 47–76. ATALA.

Daniel Becquemont. 1996. Le sous-titrage cinématographique : contraintes, sens, servitudes. In Yves Gambier, editor, *Les transferts linguistiques dans les médias audiovisuels*, pages 145–155. Presses universitaires du Septentrion.

Christophe Benzitoun, Jeanne-Marie Debaisieux, and Henri-José Deulofeu. 2016. Le projet ORFÉO : un corpus d'étude pour le français contemporain. *Corpus, 15*.

Rocío Bernabé and Piero Cavallo. 2021. Easy-to-Understand Access Services: Easy Subtitles. In *Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments*, pages 241–254. Springer International Publishing.

Claire Blanche-Benveniste. 1997. *Approches de la langue parlée en français*. Ophrys.

Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2022. Linguistically-Based Comparison of Different Approaches to Building Corpora for Text Simplification: A Case Study on Italian. *Frontiers in Psychology*, 13.

François Buet and François Yvon. 2021. Toward Genre Adapted Closed Captioning. In *Interspeech 2021*, pages 4403–4407. ISCA.

Andrew Caines, Michael McCarthy, and Paula Buttery. 2017. Parsing Transcripts of Speech. In *Proceedings of the Workshop on Speech-Centric Natural Language Processing*, pages 27–36. Association for Computational Linguistics.

Arnaldo Candido, Erick Maziero, Lucia Specia, Caroline Gasperin, Thiago Pardo, and Sandra Aluisio. 2009. Supporting the Adaptation of Texts for Poor Literacy Readers: A Text Simplification Editor for Brazilian Portuguese. In *NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42.

Ronald Carter and Michael McCarthy. 2017. Spoken Grammar: Where Are We and Where Are We Going? *Applied Linguistics*, 38(1):1–20.

Walter Daelemans, Anja Höthker, and Erik Tjong Kim Sang. 2004. Automatic Sentence Simplification for Subtitling in Dutch and English. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Jan De Belder and Marie-Francine Moens. 2010. Text Simplification for Children. In *Workshop on Accessible Search Systems*, pages 19–26.

Anna Dmitrieva, Antonina Laposhina, and Maria Lebedeva. 2021. A Comparative Study of Educational Texts for Native, Foreign, and Bilingual Young Speakers of Russian: Are Simplified Texts Equally Simple? *Frontiers in Psychology*, 12.

Solene Evain, Solange Rossato, Benjamin Lecouteux, and François Portet. 2022. Typologie de la parole spontanée à des fins d'analyse linguistique et de développement de systèmes de reconnaissance automatique de la parole. In *Proc. XXXIVe Journées d'Études sur la Parole*, pages 212–221.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C. Ziegler. 2020. ALECTOR: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1353–1361.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Michael Alexander Kirkwood Halliday. 1985. *Spoken and written language*. Oxford University Press.

David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. 2023. Dancing Between Success and Failure: Edit-level Simplification Evaluation using SALSA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3495. Association for Computational Linguistics.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a Lexical Simplifier Using Wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 458–463.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217.

Juhani Luotolahti and Filip Ginter. 2015. Sentence Compression For Automatic Subtitling. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 135–143. Linköping University Electronic Press, Sweden.

Daniel Luzzati. 1998. Rhétorique et description de l'oral. *VERBA*, 25:7–30.

Daniel Luzzati. 2013. Enseigner l'oral spontané ? In Beacco J.C., editor, *Ethique et politique en didactique des langues*, pages 188–207. Éditions Didier.

Anna Matamala. 2022. Easy-to-understand language in audiovisual translation and accessibility: State of the art and future challenges. *XLinguae*, 15(2):130–144.

Sneha Mehta, Bahareh Azarnoush, Boris Chen, Avneesh Singh Saluja, Vinith Misra, Ballav Bihani, and Ritwik K. Kumar. 2020. Simplify-Then-Translate: Automatic Preprocessing for Black-Box Translation. In *AAAI Conference on Artificial Intelligence*.

Lucía Ormaechea, Pierrette Bouillon, Maximin Coavoux, Emmanuelle Esperança-Rodier, Johanna Gerlach, Jerôme Goulian, Benjamin Lecouteux, Cécile Macaire, Jonathan Mutal, Magali Norré, Adrien Pupier, and Didier Schwab. 2023a. PROPICTO: Developing Speech-to-Pictograph Translation Systems to Enhance Communication Accessibility. In *Proceedings of the 24th Annual Conference of the European*

Association for Machine Translation, pages 515–516. European Association for Machine Translation.

Lucía Ormaechea and Nikos Tsourakis. 2023. Extracting Sentence Simplification Pairs from French Comparable Corpora Using a Two-Step Filtering Method. In *Proceedings of the 8th Swiss Text Analytics Conference 2023*. Association for Computational Linguistics.

Lucía Ormaechea, Nikos Tsourakis, Didier Schwab, Pierrette Bouillon, and Benjamin Lecouteux. 2023b. Simple, Simpler and Beyond: A Fine-Tuning BERT-Based Approach to Enhance Sentence Complexity Assessment for Text Simplification. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 120—133.

Adrien Pupier, Maximin Coavoux, Benjamin Lecouteux, and Jerome Goulian. 2022. End-to-End Dependency Parsing of Spoken French. In *Proc. Interspeech 2022*, pages 1816–1820.

Luz Rello, Ricardo Baeza-Yates, and Horacio Saggion. 2013. DysWebxia: Textos Más Accesibles Para Personas con Dislexia. *Procesamiento del Lenguaje Natural*, 51.

Horacio Saggion. 2017. *Automatic Text Simplification*. Morgan & Claypool Publishers.

Matthew Shardlow and Raheel Nawaz. 2019. Neural Text Simplification of Clinical Letters with a Domain Specific Phrase Table. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 380–389. Association for Computational Linguistics.

Sanja Stajner and Maja Popovic. 2016. Can Text Simplification Help Machine Translation? *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242.

Laurens Van den Bercken, R. H. J. Sips, and C. Lofi. 2019. Evaluating Neural Text Simplification in the Medical Domain. *The World Wide Web Conference*.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Daichi Yamaguchi, Rei Miyata, Sayuka Shimada, and Satoshi Sato. 2023. Gauging the Gap Between Human and Machine Text Simplification Through Analytical Evaluation of Simplification Strategies and Errors. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 359–375. Association for Computational Linguistics.

Xingxing Zhang and Mirella Lapata. 2017. Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 584–594.

7. Language Resource References

Benzitoun, Christophe and Debaisieux, Jeanne-Marie and Deulofeu, Henri-José. 2016. *Corpus d'Étude pour le Français Contemporain (CEFC)*. Distributed via ORTOLANG. PID https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo/index.html.

A. Expert Simplification Task: LimeSurvey Example

Figure 7: A sample of a spontaneous utterance, along with the instructions for its simplification, on the LimeSurvey platform.

B. Simplification Examples: Respondents' Proposed Answers

Ex. I	**Input**	*ouais c'est ça sauf que moi on m'a jamais expliqué le rythme du coup*
	Expert1	On ne m'a pas expliqué le rythme
	Expert2	Oui, c'est ça, sauf qu'on ne m'a jamais expliqué le rythme.
	ChatGPT	Oui, c'est vrai, mais personne ne m'a jamais expliqué le rythme.
Ex. II	**Input**	*donc il a adapté euh cet univers cette légende à sa façon le réalisateur qui s' appelle Alexandre Astier*
	Expert1	Le réalisateur Alexandre Astier a adapté cette légende à sa manière
	Expert2	Le réalisateur, qui s' appelle Alexandre Astier, a adapté cet univers à sa façon.
	ChatGPT	Donc, Alexandre Astier, le réalisateur, a changé cette légende, cet univers, à sa manière.
Ex. III	**Input**	*mais enfin cette confrérie elle avait quand même un un avantage c' est que elle venait en aide euh aux malheureux*
	Expert1	Cette association avait un avantage, elle a aidé les démunis
	Expert2	Cette association avait l'avantage de venir en aide aux malheureux
	ChatGPT	Mais cette confrérie aidait quand même les malheureux.
Ex. IV	**Input**	*et c' était pas son style de bouffer de la merde*
	Expert1	c'était pas à son goût de manger mal
	Expert2	Et il/elle n'est pas habitué à manger des aliments de mauvaise qualité
	ChatGPT	Il n'aimait pas manger de mauvaises choses.
Ex. V	**Input**	*on sent que la prise de conscience de ce genre de choses elle s' est faite tard*
	Expert1	Nous pensons que la compréhension de ce problème est arrivée tard
	Expert2	La prise de conscience de ces choses-là est arrivée tard
	ChatGPT	Les gens ont commencé à comprendre ces choses tard.

Table 4: A set of examples extracted from the data sample along with the proposed simplifications.

C. Qualitative Evaluation: LimeSurvey Example

Figure 8: An example (comprising the original spontaneous transcript and a candidate simplification) of the qualitative evaluation task on the LimeSurvey platform.

DARES: Dataset for Arabic Readability Estimation of School Materials

Mo El-Haj[1], Sultan Almujaiwel[2], Damith Premasiri[1], Tharindu Ranasinghe[3], Ruslan Mitkov[1]
[1]School of Computing and Communications, Lancaster University, UK
[2]College of Humanities and Social Sciences, King Saud University, KSA
[3]School of Computer Science and Digital Technologies, Aston University, UK
{m.el-haj, d.dolamullage, r.mitkov}@lancaster.ac.uk, salmujaiwel@ksu.edu.sa, t.ranasinghe@aston.ac.uk

Abstract

This research introduces DARES, a dataset for assessing the readability of Arabic text in Saudi school materials. DARES compromise of 13,335 instances from textbooks used in 2021 and contains two subtasks; **(a)** Coarse-grained readability assessment where the text is classified into different educational levels such as primary and secondary. **(b)** Fine-grained readability assessment where the text is classified into individual grades. We fine-tuned five transformer models that support Arabic and found that `CAMeLBERTmix` performed better in all input settings. Evaluation results showed high performance for the coarse-grained readability assessment task, achieving a weighted F1 score of 0.91 and a macro F1 score of 0.79. The fine-grained task achieved a weighted F1 score of 0.68 and a macro F1 score of 0.55. These findings demonstrate the potential of our approach for advancing Arabic text readability assessment in education, with implications for future innovations in the field.

Keywords: Arabic, text readability, LLMs, NLU, Saudi school textbooks.

1. Introduction

Text readability refers to the measure of how easily a piece of text can be understood by its readers (Dale and Chall, 1949). Assessing text readability is important for both educators and learners, as it helps improve the readability levels of educational materials (Zamanian and Heydari, 2012). As a result, automatic readability assessment tools have been developed in recent years to automate the process of selecting reading materials and assessing reading ability. Furthermore, automatic readability assessment tools have proven useful in other natural language processing (NLP) applications such as machine translation (Alva-Manchego and Shardlow, 2022) and text simplification (Aluisio et al., 2010; North et al., 2022, 2023, 2024).

Earlier automatic readability assessment tools depended on classical formulas incorporating values such as average word length and average sentence length (Flesch, 1948). However, supervised machine learning (ML) methods have recently proved successful in assessing readability (Imperial, 2021; Qiu et al., 2021). ML-based methods can consider a broader range of text features than classical formulas, such as sentence complexity, vocabulary difficulty, and the cohesion and consistency of the texts. Very recently, deep learning-based ML models have helped automate feature extraction and loosen the dependence on language specificities of automatic readability assessment (Martinc et al., 2021; Imperial, 2021).

Supervised ML models that we described before typically require a training dataset to train the models. Particularly, deep learning models would require a more extensive training set as these models fine-tune thousands of parameters in the training process (Devlin et al., 2019). To address this need, the NLP community has shown significant interest in constructing readability datasets that can be used to train the ML models (Imperial, 2021). Several datasets have been developed for high-resource languages such as English (Xia et al., 2016), Spanish (Morato et al., 2021), German (Naderi et al., 2019) and Portuguese (Leal et al., 2018). For Arabic also, there exist several datasets and methods which aim to develop readability estimation applications (Baazeem et al., 2021; Berrichi et al., 2022).

In this research, we revisit the task of Arabic readability assessment in school textbooks. While there exist several datasets for readability assessment in Arabic, we argue that these datasets have limited practical relevance in real-world scenarios. For example, Al Khalil et al. (2018) introduced a large corpus consisting of texts randomly selected from the school textbooks of the United Arab Emirates and trained different ML models to predict the grade given a text. While this approach can produce a large number of training instances, the texts do not contain information about which concepts a particular text is trying to describe in the textbook. Therefore, with such a corpus, it is challenging to discern whether a certain description of a concept is readable and consequently understandable for a given grade level, limiting its practical relevance. In this research, we address this limitation by introducing DARES, which diverges from the practice of

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 103–113

randomly collecting text from school textbooks. Instead, DARES only consists of texts that describe certain concepts. As far as we know, this is the first readability dataset that contains information about concepts. We also introduce novel neural network architectures that incorporate concepts in the readability measure.

The main contributions of this research are;

1. We introduce DARES: A dataset for Arabic readability estimation based on Saudi school material. DARES has two subtasks; **(a)** Coarse-grained readability assessment where the text is classified into different educational levels such as primary and secondary. **(b)** Fine-grained readability assessment where the text is classified into individual grades.

2. We trained multiple transformer models on both subtasks of DARES that support Arabic with different input settings and evaluated the results. We also conducted a detailed error analysis.

3. We released DARES[1], as an open-access dataset alongside the trained machine-learning models.

2. Related Work

Text readability assessment has been an active area of research across various languages for the past decade, with initial methods proposing metric formulas based on factors like sentence length and word syllable count (Crossley et al., 2011; Pitler and Nenkova, 2008). Subsequently, machine learning approaches emerged, leveraging features extracted from the text at different levels, such as words, phrases, and sentences (François and Miltsakaki, 2012). The advent of Transformer models, particularly those stemming from the BERT architecture, in the last five years revolutionised the field by employing self-attention mechanisms to grasp word context, thereby advancing the state-of-the-art in various NLU tasks (Devlin et al., 2019). Despite advancements, the development of more sophisticated techniques and language models tailored for Arabic NLU is ongoing, necessitating greater attention to custom data to accommodate the diversity of Arabic text-level readability (El-Haj and Rayson, 2016).

However, it is still not as efficient as the state-of-the-art models built for English (El-Haj et al., 2018). The work of (Tanaka-Ishii et al., 2010) sorted the readability using SVM with insufficient training data. François (2015) conducted a study on the intersection of readability and computational linguistics, applying NLP-based historical readability research. That same year, (Saddiki et al., 2015) researched Arabic as a Foreign Language using a public corpus and NLP techniques. The focus on Arabic continued with (Alotaibi et al., 2016) work on the readability of medicine leaflets and (Malik et al., 2019; El-Haj et al., 2018) introduction of an Arabic-specific readability assessment. The experiments on readability assessment in Arabic have been growing, with a number of studies published in recent years and reviewed by some studies (Cavalli-Sforza et al., 2018; Nassiri et al., 2023; El-Haj and Rayson, 2016; Bessou and Chenni, 2021; Khallaf and Sharoff, 2021). Al Khalil et al. (2018) describe a reading corpus in Modern Standard Arabic where the authors select random texts for each grade to compile a corpus.

Previously, readability assessments have been conducted using various approaches. (Bessou and Chenni, 2021; Saddiki et al., 2015; Khallaf and Sharoff, 2021) categorised documents into different readability levels, ranging from 'easy' to 'very difficult'. The study by (Vajjala, 2022) addressed the scarcity of resources for readability assessment across languages, including Arabic (Vajjala, 2022; Vajjala and Lučić, 2018). (Cavalli-Sforza et al., 2018) emphasised the need for more tools and resources in Arabic readability research. Additionally, (Dalvean and Enkhbayar, 2018) proposed a new readability measure for fiction texts, while (Al Khalil et al., 2018) introduced a levelled reading corpus for Arabic text readability estimation based on the UAE curriculum and fiction. (Malik et al., 2019) highlighted the necessity for improved Arabic readability tools in patient educational materials, and (Benzahra and Yvon, 2019) examined readability and comprehension in journalistic texts.

Machine learning techniques have also been applied in Arabic text classification. (Bessou and Chenni, 2021) explored this area, while (Khallaf and Sharoff, 2021) utilised Arabic-BERT and XLM-R for Arabic sentence difficulty classification. Furthermore, (Vajjala, 2022) provided a comprehensive review of readability assessment trends, focusing on traditional readability formulas.

In 2023, significant advancements were made. (Nassiri et al., 2023) delved into Arabic readability approaches, while (Crossley et al., 2023; Vajjala, 2022) investigated the use of transformers for readability assessment and highlighted open challenges in the field, respectively. Finally, (Hazim et al., 2023) introduced a practical application: a Google Docs add-on for Arabic readability, featuring lemmatisation and a readability lexicon.

Our approach diverges from prior research. We emphasise the extraction of texts based on con-

[1] https://github.com/DamithDR/arabic-readability-assessment

cepts (a specific word accompanied by descriptive text that explicates its meaning), a departure from traditional methods as it enables us to gauge readability in relation to specific concepts and assess comprehension levels across different grade levels, a capability lacking in previous studies, e.g. (El-Haj and Rayson, 2016).

3. DARES Dataset

The DARES dataset is sourced from the books from the Saudi Education school system. The dataset includes schoolbooks from grades 1 to 12, aligning with the educational framework set by the Ministry of Education in Saudi Arabia[2]. This dataset is derived from the new literacy plan introduced in 2021 by the Saudi Ministry of Education, incorporating the latest educational content updates for students across these grades. The curriculum covers a wide range of subjects, including religious and social studies, languages, sciences, technology, physical education, life skills, activity classes, and artistic pursuits.

3.1. Dataset Preparation

We first selected 307 books authored in Arabic for the 1-12 grades in Saudi schools for 116 subjects. Out of them, 48 were from the early elementary level (Grades 1-3), 62 were taken from the upper elementary level (Grades 4-6), 86 were from the intermediate level (Grades 7-9), and 111 were from the high school level (Grades 10-12). Some schoolbooks are published in English, and we did not include them in this research. The statistics about subjects and number of books are shown in Table 1.

Grade	Books	Words	Subjects
1	18	64,590	7
2	15	71,594	5
3	15	104,357	5
4	21	294,704	7
5	23	387,750	7
6	18	337,551	7
7	28	619,777	8
8	24	488,841	8
9	34	885,880	11
10	65	2,106,350	26
11	33	1,237,985	16
12	13	572,478	10
Total	307	7,171,857	116

Table 1: Dataset Statistics for each tier and grade with respect to number of books, words and subjects.

[2] https://moe.gov.sa/

Subject	Books	Words
AI	1	53,314
Arabic Language	67	645,855
Artistic Education	21	516,448
Arts	1	19,208
Athletics	1	44,164
Biology	7	519,878
Business	2	101,787
Chemistry	9	407,466
Computer Science	4	105,997
Critical Thinking	7	110,260
Data Science	1	32,843
Decision Making	1	113,728
Digital Skills	15	570,145
Ecology	3	95,797
Economics	1	27,013
Finance	4	93,315
Geography	3	64,671
Geology	1	60,799
Hadith	1	14,909
Health	3	112,699
History	3	68,976
IoT	2	44,966
Islamic Studies	37	574,684
Law	1	23837
Life and Family Skills	23	290,802
Life Skills	4	49,303
Management	1	101,516
Math	5	137,955
Physics	7	551,156
Professional Skills	2	25,711
Psychology	1	46,683
Quran Sciences	1	23,648
Research Skills	5	139,526
Science	28	637,783
Sociology	18	538,794
Software Eng	1	27,870
Tech	6	178,341
Total	307	7,171,857

Table 2: Dataset statistics for each subject with respect to number of books and words.

3.2. PDF to Text Conversion, OCR Processing, and Post-Editing

As the first step, we converted the original educational materials, provided in PDF format, into plain text files. We utilised tools specifically designed for PDF-to-Text conversion. In order to handle instances where the text was embedded within images, we used the open-source Arabic-trained OCR from Tesseract OCR[3]. Table 2 lists the names of the subjects, the number of textbooks, and the count of running tokens in each.

The process of extracting accurate texts proved

[3] https://github.com/tesseract-ocr/tessdata

to be less efficient than anticipated due to the variety of Arabic fonts used in the PDF files, such as AXtManal, GESSTwoLight, Helvetica, and Lotus. These fonts introduced an added complexity for the OCR. Therefore, the text obtained through OCR and subsequent conversion underwent a post-editing phase. This step was conducted by an Arabic language linguist (also a co-author of this paper) who meticulously reviewed and refined the dataset, ensuring that the 13,335 extracted key words, along with their corresponding texts, were accurately represented. This process guaranteed both syntactic accuracy and semantic coherence within the dataset, which was derived from the 307 textbooks.

3.3. Text Pre-processing

As the first pre-processing step, we used sentence segmentation to divide the text into discrete sentences. We also used the Arabic tokenisation framework[4] to perform text tokenisation and Part-of-Speech (POS) tagging.

As we mentioned before, the DARES dataset focused only on the sentences that describe concepts. Therefore, we selected sentences beginning with a 'DET NOUN' POS tag and grouped them by grade level, focusing specifically on sentences that start with the Arabic definite article ' ' at the beginning of texts in the post-processed dataset. This technique was employed because words starting with ' ' are often keywords that are defined and explicated in the curriculum. Subsequently, we carefully reviewed the extracted words, along with their corresponding texts and subjects, and removed instances where the context did not serve to define the concepts of the words. This refinement process ensured that our dataset was not only accurately tagged but also contextually coherent and relevant to the concepts and subjects under consideration. The final dataset had 13,335 instances describing concepts. Several samples of the dataset is available on Table 3.

3.4. Tasks

In the DARES dataset, we used a hierarchical labelling schema that contains two tasks, which we describe below.

(I) Coarse-grained readability assessment In this task, we grouped the grades into four levels: early elementary level (Grades 1-3), upper elementary level (Grades 4-6), intermediate level (Grades 7-9), and high school level (Grades 10-12) aligning with the Saudi school's system and used them as the labels. Figure 1 shows the number of concepts and the token distribution of each level.

[4]https://github.com/CAMeL-Lab/camel_tools

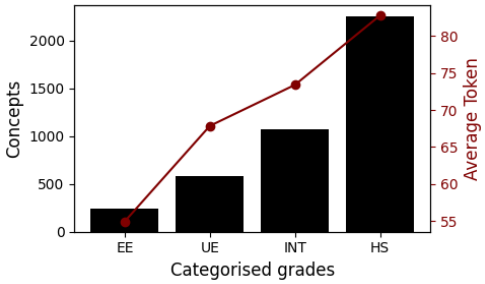


Figure 1: Concept and token distribution for the Coarse-grained level in DARES dataset. The labels are early elementary (EE), upper elementary (UE), intermediate (INT), and high school (HS).

(II) Fine-grained readability assessment For this task, we employed the original grades as the labels, resulting in a total of 12 distinct labels. Figure 2 shows the number of concepts and the token distribution of each level.

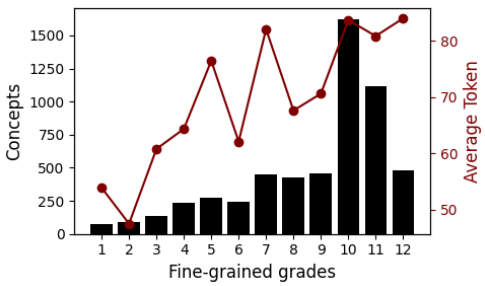


Figure 2: Concept and token distribution for the Coarse-grained level in DARES dataset.

Our methodology is based on neural transformers, which have provided state-of-the-art results in many NLP tasks, including readability assessment. We experimented with several transformer models that support Arabic; `XLM-R Large` (Conneau et al., 2020), `mBERT` (Devlin et al., 2019), `AraELECTRA` (Antoun et al., 2021), `AraBERTv2` (Antoun et al., 2020) and `CAMeLBERTmix` (Inoue et al., 2021). These models have performed well in different Arabic NLP tasks (Premasiri et al., 2022).

With each transformer model, we experimented with three input settings.

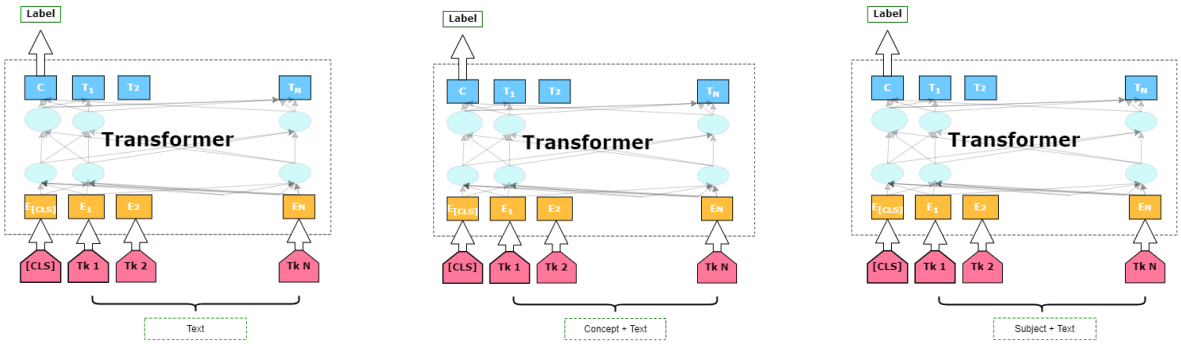
1. **text**; where we only feed the text as the input to the transformer model.

2. **concept + text**; where we concatenate the concept to the text and provide as the input to the transformer model.

3. **subject + text**; where we concatenate the subject to the text and provide as the input to the transformer model.

Subject	Concept	Arabic Text	Label(s) CG	FG
الأحياء (Biology)	الغذاء (Food)	الغذاء من الطاقة ، وهو كمية الحرارة اللازمة لرفع درجة العضلات القلبية عضلات لا إرادية حرارة الماء درجة سيليزية واحدة (Food is a form of energy, which is the amount of heat needed to raise the temperature of the involuntary cardiac muscles by one Celsius degree of water heat.)	HS	G10
العلوم (Science)	الخلية (Cell)	الخلية المجهرية تتكون جميع المخلوقات الحية من خلايا ، انظر الشكل وتعد البكتيريا أصغر المخلوقات الحية . ويتكون جسمها من خلية واحدة فقط (All living creatures are composed of microscopic cells, see the figure. Bacteria are the smallest of living organisms and consist of only one cell.)	INT	G7
العلوم (Science)	البذرة (Seed)	البذرة جزء النبات الذي ينمو ليعطي نباتا جديدا . البذرة داخل ثمرة الخوخ يمكن أن تنمو فتصير شجرة خوخ (A seed is a part of the plant that grows to produce a new plant. The seed inside a peach fruit can grow into a peach tree.)	EE	G1

Table 3: Example data instances. The column Subject represents the relevant subject the text was extracted from, and the column Concept indicates the sub-area in the subject which the text was extracted from while Text shows the extracted text. CG shows the course-grained label. The labels are early elementary (EE), upper elementary (UE), intermediate (INT), and high school (HS). FG shows the fine-grained label to the text. English translations are in green.



(a) Text (b) Concept + Text (c) Subject + Text

Figure 3: The input setting used for experiments

From an input sentence, transformers compute a feature vector $\boldsymbol{h} \in \mathbb{R}^d$, upon which we build a classifier for the task. For this task, we implemented a softmax layer, i.e., the predicted probabilities are $\boldsymbol{y}^{(B)} = \text{softmax}(W\boldsymbol{h} + b)$, where $W \in \mathbb{R}^{k \times d}$ is the softmax weight matrix, and k is the number of labels.

For all the experiments, we used a batch size of eight, Adam optimiser with learning rate $2\text{e}{-5}$, and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of the transformer model, as well as the parameters of the softmax layer, were updated. All the models were trained for five epochs.

4. Results and Evaluation

We evaluated all of our models and their variations in both tasks in DARES separately. We first divided the dataset into training sets (70%), testing sets (20%) and validation sets (10%). We trained the model on the training set and fine-tuned it on the validation set. Finally, we evaluated the performance on the testing set. For both subtasks, we used Macro F1 and Weighted F1 as the evaluation metrics to compare different models. We ran each experiment five times with five different random seeds and reported the mean. We also report the standard deviation.

4.1. Coarse-grained Readability Assessment

Table 4 shows the results for coarse-grained readability assessment. The CAMeLBERTmix model with the 'Subject+Text' input setting provided the best result, achieving a Weighted F1 score of 0.91 and a Macro F1 score of 0.79. AraELECTRA and AraBERTv2 with the same input setting followed closely to the best result, providing 0.89 Weighted F1 scores.

Input Setting	Model Name	Weighted F1	Macro F1
Text	XLM-R Large	0.53±0.13	0.32±0.15
	mBERT	0.66±0.17	0.47±0.21
	AraELECTRA	0.82±0.01	0.69±0.01
	AraBERTv2	0.81±0.00	0.70±0.01
	CAMeLBERTmix	0.84±0.00*	0.74±0.01*
Concept + Text	XLM-R Large	0.56±0.15	0.36±0.18
	mBERT	0.70±0.14	0.52±0.17
	AraELECTRA	0.82±0.00	0.70±0.01
	AraBERTv2	0.74±0.16	0.59±0.21
	CAMeLBERTmix	0.84±0.00*	0.75±0.01*
Subject + Text	XLM-R Large	0.80±0.02	0.59±0.04
	mBERT	0.85±0.03	0.65±0.06
	AraELECTRA	0.89±0.01	0.72±0.05
	AraBERTv2	0.89±0.00	0.75±0.01
	CAMeLBERTmix	**0.91±0.00***	**0.79±0.01***

Table 4: Test set results for coarse-grained readability assessment. We report **Weighted F1** and **Macro F1** for all the models and input settings. The best result from all the experiments are highlighted in **bold**.

It is also noticeable that the multilingual models such as XLM-R Large and mBERT are outperformed by Arabic specific transformer models such as AraBERTv2, AraELECTRA and CAMeLBERTmix in all the input settings. This highlights the effectiveness of language-specific transformer models in readability assessment tasks.

Overall, the 'Subject+Text' setting improved the results of all the transformer results. However, it should be noted that the 'Text' setting also provides close results, especially for Arabic-specific transformer models.

4.2. Fine-grained Readability Assessment

Table 5 presents the results for coarse-grained readability assessment. As shown in the results, the 'Subject+Text' settings with the CAMeLBERTmix model also provided the best results for the fine-grained readability assessment task, achieving a Weighted F1 score of 0.68 and a Macro F1 score of 0.55. Similar to the previous task, all the models demonstrated high performance in the 'Subject+Text' setting. Furthermore, Arabic-specific transformer models produced superior results than the multilingual transformer models.

Input Setting	Model Name	Weighted F1	Macro F1
Text	XLM-R Large	0.29 ±0.12	0.15 ±0.10
	mBERT	0.51 ±0.06	0.37 ±0.06
	AraELECTRA	0.56 ±0.01	0.42 ±0.01
	AraBERTv2	0.40 ±0.20	0.28 ±0.20
	CAMeLBERTmix	0.59 ±0.01	0.49 ±0.01
Concept + Text	XLM-R Large	0.25 ±0.13	0.12 ±0.11
	mBERT	0.53 ±0.02	0.39 ±0.03
	AraELECTRA	0.56 ±0.01	0.41 ±0.02
	AraBERTv2	0.56 ±0.01	0.44 ±0.01
	CAMeLBERTmix	0.60 ±0.01	0.51 ±0.01
Subject + Text	XLM-R Large	0.51 ±0.02	0.30 ±0.03
	mBERT	0.59 ±0.02	0.41 ±0.04
	AraELECTRA	0.63 ±0.00	0.44 ±0.01
	AraBERTv2	0.61 ±0.02	0.44 ±0.02
	CAMeLBERTmix	**0.68 ±0.00**	**0.55 ±0.01**

Table 5: Test set results for fine-grained readability assessment. We report **Weighted F1** and **Macro F1** for all the models and input settings. The best result for each input setting is marked as *, and the best result from all the experiments are highlighted in **bold**.

It should also be noted that the F1 scores for the fine-grained task are lower than the coarse-grained task. However, this is expected since the fine-grained task has more classes compared to the coarse-grained task.

5. Error Analysis

In this section, we provide a detailed error analysis of the two tasks. For the error analysis, we only use the best model and the input setting from the previous section, CAMeLBERTmix, with the 'Subject+Text' setting. The error analysis is conducted with the confusion matrix and the misclassified instances in the test set.

5.1. Coarse-grained Readability Assessment

Figure 4 illustrates the confusion matrix for coarse-grained readability assessment. Overall, the testing dataset comprises 2681 instances, among which only 252 were misclassified, indicating a relatively low error rate.

As shown in Figure 4, notable misclassifications happen between close levels such as UE and EE, where 36 UE texts were occasionally mistaken as EE. However, misclassification between distant levels such as EE and HS, are very rare.

In the following list, we show some misclassified instances with their translations in the coarse-grained task.

1. True label: EE, Predicted label: UE

Sample texts:

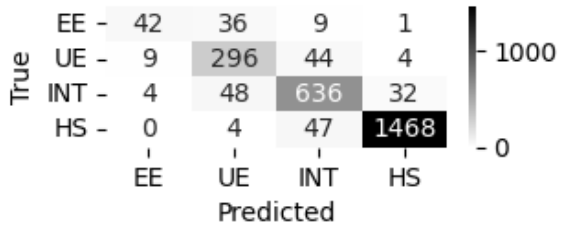


Figure 4: Confusion matrix for coarse-grained text readability estimation. The labels are early elementary (EE), upper elementary (UE), intermediate (INT), and high school (HS).

المهارات الرقمية : القواعد التي عليك اتباعها أثناء استخدام وسائل التواصل الاجتماعي يجب ألا تشارك المعلومات الشخصية مطلقا مع الأشخاص الذين تتعرف عليهم عبر الإنترنت ، ويشمل ذلك اسمك وعنوانك ورقم هاتفك ، وكذلك بريدك الإلكتروني وكلمات المرور .

Translation: Digital skills: The rules you must follow while using social media include never sharing your personal information with people you meet online. This includes your name, address, phone number, as well as your email and passwords.

2. True label: HS, Predicted label: INT

Sample texts:

الرياضيات : الاهتمام بالمهارات الرياضية ، والتي تعمل على ترابط المحتوى الرياضي وتجعل منه كلا متكاملا ومن بينها مهارات التواصل الرياضي ، ومهارات الحس الرياضي ، ومهارات جمع البيانات وتنظيمها وتفسيرها ، ومهارات التفكير العليا .

Translation: Mathematics: It is important to pay attention to mathematical skills, which interconnect mathematical content, making it an integrated whole. These skills include mathematical communication, sense of maths, data collection, organisation and interpretation skills, and higher-order thinking skills.

3. True label: UE, Predicted label: HS

Sample texts:

اللغة العربية : الترادف هو ما اختلف لفظه واتفق معناه ، أو هو إطلاق عدة كلمات على مدلول واحد ، كالأسد والليث وأسامة التي تعني مسمى واحدا ، والحسام والسيف والمهند معنى .

Translation: Arabic Language: Synonymy is when different words have the same meaning, or when several words refer to the same signified thing, such as "أسد", "ليث", and "أسامة" which all mean 'lion, and "حسام", "سيف", "مهند" which carry the same meaning for 'sword'.

4. True label: UE, Predicted label: INT

Sample texts:

العلوم : البلاستيدات الخضراء ، وهي مملوءة بمادة خضراء تسمى الميتوكندريا يحرق الغذاء في هذا الجزء . أما الخلية الحيوانية فلا تحتوي على البلاستيدات أو الكلوروفيل . الخلايا النباتية لها جدار خلوي هناك جدار صلب يحيط بالخلية النباتية يسمى الجدار الخلوي ، يعطيها شكلا يشبه الصندوق . أما الخلايا البلاستيدات الخضراء تعد مصانع الغذاء في الخلية ، وتحتوي على مادة الكلوروفيل .

Translation: Science: Green plastids are filled with a green substance called mitochondria that burns food in this part. As for animal cells, they do not contain plastids or chlorophyll. Plant cells have a cell wall, there is a hard wall surrounding the plant cell called the cell wall, which gives it a box-like shape. As for the green plastids, they are the food factories in the cell and contain chlorophyll.

5. True label: HS, Predicted label: UE

Sample texts:

المهارات الرقمية : الإنترنت شبكة عالمية تتيح لأي حاسب متصل بها الاتصال بالحاسبات الأخرى . تقدم خدمات منها الشبكة العنكبوتية العالمية تعد أحد خدمات الإنترنت وهي نظام من المستندات المترابطة تسمى صفحات الويب ويمكن لكل صفحة ويب الارتباط بواحدة أو أكثر من الصفحات الأخرى . للوصول إلى صفحات الويب نستخدم برامج تسمى متصفحات الويب تتيح لنا تصفح هذه الصفحات والضغط على الروابط للانتقال إلى صفحات أخرى . تسمى هذه الروابط ارتباطات تشعبية . تعد كل صفحة ويب فريدة ويمكن التعرف عليها من خلال عنوان يسمى محدد مواقع الويب . لاحظ أن العنوان هنا يحتوي على اسم المضيف . بالإضافة إلى معلومات أخرى تستخدم للوصول إلى مستند معين لدى مضيف محدد .

Translation: Digital Skills: The internet is a global network that allows any computer connected to it to communicate with other computers. It offers services, one of which is the World Wide Web, a system of interlinked documents called web pages. Each web page can link to one or more other pages. To access web pages, we use programs called web browsers that allow us to browse these pages and click on links to go to other pages. These links are called hyperlinks. Each web page is unique and can be identified by an address called a URL. Note that the address here contains the host name, as well as other information used to access a specific document on a specific host.

6. True label: INT, Predicted label: UE

Sample texts:

المهارات الرقمية : البحث عن مجلد أو ملف عندما يكون لديك الكثير من الملفات على جهاز الحاسب الخاص بك ، فمن الطبيعي أن تنسى المكان الذي حفظتها فيه ، لذلك إذا كنت بحاجة إلى ملف ، فيمكنك البحث عنه .

Translation: Digital Skills: When you have many files on your computer, it is normal that you might forget where you saved them. Therefore, if you need a file, you can search for it.

7. True label: INT, Predicted label: HS

Sample texts:

علوم الحاسب : أشكال اللبنات : القبعات بدء المقاطع البرمجية واقتناص الأحداث . اللبنات القابلة للتكديس تكوين الخطوات البرمجية عبر صفها (تكديسها) مع بعضها . الكل حاوية للبنات الأخرى لتطبيق التأثير (تكرار ، تحقق) على محتوياتها من اللبنات . الشروط : تعيد قيم

منطقية (صواب / خطأ) يمكن استخدامها في كل الاختيار والتكرار . القيم: الحصول على البيانات بعد إجراء العمليات عليها . مثلاً : ضم سلسلتين من النصوص . توليد رقم عشوائي ، مدخلات المستخدم بعد إجابته على سؤال ما ، إلخ .

Translation: Computer Science: Types of building blocks: Start blocks for software pieces and capturing events. Stackable blocks compile the programming steps by lining them up (stacking) together. Container blocks apply effects (repeat, check) on their contained blocks. Conditions: Return logical values (true / false) that can be used in choice and repetition blocks. Values: Obtain data after performing operations on it. For example: concatenating two strings, generating a random number, user inputs after answering a question, etc.

Misclassifications naturally occur for texts that lie on the boundary between the later stages of EE and the early stages of UE within individual subjects. This is evident in cases 1, 2, 4, and 6 for 'Digital Skills', and in case 7 for 'Computer Science'.

5.2. Fine-grained Readability Assessment

Figure 4 illustrates the confusion matrix for coarse-grained readability assessment. Among the 2681 test instances, 871 instances were misclassified, which is higher than the coarse-grained. According to the confusion matrix, majority of the misclassification occur between the close grades which also belonged to the same label in the coarse-grained level. For example, 146 Grade 10 instances were misclassified as Grade 11 and 133 Grade 10 instances were misclassified as Grade 11. This illustrates that the model may struggle with distinguishing these grades. Furthermore, misclassifications are higher among the Grade 10,11 and 12, suggesting that better models should be deployed when assessing readability in these grades.

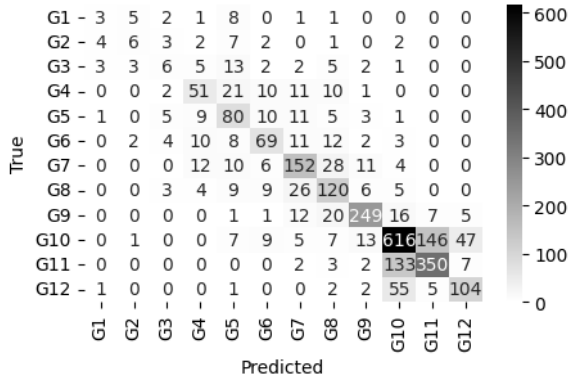


Figure 5: Confusion matrix for fine-grained text readability estimation. The labels are the different grades

6. Conclusions and Future Work

In this paper, we introduced DARES, a dataset for Arabic readability estimation based on Saudi school material. DARES has two subtasks; 1. Coarse-grained readability assessment where the text is classified into different educational levels such as primary and secondary. 2. Fine-grained readability assessment where the text is classified into individual grades.. To the best of our knowledge, DARES is the first readability assessment dataset based on concepts. We trained several transformer models that support Arabic under different input settings. The results showed that `CAMeLBERTmix` model provided the best results in both subtasks under the 'Subject + Text' setting. Furthermore, the results showed that multilingual models do not show competitive results compared to the Arabic specific models. In terms of error analysis, the majority of errors in the coarse-grained set were found in the 'Science' subject, followed by 'Arabic Language', 'Artistic Education', and 'Islamic Studies'. The fine-grained set also showed the highest number of errors in the same subjects, except for 'Artistic Education', with 'Chemistry' and 'Physics' adding to the error count as well.

The outcomes of this research hold significant implications for Arabic language education. DARES dataset can be used to The proposed readability assessment models offer educators a reliable means to prepare appropriate reading materials, enhancing the learning experience. Our research addresses the challenge of making complex concepts accessible to a wider range of students.

In future work, we hope to extend the dataset into more concepts and involve more school material. We would also like to incorporate large language models particularly trained in Arabic, such as Jais (Sengupta et al., 2023) in our methods as they have shown state-of-the-art results in many NLP tasks. Finally, we would like to develop a text summarisation pipeline for Arabic, which will have the capability to summarise the text, which has a high readability for a particular grade.

Limitations

While this study aims to advance Arabic text readability understanding, we have identified the following limitations.

1. Limited dataset size - We accept that DARES only has 13335 instances and is limited in size compared to other readability datasets. However, as we explained before, this is due to the unique nature of the way we collected DARES focusing on concepts.

2. Involvement of other readability datasets - As a

language resources paper, we did not focus on techniques such as transfer learning from other readability datasets that could have improved the results. In this paper, we focus more on the dataset collection.

3. Involvement of large language models - As we mentioned before, we did not experiment with any large language model. The models we experimented will serve as a baseline for the dataset.

Ethical Considerations

This research adheres to strict ethical standards throughout the data collection, analysis, and interpretation processes. We have taken careful measures to ensure compliance with ethical guidelines regarding educational materials, including copyright and intellectual property rights. It is important to note that the curriculum used in this research is not distributed or reused; rather, it is processed and produced solely as a training dataset for research purposes. This approach aligns with the policies outlined by the Saudi Authority for Intellectual Property and ensures the responsible use of educational materials. Additionally, our data preparation procedures prioritise transparency, integrity, spell-checking, and expert review to maintain accuracy and fidelity in our research outcomes.

References

Muhamed Al Khalil, Hind Saddiki, Nizar Habash, and Latifa Alfalasi. 2018. A leveled reading corpus of Modern Standard Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sihaam Alotaibi, Maha Alyahya, Hend Al-Khalifa, Sinaa Alageel, and Nora Abanmy. 2016. Readability of arabic medicine information leaflets: a machine learning approach. *Procedia Computer Science*, 82:122–126.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California. Association for Computational Linguistics.

Fernando Alva-Manchego and Matthew Shardlow. 2022. Towards readability-controlled machine translation of COVID-19 texts. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 287–288, Ghent, Belgium. European Association for Machine Translation.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Ibtehal Baazeem, Hend Al-Khalifa, and Abdulmalik Al-Salman. 2021. Cognitively driven arabic text readability assessment using eye-tracking. *Applied Sciences*, 11(18):8607.

Marc Benzahra and François Yvon. 2019. Measuring text readability with machine comprehension: a pilot study. In *Workshop on Building Educational Applications Using NLP*, pages 412–422, Florence, Italy.

Safae Berrichi, Naoual Nassiri, Azzeddine Mazroui, and Abdelhak Lakhouaja. 2022. Impact of feature vectorization methods on arabic text readability assessment. In *The International Conference on Artificial Intelligence and Smart Environment*, pages 504–510. Springer.

Sadik Bessou and Ghozlane Chenni. 2021. Efficient measuring of readability to improve documents accessibility for arabic language learners.

Violetta Cavalli-Sforza, Hind Saddiki, and Naoual Nassiri. 2018. Arabic readability research: Current state and future directions. *Procedia Computer Science*, 142:38–49.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2023. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55(2):491–507.

Scott A Crossley, David B Allen, and Danielle S McNamara. 2011. Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a foreign language*, 23(1):84–101.

Edgar Dale and Jeanne S Chall. 1949. The concept of readability. *Elementary English*, 26(1):19–26.

Michael Dalvean and Galbadrakh Enkhbayar. 2018. Assessing the readability of fiction: A corpus analysis and readability ranking of 200 english fiction texts. *Linguistic Research*, 35:137–170.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mahmoud El-Haj, Abdulaziz Malik, and Michael K Paasche-Orlow. 2018. Readability of arabic vs english patient educational materials. In *2018 SGIM Annual Meeting*.

Mahmoud El-Haj and Paul Edward Rayson. 2016. Osman: A novel arabic readability metric. In *Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Thomas François and Eleni Miltsakaki. 2012. Do nlp and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57.

Thomas François. 2015. When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, 2:79–97.

Reem Hazim, Hind Saddiki, Bashar Alhafni, Muhamed Al Khalil, and Nizar Habash. 2023. Arabic word-level readability visualization for assisted text simplification. *Computational Approaches to Modeling Language Lab*.

Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618, Held Online. INCOMA Ltd.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Nouran Khallaf and Serge Sharoff. 2021. Automatic difficulty classification of Arabic sentences. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 105–114, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Sidney Evaldo Leal, Magali Sanches Duran, and Sandra Aluísio. 2018. A nontrivial sentence corpus for the task of sentence readability assessment in portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413.

Abdulaziz Malik, Mahmoud El-Haj, and Michael K Paasche-Orlow. 2019. Readability of patient educational materials in english versus arabic. *HLRP: Health Literacy Research and Practice*, 3(3):e170–e173.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Jorge Morato, Ana Iglesias, Adrián Campillo, and Sonia Sanchez-Cuadrado. 2021. Automated readability assessment for spanish e-government information. *Journal of Information Systems Engineering and Management*, 6(2):em0137.

Babak Naderi, Salar Mohtaj, Karan Karan, and Sebastian Möller. 2019. Automated text readability assessment for german language: a quality of experience approach. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE.

Naoual Nassiri, Violetta Cavalli-Sforza, and Abdelhak Lakhouaja. 2023. Approaches, methods, and resources for assessing the readability of arabic texts. *ACM Transactions on Asian Low-Resource Language Information Processing (TALLIP)*, 22(4):95.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023. Deep learning approaches to lexical simplification: A survey. *arXiv preprint arXiv:2305.12000*.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2024. Multils: A

multi-task lexical simplification framework. *arXiv preprint arXiv:2402.14972*.

Kai North, Marcos Zampieri, and Tharindu Ranasinghe. 2022. ALEXSIS-PT: A new resource for Portuguese lexical simplification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6057–6062, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195.

Damith Premasiri, Tharindu Ranasinghe, Wajdi Zaghouani, and Ruslan Mitkov. 2022. DTW at qur'an QA 2022: Utilising transfer learning with transformers for question answering in a low-resource domain. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 88–95, Marseille, France. European Language Resources Association.

Xinying Qiu, Yuan Chen, Hanwu Chen, Jian-Yun Nie, Yuming Shen, and Dawei Lu. 2021. Learning syntactic dense embedding with correlation graph for automatic readability assessment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3013–3025, Online. Association for Computational Linguistics.

Hind Saddiki, Karim Bouzoubaa, and Violetta Cavalli-Sforza. 2015. Text readability for arabic as a foreign language. In *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. 2010. Sorting texts by readability. *Computational Linguistics*, 36(2).

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377", Marseille, France. European Language Resources Association.

Sowmya Vajjala and Ivana Lučić. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.

Mostafa Zamanian and Pooneh Heydari. 2012. Readability of texts: State of the art. *Theory & Practice in Language Studies*, 2(1).

Legal Text Reader Profiling: Evidences from Eye Tracking and Surprisal Based Analysis

Calogero J. Scozzaro°, Davide Colla°, Matteo Delsanto°, Antonio Mastropaolo•, Enrico Mensa°, Luisa Revelli•, Daniele P. Radicioni°
°Università degli Studi di Torino, Italy; •Università della Valle d'Aosta, Italy
°{calogero.scozzaro47@edu.unito.it; first-name.surname@unito.it}
•{initial.surname@univda.it}

Abstract

Reading movements and times are a precious cue to follow reader's strategy, and to track the underlying effort in text processing. To date, many approaches are being devised to simplify texts to overcome difficulties stemming from sentences obscure, ambiguous or deserving clarification. In the legal domain, ensuring the clarity of norms and regulations is of the utmost importance, as the full understanding of such documents lies at the foundation of core social obligations and rights. This task requires determining which utterances and text excerpts are difficult for which (sort of) reader. This investigation is the aim of the present work. We propose a preliminary study based on eye-tracking data of 61 readers, with focus on individuating different reader profiles, and on predicting reading times of our readers.

Keywords: reader profiling, eye-tracking, surprisal, legal documents, surface errors, semantic errors

1. Introduction

The certainty of law and equality in accessing legal sources are basic pillars of democratic systems: since legal and normative production is predominantly written, the analysis of these sources is crucial, and Natural Language Processing (NLP) may be also central in analyzing legal documents. Various NLP applications have been carried out in the legal domain, including summarizing legal documents, question answering systems, named entity extraction, and various types of judicial support systems. A comprehensive and detailed review and discussion of the relationship between AI (at large, but also including NLP applications) and law has been recently proposed by Villata et al. (2022).

Legislative and regulatory production may contain complex, highly specialized language, lengthy and convoluted sentences that are challenging to grasp. It is featured by specific semiotic and linguistic conventions, vocabulary, semantics, syntax and morphology that may result as difficult to understand by laypeople with no domain expertise. It is thus inherently harder to process than ordinary language: for example, legal documents such as SEC contract clauses (Tuggener et al., 2020) were compared to Simple English Wikipedia (Coster and Kauchak, 2011), and it was observed that legal clauses contain seven times as many tokens than those from Wikipedia, are featured by sentences over three times longer, and by more complex parse trees, as reported by Garimella et al. (2022). Text simplification may then provide valuable insights to legal professionals, and to laypeople lacking of domain expertise, as well. A preliminary issue, connected to textual simplification, is that of characterizing what is either obscure, ambiguous or deserving clarification, thereby needing to be reformulated. Some general readability indexes exist, building on basic parameters such as the number of sentences, the number of words, and the number of syllables, such as, e.g., the Flesch–Kincaid Grade Level (Kincaid et al., 1975; Leroy and Endicott, 2012) —which was also adapted to the Italian language (Piemontese et al., 1996)—, the Dale-Chall scores (Williams, 1972), and more global scoring approaches jointly considering lexical, morpho–syntactic and syntactic features (Dell'Orletta et al., 2011). However, no decisive evidences have been reported, nor models have been proposed able to explain the mechanisms underlying reading comprehension, to predict which elements are most disturbing and undermining for human comprehension, and whether these allow to characterize different classes of readers, e.g., differentiating between expert and non-expert reading performance.

Being able to profile readers, acquiring information on which phrases and sentences mostly impact on texts readability, and whether all readers are equally affected by such sources of difficulty would be therefore highly beneficial for text simplification, and would also allow delivering *ad hoc* paraphrases and rewriting tailored to specific reader groups or user needs.

Rich instruments are to date available to investigate language processing and comprehension in the reading task, by analyzing both readers response and internal properties of texts employed in the reading tasks: in the former case (investigating readers response) we may employ eye-tracking data, and in the latter one (focused on inherent

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 114–124

textual properties) we can analyze texts through language models. Eye tracking allows collecting precise data in form of timestamped fixations that describe and to a good extent allow to reconstruct readers' behavior. On the other side, the refinement and spread of language models allows to automatically perform subtle forms of linguistic analysis, such as determining the semantic coherence between a term and its surrounding context, thereby determining the predictability of words given their preceding context.

Several metrics have been proposed to analyze text reading and processing times. While the total reading time (TRT) —the overall duration of eye fixations for each word, including the backward regression movements— is supposed to grasp the time taken by the overall semantic integration (Radach and Kennedy, 2013), two partial and finer-grained measures have been also proposed: the duration of the first fixation (FFD) that allows estimating the cost underlying lexical access (Hofmann et al., 2022), and the number of fixations (NF), which is deemed to report about words integration in the context of what has been read so far (Frazier and Rayner, 1982).

This paper introduces the preliminary results of an experiment targeted at profiling reader's response while dealing with legal texts in Italian. To these ends we collected a corpus containing the normative production from the Aosta Valley Italian Region, composed by the Regional laws dating to the years 1960-2022 and the Regional regulations from the years between 1979 and 2022. In order to be able to gain insights on reader effort in both lexical access and semantic integration, the original utterances were manipulated and two different sorts of errors introduced: surface errors (consisting of morphological variations of terms) and semantic errors (through the introduction of unrelated terms). We present the results of a twofold experimentation: *i*) we report evidences from an eye tracking study involving 61 subjects who read a Law enacted by the Aosta Valley Region. In this setting, based on the analysis of FFDs and NFs we were able to discriminate two reader profiles exhibiting different reading strategies; and *ii*) we report a study targeted at predicting the associated reading times.

2. Background and Related Work

Two main eye movements are commonly individuated throughout the reading task, *fixations* and *saccades*. Fixations are brief stops (whose duration ranges from 50 to 1500 ms) that typically occur at each word; sometimes even more stops are needed, depending on words length and difficulty. A saccade is a fast (ranging from 10 to 100 ms) movement between each two fixations, that is used in repositioning the point of focus. In general, it is known from pioneering research in eye-tracking that individual words are fixated differently: e.g., Carpenter and Just (1983) reported that 85% content words and 35% function word get fixations. Among the main variables that impact on eye movements, one must additionally consider *i*) words length: shorter (2-3 letter) words are skipped 75% of the time, while longer (8 letter) words are fixated almost always (Rayner, 1978); and *ii*) syntactic and conceptual difficulty of the text at hand (Jacobson and Dodwell, 1979).

Eye tracking has been exploited to investigate reading at different levels, such as individual words or sentences and whole texts (Jarodzka and Brand-Gruwel, 2017). At the base level, the reading of words/sentences, regressions (backward eye movements) occurring within a single word indicate a processing problem with that word, while regressions between-words indicate comprehension problems at larger scale. A popular experimental technique employs a sliding window where parts of the text are masked (McConkie and Rayner, 1975): on such bases, different processing steps ('first pass' and 'second pass', and 'total reading times') have been hypothesized to underlie fixations and semantic processing (Rayner, 2009). Further cognitive phenomena have been also observed, such as the so-called spill-over effect (the word following an infrequent word is fixated for a longer time, while the previous word is still being processed), and the peripheral vision, that allows to perceive words that are not actually fixated. As regards as the second level, considering whole texts, the analysis typically considers sub-words or words (also AOIs, 'areas of interest') that convey specifically relevant information. An interesting measure in this setting is the 'reading depth', that measures quantities such as how much text is skipped by readers, the width of saccadic movements, and investigates strategies aimed at differentiating reading and scanning texts (such as to search for specific information). Situational models have been proposed to account for the inferential steps performed by readers and for the enrichment of read statements with prior knowledge to enforce semantic coherence (Zwaan et al., 1995). Consistent individual differences between readers also exist, associated to both lexical access and semantic integration. For instance, factors such as previous knowledge and reading expertise/ability are known to affect reading times. At the word/sentence level, good readers are more precise in targeting their regressions to the specific points that caused difficulties in comprehension; while employing prior knowledge proved beneficial for semantic integration purposes.

Most work focused on the processes underlying

lexical access and semantic integration falls into two broad approaches to model context. In the first case we have models concerned with the semantic relatedness between words and their context: in this setting, reading times are predicted based on the similarity between embeddings describing words and their context. Works adopting the second approach mostly rely on a probabilistic framework whereby words may be predicted based on their (left) context. In this view, words predictability should be intended as a function of the probability of a word given the context, and the probability of that word may work, in turn, as a main predictor of reading times: in essence, the less likely the emission of a word, the higher the *surprisal* associated to that word, and the longer the time it requires for readers to process it. Both the approaches based on relatedness and those relying on surprisal are surveyed in detail in (Salicchi et al., 2023).

In the last few years neural language models gained a central role in analyzing reading as well, since they are able to acquire conditional probability distributions over the lexicon that are also predictive of human processing times. While word length and frequency are widely acknowledged as predictors for determining lexical access, different sorts of language models have been recently compared to analyze and explain syntactic and semantic factors (Hofmann et al., 2022): N-gram models have been found to succeed in capturing short-range lexical access, while models based on recurrent neural networks show better fit in predicting the next-word. The role of model features (with focus on parameter size, spanning from 564M to 4.5B parameters) has been investigated in its impact on psychometric quality by de Varda and Marelli (2023), that challenge a widely accepted assumption postulating that the quality of predictions increases as the number of parameters grows. More specifically, also building on previous findings, such as by Shain et al. (2022), de Varda and Marelli (2023) observe that large multilingual Transformer-based models are outperformed by their smaller variants in predicting fixations, and thus are more suited to analyze lexical access and early semantic integration. Importantly enough, the authors make use of a masked language model rather than autoregressive models such as GPT (Devlin et al., 2018), thus accessing to both left and right context. Other studies found that the surprisal scores are strong predictors of reading times and eye fixations obtained through eye-tracking (Smith and Levy, 2008; Goodkind and Bicknell, 2018), along with a substantial linear relationship between models' next-word prediction accuracy and their ability in predicting reading times (Goodkind and Bicknell, 2018; Wilcox et al., 2020, 2023).

The issue of learners' reading ability has been addressed by Paracha et al. (2018), that investigated whether eye-tracking allows discriminating fluent and non-fluent students: skilled readers scan the text quickly, continuously and consistently from comprehension questions to the text, while weak readers read linearly, renouncing to select the most meaningful text elements.

3. Experiment

We start by introducing the data collected for our experiments, and then report about the experimentation: in the first experiment, we present a study on eye-tracking data of 61 persons reading a law from the Italian Region Aosta Valley and investigate their reading style when dealing with regular text, and in response to specific errors. In the second experiment we investigated whether and to what extent the fixations recorded in the former step can be predicted.

3.1. Data Collection: the Aosta Corpus

For our experiments the Aosta corpus was compiled; the corpus is composed of norms and regulations enacted by the Aosta Valley Italian Region. It contains $2,950$ Regional laws dating back to the years between 1960 and 2022, and 131 Regional regulations produced in the year between 1979 and 2022. Laws herein contain on the whole $172,669$ sentences (on average 58.53 sentences per law), $3,462,931$ tokens (on average, $1,173.87$ tokens per law), the Type-Token Ratio (TTR) is 0.546. Regulations contain on the whole $16,009$ sentences (on average 122.21 sentences per regulation), $328,931$ tokens (on average $2,510.92$ tokens per regulation), and the Type-Token Ratio (TTR) is 0.358.

From this corpus we chose the Regional Law 11/2021, 'Measures for prevention and intervention concerning the wolf species'. The choice of the Law was based on the following criteria: *i)* textual structure representative of Regional laws; *ii)* a good deal of linguistic variety ensuring the alternation of long and complex sentences and short and linear sentences; *iii)* reduced length, in order to allow for shorter reading times. By selecting a text of standard length, we would have had to present an extract, and this would have undermined the investigation of the overall understanding with post-reading questions; *iv)* the topic had to be related to a widely and socially relevant subject, rather than targeted to specific social groups. This document contains 3 articles that are further divided into 6 paragraphs, overall 32 sentences, 488 tokens, amounting to $2,783$ characters ($3,240$ including space chars), and its TTR is 0.591. Notably, the tokens were split in the same manner as they were presented to the participants during the reading

experiments, namely based on the AOIs (areas of interest: the areas actually targeted by readers fixations; more on this in Section 3.1). For example, a token such as 'finanziaria)', *financial*, was not split into 'finanziaria' and ')', but was kept as a single token.

The original text was altered to study the response of readers when dealing with errors. Overall 8 words were modified: namely, 4 errors were introduced at the surface level (e.g., a term such as 'urgenza', *urgency*, was changed to 'urrgenza'); and 4 words were replaced with existing words, such that the underlying semantics was affected by the replacement (e.g., in the phrase 'fauna selvatica', *wildlife*, 'selvatica' was changed into 'marina', with the whole meaning turning to *marine fauna*). The resulting expression is loosely related to the context of this regulation, referring to the woodland context, and more generally to the Aosta Valley Region, which is a mountainous region, far from the sea. The former modifications were expected to impact on lexical access, and the latter ones on the semantic integration.

Eye movements were recorded via an SR Research EyeLink 1000 Plus eye-tracker (spatial resolution of $0.01°$), with sampling at 1000 Hz. Participants were seated 60 cm away from a monitor with a display resolution of $1,600 \times 900$, so that approximately three characters subtended $1°$ of visual angle (the monitor was 40×24 deg of visual angle). Although viewing was binocular, eye movements were recorded from the right eye. The experiment was controlled with the SR Research Experiment Builder software.

To collect eye-tracking data 61 participants were recruited on voluntary bases, all native Italian speakers. For each participant we recorded age, level of education, occupation, region of birth/origin, mother tongue, and gender. Neither names nor other private information was asked, so that the authors had no access to information that would allow identifying the individual participants during or after data collection. The gender distribution among participants shows 23 male readers and 38 female readers; their mean age is 40.20 ± 14.70. On average, our participants received 16.41 ± 3.23 years of education. They were all informed about the aim of the eye-tracking experiment, as targeted to investigate readability issues possibly afflicting legal texts, and to individuate specific elements contributing to the difficulty of such text documents. Participants were warned to pay attention to the text meaning, and to try to understand its content, since after the reading phase they would have been interviewed about that text. Before starting they also were informed that the law text had been previously modified, with no further detail. In the first stage, after a brief training step required to calibrate the eye-tracking machinery, they started reading the aforementioned Regional Law 11/2021 from 6 slides employed to display the text through a laptop computer with 16-inch monitor, and their eye movements were recorded. After the recording of participant's eye movements, geometric areas of interest (AOI) were defined using the eye-tracking software. Each AOI is a polygon encompassing an attribute of interest within the image. In the second stage readers were asked whether they had detected any error throughout the reading, and to list the errors they could remember. The interviews were audio-recorded, and meanwhile their answers were collected in structured fashion.

3.2. Reader Profiling

3.2.1. Results

The total number of recorded fixations amounts to $38,022$. Fixations lasting less than 100 milliseconds were removed, as is customarily done in literature (Reisen et al., 2008; Salicchi et al., 2023). Specifically, $2,226$ fixations with a duration of less than 100 milliseconds were filtered out. The final number of fixations considered after the filtering process is $35,796$. Outlier readers were removed from the dataset based on the distribution of gaze plots: three readers were excluded due to an unusually low number of fixations, likely attributed to device errors, while one reader was dropped due to an exceptionally high number of fixations.

On average over AOIs, recorded total reading time (TRT) amounts to 276.64 ms, the mean number of fixations (NF) is 1.21, while the mean first fixation duration (FFD) lasted 159.77 ms; the standard deviations complementing these data are 234.18 (TRT), 0.96 (NF) and 118.64 (FFD). Such values are comparable to those in the Provo Corpus (Luke and Christianson, 2018), whose mean values (standard deviations) are 198.14 (173.03) for TRT, 0.95 (0.76) for NF, and 139.80 (107.11) for FFD (Luke and Christianson, 2018). The reliability of recorded data is also supported by the ratio between standard deviation and mean values: for our dataset these are 84.65%, 79.34%, 74.26% (for TRT, NF and FFD, respectively), and 87.33%, 80.00%, 76.62% for the Provo data. The slight increase in the average values of our dataset is likely influenced by the specialized nature of the text and the particularity of the legal domain, while the Provo Dataset contains 55 short English texts covering various topics.

By inspecting NF and FFD data —TRT was considered as a measure dependent on the previous ones—, readers can be categorized into four classes based on their mean NF and FFD values:

- class 1: readers with FFD above average and NF below average (10 subjects);

	TRT (std)	NF (std)	FFD (std)
average	276.64 (234.18)	1.21 (0.96)	159.77 (118.64)
class 1	281.23 (191.59)	1.10 (0.70)	183.32 (119.59)
class 2	371.67 (264.23)	1.57 (1.06)	186.54 (120.00)
class 3	200.74 (168.22)	0.94 (0.76)	133.21 (98.48)
class 4	274.83 (193.54)	1.34 (0.91)	142.00 (84.71)

Table 1: Mean values (and standard deviations) for total reading times (TRT), number of fixations (NF) and first fixation durations (FFD) featuring our corpus.

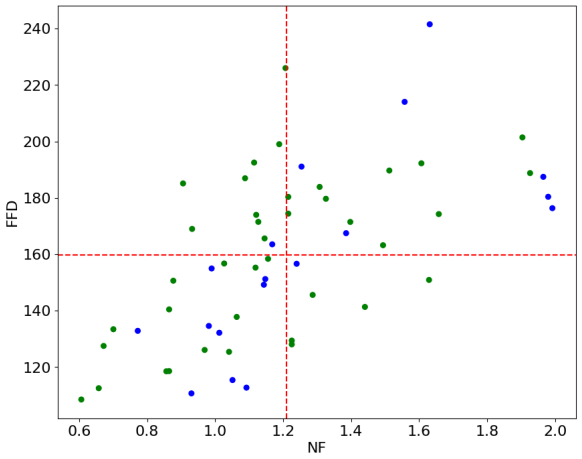


Figure 1: Plot of readers based a two-dimensional space representing NF and FFD values, with red lines indicating mean values. Class 1 is positioned on the top-left, class 2 on the top-right, class 3 on the bottom-left, and class 4 on the bottom-right. Blue points indicate readers that individuated at least 2 errors, green points those that found at most 1 error.

- class 2: with above-average FFD and above-average NF (18 subjects);
- class 3: with below-average FFD and below-average NF (23 subjects);
- class 4: with below-average FFD and above-average NF (6 subjects);

The mean values for the whole dataset and all classes are presented in Table 1; in Figure 1 we provide the plot of our readers arranged into the four classes. Classes 2 and 3 are of particular interest: class 2 identifies readers whose strategy involves higher number of fixations and longer first fixation times, while class 3 identifies readers spending less time for first fixations, and employing less fixations to read the text. After the eye-tracking session, readers were interviewed and requested to report about any errors: in this introspective effort participants were able to remember from 0 to 4 errors. Remarkably, readers that individuated at least 2 errors are mostly located either in class 2 or 3 (39% and 50%, respectively): this datum seems to suggest that the shorter the first fixation and the fewer the number of fixations, the greater the ability to identify errors. Also, 64% readers aged over 40 belong to either class 3 or 4 —thus featured by smaller FFD—, while readers under 40 are mainly (62%) found in classes 1 and 2. A correlation test was run to check whether FFD and age are (inversely) correlated, obtaining a limited Pearson correlation $\rho = -0.25$, $p < 0.058$ and a Spearman correlation $r = -0.29$, $p < 0.029$.

	CONTENT	FUNCTION
average	38.96 (21.79)	110.04 (23.33)
class 1	32.90 (21.78)	110.90 (12.85)
class 2	22.94 (10.90)	87.67 (20.07)
class 3	52.70 (19.05)	129.57 (11.66)
class 4	44.50 (19.70)	100.83 (10.75)

Table 2: Average number (std) of skips recorded in correspondence of AOIs containing content and function words.

Our categorization seems to be corroborated by the analysis of skipped AOIs: while readers from class 2 skip few (less than average) function words and few content words, almost all class 3 readers skip more function words than readers from other classes, and most of them are above average also for skipping AOIs associated to content words. By considering the number of skips, we observe that readers from class 2 consistently skip less function and content words, while those in class 3 are well above the average, as illustrated in Table 2. The regression analysis also supports our categorization: on average, we recorded 110 regressions per reader, lasting around 219 ms. The reading strategy of class 3 readers involves less (below average) and shorter (also below average) regressions, while conversely class 2 readers are featured by more and longer regressions. To complete the picture, readers from class 1 exhibit below average regressions, but lasting above average, while class 4 readers are featured by shorter but numerous regressions. These data, paired with the higher success rate in recognizing errors, seem to qualify readers from class 3 as expert readers.

The differential behavior of readers on content and function words shows that the total reading times for class 1 and 4 readers are close to the average values over all classes (which is 123.9 ms per content word syllable, and 101.86 ms per function word syllable). Readers from class 2 employ some 30% longer time than average readers to read content words and 46% on function words. Readers from class 3 save around 25% reading time on content words and 37% on function words. Detailed figures are reported in Table 3.

We investigated the response of readers when dealing with errors: for both surface and semantic errors, we observe total reading times consistently higher than for the rest of the text (please refer to Table 4). Mean total reading times are similar for both

content w	TRT (std)	NF (std)	FFD (std)
average	123.89 (93.66)	0.54 (0.38)	70.29 (43.24)
class 1	128.04 (78.41)	0.49 (0.28)	82.26 (45.56)
class 2	161.29(103.21)	0.67(0.41)	78.51 (42.21)
class 3	93.40 (70.29)	0.43 (0.31)	61.33 (37.86)
class 4	121.68 (80.77)	0.59 (0.37)	59.99 (30.42)
function w	**TRT (std)**	**NF (std)**	**FFD (std)**
average	101.86 (130.35)	0.46 (0.54)	73.24 (91.27)
class 1	102.66 (108.08)	0.42 (0.40)	84.08 (87.14)
class 2	148.94 (150.89)	0.64 (0.61)	100.51 (97.19)
class 3	64.09 (88.19)	0.31 (0.42)	51.78 (69.61)
class 4	104.10 (104.31)	0.52 (0.50)	74.60 (70.09)

Table 3: Mean values (and standard deviations), expressed in ms for TRT and FFD, characterizing fixations for content words (top) and function words (bottom); reported figures are normalized by the number of syllables.

surface	TRT (std)	NF (std)	FFD (std)
average	608.73 (468.31)	2.19 (1.64)	213.88 (149.06)
class 1	598.18 (463.38)	1.88 (1.55)	247.48 (180.93)
class 2	789.64 (519.72)	2.68 (1.70)	245.82 (173.62)
class 3	481.74 (353.50)	1.89 (1.44)	186.41 (98.64)
class 4	570.42 (356.12)	2.42 (1.55)	167.33 (64.93)
semantic	**TRT (std)**	**NF (std)**	**FFD (std)**
average	613.31 (498.64)	2.51 (2.00)	207.80 (110.43)
class 1	670.23 (438.33)	2.48 (1.59)	234.85 (133.65)
class 2	782.83 (579.59)	3.28 (2.52)	221.97 (105.42)
class 3	418.89 (319.64)	1.82 (1.17)	180.28 (94.56)
class 4	755.17 (590.17)	2.92 (2.20)	225.66 (82.11)

Table 4: Reading times relative to words containing surface (on top, tagged as 'morph.') or semantic (bottom, 'sem.') errors. Values averaged over all readers and over the four reader classes are reported.

kinds of error for the average reader: more specifically, dealing with both surface and semantic errors involved higher FFD and more fixations (NF), resulting in twice as longer total reading times (TRT) with respect to the average over the whole text (please refer to Table 1). As expected, the growth of average FFD (which is mostly concerned with lexical access) is in percentage analogous for both kinds of error; conversely, semantic errors were responsible for more consistent growth in the NF value: we recorded on average 1.21 NF per word in the overall data, which raises to 2.19 for words with surface errors, and to 2.51 for words violating the semantic/contextual integrity of the surrounding sentence. As regards as the response of readers in the four classes to the introduced errors, readers from class 3 dealing with surface errors reveal the most consistent increase over the four classes, both in the FFD values and in the average NF. It is noteworthy that half readers that correctly individuated at least 2 errors belong to this class: so readers that in general are featured by smallest FFD and NF (placed in the bottom-left corner in Figure 1) are also those with highest accuracy in identifying errors, and whose reading strategy was influenced most by errors. By recording the average number of regressions to AOIs containing errors, we observe that class 2 readers conduct an equal number of regressions compared to average readers on surface errors, and 17% more regressions on semantic errors; conversely, individuals from class 3 perform 9% more regressions than average on surface errors, and 10% less than average on semantic errors. By computing the ratio between the average number of regressions associated to AOIs containing words with errors and the average number of regressions in all other AOIs we create an index to analyze the growth of regressions corresponding to words with errors. Looking at such index, we realize that readers from class 2 conduct 1.23 (1.80) as many regressions on surface (semantic) errors, while those in class 3 conduct 2.35 (2.43) as many regressions on surface (semantic) errors.

	TRT (std)	NF (std)	FFD (std)
average	1,077.35 (816.44)	3.12 (2.34)	244.44 (226.08)
class 1	1,259.20 (1,032.94)	3.20 (2.82)	338.30 (294.52)
class 2	1,324.72 (855.07)	3.56 (2.29)	282.11 (297.22)
class 3	805.04 (550.71)	2.57 (1.66)	172.78 (88.38)
class 4	1,076.00 (821.70)	3.83 (3.18)	249.67 (85.90)

Table 5: Reading times recorded for the token 'd'urrgenza' for all readers, and the four reader classes.

In Table 5 we present the values relating to the impact of one of the four surface anomalies introduced *ad hoc*: the orthographic rendering of the 'd'urrgenza' syntagm in which the double 'r' was unduly introduced. While on average, Classes $1, 2$ and $3, 4$ exhibit comparable first fixation time duration (by construction: please refer to Table 1), in correspondence of such error, readers from classes 2 and 3 show —over the four classes— the smallest increase in their FFD, which was 1.5 times longer than for the rest of text for Class 2, and 1.3 times longer for Class 3.

3.3. Prediction of Reading Times

In this Section we describe the different models devised for the regression task aimed at predicting the three metrics TRT, NF, and FFD, and provide the obtained results.

3.3.1. Procedure

We implemented three different regression models.

– The first one is our baseline model (BL) with word-related statistics that are known to influence sentence and word processing (i.e., word frequency, word length, word position within the sentence, previous word frequency, previous word length), similar to the approach adopted by Salicchi et al. (2023).

– The second model (BL-SUR) also includes baseline features and adds surprisal scores, computed by employing a language model which is an adaptation to Italian of an English GPT-2 model (de Vries and Nissim, 2021).[1] Surprisal associated to a word w_n is defined as the negative logarithm of the probability of emitting w_n given its history $h = \{w_0, w_1, \ldots, w_{n-1}\}$: $\text{SUR}(w) = -\log P(w_n|w_0, w_1, \ldots, w_{n-1})$ (Hale, 2016).
– The third model (BL-SUR-FT) incorporates baseline features along with surprisal, computed using a fine-tuned version of the GPT-2 model obtained by exposing the language model to the laws and regulations in the AOSTA corpus, excluding 'Regional Law 11/2021'.

The regressor used is the LightGBM regressor,[2] based on the gradient boosting framework, which proved successful in the CMCL 2021 Shared Task on Eye-Tracking Prediction (Hollenstein et al., 2021; Bestgen, 2021). Gradient boosting is an ensemble learning technique based on weak learners, typically decision trees, with the objective of minimizing a given loss function. Key features of LightGBM include its leaf-wise tree growth strategy, which means that the algorithm grows the tree by expanding the leaf with the maximum delta loss instead of growing it level by level. Such strategy allows the model to find optimal split points more quickly. Moreover, a binning approach was adopted, aimed at computing optimal split points: instead of evaluating every possible split point for each feature, this strategy groups together the feature values into bins, which allows for more efficient computation. To optimize the performance of the LightGBM regressor, a comprehensive search for optimal hyperparameters was performed using a grid search technique.

The hyperparameters considered for optimization include:

– `num_leaves`: The maximum number of leaves in each tree. A range of values, such as $[4, 5, 8, 10, 20, 30]$ was explored to identify the optimal balance between model complexity and generalization.
– `learning_rate`: The step size at each iteration during training. Different learning rates $(0.1, 0.05, 0.005)$ were investigated to speed up convergence.
– `n_estimators`: The number of trees to be built. Various values $(50, 100, 200, 500)$ were tested in this setting to determine the optimal number of trees to achieve a balance between underfitting and overfitting.
– `max_depth`: The maximum depth of each tree. Values such as $[-1, 3, 5]$ were explored to control the complexity of individual trees.

The optimization process specifically targeted the mean absolute error (MAE). The evaluation of different parameter combinations was performed through a 5-fold cross-validation strategy during the grid search. This approach guarantees robustness and reliability in evaluating the model's generalization capabilities, while explicitly focusing on minimizing the MAE for optimal predictive accuracy.

[1] https://huggingface.co/GroNLP/gpt2-small-italian.

[2] https://lightgbm.readthedocs.io

	TRT (std)	NF (std)	FFD (std)
avg	20.80 (17.61)	25.20 (19.99)	12.00 (8.91)
class 1	21.38 (14.57)	23.22 (14.78)	13.94 (9.09)
class 2	28.26 (20.09)	33.22 (22.43)	14.18 (9.12)
class 3	14.69 (12.31)	18.97 (15.34)	9.76 (7.22)
class 4	20.90 (14.72)	28.31 (19.23)	10.80 (6.44)

Table 6: Figures obtained after scaling the data reported in Table 1: TRT and FFD (that are expressed as ms) were scaled based on the maximum value of TRT, while NF values were scaled based on their maximum.

3.3.2. Results

To evaluate our models we computed the Mean Absolute Error (MAE), which is a standard measure in this setting. That is, given n as the number of tokens, y_i as the actual value for i, and $\hat{y}_i$ as the predicted value for i, $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$. We also report MAE/mean scores. In fact, while MAE grasps the average difference between predicted and actual values, which is an absolute value, the latter metric scales such figures with respect to mean values, thus informing on the proportional magnitude of the error. Before computing the MAE, our features were scaled between 0 and 100, following the methodology described by Hollenstein et al. (2021).[3] The final scaled values are provided in Table 6.

We found that our best-performing model is the BL-surprisal with fine-tuning (BL-SUR-FT), whose error estimates are presented in Table 7.

By looking at the four reader classes, we obtained most favorable prediction of reading times on class 3, where we observe lowest MAE through all three measures, with MAE/mean also confirming that the predictions on readers from this class are more reliable than those on subjects from other classes. Through all classes lexical access seems to be more easily predicted than the semantic integration: consistent with the findings by Hollenstein

[3] TRT and FFD were jointly scaled as they are both measured in milliseconds (but we diverged from the approach used in the aforementioned study, due to the absence of the "go-past-time" (GPT) feature in the present setting, where we used TRT), while NF was independently scaled.

	TRT	NF	FFD
	MAE ($\frac{\text{MAE}}{\text{mean}}$)	MAE ($\frac{\text{MAE}}{\text{mean}}$)	MAE ($\frac{\text{MAE}}{\text{mean}}$)
average	4.14 (0.20)	4.52 (0.18)	1.81 (0.15)
class 1	5.90 (0.28)	5.70 (0.25)	3.25 (0.23)
class 2	6.64 (0.24)	6.77 (0.20)	2.53 (0.18)
class 3	3.43 (0.23)	4.29 (0.23)	1.84 (0.19)
class 4	6.94 (0.33)	8.54 (0.30)	2.81 (0.26)

Table 7: MAE (MAE/mean) values obtained through the BL-SUR-FT model implementing the baseline enriched with surprisal scores computed through a model fine-tuned on the AOSTA corpus.

et al. (2021), FFD confirms to be more accurately predicted than TRT and NF, that are acknowledged to grasp reader's effort throughout the semantic processing stage.

3.4. Discussion

A basic reader profiling was performed by partitioning readers based on their average number of fixations and on the duration of their first fixations. It is known that such measures can be considered as a proxy for different significant stages in linguistic processing.

As regards as the first task, aimed at reader profiling, two main reader classes were identified, that cover around 72% of those who participated in our experiments: if we wanted to resort to simplistic labels, we found fast and slow readers. We closely examined our data, and found that different views on data suggest that two main approaches to reading may be individuated: those employing less and faster fixations, slightly more accurate in individuating errors, skipping more words than average reader (possibly adapting skips to function and content words), employing less and shorter regressions even when dealing with errors in the text. In the other class we have a reading style involving more and longer fixations, less accurate in individuating errors, that are not familiar with skipping words, employing more and longer regressions, with reduced differences between content and function words, less sensitive to errors, and to the different types of error. Furthermore, we found an interesting (though weak) correlation of some variables with socio-demographic descriptors, such as that between FFD and readers age. Such elements might be helpful in refining reader profiles, and in investigating reading effort: such investigation will be addressed in future work.

As regards as the second task, aimed at predicting reading times, a thorough comparison with results available in literature can be hardly obtained, since differences may stem from factors that cannot be accounted for, such as the intrinsic properties of texts at hand. The recorded error on the number of fixations prediction is in line with the results in literature, e.g. by Hollenstein et al. (2021), but the documents in our corpus differ from those employed in the cited work: we dealt with the Italian Language (whose structure differs from English, with longer sentences and even different word lengths (Smith, 2012)), and our corpus includes Italian laws and regulations, against sentences from movie reviews borrowed from the Stanford Sentiment Treebank (Socher et al., 2013) and Wikipedia (Culotta et al., 2006). Additionally, our documents contain both surface and semantic errors that made more complex the task of predicting reading times, and individuals not necessarily expert in legal language were recruited. The greater difficulty of these texts is evidenced by the average NF featuring our data: after scaling this amounts to 25.2 (please refer to Table 6), while in the paper by Hollenstein et al. (2021) this datum is 15.1. Predicting reading times for the four reader classes turned out to be very challenging: MAE (and MAE/mean, too) is always higher than for average readers. Among classes, reading times of subjects in class 3 were those predicted with minimum error. Probabilistic language modeling, as a device able to describe the incremental mechanisms underlying language processing should be helpful to investigate the different reading strategies. Such strategies are basically concerned with planning and handling expectations on what follows, and on evaluating how these match with actual stimuli (Levy, 2008); surprisal was plugged into our models to support the prediction of reading times by also accounting for the difficulty of predicting words. Although it contributed to refining the baseline model, especially after the fine-tuning step, further work is needed to further improve the accuracy in the prediction of reading times.

4. Conclusions

In this work we have introduced a new dataset collecting Regional laws and regulations in Italian. One of these laws was modified by inserting 8 errors, and used for an eye-tracking experiment in which 61 readers were tracked. Collected data were utilized for reader profiling purposes and to predict their reading times. In the former case we individuated two main groups exhibiting rather different reading styles to cope with general text and with errors therein. In the latter experiment we applied an approach based on the gradient boosting framework; our best performing model also makes use of surprisal scores obtained through an Italian porting of a GPT-2 model fine-tuned on the set of Regional legal documents, consistent with the document used for experimentation. While the prediction of reading proved to be in line with results reported in literature, predicting the reading times of

the subjects in the two main classes individuated in the former experiment revealed a very challenging task.

Since the Aosta Valley is a bilingual (Italian and French) Region, and its body of regulations and laws is thus a naturally parallel corpus, in future work we will collect French documents and eye-tracking data on these. We will also investigate whether text difficulty and errors interact with cognitive load and how such temporal factors affect readers' performance, by examining how fixations and regressions vary through time. Finally, by considering the entire Aosta Corpus from 1960 to 2022, it would be interesting to analyze the evolution of the legal lexicon and language from a diachronic perspective, and to investigate whether older and more recent language differently impact on the reading task.

Acknowledgments

This work was carried out in the frame of the project 'The accessibility of regulatory texts as a tool for inclusion: case study and applicative tools in Valle d'Aosta', based at the University of Valle d'Aosta,[4] financed by the CRT Foundation, 2021 and 2022.

The eye-tracking data collection was made possible by the Human Science and Technologies laboratories from the University of Turin;[5] we are specially grateful to Prof. Francesca Garbarini, Prof. Olga Dal Monte and Dr. Monia Cariola for their keen and generous support.

5. Bibliographical References

Yves Bestgen. 2021. LAST at CMCL 2021 shared task: Predicting gaze data during reading with a gradient boosting decision tree approach. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 90–96, Online. Association for Computational Linguistics.

Patricia A Carpenter and Marcel Adam Just. 1983. What your eyes do while your mind is reading11this research was partially supported by grant g-79-0119 from the national institute of education and grant mh-29617 from the national institute of mental health. *Eye movements in reading*, pages 275–307.

William Coster and David Kauchak. 2011. Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669.

Aron Culotta, Andrew McCallum, and Jonathan Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 296–303.

Andrea de Varda and Marco Marelli. 2023. Scaling in cognitive modelling: a multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149, Toronto, Canada. Association for Computational Linguistics.

Wietse de Vries and Malvina Nissim. 2021. As good as new. how to successfully recycle english gpt-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read–it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lyn Frazier and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, 14(2):178–210.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nanda Kambhatla. 2022. Text simplification for legal domain: Insights and Challenges. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 296–304.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

[4] Original title: 'L'accessibilità dei testi normativi come dispositivo di inclusione: studio di caso e strumenti applicativi in Valle d'Aosta con sede presso l'Università della Valle d'Aosta'.

[5] https://www.hst.unito.it

Markus J Hofmann, Steffen Remus, Chris Biemann, Ralph Radach, and Lars Kuchinke. 2022. Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, 4:730570.

Nora Hollenstein, Emmanuele Chersoni, Cassandra L. Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021. CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–78, Online. Association for Computational Linguistics.

J Zachary Jacobson and Peter C Dodwell. 1979. Saccadic eye movements during reading. *Brain and Language*, 8(3):303–314.

Halszka Jarodzka and Saskia Brand-Gruwel. 2017. Tracking the reading eye: Towards a model of real-world reading.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel.

Gondy Leroy and James E Endicott. 2012. Combining nlp with evidence-based methods to find text metrics related to perceived and actual text difficulty. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 749–754.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

George W McConkie and Keith Rayner. 1975. The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, 17:578–586.

Samiullah Paracha, Ayaka Inuoue, and Sania Jehanzeb. 2018. Detecting online learners' reading ability via eye-tracking. In *Optimizing Student Engagement in Online Learning Environments*, pages 163–185. IGI Global.

Maria Emanuela Piemontese, M Piemontese, et al. 1996. Capire e farsi capire. teorie e tecniche della scrittura controllata.

Ralph Radach and Alan Kennedy. 2013. Eye movements in reading: Some theoretical context. *The Quarterly journal of experimental psychology*, 66(3):429–452.

Keith Rayner. 1978. Eye movements in reading and information processing. *Psychological bulletin*, 85(3):618.

Keith Rayner. 2009. The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology*, 62(8):1457–1506.

Nils Reisen, Ulrich Hoffrage, and Fred W Mast. 2008. Identifying decision strategies in a consumer choice situation. *Judgment and decision making*, 3(8):641–658.

Lavinia Salicchi, Emmanuele Chersoni, and Alessandro Lenci. 2023. A study on surprisal and semantic relatedness for eye-tracking data prediction. *Frontiers in Psychology*, 14:1112365.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Philip Levy. 2022. Large-scale evidence for logarithmic effects of word predictability on reading time.

Nathaniel J Smith and Roger Levy. 2008. Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

Reginald Smith. 2012. Distinct word length frequencies: distributions and symbol entropies. *Glottometrics*, 23:7–22.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Don Tuggener, Pius Von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. Ledgar: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1235–1241.

Serena Villata, Michal Araszkiewicz, Kevin Ashley, Trevor Bench-Capon, L Karl Branting, Jack G Conrad, and Adam Wyner. 2022. Thirty years of artificial intelligence and law: the third decade. *Artificial Intelligence and Law*, 30(4):561–591.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *CoRR*, abs/2006.01912.

Robert T Williams. 1972. A table for rapid determination of revised dale-chall readability scores. *The Reading Teacher*, 26(2):158–165.

Rolf A Zwaan, Mark C Langston, and Arthur C Graesser. 1995. The construction of situation models in narrative comprehension: An event-indexing model. *Psychological science*, 6(5):292–297.

The Simplification of the Language of Public Administration: The Case of Ombudsman Institutions

Gabriel González-Delgado, Borja Navarro-Colorado
University of Alicante
Alicante, Spain
ggd7@gcloud.ua.es, borja@dlsi.ua.es

Abstract

Language produced by Public Administrations has crucial implications in citizens' lives. However, its syntactic complexity and the use of legal jargon, among other factors, make it difficult to be understood for laypeople and certain target audiences. The NLP task of Automatic Text Simplification (ATS) can help to the necessary simplification of this technical language. For that purpose, specialized parallel datasets of complex-simple pairs need to be developed for the training of these ATS systems. In this position paper, an on-going project is presented, whose main objectives are (a) to extensively analyze the syntactical, lexical, and discursive features of the language of English-speaking ombudsmen, as samples of public administrative language, with special attention to those characteristics that pose a threat to comprehension, and (b) to develop the OmbudsCorpus, a parallel corpus of complex-simple supra-sentential fragments from ombudsmen's case reports that have been manually simplified by professionals and annotated with standardized simplification operations. This research endeavor aims to provide a deeper understanding of the simplification process and to enhance the training of ATS systems specialized in administrative texts.

Keywords: text simplification, public administrative language, parallel corpus

1. Introduction

Legal language, when it addresses laypeople, may be difficult to be understood. This lack of understanding in asymmetrical communication between experts and non-experts may lead to negative consequences in people's lives. Within the legal domain, administrative language is the one citizenship has the most relation with. It is the language produced by public bodies for the implementation of laws and legal regulations. However, obscure structures and complex terminology can pose a threat to the comprehension of its meaning, preventing people from being able to complete administrative procedures. For that reason, there exist some civil movements, such as the Plain English campaign, that advocate for the right to be addressed in a clear and understandable way by Public Administrations.

Ombudsman institutions, as whistler-blowers and guarantors of citizens' rights against maladministration, can also play an evangelization role by putting forward good-practice recommendations, denouncing abusive practices, and training public servants in charge of writing this type of texts. Nevertheless, this is a costly and time-consuming task. It is in this context that automatic text simplification (ATS) can be of use to make technical language clearer and more comprehensible.

This paper is framed within one of the author's on-going research project. Its first main objective is to analyze the linguistic features of the language of ombudsman offices as an example of public administrative language. A deeper understanding of this specialized language can contribute to advance in its necessary simplification. To that end, corpus linguistics enables the processing of large amounts of texts to obtain quantitative results. The choice of compiling a corpus out of texts produced by English-speaking ombudsmen (named the OmbudsCorpus) is not arbitrary. All the ombudsmen's case reports and annual reports are publicly available on their websites, making it an abundant source of linguistic evidence in the domain of administrative language. Besides, they are present in almost every English-speaking country, which allows for variational studies. The second objective of this project is the creation of a parallel corpus of original fragments from ombudsmen's case reports and their manually simplified version. These simplified fragments include standardized annotations on every simplification operation performed, so the parallel corpus can be used as reference data for the training and evaluation of ATS systems specialized in public administrative language.

This paper will be structured as it follows: Section 2 will discuss the main issues regarding the simplification of legal language. In Section 3, the different methodological frameworks for the creation of reference data for ATS systems will be explained. An account of the OmbudsCorpus, including its sources and the methodology followed for its development, will be provided in Section 4. Finally, some conclusions and insights on the contributions this research project aims at will be put forward in Section 5.

2. Issues about the simplification of legal language

The first and main issue about simplification is the notion of simplicity itself. When can an utterance

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 125–133

be considered complex or simple? At what point does a text stop being complex in the process of simplification? Is a complex text equally difficult to everyone? To be able to answer these questions, one must firstly approach the issue of complexity/simplicity as a continuum. We cannot say that a text is complex or simple *per se*, but that some of its components may entail some complexity and others that may contribute to an easy understanding. That is, the difficulty or ease of a text is conditioned by several internal and external factors. The familiarity of its lexicon or the syntactic complexity are some instances of internal factor than can influence comprehensibility. But even the complexity of these internal factors may be differently perceived from reader to reader. Thus, what ultimately determines the comprehensibility of a text is its target audience. Different addressee profiles (children, people with low literacy levels or lay-people, the elderly, foreigners, people with intellectual or speech disabilities, etc.) and with different backgrounds (for instance, familiarity with technical jargon and discursive genres) may present different needs to understand the content of legal documents (Garimella et al., 2022).

Legal language is characterized by the overuse of formulaic and archaic language (e.g. Latinisms), passivity and impersonal structures, abbreviations, non-finite clauses headed by gerunds, among others (see Alcaraz et al., 2013; Bhatia, 1987; Charrow et al. (2015); Danet, 1980, 1983, 1985; Gustafsson, 1983; Maley, 1987; Mellinkoff, 1963). All these features together result in dense and complicated texts that could be written in a more user-friendly manner while preserving its intended meaning. That is what plain language recommendations aim at. Throughout all the English-speaking countries that have joined this movement, it is possible to find the following ten common recommendations (see Section 8 for references to Plain Language manuals):

1. Keep your sentences short (between 15-20 words).
2. Use simple, clear words.
3. Avoid complex, technical words and choose a simpler synonym.
4. Take care when using foreign expressions, namely from French and Latin origin.
5. Take care when using initials and acronyms.
6. Avoid chains of nouns, also known as *nouns strings* (“nouns strung together to act as adjectives”).
7. Construct sentences following the order: Subject + Verb + Objects.
8. Use active voice instead of passive constructions and impersonality.
9. Address the receiver directly.
10. Consider using illustrations, tables and lists to make complex material easier to understand.

As it can be seen, these recommendations try to tackle some of the main features of legal language that make a text complex. However, some of their propositions are too vague and generic, and they fail to take into account some crucial factors that influence comprehensibility.

Simplification, whether in general or in technical contexts, is sometimes seen as the replacement of a long word for a shorter one, or the shortening of long sentences, but that does not necessarily lead to better comprehension (McNamara et al., 2014). In fact, according to Brysbaert et al. (2011), the variable ‘word length’ only correlates to 1.2% of the reading processing time. On the contrary, what really influences the degree of complexity of a word is its *frequency* (van Heuven et al., 2014). Other psycholinguistic parameters that have an impact on the lexical decision time (Brysbaert et al., 2011) are the degree of *concreteness* (Brysbaert et al., 2014) (also referred to as *sensi-motor content* (Lynott et al., 2020)), *age of acquisition* (Kuperman et al., 2012), the *semantic density* (Hoffman et al., 2013), and the *local coherence* (Hoffman et al., 2018).

In the same vein, syntactic complexity is not just a matter of length. It can be better explained by the analysis of the frequency of certain Universal Dependencies (De Marneffe et al., 2021), as explored by Deilen et al. (2023): acl (adnominal clause or clausal modifier of noun), advcl (adverbial clause modifier), ccomp (clausal component), csubj (clausal subject), xcomp (open clausal element) or parataxis (parataxis relation).

For that reason, it is necessary to implement these variables when determining the degree of complexity of a text, as it will be shown in Section 4.3, so that the simplification of legal language, either manually performed by professional or automatized by a NLP tool, can produce objectively clearer and simpler outputs that take into consideration the subjective needs of the target population.

3. Datasets for text simplification and evaluation issues

3.1 Reference data

The lack of complex-simple parallel corpora developed from legal texts is one of the main problems for the task of ATS in this domain (Garimella et al., 2022). These parallel datasets are the reference data that ATS systems are trained on. Besides, to evaluate the performance of an ATS system, outputs need to be compared to that reference data (Cardon et al., 2022, p. 1842). Thus, the approach taken to determine

what reference data an ATS system will be trained with crucially impacts the outputs produced.

Various methodologies for the creation of reference data have been reviewed by Grabar and Saggion (2022). While *expert judgment* or content extracted from *textbooks* may be established as reference data, these methods are heavily reliant on the theoretical comprehension of the producers regarding the requirements of the target audience. To address this constraint, *crowd-sourced* simplifications are used to gather extensive reference data based on the target population's judgement. However, as an online process, it is difficult to fully verify whether contributors fit in that aimed audience. An alternative method involves the application of *eye-tracking*, wherein the eye movements of readers are monitored as they engage with a reference text, enabling the quantification of attention allocation. Prolonged fixation on specific lexical units indicates higher complexity. As a drawback, this approach demands meticulous control and technical support.

Annotated reference data curated by professionals appears as another prevalent technique. Human annotation enhances the efficacy of Automatic Text Simplification (ATS) systems, particularly for rule-based systems, by elucidating the intricacies of lexical, syntactic, and even pragmatic simplification processes. However, this approach needs substantial efforts and is subject to the limitations of time and resources. Moreover, it retains a subjective element influenced by annotators' comprehension of simplification rules (Shardlow, 2014).

Newsela (Xu et al., 2015; 1,130 sentences and 5 simplified versions per sentence) and TurkCorpus (Xu et al., 2016; 2,350 sentences with 8 simplified references each) are the main reference data produced by human simplification and annotation used for ATS evaluation. They are in English and do not focus on any specific domain. The $ASSET_{ann}$ corpus (Cardon et al., 2022) has recently been proposed as an attempt to standardize the annotation process in the simplification task.

In other languages, it is possible to find the *Dsim* corpus (Klerke and Søgaard, 2012), in Danish, with roughly 50,000 sentences pairs simplified from news telegrams by trained journalists; in Brazilian Portuguese, Specia et al. (2008) crafted a manual based on the simplification and annotation of ca. 2,000 sentences extracted from news articles; in Japanese, see Goto et al. (2015), who combined automatic alignment for training data (~10,000 pairs) and manual alignment for validation (~700) and testing (~2,000); in Italian, *Terence* (Brunato et al., 2014) was developed for the simplification of texts targeting children and it contains approximately 1,000 manually aligned pairs. *SIMPITIKI* (Tonelli et al., 2016) is another corpus in Italian compiled from Wikipedia, which contains 345 sentence pairs and 575 annotations of simplification operations. Battisti et al. (2020) presented a parallel corpus in German which included annotation on text structure, typography, and images. In this same language, Spring et al. (2021) reported their work on a corpus in which simplifications were classified within A1, A2, and B1 levels of the Common European Framework of Reference for Languages. In French, *CLEAR* has progressively been developed (Cardon & Grabar, 2018, 2020; Koptient et al., 2019) as a specialized corpus in the biomedical domain with more than 4,500 parallel sentences in its latest version (2020). In Spanish, CLARA-MeD (Campillos-Llanos et al., 2022) is also a medical-domain corpus made up of about 25,000 pairs. EASIER (Alarcon et al., 2023) is a domain-independent corpus in Spanish with only lexical annotations.

For the task of ATS of legal documents, some specific corpora exist. *SimPA* (Scarton, et al., 2018) is a corpus in English extracted from the Sheffield City Council's website. Through crowdsourcing, it is made up of 1,100 original sentences with 3 lexically simplified versions and one syntactical simplified pair. *SIMPITIKI* (Tonelli et al., 2016) also contains a defined selection of 591 simplified sentences from the Public Administration domain that were manually created and annotated.

3.2 Evaluation

The optimal approach for evaluating the performance of Automated Text Simplification (ATS) systems is through human assessment, which can be conducted either by expert linguists or by a diverse sample from the target population (Alva-Manchego et al., 2020). Within this methodology, evaluators typically employ a Likert scale ranging from 1 to 5 to rate outputs across three key criteria, namely *fluency* (grammatical correctness), *adequacy* (preservation of meaning), and *simplicity* (Štajner et al., 2016).

However, this method requires substantial human and time resources. Consequently, automatic evaluation metrics have been developed, with BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016) being the most widely used. It is essential to note, nonetheless, that these metrics have not escaped criticism (Grabar & Saggion, 2022), as their primary focus lies in measuring lexical similarity rather than simplicity.

Moreover, classical readability metrics, including the Flesch Reading Ease (Flesch, 1948), Gunning Fog Index (Gunning, 1952), Automatic Readability Index (ARI) (Senter and Smith, 1967), and particularly Flesch-Kincaid Grade Level (Flesch, 1975) are employed to automatically evaluate ATS systems (Alva-Manchego et al., 2020). Nevertheless, there has been significant

criticism regarding the use of these metrics to gauge text simplicity, as they predominantly consider factors such as word and sentence length (Crossley et al., 2008), which are deemed but superfluous factors to simplicity. Traditional readability metrics do not encompass psycholinguistic factors that contribute to text complexity. McNamara et al. (2014) observe a strong correlation between psycholinguistic features such as word frequency, familiarity, age of acquisition, concreteness, and imageability, and lexical decision time, suggesting they offer a more truthful measure of lexical complexity.

4. The OmbudsCorpus

4.1 Corpus compilation

The corpus that has been compiled for the characterization of this language of the public administration is composed of Annual Reports and Case Reports from English-speaking ombudsmen (Appendix A provides a detailed list of the country these institutions are from). Texts' date of production ranges from 1992 to 2022. This information is annotated in each document so variational factors in terms of diachronic and diatopic variation may also be explored.

Ombudsman offices assign complains to an area and then publish on their websites the result in a case report. To allow for a homogeneous composition, texts were selected from three thematic areas which are shared across all Offices: Education, Health, and Housing.

Besides, the volume of workload in each ombudsman is different mainly due to demographic reasons, and so is the amount of available documentation. If all samples from all the ombudsmen were analyzed at once in a single corpus, to guarantee the representativeness of all the sources, the proportion of words per ombudsman would be limited to the one with less available information. Therefore, different subcorpora including the maximum amount of information within each country, while keeping each area proportionally represented as far as size is concerned, have been established. Thus, the linguistic analysis will be performed separately for each subcorpus, and results will be compared among them to extract common features. The table in Appendix A also includes figures on the number of tokens per country and thematic area. The overall size of the OmbudsCorpus is ca. 12,600,000 tokens (~11.7M from Annual Reports and ~950K from Case Reports).

4.2 Corpus simplification and annotation

The parallel OmbudsCorpus is composed of original fragments from case reports and its simplified counterpart. The simplification was performed by two professionals (expert linguists in the field of simplification of languages for specific purposes), who also included the annotation of each transformation operation that fragments had undergone to be rendered simpler.

To select the original fragments, each text was analyzed in terms of complexity. To determine lexical complexity, the variables of 'word frequency', 'familiarity', 'concreteness', and 'imageability' of content words were measured by means of TAALES 2.2 tool (Kyle et al., 2018). The variables under consideration in measuring syntactic complexity were 'subordinating conjunctions per clause' (mark_per_cl), 'passive auxiliary verbs per clause' (auxpass_per_cl), 'dependents per clause' (cl_av_deps), and 'clausal complements per clause' (ccomp_per_cl). TAASSC 1.3.8 (Kyle, 2016) was used for that purpose. Even though some scholars (see Alva-Manchego et al., 2020, p. 40; Crossley et al., 2008) advice against readability metrics to assess actual complexity, the Flesch-Kincaid grade level, as a well-established metric, was applied to allow for comparison with other state-of-the-art datasets which include it.

The most complex texts according to these metrics were selected, so the resulting simplification and annotation will present the most paradigmatic instances. These fragments contain more than one sentence, so simplification was performed at a supra-sentential level. Thus, the limitation of evaluation measures only being based at the level of sentence (Todirascu et al., 2013) is meant to be overcome. Almost half of these texts had been produced by the Scottish Public Service Ombudsman. Regarding thematic areas, Housing is the most complex one. In terms of the date of publication, the vast majority of them belong to the last decade.

Texts simplified include the type of simplification applied. Only simplified fragments were annotated to reflect the different simplification operations they had undergone in comparison to the original fragment. It has been represented with XML tags (see Appendix B) following the formalization proposed by Cardon et al. (2022). The main tags correspond to common general operations: insert, delete, replace. Each one has its own subtypes, as insert or delete modifiers or replace with synonyms. For other actions there are also specific tags, as <move> or <verbf/> (when there is a modification of a verbal feature, like tense or modality). Finally, some specific tags have been defined for very common specific actions such as "to personal form" (<fromImp/>). In so doing, the recommendations of Plain Language can be translated to standardized simplification operations, as presented in Appendix C.

4.3 Evaluation

The OmbudsCorpus is evaluated at different instances. Regarding the annotation of the

simplification operations, a parallel comparison of the tags used in each fragment by both annotators will be performed. It is important to bear in mind, as Stodden & Kallmeyer (2022) warn, that disagreement "does not always indicate a bad quality of the annotations, (…) it can be due to different subjective perspectives on the task". In fact, because there is no "perfect" simplification, two humans can create different simplified texts from the same textual source, both correct. For this reason, we think that the evaluation of simplification should be framed within the perspectivist approach to corpus annotation (Cabitza et al., 2023). This approach considers that the disagreement between two annotators is not an error, but rather different visions (or interpretations) of the same phenomenon, both correct.

As far as the assessment of the simplified versions is concerned, the same complexity variables as the ones applied to the original fragments (see Section 4.2) are analyzed to establish the extent of the simplification. T-tests are performed for each parameter to compare if there is a statistically significant improvement (p-value <0.05).

Metric	Original	Simp	p-value
KF_Freq_CW_Log	2.142	2.301	< 0.05
TL_Freq_CW_Log	2.785	2.983	< 0.05
Brown_Freq_CW_Log	1.299	1.494	< 0.05
Familiarity	559.43	565.87	< 0.05
Concreteness	345.59	342.66	> 0.05
Imageability	368.14	370.48	> 0.05
ccomp_per_cl	0.216	0.164	< 0.05
mark_per_cl	0.210	0.179	> 0.05
auxpass_per_cl	0.126	0.084	< 0.05
Flesch-Kincaid Grade Level	14.960	8.857	< 0.05

Table 1: Comparison of metrics between original and simplified versions.

All metrics improved in the simplified version, except for *concreteness.* The replacement of complex words has been done by more frequent and familiar words, as the metrics on logarithmic *frequencies* and the *familiarity* metric indicate. All of them with a p-value <0.05. However, the lexicon chosen for the substitution of complex words still retains high levels of abstraction. *Imageability*, which usually correlates with concreteness, shows some improvement, even though the difference is not statistically significant either.

Regarding the syntactic metrics analyzed, the average of clausal components per clause (*ccomp_per_cl*) and passive verbs per clause (*auxpass_per_cl*) was reduced significantly. In other words, simplified fragments contain fewer subordinate clauses and more sentences in the active voice. Despite the reduction in subordinate clauses, the difference in the number of subordinate conjunctions per clause (*mark_per_cl*) is not statistically significant.

This analysis allows us to identify specific pairs of fragments within the parallel corpus which may require further simplification so that an optimal simplification may be reached.

5. Conclusions

In this paper, we have presented the main objectives of this on-going research project and the research needs it targets. Previous to the task of the simplification of the public administrative language, it is necessary to know the stylistic features that are present in this register and that convey the most complexity to citizenship. The compilation and analysis of a specialized corpus from ombudsmen's text will fill in this knowledge gap.

Regarding the notion of simplicity itself, it is necessary to approach this issue from the concept of comprehensibility, instead of that of readability, as it is often done. As it has been explained, quantitative indices such word or sentence length cannot determine by themselves the complexity of a text. Psycholinguistic studies on the parameters influencing comprehension and more sophisticated metrics on syntactic structures can shed some light on this regard. These are the metrics that have been implemented in the evaluation of the OmbudsCorpus. It is important to bear in mind that the psycholinguistic parameters included in the tool TAALES (i.e., familiarity, concreteness, age of acquisition, etc.) are based on human ratings. That is where the key to determine simplicity/complexity lies.

Literature on automatic text simplification of specialized domains highlights the need for the creation of parallel datasets that serve as reference data for the training of ATS systems. Annotated reference data have proved to achieve the best state-of-the-art results. The parallel OmbudsCorpus has been developed following this methodology, incorporating the annotation of all the simplification operations applied to the original fragment. It is composed of supra-sentential pairs, in an attempt to overcome the limitations of previous datasets which remain at sentential level. It has also been enriched with syntactic and lexical parameters so the degree of complexity can objectively be compared from the original fragments to its simplified version. The intended enrichment with ratings by target audiences is an additional measure that would definitely establish a benchmark in the validation and assessment of reference data in the legal domain. A test with different ATS systems will determine the usefulness of all this annotated information in a parallel corpus.

6. Acknowledgements

This paper has been partially funded by the Spanish Government through the R&D projects "CORTEX: Conscious Text Generation" (PID2021-123956OB-I00, funded by MCIN/ AEI/10.13039/501100011033/ and by "ERDF A way of making Europe") and "CLEAR.TEXT: Enhancing the modernization public sector organizations by deploying Natural Language Processing to make their digital content CLEARER to those with cognitive disabilities" (TED2021-130707B-I00), and by the Generalitat Valenciana through the project "NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/21)".

7. Bibliographical References

Alarcon R, Moreno L, Martínez P (2023) EASIER corpus: A lexical simplification resource for people with cognitive impairments. PLoS ONE, 18(4). https://doi.org/10.1371/journal.pone.0283622

Alcaraz, E., Campos, M. A. y Miguélez, C. (2013). *El inglés jurídico norteamericano*. Barcelona, Ariel.

Alva-Manchego, F., Scarton, C., & Specia, L. (2020). Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, *46*(1), 135-187.

Battisti, A., Pfütze, D., Säuberli, A., Kostrzewa, M., & Ebling, S. (2020). A Corpus for Automatic Readability Assessment and Text Simplification of German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3302–3311), Marseille, France. European Language Resources Association.

Bhatia, V. K. (1987). Language of the law. *Language Teaching, 20*(4), 227–234. https://doi.org/10.1017/S026144480000464X

Bott, S., & Saggion, H. (2014). Text simplification resources for Spanish. *Language Resources and Evaluation*, *48*(1), 93-120.

Brunato, D., Dell'Orletta, F., Venturi, G., & Montemagni, S. (2014). Defining an annotation scheme with a view to automatic text simplification. *Defining an annotation scheme with a view to automatic text simplification* (pp. 87-92).

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental psychology*.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, *46*, 904-911.

Cabitza, F., Campagner, A., Basile, V. (2023). Toward a Perspectivist Turn in Ground Truthing for Predictive Computing, *Proceedings of the AAAI Conference on Artificial Intelligence.*

Campillos Llanos, L., Terroba Reinares, A. R., Zakhir Puig, S., Valverde, A., & Capllonch-Carrión, A. (2022). Building a comparable corpus and a benchmark for Spanish medical text simplification.

Cardon, R., & Grabar, N. (2020). French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 710-716).

Cardon, R., Bibal, A., Wilkens, R., Alfter, D., Norré, M., Müller, A., Watrin, P., & François, T. (2022). Linguistic Corpus Annotation for Automatic Text Simplification Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 1842-1866).

Charrow, V. R., Crandall, J. A., & Charrow, R. P. (2015). Characteristics and functions of legal language. In *Sublanguage* (pp. 175-190). de Gruyter.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Coster, W., & Kauchak, D. (2011, June). Simple English Wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 665-669).

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly*, *42*(3), 475-493.

Danet, B. (1980). Language in the legal process. *Law & Soc'y Rev.*, 14, 445.

Danet, B. (1983): Language in legal and bureaucratic settings. In: *Language as a social problem, special issue of Transaction-Society* (A. Grimshaw, ed.).

Danet, B. (1985). Legal discourse. *Handbook of discourse analysis*, 1, 273-291.

De Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational linguistics*, *47*(2), 255-308.

Deilen, S., Garrido, S. H., Lapshinova-Koltunski, E., & Maaß, C. Using ChatGPT as a CAT tool in Easy Language translation.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221.

Garimella, A., Sancheti, A., Aggarwal, V., Ganesh, A., Chhaya, N., & Kambhatla, N. (2022). Text Simplification for Legal Domain: Insights and Challenges. In *Proceedings of the Natural Legal Language Processing Workshop 2022* (pp. 296-304).

Goto, I., Tanaka, H., & Kumano, T. (2015). Japanese news simplification: tak design, data set construction, and analysis of simplified text.

In *Proceedings of Machine Translation Summit XV: Papers*.

Grabar, N. & Cardon, R. (2018). CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)* (pp. 3–9), Tilburg, the Netherlands. Association for Computational Linguistics.

Grabar, N., & Saggion, H. (2022). Evaluation of Automatic Text Simplification: Where are we now, where should we go from here. In *Traitement Automatique des Langues Naturelles* (pp. 453-463). ATALA.

Gunning, R. (1952). "The technique of clear writing". Information Transfer and Management. McGraw-Hill.

Gustafsson, M. (1984). The syntactic features of binomial expressions in legal English. *Text-Interdisciplinary Journal for the Study of Discourse*, *4*(1-3), 123-142.

Hoffman, P., Jefferies, E., Haffey, A., Littlejohns, T., & Lambon Ralph, M. A. (2013). Domain-specific control of semantic cognition: A dissociation within patients with semantic working memory deficits. *Aphasiology*, *27*(6), 740-764.

Hoffman, P., Cogdell-Brooke, L., & Thompson, H. E. (2020). Going off the rails: Impaired coherence in the speech of patients with semantic control deficits. *Neuropsychologia*, *146*, 107516.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Flesch-kincaid grade level. *Memphis: United States Navy*.

Klerke, S., & Søgaard, A. (2012, May). DSim, a Danish Parallel Corpus for Text Simplification. In *Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC'12* (pp. 4015-4018).

Koptient, A., Cardon, R., & Grabar, N. (2019). Simplification-induced transformations: typology and some characteristics. In *BioNLP 2019*.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior research methods*, *44*, 978-990.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior research methods*, *52*, 1271-1291.

Maley, Y. (1987). *The Language of Legislation*. *16*(1), 25–48.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Mellinkoff, D. (1963). *The Language of the Law*. Little, Brown.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

Scarton, C., Paetzold, G., & Specia, L. (2018, May). Simpa: A sentence-level simplification corpus for the public administration domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 4333-4338).

Senter, R. J., & Smith, E. A. (1967). *Automated readability index* (pp. 1-14). Technical report, DTIC document.

Shardlow, M. (2014). A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications*, *4*(1), 58-70.

Siddharthan, A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, *165*(2), 259-298.

Specia, L., Aluísio, S. M., & Pardo, T. A. (2008). Manual de simplificação sintática para o português. *NILC, Sao Carlos-SP*.

Spring, N., Rios, A., & Ebling, S. (2021). Exploring German multi-level text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)* (pp. 1339–1349).

Štajner, S., Popovic, M., Saggion, H., Specia, L., & Fishel, M. (2016). Shared task on quality assessment for text simplification. In Proceedings of the Workshop on Quality Assessment for Text Simplification - LREC 2016 (pp. 22–31.

Stodden, R., & Kallmeyer, L. (2022, May). TS-ANNO: an annotation tool to build, annotate and evaluate text simplification corpora. In *Proceedings of the 60th annual meeting of the association for computational linguistics: system demonstrations* (pp. 145-155).

Todirascu, A., François, T., Gala, N., Fairon, C., Ligozat, A. L., & Bernhard, D. (2013, October). Coherence and cohesion for the assessment of text readability. In *Proceedings of 10th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2013)* (pp. 11-19).

Tonelli, S., Aprosio, A. P., & Saltori, F. (2016). SIMPITIKI: a Simplification corpus for Italian. In *CLiC-it/EVALITA* (pp. 4333-4338).

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly journal of experimental psychology*, *67*(6), 1176-1190.

Xu, W., Callison-Burch, C., & Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, *3*, 283-297.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016). Optimizing statistical machine translation for text

simplification. *Transactions of the Association for Computational Linguistics*, *4*, 401-415.
Zhu, Z., Bernhard, D., & Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)* (pp. 1353-1361).

8. Language Resource References

Plain English manuals

Australia Government Office of Parliamentary Counsel. (1993). Plain English Manual, 1–40.
European Commission. (n.d.). *Claire's Clear Writing Tips*.
National Adult Literacy Agency. (n.d.). Plain English guidelines at a glance, 1–2.
NWT Literacy Council. (2013). *A Plain Language Handnook: Write for your Reader*. https://doi.org/10.1109/tpc.1980.6501907
Plain English Campaign. (2001). *The A to Z of Alternative Words*.
Plain English Campaign. (2018). *How to write in plain English*. Retrieved from https://www.plainenglish.co.uk/files/howto.pdf
Plain Language Action and Information Network. (2011). *Federal Plain Language Guidelines*. Retrieved from https://www.plainlanguage.gov/howto/guidelines/FederalPLGuidelines/FederalPLGuidelines.pdf

Software

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* (Doctoral Dissertation). Retrieved from https://scholarworks.gsu.edu/alesl_diss/35/
Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods 50*(3), pp. 1030-1046. https://doi.org/10.3758/s13428-017-0924-4
Stodden, R., & Kallmeyer, L. (2022, May). TS-ANNO: an annotation tool to build, annotate and evaluate text simplification corpora. In *Proceedings of the 60th annual meeting of the association for computational linguistics: system demonstrations* (pp. 145-155).

9. Appendices

Appendix A: Figures of the OmbudsCorpus by countries and types of texts (annual and case reports) and thematic area (Education, Health, and Housing)

Country*	Type	Ed	He	Ho	T
Australia	Annual	-	-	-	6.7M
	Case	155K	64K	124K	343K
Canada	Annual	-	-	-	1.9M
	Case	3.5K	3.3K	3.5K	10.5K
Ireland	Annual	-	-	-	588K
	Case	6.5K	6.9K	6.3K	19.8K
New Zealand	Annual	-	-	-	743K
	Case	17K	13.5K	15K	45.5K
UK	Annual	-	-	-	1.2M
	Case	198K	179K	139K	516K
USA	Annual	-	-	-	526K
	Case	4.2K	3.9K	4K	12K
TOTAL	Annual	-	-	-	**11.7M**
	Case	384K	272K	292K	**948K**
				TOTAL	**12.6M**

* Texts for each country have been retrieved from the following sources:

- Australia:
  - New South Wales Ombudsman.
  - Northern Territory Ombudsman.
  - Queensland Ombudsman.
  - Tasmania Ombudsman.
  - Victorian Ombudsman.
  - Western Australia Ombudsman.
- Canada:
  - British Columbia Ombudsperson.
  - Manitoba Ombudsman.
  - Ombud New Brunswick.
  - Saskatchewan Ombudsman.
- Ireland:
  - Ombudsman of Ireland.
- New Zealand:
  - Ombudsman New Zealand.
- UK:
  - Local Government and Social Care Ombudsman (England).
  - Northern Ireland Public Services Ombudsman.
  - Parliamentary and Health Service Ombudsman (UK).
  - Public Services Ombudsman for Wales.
  - Scottish Public Services Ombudsman.
- USA:
  - Hawaii State Ombudsman.
  - Iowa Office of Ombudsman.

Appendix B: List of tags used for annotation

Action	Tag
Delete proposition	<delete type="prop"/>
Delete modifier	<delete type="mod"/>
Delete for consistency	<delete type="cst"/>
Delete other	<delete type="other" subtype="…"/>
Replace with synonym (word-to-word)	<replace type="synonym" subtype="w2w">…</replace>
Replace with synonym (word-to-phrase)	<replace type="synonym" subtype="w2ph">…</replace>
Replace with synonym (phrase-to-word)	<replace type="synonym" subtype="ph2w">…</replace>
Replace with synonym (phrase-to-phrase)	<replace type="synonym" subtype="ph2ph">…</replace>
Replace with hypernym	<replace type="hypernym">…</replace>
Replace with hyponym	<replace type="hyponym">…</replace>
Replace segment with a pronoun	<replace type="pron">…</replace>
Replace singular with plural	<replace type="s2p">…</replace>
Replace plural with singular	<replace type="p2s">…</replace>
Modify verbal features	<verbf/>
Active to passive	<replace type="a2p">…</replace>
Passive to active	<replace type="p2a">…</replace>
Part-of-speech change	<POSchange/>
Split	<split/>
Merge	<merge/>
To impersonal form	<toImp/>
To personal form	<fromImp/>
Affirmation to negation	<replace type="a2n"/> …</replace>
Negation to affirmation	<replace type="n2a"/> …</replace>

Appendix C: "translation" of Plain Language recommendations to formalized simplification operations

Recommendation	Simplification operation
Eliminate unnecessary words or phrases	Delete modifier
	Delete proposition
Avoid complex words	Delete modifier
	Delete proposition
	Replace with synonym
	Replace with hypernym
	Replace with hyponym
	Specification
Take care when using foreign expressions	Replace with synonym
	Specification
Use terms consistently throughout the text	Replace with synonym
	Insert for consistency
Avoid nominalization	Replace noun with verb
Keep sentences short	Delete modifier
	Delete proposition
	Split
	Merge
	Replace with synonym (phrase-to-word)
Use active voice instead of passivity and impersonality	Passive to active
	Modify verbal features
	To personal form
Use simple sentences: Subject + Verb + Complements	Delete modifier
	Delete proposition
	Delete for consistency
	Insert for consistency
	Move
Try to use affirmative sentences	Negation to affirmation
Address the receiver directly	Proximization

Term Variation in Institutional Languages: Degrees of Specialization in Municipal Waste Management Terminology

Cirillo Nicola, Vellutino Daniela
University of Salerno
84084 Fisciano, SA, Italy
{nicirillo, dvellutino}@unisa.it

Abstract

Institutional Italian is a variety of Italian used in the official communications of institutions, especially in public administrations. Besides legal and administrative languages, it comprises the language used in websites, social media, and advertising material produced by public administrations. We show that standard measures of lexical complexity alone, like the percentage of basic vocabulary, may be misleading when used for delineating the lexical profile of institutional languages and should be complemented with the examination of terminological variants. This study compares the terminology of three types of institutional texts: administrative acts, technical-operational texts, and informative texts. In particular, we collected 82 terms with various degrees of specialization and analysed their distribution within the subcorpora of ItaIst-DdAC_GRU, a corpus composed of institutional texts drafted by Italian municipalities about municipal waste management. Results suggest that administrative acts employ high-specialization terms compliant with the law, often in the form of acronyms. Conversely, informative texts contain more low-specialization terms, privileging single-word terms to remain self-contained. Finally, the terminology of technical-operational texts is characterised by standardized and formulaic phrases.

Keywords: institutional languages, terminological variation, text simplification

1. Introduction

Information and communication activities of the institutions have reshaped the sociolinguistic space of contemporary Italian. In recent years, a new variety of Italian language emerged: institutional Italian (Vellutino et al., 2012; Vellutino, 2018). In public administrations, this linguistic variety incorporates and redefines the historically attested variety of administrative bureaucratic Italian (Sobrero, 1993; Piemontese, 1999; Raso, 2005; Cortelazzo, 2021). Institutional Italian is used within the official communications of institutions in Italy and other countries that have Italian as their official language, e.g., the Swiss Confederation (Ferrari and Pecorari, 2022).

Vellutino et al. (2012); Vellutino (2018) represent the uses of institutional Italian, revisiting the model of sociolinguistic variation of contemporary Italian, originally proposed by Berruto (1987).

In public administrations, institutional Italian has different socio-pragmatic uses as displayed in Figure 1. They range from the specialized communication of the institutional languages of law and administration (i.e., special institutional languages) to institutional languages that use the media for conveying public and institutional information and communications (i.e., media institutional languages).

Vellutino et al. (2012); Vellutino (2018) proposed a classification model of institutional texts – CPI model (Comunicazione Pubblica e Informazione istituzionale 'public communication and institutional information') – which distinguishes the texts of the special institutional languages of law and administration from the texts of the institutional languages for information and communication, considering the different pragmatic-communicative contexts linked to the purposes of the discipline of public information and communication activities of administrations, defined by the relevant Italian law (Legge 150/2000; Legge 69/2009; D.lgs 33/2013; D.lgs 96/2017) and European Union regulations, in particular, about structural funds and recovery and resilience facility (EU Regulation 2021/1060, Next Generation Europe).

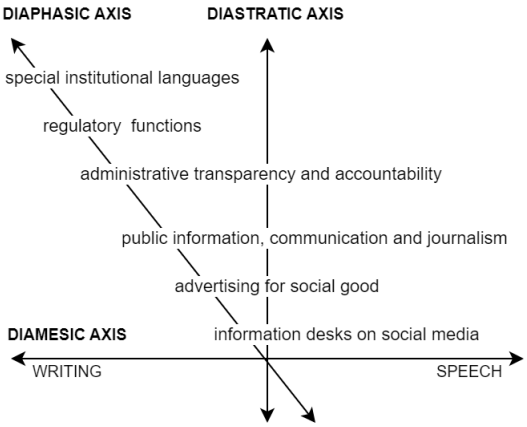


Figure 1: Socio-pragmatic uses of institutional Italian (Vellutino, 2018).

From a typological-structural point of view, the

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 134–140

linguistic variety of institutional Italian is characterized by a rich textual repertoire and an endless neological dynamism due to the ongoing entry of specialized terminologies, often multiword expressions, which can also be reduced to acronyms, giving rise to lexical variants with different degrees of specialization (Serianni, 2007; Vellutino, 2018). High-specialization terms are known only to a close circle of specialists while low-specialization terms, being known to well-educated speakers, mix with the general lexicon and form a grey area between general and special languages (Gualdo and Telve, 2011).

An example of the mechanisms that form institutional terms involves the term *credito formativo* 'training credit'. This institutional term can further specialize for a specific domain of knowledge through an adjective: *credito formativo universitario* 'university training credit'. This second terminological formation can then be reduced to the acronym *CFU*, which is part of a jargon, from a sociolinguistic point of view.

In institutional texts, multi-word terms are phrases carrying a specific meaning. They can be considered a signal not only of the use of terminology but also of the transition from the "rigidity" of the text types of legal advertising, characterized by a special lexicon, to the"flexibility" of the text types of public and institutional information and communication.

This study aims to delineate the lexical profile of the text types defined in the CPI Model. Namely, we try to answer the following research questions.

- How complex is the lexicon of the different institutional text types?
- Is the percentage of basic vocabulary alone a good indicator of lexical complexity in institutional Italian?

To answer these questions, we examine the distribution of term variants, with different degrees of specialization, in a corpus of institutional texts about municipal waste management, produced by Italian municipalities

The remainder of this paper is organized as follows. Section 2 illustrates previous studies about administrative Italian. Section 3 outlines the methodology and the language resources used in the study. Within Section 4, we present and discuss experimental results. Section 5 provides conclusions.

2. Related Work

Administrative Italian has always been known for posing readability issues that hinder citizens from accessing public information. Nevertheless, despite the numerous simplification efforts, the problem is far from being solved (Lubello, 2018).

Attempting to improve the communication between public administrations and citizens, many authors provided essential guidelines addressing the simplification of administrative texts (Fioritto, 1997; Vellutino, 2018; Cortelazzo, 2021). Their key suggestions are the following:

- Use short sentences.
- Respect the subject-verb-object order.
- Avoid subordinate clauses, preferring coordination.
- Avoid the passive voice.
- Use common tenses.
- Use a basic vocabulary.
- Avoid technical terms when possible, otherwise, explain them.

The *Vocabolario di Base* 'basic vocabulary' VdB (De Mauro and Chiari, 2016) categorized Italian words based on their accessibility to speakers, defining three distinct classes: fundamental lexicon (approximately 2,000 lexemes); high-usage lexicon (approximately 3,000 lexemes); and high-availability lexicon (approximately 2,500 lexemes). The fundamental lexicon covers 86% of the word occurrences, the high-usage lexicon accounts for 6%, and the remaining 28,000 lexemes collectively contribute 8%.

From the perspective of natural language processing, various strategies and tools automatically measure text complexity and assign readability scores by analysing lexical and syntactic features. The GULPEASE index (Lucisano and Piemontese, 1988) exploits the length of words (in character) and the length of sentences (in words) to estimate the readability of a text. It also includes an interpretation scale, based on empirical tests. In addition, the Read-It tool (Dell'Orletta et al., 2011) combines statistical text features with lexical and syntactic information obtained from the VdB and the dependency graph of a sentence.

Corpora are another essential resource for the study of institutional languages. PAWaC (Passaro and Lenci, 2019) is a web corpus composed of administrative documents from the websites of Tuscan municipalities. SIMPITIKI (Tonelli et al., 2019) and Admin-It (Miliani et al., 2022) are parallel corpora containing original sentences and related simplified versions, obtained with various simplification strategies.

3. Material and Methods

3.1. ItaIst-DdAC_GRU corpus

The corpus employed in this study is ItaIst-DdAC_GRU (Vellutino and Cirillo, 2024), a corpus of administrative, technical and informative texts drafted by Italian municipalities about municipal waste management.

The texts were collected by the students of the course "Public Communication and Institutional Languages" at the University of Salerno. They collected the documents from the website of their municipality of residence or, when not available, requested them, exercising the right of simple civic access. Then, the documents were classified according to the CPI Model (Vellutino et al., 2012; Vellutino, 2018).

ItaIst-DdAC_GRU is divided into four subcorpora: *admin*, *tech*, *acc*, and *info*. Table 1 summarises the corpus composition. Being too small, the *acc* subcorpus has not been considered in this study.

The *admin* subcorpus is composed of administrative acts, mainly resolutions, forms and ordinances. The *tech* subcorpus includes technical-operational texts like MUD[1] and PEF[2] documents. The *info* subcorpus comprises informative texts like public notices, calendars and guides for the separate collection.

3.2. List of term variants

To select the terms for the analysis, we started from a list of words and phrases automatically extracted from the ItaIst-DdAC_GRU corpus through the Sketch Engine[3] keyword extraction tool (Kilgarriff, 2009). From this list, we selected only the terms with a consistent number of variants.

Moreover, the list was enriched by finding longer phrases derived from known terms with the aid of the collocation tool of Sketch Engine. E.g., from the term *centro comunale di raccolta 'municipal recycling centre'* we found its variant *centro comunale di raccolta dei rifiuti* 'municipal waste recycling centre'.

The final list contains 82 terms, expressing 6 concepts (see Appendix A).

3.3. Experimental tests

For the purpose of delineating the lexical profile of institutional languages, we conducted three experiments on the *admin*, *tech* and *info* subcorpora of ItaIst-DdAC_GRU.

3.3.1. Experiment 1

Experiment 1 aims to assess the complexity of the lexicon of each subcorpus, without considering terminology. In this experiment, lexicon complexity is modelled as the percentage of words from the basic vocabulary (VdB). The fewer VdB words a corpus contains, the more complex its lexicon. Moreover, the inner composition of the VdB also plays a role, high-availability words are more complex than high-usage words while fundamental words are the simplest. We also compared the percentage of VdB words with another index of lexical complexity: the Type-Token Ratio (TTR), which measures the richness of vocabulary.

The metrics mentioned above are computed via the Read-It tool[4] (Dell'Orletta et al., 2011). Being the full corpus too big to be processed by Read-It, this test was conducted on a simple random sample of 100 sentences from each subcorpus. In addition, we compared the results with a baseline extracted from the web corpus itTenTen20 (Jakubíček et al., 2013) by selecting 100 random sentences containing the article *il* 'the'.

3.3.2. Experiment 2

In experiment 2, we analysed the distribution of single-word terms, multi-word terms, and acronyms throughout the subcorpora.

Therefore, for each subcorpus, we calculated the relative frequency of the collected terms, grouping them by structure (i.e., single-word, multi-word, acronym). Moreover, we determined the significance of the observed association between term structures and text types through the chi-square test of independence.

3.3.3. Experiment 3

The goal of experiment 3 is to identify the features of the terminology used in each institutional text type

To this end, we computed the frequency of the collected terms in each subcorpus and, from the contingency table, we calculated the difference between observed and expected frequency.[5] Finally,

[1] *Modello Unico di Dichiarazione Ambientale* 'unified model for environmental declaration'

[2] *Piano Economico Finanziario* 'Economic and financial plan (of the separate collection service)'

[3] https://www.sketchengine.eu/ accessed on 6 March 2024

[4] https://www.ilc.cnr.it/dylanlab/apps/texttools/?tt_user=guest accessed on 6 March 2024.

[5] A positive value means that the term occurs in a subcorpus more times than expected under the null hypothesis (i.e., the hypothesis that a term is evenly distributed across the subcorpora). Conversely, a negative value means the term occurs fewer times than expected.

Subcorpus	Text type	Documents	Sentences	Tokens
admin	Administrative acts	140	24,193	1,021,131
tech	Technical-operational texts	26	4,279	183,773
acc	Texts for accountability	13	451	22,133
info	Informative texts	126	5,045	152,806
TOT		306	33,959	1,379,843

Table 1: Itaist-DdAC_GRU corpus.

we qualitatively analyzed, for each subcorpus, the most associated terms expressing a given concept.

4. Results ad Discussion

The results of experiment 1 are shown in Table 2. They seem to indicate that the lexicon of the *info* subcorpus is the most complex. It has a lower percentage of VdB than *admin*, the lowest percentage of fundamental lexicon and the highest percentage of high-availability lexicon.

Nevertheless, the TTR does not support this hypothesis. The *info* subcorpus has the lowest TTR, even lower than the baseline. The reason may be that a specialized corpus theoretically needs fewer lexemes than a web one since it is about a single topic. However, administrative acts and technical-operational texts compensate by employing a more sophisticated vocabulary, while the vocabulary of informative texts is relatively simple.

If we interpret the results of experiment 1 considering that terminology plays a significant role in specialized corpora, the high percentage of VdB in administrative acts may be attributed to their verbose nature. Conversely, informative and technical-operational texts contain fewer regular words and more terms, because they express concepts more concisely. Moreover, the fact that the *info* subcorpus contains many high-availability words suggests that informative texts use more low-specialization terms, some of which fall within the high-availability lexicon. From this perspective, informative texts have the simplest lexicon.

Figure 2 shows the results of experiment 2. There is a significant difference in the distribution of term structures throughout the text types (df=4, χ^2=520.33, p<0.001): single-word terms are preferred in informative texts; multi-word terms appear mostly in technical-operational texts and administrative acts; and acronyms are more frequent in administrative acts and informative texts.

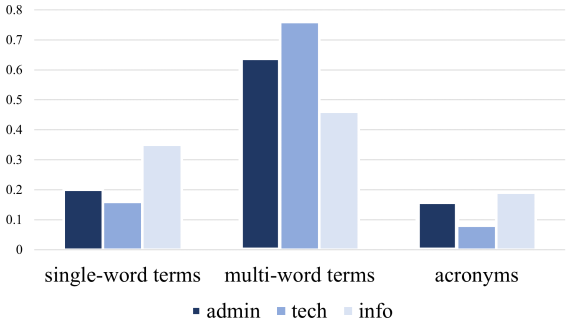


Figure 2: Distribution of term structures in the subcorpora of Itaist-DdAC_GRU.

4.1. Results of experiment 3

The concept <*centro di raccolta*> 'waste recycling centre', in administrative acts is mostly conveyed through the acronym *CRC* (+137) and the term *centro di raccolta* 'recycling centre' (+110), as defined in the Italian legislation. Widely used are also the acronym *CdR* (+74), and the variant *centro di raccolta comunale* 'municipal recycling centre' (+39). In contrast, informative texts are characterised by the more colloquial variants *isola ecologica* lit. 'ecological island' (+132) and *ecocentro* 'ecocentre' (+46). Technical-operational texts do not possess any strong relationship with any term expressing this concept.

The concept <*rifiuto organico*> 'organic waste' is mostly conveyed in administrative acts through the term *frazione organica* 'organic fraction' (+31). In informative texts, the preferred variants are the singe-word terms *umido* 'wet waste' (+138) and *organico* 'organic waste' (+63). Technical-operational texts extensively use the term *rifiuto biodegradabile* 'biodegradable waste' (+48) In particular, it appears mostly in MUD documents (*Modello Unico di Dichiarazione Ambientale* 'unified model for environmental declaration'), inside the EWC code[6] 20.01.08 *rifiuti biodegradabili di cucine e mense* 'biodegradable kitchen and canteen waste'.

While no term expressing the concept <*rifiuto indifferenziato*> 'mixed waste' is particularly associated with administrative acts, in informative texts it is mostly referred to as *indifferenziato* 'undifferentiated waste' (+179) and *secco residuo* 'dry residual waste' (+49). In technical-operational texts, the preferred term is *rifiuti urbani non differenziati* 'general mixed waste' (+23), which corresponds to the EWC code 20.03.01.

[6]European Waste Catalogue

Subcorpus	VdB	fu	hu	ha	TTR
admin	41.9%	71.1%	23.1%	5.7%	0.79
tech	36.0%	71.6%	22.6%	5.8%	0.80
info	37.6%	67.8%	23.0%	9.3%	0.70
baseline (itTenTen)	60.3%	73.9%	22.4%	3.8%	0.74

Table 2: Type-token ratio (TTR) and percentage of words from the basic vocabulary (Vdb), further divided by repertoire of use. I.e., fundamental (fu); high-usage (hi); and high-availability (ha).

The concept <*rifiuti urbani*> 'municipal waste' is expressed in administrative acts mainly through the phrase *rifiuti solidi urbani* 'municipal solid waste' (+58) and the acronyms *RU* (+34) and *RSU* (+24). No term expressing this concept has a positive relationship with informative texts while technical-operational texts are strongly associated with the term *rifiuti urbani* 'municipal waste' (+667).

The concept <*raccolta porta a porta*> 'door-to-door waste collection' is mostly conveyed in administrative acts through the terms *raccolta domiciliare* lit. 'domestic collection' and *servizio di raccolta domiciliare* lit. 'domestic collection service'. Conversely, in informative texts, the preferred variants are the terms *porta a porta* 'door-to-door' (+177) and its acronym *PAP* (+142).

5. Conclusion

Socio-pragmatic uses of institutional Italian comprise special and media institutional languages. The former is used to legislate and administrate and the latter to communicate with the general public through various media: newspapers, websites and advertising material. For these uses, institutional Italian has different lexica and employs different terms, with various degrees of specialization, to refer to similar concepts.

In order to define the lexical profile of institutional Italian, we collected 82 different terms expressing 6 concepts and examined their distribution across the three subcorpora of the ItaIst-DdAC_GRU corpus, namely administrative acts, informative texts and technical-operational texts.

Results show that administrative acts employ high-specialization terms compliant with the law, often in the form of acronyms. Conversely, informative texts contain more low-specialization terms and make extensive use of single-word terms and acronyms to remain self-contained. The terminology of technical-operational texts is largely composed of standardized and formulaic phrases.

Furthermore, results also suggest that standard metrics of lexicon complexity that do not consider terminology may lead to erroneous conclusions when applied to specialized corpora and should therefore be carefully interpreted and preferably complemented with the analysis of terminological variation.

In the future, we aim to develop an index of terminological specialization and a method to accurately measure the lexical and terminological complexity of specialized corpora.

6. Acknowledgements

This work was conducted within the PRIN 2020 Project *VerbACxSS: su verbi analitici, complessità, verbi sintetici, e semplificazione. Per l'accessibilità*, funded by Ministero dell'Università e della Ricerca (PRIN 2020_2020BJKB9M).

7. Bibliographical References

Gaetano Berruto. 1987. *Sociolinguistica dell'italiano contemporaneo*. Carocci. 2ª ed. 2012.

Maria Teresa Cabré. 1999. *Terminology: Theory, methods, and applications*, volume 1. John Benjamins Publishing.

Michele Cortelazzo. 2021. *Il linguaggio amministrativo. Principi e pratiche di modernizzazione*. Carocci.

Tullio De Mauro and I Chiari. 2016. Il nuovo vocabolario di base della lingua italiana. *Internazionale*.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read–it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83.

Annibale Elia, Alessandro Maisto, Lorenza Melillo, and Serena Pelosi. 2021. The lexical complexity and basic vocabulary of the italian language. In *Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities: 14th International Conference, NooJ 2020, Zagreb, Croatia, June 5–7, 2020, Revised Selected Papers 14*, pages 14–23. Springer.

Angela Ferrari and Filippo Pecorari. 2022. *Le buone pratiche redazionali nei testi istituzionali svizzeri in lingua italiana*. Franco Cesati.

Alfredo Fioritto. 1997. *Manuale di stile. Strumenti per semplificare il linguaggio delle amministrazioni pubbliche*. Il Mulino.

Riccardo Gualdo and Stefano Telve. 2011. *Linguaggi specialistici dell'italiano*. Carocci.

M. Jakubíček, A. Kilgarriff, V. Kovář, P. Rychlý, and V. Suchomel. 2013. The tenten corpus family. In *7th International Corpus Linguistics Conference CL*, pages 125–127.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proc. Corpus Linguistics*, volume 6.

Sergio Lubello. 2018. L'antilingua gode di buona salute: nuove forme, vecchi vizi. In *Comunicare cittadinanza nell'era digitale Saggi sul linguaggio burocratico 2.0*, pages 31–43. FrancoAngeli.

Pietro Lucisano and Maria Emanuela Piemontese. 1988. Gulpease: una formula per la predizione della leggibilita di testi in lingua italiana. *Scuola e città*, pages 110–124.

Martina Miliani, Serena Auriemma, Fernando Alva-Manchego, and Alessandro Lenci. 2022. Neural readability pairwise ranking for sentences in italian administrative language. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 849–866.

Lucia Passaro and Alessandro Lenci. 2016. Extracting terms with extra. In *Computerised and corpus-based approaches to phraseology: Monolingual and multilingual perspectives*, pages 188–196. Tradulex.

Maria Emanuela Piemontese. 1996. *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Tecnodid.

Maria Emanuela Piemontese. 1999. La comunicazione pubblica e istituzionale. il punto di vista linguistico. In Stefano Gensini, editor, *Manuale della comunicazione*. Carocci.

Tommaso Raso. 2005. *La scrittura burocratica. La lingua e l'organizzazione del testo*. Carocci.

Luca Serianni. 2007. *Italiani scritti*. Il Mulino.

Alberto A. Sobrero. 1993. *Introduzione all'italiano contemporaneo. La variazione e gli usi*. Laterza.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. Simpitiki: a simplification corpus for italian. In *CLiC-it/EVALITA*, pages 4333–4338.

Daniela Vellutino. 2018. *L'italiano istituzionale per la comunicazione pubblica*. Il Mulino.

Daniela Vellutino and Nicola Cirillo. 2024. Corpus «itaist»: Note per lo sviluppo di una risorsa linguistica per lo studio dell'italiano istituzionale per il diritto di accesso civico. [Manuscript submitted for publication].

Daniela Vellutino, Federica Marano, and Annibale Elia. 2012. L'italiano istituzionale e le sue varietà d'uso pubblico. aspetti lessicali nei tipi di testo d'informazione e comunicazione delle pubbliche amministrazioni. In *Atti XI Congresso SILFI*.

8. Language Resource References

Passaro, Lucia C. and Lenci, Alessandro. 2019. *PaWaC - Public Administration Web as Corpus Corpus*. distributed via European Commission, Directorate-General for Communications Networks, Content and Technology. PID http://data.europa.eu/88u/dataset/elrc_1282.

Tonelli, Sara and Palmero Aprosio, Alessio and Saltori, Francesca. 2019. *SIMPITIKI corpus for simplification in Italian*. distributed via Zenodo. PID https://doi.org/10.5281/zenodo.2535632.

A. Terms selected for the study

Centro di raccolta CCR; CdR; centro comunale di raccolta; centro comunale di raccolta rifiuti; centro di raccolta; centro di raccolta comunale; centro di raccolta dei rifiuti urbani; centro di raccolta intercomunale; centro di raccolta rifiuti; centro di raccolta rifiuti solidi urbani; centro di raccolta rifiuti urbani; centro di raccolta temporaneo; CRC; eco isola; ecocentro; ecocentro comunale; ecopiazzola; isola ecologica; isola ecologica comunale; isola ecologica itinerante.

Rifiuto organico FORSU; frazione biodegradabile; frazione organica; frazione organica di rifiuti; frazione organica umida; frazione umida; organico; rifiuto biodegradabile; rifiuto organico; rifiuto umido; umido.

Rifiuto indifferenziato frazione indifferenziata; frazione indifferenziato residuale; frazione non riciclabile; frazione residua; frazione rifiuti indifferenziati; frazione secca indifferenziata; frazione secca non differenziata; frazione secca non riciclabile; frazione secca residua; frazione secca residua indifferenziata; indifferenziato; materiale non riciclabile; residuo indifferenziato; residuo secco; rifiuti domestici indifferenziati; rifiuti urbani indifferenziati; rifiuti

urbani non differenziati; rifiuto indifferenziato; rifiuto indifferenziato residuale; rifiuto residuo; rifiuto secco indifferenziato; rifiuto secco non riciclabile; rifiuto secco residuo; RSU indifferenziati; secco indifferenziato; secco non riciclabile; secco residuo; secco residuo indifferenziato.

Rifiuti urbani RSU; RU; rifiuti solidi urbani; rifiuti urbani.

Raccolta differenziata differenziata; raccolta differenziata; raccolta differenziata dei rifiuti; raccolta differenziata dei RSU.

Raccolta porta a porta PAP; porta a porta; raccolta differenziata domiciliare; raccolta differenziata porta a porta; raccolta domiciliare; raccolta porta a porta; raccolta rifiuti porta a porta; servizio di raccolta domiciliare; servizio porta a porta; sistema di raccolta differenziata porta a porta; sistema di raccolta domiciliare; sistema porta a porta.

Concept	**Term structure**		
	sw	mw	acronym
Centro di raccolta waste *recycling centre*	2	15	3
Rifiuto organico *organic waste*	2	8	1
Rifiuto indifferenziato *mixed waste*	1	28	0
Rifiuti urbani *municipal waste*	0	2	2
Raccolta differenziata *separate collection*	1	4	1
Raccolta porta a porta *door-to-door waste collection*	0	11	1
TOT	6	68	8

Table 3: Terms selected for the study, divided into single-word terms (sw), multi-word terms (mw) and acronyms.

LARGEMED: a Resource for Identifying and Generating Paraphrases for French Medical Terms

Ioana Buhnila, Amalia Todirascu
ATILF UMR 7118 (CNRS-University of Lorraine), LiLPa UR 1339 (University of Strasbourg)
Nancy, Strasbourg (France)
ioana.buhnila@univ-lorraine.fr, todiras@unistra.fr

Abstract

This article presents a method extending an existing French corpus of paraphrases of medical terms RefoMed (Buhnila, 2023) with new data from Web archives created during the Covid-19 pandemic. Our method semi-automatically detects new terms and paraphrase markers introducing paraphrases from these Web archives, followed by a manual annotation step to identify paraphrases and their lexical and semantic properties. The extended large corpus LARGEMED could be used for automatic medical text simplification for patients and their families. To automatise data collection, we propose two experiments. The first experiment uses the new LARGEMED dataset to train a binary classifier aiming to detect new sentences containing possible paraphrases. The second experiment aims to use correct paraphrases to train a model for paraphrase generation, by adapting T5 Language Model to the paraphrase generation task using an adversarial algorithm.

Keywords: medical terms, paraphrases, automatic paraphrase generation

1. Introduction

Text adaptation aims to produce a simplified version (for example at lexical level) of the original document for a specific target audience with reading difficulties or insufficient knowledge. In the medical domain, text adaptation for patients or patients' families helps them to better understand their illness and to better fight against it. Medical knowledge is shared by health specialists and experts, but lay people have difficulties to understand the content of medical texts, due to the high density of scientific terms with opaque meaning. Terms are lexical units identifying a concept from a specialised domain (Condamines, 1997). Thus, text adaptation systems propose synonyms, alternative explanations, definitions or paraphrases of difficult medical terms for the target audience (patients or people with shallow medical knowledge).

However, automatic text adaptation requires large corpora or paraphrase datasets. Few French NLP resources are available for the medical domain, such as the parallel medical corpus **CLEAR**, containing aligned scientific and simplified medical abstracts (Grabar and Cardon, 2018), or the **RefoMed** dataset (Buhnila, 2023) containing pairs of medical terms and their paraphrases.

Thus, we propose a method for building a large corpus, containing medical terms and their various paraphrases, useful for automatic text simplification. **Paraphrases** are considered to be sequences of words aiming to preserve the sense of the paraphrased term (Fuchs, 1982; Vassiliadou, 2020), with various surface forms: simple words, phrases, sentences. Building such datasets is a difficult task, due to the various lexical and syntactic forms of the paraphrases. In this article, we adopt the definition proposed by Eshkol-Taravella and Grabar (2017): we consider that definitions, exemplifications and explanations represent various forms of **subsentential paraphrases** (paraphrases identified in the same sentence as the term). We aim to build a large resource with various forms of subsentential paraphrases for medical terms that might enhance the accessibility of medical knowledge to a non-specialist audience.

In this context, we propose two main contributions: **(1)** a large corpus **LARGEMED**, containing French terms and their subsentential paraphrases semi-automatically extracted from medical texts. The resource is annotated with lexical relations and semantico-pragmatic functions of the paraphrases; **(2)** some experiments aiming to extend LARGEMED by automatic paraphrase classification and generation;

Firstly, we present the concept of paraphrase in linguistics and NLP followed by our own definition. We continue with the state-of-the-art methods of classification and paraphrase generation, as well as the few French NLP resources available for medical domain used for automatic paraphrase classification, generation or text adaptation. Then, we describe the data found in the RefoMed corpus and the annotation guidelines applied to our own corpus containing Covid-19 terms and their paraphrases. In the next section we detail our method to collect data from Web archives used to complete RefoMed. Subsequently, we detail the classification and the generation experiments, based on LARGEMED, in order to eventually collect more data. We discuss our results and

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 141–151

conclude with future perspectives for our work.

2. Background

No unique definition of the notion of paraphrase is available in linguistics, computational linguistics and NLP. Fuchs (1982; 2020) considers that the paraphrase should be semantically equivalent to the paraphrased word or term. Eshkol-Taravella and Grabar (2017) adopt a broader point of view of the concept of paraphrase, assuming that it can have various lexical or syntactic forms while preserving similar or same meaning. Between the terms and their paraphrases, several lexical relations could be established: *synonymy*, *hypernymy*, *hyponymy* (2). Eshkol-Taravella and Grabar (2017) assume that the intention behind the usage of paraphrases in discours can exhibit several semantico-pragmatic functions, such as *definition* (1), *explanation*, *exemplification*, or *rephrasing*. We illustrate this linguistic variety with some examples extracted from the CLEAR corpus (Grabar and Cardon, 2018) (where the medical term is in **bold** font and the paraphrase in *italic*):

1. **Les troubles de l'équilibre** étaient définis si *le patient n'était pas en mesure de rester au moins cinq secondes en appui unipodal*.

   (**The equilibrium troubles** are defined as *the patient is not able to stay at least 5 seconds in single-leg support*.)

2. Les autres **traitements immunosuppresseurs** (*mycophénolate mofétil, cyclophosphamide, méthotrexate, azathioprine*) [...] sont discutés (The other **immunosuppressors treatments** (*mycophénolate mofétil, cyclophosphamide, méthotrexate, azathioprine*) [...] are discussed)

In NLP, two segments of text are considered paraphrases if similarity measures are high (such as cosine similarity or BLEU (Reiter, 2018), but these scores use only morphological or syntactic cues. Adversative paraphrases (with different lexical or syntactic forms, but with similar meaning) are more difficult to detect than paraphrases with few syntactic variations (Nighojkar and Licato, 2021). Paraphrase markers such as multi-word expressions (*c'est-à-dire* - 'that is to say', *signifie* - 'means', *est un/une* - 'is a') or punctuation signs, are often used to introduce paraphrases and they could help paraphrase automatic identification (Grabar and Hamon, 2015).

In our paper, we define **medical paraphrases** as different lexical representations that designate, simplify, or explain medical terms, while keeping a similar meaning (Fuchs, 2020; Vassiliadou, 2020; Buhnila, 2023). Our definition of the linguistic concept of paraphrase includes different types of word sequences, such as *definitions, rephrasing, exemplifications, explanations or abbreviations* (Eshkol-Taravella and Grabar, 2017; Buhnila, 2022b). We build a dataset of simple and multi-word terms linked to their **subsentential paraphrases**.The paraphrases could be simpler words or expressions, noun or verbal structures or simple enumerations of examples, often introduced by an explicit paraphrase marker. To illustrate our definition, we present some examples identified in our corpus. The term is displayed in **bold**, the paraphrase is in *italic* and the paraphrase marker that introduces the paraphrase is tagged with $< m >< /m >$:

- **distanciation physique** d'autres que cela <m>signifie</m> *couper les contacts sociaux* (**physical distancing** from others, which <m>means</m> *cutting off social contacts*);
- **l'anosmie**, <m>c'est-à-dire</m> *une perte totale de l'odorat* (**anosmia**, <m>meaning</m> *a total loss of sense of smell*).

We consider that medical paraphrases are useful for text simplification or adaptation. Simpler synonyms or hyperonyms might simplify the comprehension of the target audience, as well as definitions or exemplifications. Complex resources are required for such systems, but also various methods for producing them. Thus, we present related work on medical text simplification, paraphrase datasets or corpora and paraphrase identification or generation.

3. Related Work

Text simplification in the medical domain aims to explain or to replace scientific terms with simple words or paraphrases in order to enhance information accessibility to lay people (Grabar and Hamon, 2015, 2016; Cardon and Grabar, 2018; Koptient et al., 2019; Cardon and Grabar, 2021; Buhnila, 2022a). This simplified medical content might also be used to facilitate communication with patients (Pecout et al. 2019; Koptient and Grabar 2020). To simplify a medical text, two steps are necessary. Firstly, we identify medical terms, and secondly, we find the appropriate paraphrases for these terms. Both tasks are difficult. Automatic term identification based on terminological databases or ontologies with large coverage (such as **SNOMED** (Cote, 1998)) will not be able to identify newly created terms. For example, the Covid-19 pandemic created a large number of new terms, but they are not all included in the existing knowledge bases.[1] Tools for term

[1] After the end of this study, we came across a bilingual (French-English) ontology with Covid-19 terms accessi-

identification extract candidates from open-source texts and are more reliable, but the output has to be manually filtered (Rigouts Terryn et al., 2020).

For the task of text simplification, paraphrase resources should relate terms to their paraphrases. Most of the large paraphrase datasets contain sentential paraphrases from general language available in English: **MSRP** (*The Microsoft Research Paraphrase Corpus*) (Dolan et al., 2004), **PPDB** (*ParaPhrase DataBase*) (Ganitkevitch and Callison-Burch, 2014), **PAWS** (Zhang et al., 2019), (*Paraphrase Adversaries from Word Scrambling*). French language is represented in few resources (mostly multilingual), and only for the general domain, such as **PPDB**, **TaPaCo** (Scherrer, 2020) or **ParaCotta** (Aji et al., 2022). Subsentential paraphrases might be more appropriate to provide explanations or definitions for the terms, but few datasets containing subsentential paraphrases are available. One such resource is **PARADE** (He et al., 2020), a computer science dataset of definition-style paraphrases for English technical concepts extracted from online user-generated flashcards. These paraphrase datasets were built from general or computer science corpora, but they do not cover data from the field of medicine.

Due to the lack or medical paraphrase datasets or parallel corpora (original and paraphrases), NLP systems were developed for paraphrases identification or generation. Various statistical or deep learning methods were tested on paraphrase identification. Methods based on similarity measures, such as **Textual Semantic Similarity** (**STS**) (Agirre et al., 2016) or **Paraphrase Identification** (**PI**) (Brockett and Dolan 2005; Xu et al. 2015) identify paraphrases by counting words that have a certain degree of semantic equivalence and a similar lexical surface form. Various classifiers identify specific types of paraphrases based on syntactic criteria (Zhou et al., 2022). Sentence-level paraphrase identification methods are very effective for English datasets Peng et al. (2023) using BERT language model (Devlin et al., 2018). Again, few methods are designed to detect subsentential paraphrases. Linguistic patterns and n-grams are used to extract subsentential paraphrases from large medical comparable corpora (Cartoni and Deléger, 2011). Some methods use comparable corpora and **Abstract Meaning Representation (AMR)** (Bouamor et al., 2013) to detect subsentential paraphrases. These methods have some drawbacks when it comes to identify paraphrases with various surface forms for specific medical terms. Subsentential paraphrases, such as short definitions, exemplifications, explanations, or abbreviations might take different surface forms, but helps user's comprehension. Semantic similarity techniques fail to identify these types of paraphrases.

To avoid these drawbacks and to be able to create new paraphrase datasets, alternative methods, such as **Paraphrase Generation Method** (**PG**) (Gupta et al. 2018; Bowman et al. 2015) are employed to generate paraphrases with various forms, but similar meaning. Among these, the **APT** (*Adversarial Paraphrasing Task*) neural architecture (Nighojkar and Licato, 2021) uses a method for generating paraphrases with equivalent meanings and lexical and syntactic differences. This model identifies the general meaning of a sentence, not just the meaning of individual words. It is possible to infer the meaning from the term to the paraphrase and vice-versa.

In this paper, we present a dataset of subsentential paraphrases, as this type of paraphrase is not much exploited in the NLP community for the medical domain. In the next section, we present our project and our method used to create a large subsentential medical paraphrases dataset in French, LARGEMED. Moreover, we use this corpus as a resource for experimenting several methods for paraphrase classification and generation.

4. Method

The **ADAPTMED project** aims to create a large collection of terms and their paraphrases, by extending an existing subsentential paraphrase corpus **RefoMed** (Buhnila, 2023) with new terms from the Covid-19 pandemic and related topics such as social measures and vaccine campaign. Indeed, the Covid-19 pandemic generated a lot of new terms and paraphrases, frequently found in the Web archives created by the **National French Library (NFL)**[2].

We represent graphically our method in **Figure 1**. Firstly, to build a large paraphrase corpus, we identified the Web archives about the Covid-19 pandemic (a collection of Web pages dated from March 2020 to July 2020) maintained by the **NFL**. The archive contains a large number of new terms related to Covid-19, but also various paraphrases of this new terms, as people needed to better understand this new disease. The Web pages are available in several versions, due to frequent updates of the information during the pandemic. The pages are indexed with Apache Solr and the archives were manually explored with a specific query language. This query language is very complex and the requests had to be manually checked to identify the term and its paraphrase on the last version of the Web site. This step

ble here: https://www.hetop.eu/hetop/rep/fr/COVID/

[2]Bibliothèque Nationale de France (BNF)

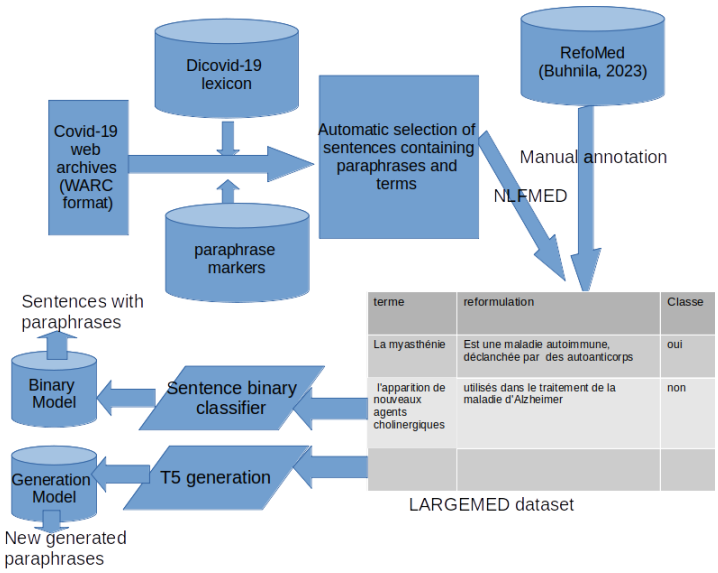


Figure 1: The method used to collect data and to develop the paraphrase classification and generation system.

is time-consuming, so the requests are send automatically to the Solr search engine to obtain a complete list of Web pages containing potentially terms and paraphrases, possibly accompanied by paraphrase markers.

From the Web pages which potentially contained paraphrases, we pre-selected the sentences that had both Covid-19 terms from the **Dicovid-19** dictionary [3] and paraphrase markers (expressions: *c'est à dire* - 'that is to say', *autrement dit* - 'in other terms'). These sentences were manually annotated and linguistically analysed. We present the annotation process in detail in section 5.3.2. The pre-selection of sentences containing paraphrases might reduce the number of sentences that should be manually annotated. Firstly, we set up a sentence binary classification using **CamemBERT** (Martin et al., 2019) to detect if the sentence contains a paraphrase in order to generalize the process for future experiments. Secondly, the paraphrases from our corpus are used to adapt a generation model **t5-base** (Raffel et al., 2020) for text generation to medical domain.

The next section presents the medical resources and the method of collecting subsentential paraphrases.

5. Resources

To compile and extend a large dataset containing terms and their subsentential paraphrases, we use a term list to select sentences containing terms, and therefore complete an existing paraphrase corpus RefoMed (Buhnila, 2023). We present this paraphrase corpus in the next section.

5.1. Existing Medical Paraphrase Corpus

RefoMed (Buhnila, 2023)[4] is a corpus of medical subsentential paraphrases in French and Romanian. The RefoMed corpus contains 11,653 pairs of medical terms and their medical paraphrases, 8,626 pairs in French and 3,027 pairs in Romanian. For this study we use only the French sub-corpus. The source corpora for French are **ClassYN** (Todirascu et al., 2012) and **CLEAR Cochrane** (Grabar and Cardon, 2018), both comparable corpora of scientific and simplified medical texts and abstracts. The RefoMed dataset was built by automatically extracting sentences that contain medical terms with the **SIFR-BioPortal** annotator (Tchechmedjiev et al., 2018), using the **SNOMED-3.5VF** medical ontology (Cote, 1998) (150,906 medical concepts). The sentences included in this corpus were selected if they contained both terms and paraphrase markers, such as *c'est-à-dire* ("so called"), *autrement dit*("in other words"), *également appelé*("also called"), *est une maladie*("is a disease"), *signifie* ("signifies / means") and punctuation signs, such as colons and brackets (Grabar and Hamon 2015; Antoine and Grabar 2016; Buhnila 2022b). The sentences were manually annotated and validated by 2 non-expert human coders. The coders follow the specific guidelines to annotate the status (if the sentence contains a paraphrase or not), the term, the paraphrase markers, the para-

[3] https://dicovid19.com

[4] https://github.com/ibuhnila/refomed

Lang	Term	Mkr	Paraphase
fr	dyspnée	c'est-à-dire	gêne respiratoire
en-tr	dyspnea	i.e	difficulty breathing

Table 1: An example of annotated paraphrase in French (**fr**) and its translation in English (**en-tr**). **Term** is the medical term automatically identified with Snomed, **Mkr** is the paraphrase marker that helps identifying the paraphrase, and **Paraphrase** represents a subsentential paraphrase.

phrase, the lexical relations (the paraphrase is a hyponym, hypernym or synonym to the term) and their semantico-pragmatic functions (definition, explanation, exemplification). The inter-coder agreement, computed as Krippendorff's α is moderate (0.61) for the paraphrase class. The validated term-marker-paraphrase pairs were included into **RefoMed**.
We build the new Covid-19 paraphrase corpus following the same method of selection of sentences containing a term from a lexical resource (Dicovid-19 in our case), and we follow the same annotation guidelines from (Buhnila, 2023), as explained in the section 5.3.2.

5.2. The Dicovid-19 dictionary

Several large coverage medical term databases are available, such as **UMLS** for English (Bodenreider, 2004) or **SNOMED International** for French (Cote, 1998), but they do not contain newly created terms related to the Covid-19 pandemic. Thus, we used the French **Dicovid-19** dictionary which contains 296 terms used or formed during the Covid-19 pandemic, such as *super spreader, vaccinodrome - 'mega vaccine center', N92 mask, distanciation sociale - 'social distancing', antivax - 'anti-vaccine collaborator'*. This dictionary is a key resource to select sentences containing Covid-19 terms and has been manually defined during the Covid-19 pandemic by a French lexicographer.

5.3. A new corpus NLF: Covid-19 Terms and Paraphrases

5.3.1. Data collection

The NLF Web archives contain 15TB of data and was build by automatic indexation of French Web pages such as newspapers, scientific blogs, popularisation blogs containing at least one mention of Covid-19 pandemic. Due to its size and the risk of incorrectly indexing web pages, functional words (such as punctuation, prepositions, conjunctions, simple verbs like *to be*, *to have*) were not included in the Solr search engine. Thus, we adapt our queries considering these constraints.
To collect the data we use expert queries including a term, a marker and a span window between them. Indeed, the query *text: "distanciation physique signifie" 7 AND (collections:"épidémie Covid-19")*, helps us to find the term **distanciation physique 'social distanciation'** along with the paraphrase marker **signifie - 'means'** (the number *7* indicates the word span). This query detected a paraphrase for the Covid-19 term *distanciation physique d'autres que cela signifie couper les contacts sociaux* (**physical distancing**: "physical distancing from others, which means cutting off social contacts"). The queries were manually written using Solr's interface.
Then, we manually selected Web pages and check if the page contained at least one Covid-19 term and its paraphrase in the same sentence. Afterwards, the url addresses were used to extract the text contained in the pages, by using the instance of the Apache Solr search engine.

The next step was the semi-automatically extraction of sentences with term-paraphrase pairs, introduced by paraphrase markers identified in the literature (Eshkol-Taravella and Grabar 2017; Buhnila 2022b). We asked the coders to identify the term, the paraphrase marker and the paraphrase as shown in **Table 1**.

We extracted 8,565 sentences containing at least a term and a paraphrase marker (out of 25,644 selected sentences). Through automatic annotation, we identified 893 pairs of terms and paraphrases in the same sentence (data is showed in **Table 2**). Only 10.42 % of sentences contained real paraphrases, manually validated. Additionally, we selected some definitions and paraphrases from Wikipedia Web pages of Covid-19 terms (140 sentences contain terms and their definitions or explanation). Then, we manually annotated them with lexical relations and semantic-pragmatic functions. We present the annotation process in the next section.

Sent	Term	T-M Sent	C-Para	M-Para	Total Para
25,644	8,565	1,725	637	176	**893**

Table 2: Quantitative data extracted from the url of the Covid-19 NLF archive collection. **T-M Sent** represents the number of sentences containing at least one term (**T**) and a marker (**M**); **C-Para** states the number of correct paraphrases (one per sentence); **M-Para** indicates the number of multiples paraphrases per sentence; **Total** represent the number of correct paraphrases.

5.3.2. Annotation Process

To build the corpus, we follow the annotation method used for the RefoMed corpus (Buhnila, 2023). The RefoMed corpus was automatically annotated in terms and paraphrase markers and the paraphrases of medical terms were manually analysed from a lexical and semantico-pragmatic perspective following the guidelines provided by Eshkol-Taravella and Grabar (2017).

Medical terms and paraphrase markers annotation. Sentences containing both medical terms from the DiCovid-19 dictionary and the paraphrase markers are identified automatically using a rule-based method, applying regular expressions developed in Perl. Then, these sentences are manually annotated by at least two coders. The first task is to determine whether the sentences contain valid medical paraphrases or no paraphrase at all. Additionally, the term, the paraphrase marker and the paraphrase are also annotated.

Lexical and semantico-pragmatic annotation. The second task consists on the identification of lexical relations and the semantico-pragmatic functions of the paraphrases. On one hand, the lexical relations were defined as lexical links that exist between the two segments, the medical term and its paraphrase. These lexical relations can be synonymy, hypernymy, hyponymy and meronymy, as they are frequent in medical texts (Condamines 2018; Ramadier 2016; Săpoiu 2013). On the other hand, semantico-pragmatic functions represent the reasons that drives the speaker to use paraphrases in written medical texts, such as definition, rephrasing, designation, exemplification, or explanation (Eshkol-Taravella and Grabar 2017; Buhnila 2022b).

Thus, we obtained a new dataset, **NLFMED**, containing 1,033 medical paraphrases of Covid-19 terms, and a rich annotation following the same guidelines as for **RefoMed**. The two datasets are merged together into a larger dataset **LARGEMED** (17,393 sentences, annotated with terms, markers, paraphrases, lexical relations and semantic-pragmatic functions). This corpus is available for experiments of paraphrase classification and generation, in order to automatize data collection. These experiments are presented in the next section. Afterwards, we discuss the findings and limitations of our method for data collection.

6. Results and Discussion

Firstly, we evaluate the results of the annotation process applied in the **NLFMED** dataset. Secondly, we present the results from the classification and generation experiments conducted using this augmented paraphrase dataset.

6.1. Corpus Annotation and Evaluation

Only 1,725 sentences out 8,565 sentences containing Covid-19 terms contained both terms and paraphrase markers. To these sentences, we added 140 term definitions and explanations from the Wikipedia pages presenting the Covid terms. The annotation done by the two coders resulted in 1,033 correct paraphrases. We computed the Krippendorff's α score for several tasks: **a)** classification of sentences containing paraphrases; **b)** paraphrase markers; **c)** correct paraphrases; **d)** lexical relations and **e)** semantico-pragmatic functions.

For the task of sentence classification, we used the labels *"yes"* if the sentence contains a valid subsentential paraphrase and *"no"* - if the sentence contains no valid paraphrase. For this task, the inter-coder agreement is very high (**0,95**), meaning that the coders agreed in most of the cases. Then, we computed this agreement for the subsentential paraphrases : the coders agreed on recognizing a paraphrase in the sentence. The Krippendorff's α score was still very good (**0,80**) for this task as well as for the task of finding common discourse markers that introduce a paraphrase (α=**0,82**). For the other elements that were annotated, the inter-coder agreement was good for the semantic-pragmatic functions (α=**0,77**), but weaker for lexical relations (α=**0,55**). Most cases of agreement concern the definition and the exemplification contexts, while paraphrases or explanations are more often subject of disagreement. For the lexical relation annotation, several confusions between meronymy and hyponymy or hyponymy/hypernymy (due to the reverse order of term and of the paraphrases) could be an explanation of a lower agreement score.

The existing **RefoMed** dataset and the newly built one from the Covid-19 Web archives NLF are compiled into a single medical subsentential paraphrase corpus for French **LARGEMED**. The same annotation guidelines are used to build both datasets. The method of building this corpus is mainly based on existing dictionaries (SNOMED for **RefoMed** and Dicovid-19 for the **NLFMED** dataset). If the terms are not found in the dictionary, then the sentences containing a paraphrase are not selected. In order to automate data collection, we conduct some experiments with the resulting corpus LARGEMED to build a binary model to detect if the sentence contains or not a paraphrase (section 6.2) or to adapt a generation model for creating variants of medical paraphrases (section 6.3). We present these experiments and the results obtained in the following subsections.

6.2. Binary Classification Experiments

The process of manual selection of sentences containing real paraphrases is time-consuming, but of high quality, when validated by human coders. In order to automatize the selection of sentences potentially containing paraphrases and to accelerate manual annotation, we built a binary classification model for detecting sentences containing paraphrases. For this purpose, we adapted the French **CamemBERT** language model (Martin et al., 2019) for the task of sentence classification, by pairing it with a set of 17,393 sentences manually annotated from the LARGEMED dataset. We used the information about paraphrase status (*yes* or *no*). We applied a cross-validation strategy with 5 and 10 folds, and we obtained the accuracy score of **0,84** and respectively **0,89**. From the several configurations of optimizers and loss functions, the *Adam* optimizer and the *SparseCategorialCrossentropy* loss function obtained the best results.
To compare this result with a bidirectional LSTM architecture, we use **CamemBERT** (Martin et al., 2019) to represent each sentence. The results show few variations between parameters such as the maximum length of the sentence containing or not paraphrases. However, we tried several configurations (embedding size of 150 and 200) and hyperparameters with the bidirectional LSTM architecture.

We obtained better accuracy results with CamemBERT when we used cross-validation (0,84, if we consider k=5 and 0,89 if we consider k=10) (see **Table 3**). For the bidirectional LSTM, we randomly selected 90% or 75% of the data for training, and we used several embedding size (150, 200). In this case, the accuracy was only 0,81.

Train	Test	Embd size	Embd LM	Acc
75%	25%	200	CBERT	0.81
90%	10%	150	CBERT	0.81
Cross k=5	-	150	CBERT	0.84
Cross k=10	-	200	CBERT	0.89

Table 3: Results of the classification task. **Train** represents the training split size, while **Test** is the test split size. **Embd size** is the **embeddings size** used for the experiments and **Emdb LM** represents the Language Model (LM) used for the task, which is the French LM CamemBERT (CBERT). For cross validation Cross, the values for k folds are available. We evaluate our results with accuracy (**Acc**).

We expected to obtain better result to automate the search of sentences with potential paraphrases. 11% of automatic annotation of the status of the sentences are errors, so this result should be improved. However, it is simpler to correct the automatic annotation rather than to do it from scratch. While we collected a large number of sentences from the Web archives, presumably containing terms and paraphrase markers, the sentence classifier helps reducing the time required to annotate the corpus, at least for the status task and will be useful to complete the dataset with new sentences containing potential paraphrases. For the other tasks, especially for lexical relation identification, the inter-coder agreement is too low to try to automatize the process.

6.3. Generation Experiments

As an alternative to data collection from existing Web pages, we propose to evaluate the quality of a paraphrase generation tool to obtain new paraphrases for the medical terms. Thus, we present the experiments using the new dataset LARGEMED in order to adapt a model to generate new medical paraphrases for the French Covid-19 terms. We adapt the APT neural architecture for adversative paraphrases and we use the T5 language model for generation and the dataset presented at section 5.3.

6.3.1. The APT Neural Network

The **APT** (*Adversarial Paraphrasing Task*) neural architecture (Nighojkar and Licato, 2021) uses a method for generating paraphrases with equivalent meanings but with lexical and syntactic differences at the surface level. This model identifies the general meaning of a sentence, not just the meaning of individual words. The APT architecture verifies if two sentences that are mutually implicit are also semantically equivalent. **APT** uses **BLEURT** (Sellam et al., 2020) to measure structure dissimilarity. **BLEURT** score evaluates automatically generated texts based on the word embeddings of the BERT language model (Devlin et al., 2018).
The corpus of paraphrases is used to adapt the APT paraphrase generation architecture for French medical data. **APT** generates paraphrases which have similar meanings (e.g. it is possible to infer the meaning of the term from the paraphrase and the term's meaning from the paraphrase).
The main changes of this strategy is the use of T5 model, available for French, which should be adapted for medical data, by using LARGEMED dataset including Covid-19 related terms and their paraphrases.

6.3.2. T5 Language Model

T5 (*Text-to-Text Transformer*) (Raffel et al., 2020) was pre-trained on **C4** (*Colossal Clean Crawled Corpus*), a corpus with 7 terabytes of data extracted from the Common Crawl Web corpus. T5 had been trained for several specific NLP tasks, including

paraphrase identification and sentence similarity. We adapt it for our own dataset of subsentential medical paraphrases in French.

6.3.3. Technical Aspects

We extract our experimental data from the **LARGEMED** paraphrases dataset (9,557 terms and their paraphrases from **RefoMed** and 1,033 paraphrases of Covid-19 terms from **NLF**). We fine-tuned the model t5_base with several configurations (the size of the paraphrase is 128 and 256 respectively): learning rate (3e-4), 4 epochs, the batch size (20), dropping rate (0,01), and AdamW optimiser (1e-8).

6.3.4. Generation Results

We obtained 2,372 generated paraphrases for a test set of 576 terms contained in the test file (96 terms are related to Covid-19 pandemic). For each term, we obtained at most 5 paraphrase predictions. We analysed the predictions and annotated with *1* if the generated paraphrases are correct and *0* if they are incorrect.

Predictions	Nb of terms	Percentage
At least 1 correct result	204	35.41 %
No correct result	372	64.59 %
Total	576	100 %

Table 4: The paraphrases generated (**Predictions**) by the T5 base model adapted for medical domain.

The paraphrases generated for 95 Covid-19 terms are generally quite far from the expected prediction. The few mentions of the Sars coronavirus or of the disease produce some paraphrases containing virus or disease with respiratory symptoms, but a large part of these terms do not generate valid output. We show some incorrect examples below, where *Truth* represents the initial paraphrase for the term, while *Prediction* represents the paraphrase generated by the language model.

- *Term*: **maladie à coronavirus 2019 (coronavirus disease 2019)**

  Truth: Covid-19 (Covid-19)

  Prediction: à transmission hépatique (hepatically transmitted)

- *Term*: **choc cytokinique (cytokine shock)**

  Truth: réponse exacerbée du système immunitaire inné (exacerbated response of the innate immune system)

  Prediction: une maladie de l'hémoglobine (a haemoglobin disease)

From all the predictions for the Covid-19 terms, we identify correct paraphrase predictions for 24 terms out of 96 from the Covid-19 term list. The correct paraphrases proposed are in general introduced by hypernyms: *Covid-19 longue* (long Covid-19) is paraphrased with *maladie chronique* (chronic disease); *la réplication virale* (the viral replication) is paraphrased with *une réplication de l'infection* (a replication of the infection).

We consider that the low performance of the language model in our experiments could be explained by the few occurrences of Covid terms in the training data set. Some Covid-19 terms design the measures to limit pandemic (*social distance, PCR test*) which are difficult to predict from the medical texts used to train the model. In the actual state of the model, few new paraphrases are provided if we compare with the paraphrases already available in the LARGEMED dataset.

7. Conclusion and Future Work

In this article we present a work in progress aiming to build a paraphrase corpus for medical terms collected from the Web archives of the National French Library and a method to extend this corpus by paraphrase classification and generation. Secondly, we follow the guidelines for annotating the paraphrases with lexical relations and semantico-pragmatic functions already applied for **RefoMed**. We created a new annotated resource of 1,033 Covid-19 related medical terms with their correspondent paraphrase **NLFMED** and compiled it into a larger French dataset **LARGEMED** (17,393 terms and their subsentential paraphrases). We obtained an accuracy score of 0.89 for the paraphrase classification task with CamemBERT. Still, it is possible to apply this classifier to pre-select sentences with paraphrases and then to refine by searching paraphrase markers and terms. The paraphrase generation is a difficult task. The results were not satisfactory for the Covid terms, due to the small size of our Covid-19 paraphrase dataset.

Future work includes enlarging the paraphrase Covid-19 dataset automatically with Solr extractions and then applying the binary classification to pre-select sentences containing paraphrases. Actually, the collection of new Web pages containing Dicovid terms is still in progress. The task of automatic paraphrase generation could give better results by combining APT with a language model adapted for the medical domain in French, such as CamemBERT-Bio (Martin et al., 2019) or DrBERT (Labrak et al., 2023), but also combining our dataset with other dataset available for general language. The final dataset will be used in a text simplification system for medical domain.

8. Acknowledgements

The ADAPTMED project has been funded by the National French Library (BNF) (https://www.bnf.fr) and supported by the National Library of the University of Strasbourg (BNU) (https://bnu.fr/fr), by FRLC (Research Network on Language and Communication) (https://frlc.hypotheses.org/) and the LiLPa research unit (University of Strasbourg) (https://lilpa.unistra.fr/).

9. References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Alham Fikri Aji, Tirana Noor Fatyanosa, Radityo Eko Prasojo, Philip Arthur, Suci Fitriany, Salma Qonitah, Nadhifa Zulfa, Tomi Santoso, and Mahendra Data. 2022. Paracotta: Synthetic multilingual paraphrase corpora from the most diverse translation sample pair. *arXiv preprint arXiv:2205.04651*.

Edwide Antoine and Natalia Grabar. 2016. Exploitation de reformulations pour l'acquisition d'un vocabulaire expert/non expert. In *TALN 2016: Traitement Automatique des Langues Naturelles*.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Houda Bouamor, Auréelien Max, and Anne Vilnat. 2013. Multitechnique paraphrase alignment: A contribution to pinpointing sub-sentential paraphrases. *ACM Trans. Intell. Syst. Technol.*, 4(3).

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Chris Brockett and William B Dolan. 2005. Support vector machines for paraphrase identification and corpus construction. In *Proceedings of the third international workshop on paraphrasing (IWP2005)*.

Ioana Buhnila. 2022a. Identifying medical paraphrases in scientific versus popularization texts in french for laypeople understanding. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 69–79.

Ioana Buhnila. 2022b. Le rôle des marqueurs et indicateurs dans l'analyse lexicale et sémantico-pragmatique de reformulations médicales. In *SHS Web of Conferences*, volume 138, page 10005. EDP Sciences.

Ioana Buhnila. 2023. *Une méthode automatique de construction de corpus de reformulation*. Ph.D. thesis, University of Strasbourg, France.

Rémi Cardon and Natalia Grabar. 2018. Identification of parallel sentences in comparable monolingual corpora from different registers. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 83–93.

Rémi Cardon and Natalia Grabar. 2021. Simplification automatique de textes biomédicaux en français: lorsque des données précises de petite taille aident. In *Traitement Automatique des Langues Naturelles*, pages 275–277. ATALA.

Bruno Cartoni and Louise Deléger. 2011. Découverte de patrons paraphrastiques en corpus comparable: une approche basée sur les n-grammes (extracting paraphrastic patterns comparable corpus: an approach based on n-grams). In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 182–187, Montpellier, France. ATALA.

Anne Condamines. 2018. Nouvelles perspectives pour la terminologie textuelle.

Josette Condamines, Anne ; Rebeyrolle. 1997. Point de vue en langue spécialisée. *Meta*, 42(1):174–184.

Roger A Cote. 1998. Systematized nomenclature of human and veterinary medicine: Snomed international. version 3.5. *Northfield, IL: College of American Pathologists*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

William Dolan, Chris Quirk, Chris Brockett, and Bill Dolan. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*.

Iris Eshkol-Taravella and Natalia Grabar. 2017. Taxonomy in reformulations from a corpus linguistics

perspective. *Syntaxe et semantique*, 18(1):149–184.

Catherine Fuchs. 1982. La paraphrase entre la langue et le discours. *Langue française*, La vulgarisation(53):22–33.

Catherine Fuchs. 2020. Paraphrase et reformulation: un chassé-croisé entre deux notions.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In *LREC*, pages 4276–4283. Citeseer.

Natalia Grabar and Rémi Cardon. 2018. Clear-simple corpus for medical french. In *ATA*.

Natalia Grabar and Thierry Hamon. 2015. Extraction automatique de paraphrases grand public pour les termes médicaux. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 182–195.

Natalia Grabar and Thierry Hamon. 2016. Exploitation de la morphologie pour l'extraction automatique de paraphrases grand public des termes médicaux. *Traitement Automatique des Langues*, 57(1).

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

Yun He, Zhuoer Wang, Yin Zhang, Ruihong Huang, and James Caverlee. 2020. Parade: A new dataset for paraphrase identification requiring computer science domain knowledge. *arXiv preprint arXiv:2010.03725*.

Anaïs Koptient, Rémi Cardon, and Natalia Grabar. 2019. Simplification-induced transformations: typology and some characteristics. In *BioNLP 2019*.

Anaïs Koptient and Natalia Grabar. 2020. Fine-grained text simplification in french: steps towards a better grammaticality. In *International Symposium on Health Information Management Research*.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Drbert: A robust pre-trained model in french for biomedical and clinical domains. *bioRxiv*, pages 2023–04.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Animesh Nighojkar and John Licato. 2021. Improving paraphrase detection with the adversarial paraphrasing task. *arXiv preprint arXiv:2106.07691*.

Anaïs Pecout, Thi Mai Tran, and Natalia Grabar. 2019. Améliorer la diffusion de l'information sur la maladie d'alzheimer: étude pilote sur la simplification de textes médicaux. *Ela. Etudes de linguistique appliquée*, 3(195):325–341.

Qiwei Peng, David Weir, and Julie Weeds. 2023. Testing paraphrase models on recognising sentence pairs at different degrees of semantic overlap. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 259–269, Toronto, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Lionel Ramadier. 2016. *Indexation et apprentissage de termes et de relations à partir de comptes rendus de radiologie*. Ph.D. thesis, Université Montpellier.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Ayla Rigouts Terryn, Véronique Hoste, Patrick Drouin, and Els Lefever. 2020. Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset.

Camelia Săpoiu. 2013. *Hiponimia în terminologia medicală: modalităţi de abordare în semantică şi lexicografie*. Trend.

Yves Scherrer. 2020. Tapaco: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Andon Tchechmedjiev, Amine Abdaoui, Vincent Emonet, Stella Zevio, and Clement Jonquet. 2018. Sifr annotator: ontology-based semantic annotation of french biomedical text and clinical notes. *BMC bioinformatics*, 19:1–26.

Amalia Todirascu, Sebastian Padó, Jennifer Krisch, Max Kisselew, and Ulrich Heid. 2012. French and german corpora for audience-based text type classification. In *LREC*, volume 2012, pages 1591–1597.

Hélène Vassiliadou. 2020. Peut-on aborder la notion de" reformulation" autrement que par la typologie des marqueurs? pour une analyse sémasiologique et onomasiologique.

Wei Xu, Chris Callison-Burch, and William B Dolan. 2015. Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

Chao Zhou, Cheng Qiu, and Daniel Acuna. 2022. Paraphrase identification with deep learning: A review of datasets and methods.

Clearer Governmental Communication: Text Simplification with ChatGPT Evaluated through Quantitative and Qualitative Research

Nadine Beks van Raaij, Ksenia Podoynitsyna, Daan Kolkman
Jheronimus Academy of Data Science, Vlerick Business School, Utrecht University
Den Bosch - Eindhoven - Tilburg, The Netherlands, Vlerick, Belgium, Utrecht, The Netherlands
nadine.v.raaij (at) gmail.com, ksenia.podoynitsyna (at) vlerick.com, d.a.kolkman (at) uu.nl

Abstract

This study explores the use of ChatGPT for simplifying Dutch government letters to improve their comprehensibility while preserving legal accuracy. We employed a three-stage mixed-methods evaluation approach to assess the effectiveness of a naive baseline, RoBERTa, and ChatGPT in simplifying six of the most complex letters selected from a corpus of 200. The evaluation process involved comparing the outputs using four metrics (ROUGE, BLEU, BLEURT, and LiNT), followed by reviews from legal and linguistic experts, and culminating in a randomized controlled trial with 72 participants to test comprehension. Our results indicate that ChatGPT substantially enhances the comprehension of government letters, evidenced by more than a 20% increase in comprehensibility scores and a 19% improvement in participants' ability to correctly answer questions related to follow-up actions based on the simplified texts. Additionally, our study underscores the importance of a thorough evaluation framework and advises caution in solely depending on automated metrics for assessing text simplification.

Keywords: natural language generation, text simplification, ChatGPT 3.5, prompt engineering, legal documents, real-life task, human evaluation

1. Introduction

Text simplification (TS), a Natural Language Processing (NLP) task, aims to enhance readability and comprehensibility while retaining the essence of the text (Alva-Manchego and Shardlow, 2022; Al-Thanyyan and Azmi, 2021; Shardlow, 2014). TS can help diverse audiences, from people with disabilities (Carroll et al., 1998) and non-native speakers (Stajner, 2021) to those with limited literacy (Belder et al., 2010) by ensuring text accessibility and comprehension.

The value of TS is particularly apparent in government communication. Clear communication from government bodies is vital for promoting transparency, fostering civic engagement, and facilitating informed participation (Renkema, 2013; Lentz and Pander Maat, 2011; Sanders and Jansen, 2011; Kraf and Pander Maat, 2009). Yet, many governments, including that of the Netherlands, grapple with comprehensible communication (Pander Maat and van der Geest, 2021; Lentz et al., 2017). Recent episodes in the Netherlands underscore the challenge of government communications (Amnesty, 2021), with studies such as Pander Maat and van der Geest (2021) pinpointing issues in the comprehensibility of government letters.

Recognising these challenges, the Dutch government has taken proactive steps by enlisting communication experts to revise letters to citizens (Gebruiker-Centraal, 2022) and experimenting with NLP solutions (Rijksoverheid, 2023). Exploratory work by Feng et al. (2023) and Jeblick et al. (2022) demonstrates the potential of ChatGPT for TS on several benchmark datasets and radiology reports respectively. Motivated by these developments, our paper considers the question:

> *To what extent can large language models (LLMs) improve the comprehensibility of Dutch letters sent by governmental organisations?*

We answer this question by investigating empirically three approaches to TS: a naive token-substitution model, RoBERTa (Robustly Optimized BERT Pretraining Approach), and ChatGPT. We do so by a three-step mixed-method evaluation procedure which involves: 1. A comparison of evaluation metrics (ROUGE, BLEU, BLEURT, and LiNT); 2. Qualitative assessment by a legal and linguistic expert; 3. A randomized controlled trial with 72 participants. We demonstrate the importance of a robust evaluation procedure and find that TS using ChatGPT improves the comprehensibility of Dutch letters by 20%. Since ChatGPT 3.5 and 4 can handle multiple languages (Feng et al., 2023) our results have relevance for TS at large.

2. Related work

Although alternatives such as Bidirectional Encoder Representations from Transformers (BERT) exist, Generative Pre-trained Transformer (GPT) models typically outperform these alternatives (Tan and Kieuvongngam, 2020; Eisele, 2019), which is why we set out to explore GPT models in this study. The

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 152–178

value of this architecture has been demonstrated in relation to language learning (Young and Shishido, 2023; Luo et al., 2023) and TS of medical reports (Lyu et al., 2023; Holmes et al., 2023; Jeblick et al., 2022).
Suha and Azmi (2021) provide an overview of the past research for multiple languages in the field of TS and conclude that Data-driven simplifications outperform Rule-based simplifications. Furthermore, Suha and Azmi (2021) highlight the need for further research in developing new simplification techniques and reliable evaluation methods. Therefore, this research contributes to the research of performing a hybrid evaluation.

2.1. Prompt engineering ChatGPT

The quality of prompts provided to GPTs deeply impacts their outputs, which is why others have focused on prompt engineering for TS Feng et al. (2023); Holmes et al. (2023); Lyu et al. (2023); Engelmann et al. (2023). One recommendation of these studies is to process texts one by one instead of providing multiple texts at once as input for ChatGPT to avoid model hallucinations. Therefore, in this study, we chose to focus on one letter per prompt or a related set of prompts.
In addition, these studies use prompts that explicitly ask the model to "retain the content" and mention the original author's role or intended audience in the prompt to provide extra context. Often they also provide a dataset with example classifications of difficult/complex words/texts or offer example simplifications. These studies do not delve deeper into the methodology behind the generation of these few-shot/one-shot/zero-shot prompts or comparisons of different prompts that aim for the same audience and purpose. Holmes et al. (2023); Lyu et al. (2023) show the success of TS in a medical context for different audiences having differences in education level. Others have ventured to transform texts to particular readability levels in an effort to produce educational material for language students (Young and Shishido, 2023; Alkaldi and Inkpen, 2023). However, readability and comprehensibility are not the same[1] and without labeled texts, performing these simplifications is challenging.
This study employs prompt engineering for a single audience, citizens, who do not all have the same legal background or expert knowledge and should therefore receive plain language from governmental organizations. We follow up on the best practices of the above-named studies.
Our main focus is increasing the comprehensibility of the letters in practice. The prompts we used do not contain specifications about what is complex and what constitutes an example simplification. This is because there is a gap between what should be easy to comprehend and what actually is easy to comprehend for the majority of people. Therefore, we validate our results by focusing on the evaluation by the actual readers (through the randomized controlled trial) instead of prompting an automatic evaluation metric based on assigned examples that should be easy to comprehend or difficult to comprehend.

2.2. Automatic evaluation metrics

We use four quantitative evaluation metrics that align with established evaluation methods for automatic text summarization:

2.2.1. ROUGE

In a comprehensive review of automatic text summarization by Yadav et al. 2022 Recall-Oriented Understudy for Gisting Evaluation (ROUGE) was used. Additionally, Offerijns et al. 2020 and Gao et al. 2019 employed BLEU alongside ROUGE, enriching assessment with precision and recall considerations. Building on these foundations, this research also employs the ROUGE metric, which evaluates summarization and translation quality using scores ranging from 0 to 1, wherein higher values signify enhanced summarization or translation proficiency.

2.2.2. BLEU

Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) is the second evaluation metric used in this research. BLEU is a popular automatic evaluation metric used to assess the quality of machine-translation output. It compares a machine-generated translation with one or more human reference translations and assigns a score based on how similar they are. The score ranges from 0 to 1, with 1 indicating a perfect match between the machine translation and the human reference translation.
This metric is not without problems for different text generation tasks. BLEU is not well suited, for example, for assessing simplicity from a lexical nor a structural point of view (Sulem et al., 2018). These findings indicated a weak or nonexistent correlation between BLEU and parameters related to grammaticality and meaning preservation in cases where sentence splitting is involved. Additionally, Sulem et al. (2018) found that BLEU tends to have a negative correlation with simplicity, which penalises

[1] Readability pertains to how easily a text can be read, often assessed through factors like sentence and word length (Dols, 2018; Lentz et al., 2017; Pander Maat and Dekker, 2016; Renkema, 2011). Comprehensibility relates to how well a reader can grasp a text's meaning, influenced by factors like idea complexity, text structure, and vocabulary difficulty. Comprehensibility ensures a text is not only easy to read but also easy to understand (Lentz et al., 2017; P., 2012; Renkema, 2011).

simpler sentences. They demonstrated this, via a created corpus for sentence splitting, containing multiple paraphrases, and compared it to human judgements. However, TS does not only rely on sentence splitting and other simplification studies (Xu et al., 2016; Stajner et al., 2014) have shown that it correlates with human judgements of grammaticality and meaning preservation. Therefore further research into BLEU was performed. Furthermore, BLEU was included in the end to create a benchmark for the automatic evaluation metrics. By comparing the scores of BLEU with BLEURT, the scores of the BLEURT become more valuable.

2.2.3. BLEURT

Incorporating BLEU-based Learned Evaluation for Text (BLEURT) enriches the evaluation strategy of this study. BLEURT evaluates text quality by gauging the correspondence between generated content and human assessments. Unlike BLEU, which primarily examines n-gram overlap, BLEURT delves into semantic alignment, enhancing its assessment accuracy. With a score range of -1 to 1, higher BLEURT scores signify superior performance. This approach is further validated by its alignment with human judgement, considering both surface-level and semantic similarity (Dipanjan and Parikh, 2020).

2.2.4. LiNT

Leesbaarheidsinstrument voor Nederlandse Teksten (LiNT) is the first evaluation metric that is used to evaluate the text on readability. LiNT was chosen as previous research proved this metric to be the most reliable Dutch metric to evaluate text on readability (Lentz, 2021; Pander Maat and Dekker, 2016; Kraf et al., 2011; Kraf and Pander Maat, 2009). LiNT makes calculations about the sentence structure characteristics and word characteristics and summarises this in a formula that is based on the T-scan (Pander Maat and Dekker, 2016) and outputs a LiNT score ranging from 1 to 100 (1 is the easiest, 100 the most difficult). Furthermore, LiNT categorises these scores into four levels: level one is the easiest, and level four is the most difficult. Level one holds for scores up to 36, between 36 and 51 the level is two, between 51 and 61.5 the level is three, and above 61.5 the level is four. The meaning of these levels according to Pander Maat and Ditewig 2017 is that with level one, 14% of the adult readers in the Netherlands and Flanders do not understand the text. For level two this is 30%, for level three this is 52% and for level four this is 80%.

2.3. Qualitative research with human evaluations

The cited studies underscore the importance of evaluating text summarization and simplification through a combination of quantitative and qualitative methods. Iskender et al. (2021) highlight the importance of exploring the reliability of human evaluations for text summarizations by analyzing the evaluators' characteristics. Furthermore, factors such as lexical and syntactic changes, and comprehensibility dimensions should be addressed. Notably, Nguyen et al. (2021) employ ROUGE for quantitative assessment and involve experts for qualitative evaluation, while Gosens (2008) conducted a qualitative study considering reader comprehension and analysed the results by means and standard deviation and made a comparison between the original and adjusted texts. Sikkema et al. (2017) explored comprehensibility dimensions in debt collection letters with education levels and letter volume as influential factors. Other related studies (Dols, 2018; Lentz et al., 2017; Renkema, 2011) also contribute insights into text evaluation methodologies. This research adapts best practices from previous studies, employing an expert review and a randomized controlled trial to evaluate three letters each and validate the results with regression analyses, aligning with established recommendations (Roobaea and Mayhew, 2014; Molich, 2010; Macefield, 2009; Hertzog, 2008; Faulkner, 2003).

3. Experiment

We conduct our experiment with three models: 1. a naive model that substitutes jargon with a simple explanation; 2. RoBERTa (Robustly Optimized BERT Pretraining Approach) that was finetuned with the same jargon-definition list as the naive model; 3. ChatGPT (based on GPT-3.5-turbo) with prompt engineering.

For ChatGPT, four different chats for every letter of the test data were used.[2] The simplified texts were pasted into a Word file and saved separately per letter. All the letters were manually checked for spelling and grammar mistakes. The generated letters by the naive model contained one spelling mistake which is explained further in section 4.1.2. The results of RoBERTa contain many grammatically incorrect sentences which are also further elaborated in section 4.1.2. No spelling or grammar mistakes were found in the letters simplified by ChatGPT. Furthermore, we checked if all the letters included the same contact details and, in case of a deviation, this has been adjusted to the original. Lastly, the layout of all the letters has been made equal meaning white spaces are added or deleted to comply with the layout of the original letters. This was done because previous research has shown that the layout influences the comprehension and interpretation of letters (Dols, 2018).[3]

[2]The questions asked in every chat for ChatGPT can be found in appendix B.

[3]The letters used for this research can be found in

Corsius et al. (2023) introduced a dataset of 200 letters (100 on Finance and 100 on Care) originating from multiple governmental organizations spread over the Netherlands. On average the length of these letters is 627 words. Corsius et al. (2023) identified six letters (3 on Finance and 3 on Care) that were hardest to comprehend. We use these six letters in the first stage of our evaluation procedure in which we compare the quantitative evaluation metrics (ROUGE, BLEU, BLEURT, and LiNT) for the three different TS approaches.

Given the often unreliable results of the evaluation metrics (Engelmann et al., 2023), the second stage of our evaluation procedure involves experts. More specifically, the outputs of the three models are evaluated by a legal and linguistic expert using the CCC-model (Lentz and Elling, 2003). The CCC-model is a framework for text evaluation that stands for Correspondence, Consistency, and Correctness, and that needs to be applied across five levels: text type, content, structure, formulation, and presentation. The experts discussed each simplified letter using this framework. Special consideration was given to the degree to which the simplified letters were equivalent from a legal perspective.

In the third and final stage of our evaluation procedure, we conducted a randomized controlled trial with 72 participants who all read three letters (in a random combination of original and simplified versions for each of these letters). As a result, we have 216 observations on the reader-letter level. This sample size is in line with recommendations by Fritz et al. (2012); Hertzog (2008).

Seventy-two participants were recruited online through convenience sampling. Participants were required to be at least 18 years old with a basic understanding of Dutch.[4]

As opposed to previous research on the comprehensibility of Dutch governmental texts (Corsius et al., 2023; Dols, 2018) this research distinguishes different reader' characteristics that influence the interpretation of comprehension. The participants were randomly divided into eight groups, each reading different combinations of letters in the same order. Figure 1 shows the distribution of the groups per education level. We follow a procedure where participants read three letters and answered questions about their content (understanding questions[5]), effectiveness (action questions[6]) and tone (tone questions[7]). The questionnaires were created by the linguist and legal expert and follow the guidelines of Grusky et al. (2018) and literature by Cox and Brayton (2008).[8]

4. Results

4.1. Automatic evaluation metrics

In this first stage of our evaluation, we find that both the naive approach and RoBERTa attain decent results (based on the automatic evaluation metrics[9]), while ChatGPT scores are less impressive.

4.1.1. Original letters

The LiNT score for the original letters was calculated to give an indication of the difficulty level. Five of the six original letters have a LiNT score between 36 and 51, indicating that 30% of adult readers in the Netherlands do not understand these letters. As these scores show, the letters in the theme Care are more difficult compared to the letters in the theme Finance. We will use these three letters as a critical case study in the randomized controlled trial.

4.1.2. Simplified letters

Interestingly the naive model has higher LiNT scores than the original letters, except for the $Regels\_pgb$ letter where the naive model scored 44 and the original 47. This indicates that the naive model decreased the readability. However, the LiNT scores did differ at most 4 points from the original letters and did have the same level categorisations, meaning that the difference is only minor. Looking at the other metrics, the naive model had high scores for both precision and recall. The BLEU and ROUGE scores are close to one for the naive model. This is to be expected from the fact that the ROUGE, BLEU, and BLEURT scores take the original letters as references and the naive model does not change any sentence structures or grammatical aspects. The BLEU scores decrease when the n-grams increase. This is logical as the naive model substitutes words or small parts of a sentence meaning that there is the smallest difference on the 1-gram level, and the biggest difference (lowest similarity) on the 4-gram level. However, these scores are still close to one, indicating a high similarity.

The RoBERTa model achieved the lowest LiNT scores and was able to get all letters categorised in level one. The lowest score was achieved for the letter $Betalen\_in\_delen$ with 26 points. The highest LiNT score of the RoBERTa model, being 32,

appendix A.

[4] Participants' characteristics such as education level and reading habits can be found in appendix F.

[5] Questions to test if the reader correctly understood what was meant with certain terms and statements.

[6] Questions to test if the reader knew which steps to take or what actions to do in certain situations or when encountering problems.

[7] Questions regarding the interpretation and tone of the text.

[8] The full questionnaires of the randomized controlled trial can be found in appendix D.

[9] The results of the automatic evaluation metrics of these models can be found in appendix E.

was achieved for the letter $Regels\_pgb$. These two letters are also marked as the simplest and most difficult letters based on the scores of the original letters. RoBERTa model thus scores considerably lower for the LiNT metric compared to the original letters. For the BLEU and ROUGE metrics we can see lower scores compared to the naive model. Regarding the BLEURT score, RoBERTa model achieved the lowest scores and seems to have only limited similarity with the original letters.

For four of the six letters, ChatGPT scored significantly higher compared to the original letter for the LiNT metric. This indicates that ChatGPT transformed the original letters to letters that are harder to read. For the other two letters ($Regels\_pgb$ and $Gemeentelijke\_belastingen$), ChatGPT scored lower compared to the original letter. Remarkable for these two letters is that they have the highest (BLEU_1 = 0.79, ROUGE_1 = 0.76) and lowest (BLEU_1 = 0.32, ROUGE_1 = 0.58) BLEU and ROUGE scores. This could imply that the LiNT metric encountered difficulties in evaluating these letters with the result that the scores differ from the others.

Regarding the BLEURT metric, ChatGPT scores range from 0.46 to 0.75. From these results, it seems that ChatGPT is able to simplify the letters while retaining the structure of the original letters. Taking this evidence together with the results of the expert review and the randomized controlled trial results, we can conclude that the technical metrics results should be treated with caution when evaluating the results of the TS task.

4.2. Expert review

The recommendations of the research of Cramwinckel 2014 together with the juridical background of the legal expert have been used as a guideline for the evaluation of the simplified letters in terms of juridical correctness. Below a summary of the experts' review is given.

The experts observe the letters simplified by the naive model are almost identical to the original letters. This is a result of too few words occurring in the original letters that were in the definition list of the naive model. Therefore the naive model did not find enough words to replace, which resulted in identical letters except 3 words per letter on average. Furthermore, the naive model replaced subwords which are part of a longer word. In instance where the full word is not included in the definition list, replacement of subwords results in linguistically incorrect sentences. An example is the original word "mogelijk" (possible) where "gelijk" (equal) was found in the definition list and had a definition of "nu" now). The original word was replaced by "monu", which is not a Dutch word. Therefore it was concluded that the naive model did not give the aimed simplifications and was not evaluated further.

The experts also evaluated the simplifications of the RoBERTa model. It was concluded that this model simplified the letters too much, with the result that the meaning of the text was gone. An example of an oversimplification is the original word "besluit" (decision) which was simplified to "antwoord" (response) by RoBERTa. This is neither linguistically nor juridically correct. Therefore we decided not to evaluate this model any further.

From the simplifications of ChatGPT, the linguistic expert observed that they have a shorter syntactic dependency length (SDL) compared to the original letters, which makes them easier to read. This is in line with Kleijn et al. 2016 who proved in their research that shorter SDL results in shorter processing times and positively affects the understanding of texts. Furthermore, the linguist expert concluded that the simplified texts were linguistically correct. The legal expert concluded that the important juridical information of the original letters was present in the simplified letters too. In sum, the simplifications of ChatGPT were considered sufficient in terms of linguistic and juridical correctness and were further evaluated with the randomized controlled trial.

4.3. Prompt engineering ChatGPT

Based on the results of the first two stages of our evaluation procedure, we refined the prompts.

In the first attempt, ChatGPT's ability to simplify text to Common European Framework of Reference for Languages (CEFR) levels was explored. The output was then classified by "Klinkende taal" (as Kraf et al. (2011) concluded this software performed the best for this classification task) and the experts to a CEFR language level. However, due to the contextual complexity of governmental letters, accurately determining the language level proved challenging. This aligns with prior research (Suha and Azmi, 2021) suggesting that CEFR may not be suitable for texts with specialized content, leading to the exclusion of this approach.

An effort was made to enrich ChatGPT's vocabulary and improve simplification quality by providing additional input based on a jargon definition list by Gebruiker-Centraal (2022). The jargon of this definition list did occur only limited in the tested letters and had very general explanations according to the linguistic expert. Although ChatGPT didn't directly utilize these definitions for simplification, it aided in detecting and avoiding difficult jargon. As this approach didn't contribute significantly, it wasn't included in the final prompt version.

Furthermore, a comparison was made between simple prompts and combinations of prompts consistent with the "chain-of-thought" approach, with various questioning approaches tested. All comply-

ing with the best practices of Madaan et al. (2023); Yu et al. (2023). Asking for "comprehension" rather than "simplification" yielded better results, avoiding over-simplification and information loss. The choice between "simple" and "easy" phrasing did not substantially impact outcomes. Among the different questions, using a few-shot approach consistently produced improved simplifications. Based on this, the final version of the ChatGPT prompt utilized the one-shot approach for enhanced simplification.

4.3.1. Final version: from bullet points to an easy text

From the few-shot attempts, it became clear that when was asked to rewrite the text to bullet points, all the important information was included. Since this was one of the problems with the earlier simplifications, we gave ChatGPT a prompt to first rewrite the text in bullet points and then make an easy text from these bullet points.[10]

4.4. Randomized controlled trial

Seventy-two participants (thirty-six men and thirty-six women) were recruited online between February and March 2023 for this study through convenience sampling. Participants were asked to fill in their availability and contact details. Prior to conducting the reading comprehension experiment, ethical approval from our Ethical Review Board was sought and obtained. Participants were required to be at least 18 years of age and have at least a basic understanding of Dutch. No other demographic characteristics were considered in the recruitment process.

The experiment has followed the guidelines of the ISO framework (Bevan et al., 2016). Participants were randomly divided into eight groups, with each group reading a different combination of the three letters. Figure 1 shows the number of participants per group and education level. The abbreviation "O" represents the Original version whereas "G" represents the Generated simple version by ChatGPT.

The letters were presented to participants in a predetermined order. This was done to control for order effects and to reduce potential biases. Before reading the letters, participants were given a brief introduction to the study and provided with a short scenario introduction for every letter. They were instructed to read the letters carefully and take notes if they wished. After reading a paragraph of the letter, participants were asked to answer questions about the letter.

The questionnaire consists of both closed and open-ended questions and took approximately 10-15 minutes to complete in total per letter. After reading all three letters and completing the questionnaires, participants were debriefed on the purpose of the study and thanked for their participation.

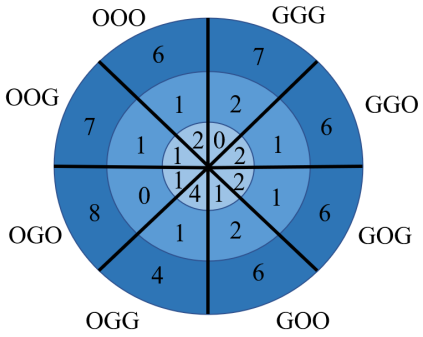


Figure 1: **Participants distribution:** Every chart represents a subgroup that reads the same letters in the same order. "O" represents the original letter and "G" represents the generated simplified letter. Every area has a number representing the number of participants. The inner circle represents the low-educated participants, the middle ring represents the middle-educated participants and the outside ring represents the high-educated participants. No distinction was made in this graph between men and women since there was an equal division within the subgroups.

4.4.1. Means comparison of original and simplified letters

Figure 2 shows the scores for the original and simplified letters. The first original letter saw notable enhancement when simplified, showing increased correct answers. This pattern persisted across subsequent letters, confirming improved comprehension. Aggregated results further confirm this, with participants scoring above 90% for both understanding and action question types for the simplified letters.

4.4.2. Regression analyses

We further investigate the difference in the performance of participants using (generalized) linear regression analyses and Multivariate Analysis of Variance (MANOVA).[11] This approach was chosen in order to make valid conclusions and investigate possible correlations that might influence the percentages and averages as seen in previous research (Dols, 2018).

Both analyses confirm that the simplified versions of the letters were better understood having significant scores for the simplified type of letter influencing the percentages of correctly answered questions for all three letters. Additionally, the age of participants

[10] This resulted in the final prompts which are shown in chat four of appendix B

[11] The full outcome of these analyses can be found in appendix G and H - values for the dummies $l1_g$,$l2_g$, and $l3_g$ represent simplified letters.

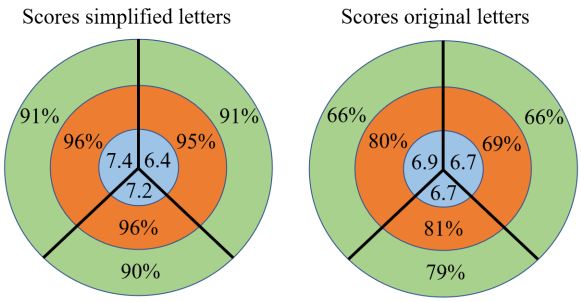


Figure 2: **Scores of the letters** The left diagram represents the scores of the simplified letters and the right diagram represents the scores of the original letters. Every chart represents a letter. The top-right represents the first letter, the top-left represents the second letter and the lowest chart represents the third letter. The surface of the chart represents the number of words that the letter contained (meaning that a bigger surface relates to a longer letter). The colours represent the type of questions. Green: understanding questions, orange: action questions and blue: tone questions. For the understanding and action questions, the percentage of correct answers is shown. For the tone questions, the average grade that participants assigned to the letters is shown.

emerged as a significant variable, demonstrating that higher ages negatively influenced the percentage of correctly answered action questions.
The MANOVA results further confirm our findings, providing an additional layer of confirmation for the positive impact of the simplified letters. The notable consistency across these types of analyses proves the robustness of our findings.

5. Discussion

This study investigated the extent to which a naive model, RoBERTa, and ChatGPT can improve the comprehensibility of formal texts written by Dutch governmental organisations. The challenge of TS in such a context is that the result needs to retain essential information allowing citizens to take actions while making the text easier to understand and act upon.
The results from multiple attempts at prompt engineering showed that it is possible to develop a one-shot learning approach (Kojima et al., 2022) to achieve excellent results, which makes scaling up this TS task easier. Initially writing the text in bullet points, followed by transforming these into easy-to-read text, proved to be the most effective prompt for this research.
Despite the evaluation metrics suggesting otherwise, the expert analysis determined that only the ChatGPT model's generated letters fulfilled the simplification criteria, maintaining all crucial information and terminology. Consequently, we proceeded with this model exclusively for the randomized controlled trial.
The results of the randomized controlled trial show that the ChatGPT model excelled in terms of enhancing the comprehensibility of the letters. An average increase of more than 20% was achieved for the percentage of correctly answered understanding questions. For the percentage of correctly answered action questions, there was an increase of 19% on average. Additionally, the grade for the tone was higher for the second and third letters, namely 0.5 on a 10-point scale. Only the first letter received on average 0.3 less compared to the tone grade for the original letter. However, the results regarding the tone were not significant because of inconsistent grading by the participants and too limited data because only three grades were solicited.
The randomized controlled trial results were further analysed with regression analyses to examine how correctly answering understanding and action questions was influenced by the simplified versions of the letters, controlling for various other variables. Three models were used to perform the analyses: Generalised Linear models, Linear Models, and MANOVA models. Across all models, the dummy variables indicating the simplified version were consistently significant, validating the results presented in Figure 2. Regarding our main research question, the machine learning model ChatGPT has demonstrated a substantial improvement in terms of the comprehensibility of the letters.

5.1. Future work and limitations

Future work could focus on improving the prompts, as initial exploration of tailoring prompts to particular audiences shows promise. Tailored prompts can make calls to action clearer and more compelling for specific audiences, thereby increasing the likelihood of the desired response, whether it's complying with regulations, or participating in civic activities.
Therefore, future work in this area could focus on developing more sophisticated techniques for audience analysis and prompt customization, thereby maximizing the impact of simplified texts for diverse audiences.

5.1.1. Evaluation methodology

TS for the purpose of general readership is a task that requires human evaluation to validate the results of the task (Engelmann et al., 2023). Numerous automatic evaluation metrics are developed to help alleviate this resource-intensive task. This study joins Young and Shishido (2023) in raising concerns with regard to the reliability of the automatic evaluation metrics for assessing simplification tasks performance of LLMs in general and ChatGPT in particular. We find that ROUGE, BLEU, and BLEURT poorly capture the quality of TS of governmental texts for general audiences. Hence they should be used with caution and ideally refined. Consensus is lacking on such metrics' suitability for TS assessment (Engelmann et al., 2023). Our results demonstrate that the models with the highest BLEU and ROUGE scores did not necessarily yield the best simplifications. We observe that BLEURT scores were not consistently 1.0 when evaluating identical reference text due to dataset limitations and model constraints such as syntactic structure: BLEURT may not fully account for changes in syntactic structure introduced by the simplification process. A simplified sentence may have a different sentence structure compared to the original, which could affect readability and clarity in a positive way but result in lower scores for the automatic evaluation metrics as the structure changes compared to the reference.

Furthermore, automatic evaluation metrics may struggle to evaluate how well the simplified text captures the intended meaning within the broader context. They primarily focus on local similarity measures and may not capture broader contextual information. However, our experience is that the original letters are not very well-structured neither coherent. Changes to both sentences structure and paragraphs placement to make it in a broader context coherent, is advisable in such cases.

For proper evaluation, combining BLEURT scores with other metrics and expert assessments is advised. Future research could consider expanding the reference texts to improve the performance of automatic evaluation metrics. The qualitative interviews and randomized controlled trials, though valuable, have limitations. Future studies should involve a wider range of experts and include for example the original letters' authors. Moreover, including more people with lower education levels and those with reading disabilities as participants could yield potentially even greater results for the TS impact in the randomized controlled trials.

In conclusion, this research offers insights into the efficacy of simplifying Dutch formal texts with ChatGPT, while at the same time underpinning the need for refinement and further exploration of evaluation methodologies.

5.1.2. Scaling up text simplification tools

The scope of this study was limited to testing multiple text simplification models and their evaluation. However, for future deployment, research needs to be performed with the stakeholders who are going to use the envisioned tool in their work as authors of letters. Therefore, it is recommended to conduct interviews with these stakeholders and find a form of implementation that suits them. An example format of implementation could be a web-based interface such as "Simpel" (Rijksoverheid, 2023) has for citizens (but then designed for personal computer use instead of smartphones) that allows the authors of the letters to input the original text and receive the simplified version straight away. The interface could also provide options for customising the level of simplification based on the target audience or purpose of the letter, as it can be fully focused on supporting the authors of the letters. By showing the original input and the generated simplified text on one screen the authors can rate the level of simplification and extent to which the essence of the text is retained, being important from the legal perspective. Correlating these ratings with new and existing TS evaluation metrics will allow the researchers to refine them further.

Bringing the results of this research into deployment requires several steps. First, a suitable model must be chosen to be able to simplify large letters at once. As alternatives for ChatGPT are popping up (Harnish, 2023), a comparison of these models should be made whereas the best model should be chosen for implementation. Furthermore, a way to check automatically for missing information must be implemented and/or a disclaimer must be provided that the author must check this him- or herself.

Once the model has been successfully deployed and proven effective for the authors, it could have a significant impact on improving the readability and comprehensibility of governmental texts. This could lead to better communication and engagement with citizens, as well as more efficient and effective use of resources by governmental organisations.

Considering the ongoing efforts to make LLMs in general and ChatGPT in particular more responsible, the performance of the next generation ChatGPT (e.g. ChatGPT 4.0) is not necessarily better than ChatGPT 3.5 (Chen et al., 2023) hence performance of TS tasks also requires a continuous re-assessment as new LLMs emerge. Scaling to different types of letters and languages also requires further investigation.

5.2. Ethical considerations

The deployment of LLMs for the simplification of formal texts from governmental organizations to citizens introduces a novel approach to enhancing accessibility and comprehension. While this technology promises significant benefits in making government communications more understandable to a broader audience, it also raises ethical considerations that must be addressed to ensure its responsible use. This section outlines the primary ethical concerns related to potential biases and harms that could arise from such automated systems and the measures taken to mitigate these risks.

5.2.1. Potential biases and harms

The use of LLMs comes with specific ethical concerns, which we tried to address using the following strategies:

1. Controlled Input Information: Unlike typical LLM applications that generate content based on provided information, our approach strictly limits the model's role to simplifying the text without altering the content. This significantly reduces the risk of introducing new biases or errors in the message content, as the original information remains intact.

2. User Oversight and Control: We emphasize the importance of human oversight in the text simplification process. By ensuring that users (government officials or designated communicators) retain full control over the output, we can mitigate risks associated with automated generation. This approach allows for the careful review and adjustment of simplified texts to ensure they accurately and effectively convey the intended message without unintended biases or simplifications that could distort the meaning.

3. Transparency and Accountability: We tried to be transparent in the use of LLMs for text simplification. Specifically, by documenting and communicating the processes involved, including how the models were trained and the criteria used for simplification.

Overall, we feel that by maintaining strict control over the input information, ensuring user oversight, promoting transparency, and committing to continuous improvement, we can leverage the benefits of this technology for TS while minimizing risks.

6. Bibliographical References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification.

Wejdan Alkaldi and Diana Inkpen. 2023. Text simplification to specific readability levels. *Mathematics*, 11.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification.

Fernando Alva-Manchego and Matthew Shardlow. 2022. Towards readability-controlled machine translation of COVID-19 texts. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 287–288, Ghent, Belgium. European Association for Machine Translation.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

International Amnesty. 2021. Xenofobe machines: Discriminatie door ongereguleerd gebruik van algoritmen in het nederlandse toeslagenschandaal.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

R. Anthony, Artino Jr., Jeffrey S. La Rochelle, Kent J. Dezee, and Hunter Gehlback. 2014. Developing questionnaires for educational research: Amee guide. *Medical Teacher*, 36(87):463–474.

Jan De Belder, Koen Deschacht, and Marie-Francine Moens. 2010. Lexical simplification.

Nigel Bevan, Jim Carter, Jonathan Earthy, Thomas Geis, and Susan Harker. 2016. New iso standards for usability, usability reports and usability measures. *HCI*, pages 268–278.

Eva Boontje. 2011. Een onderzoek naar de begrijpelijkheid van hypotheekvoorlichting. Master's thesis, Bacheloropleiding Communicatiestudies Faculteit Geesteswetenschappen Universiteit Utrecht, April.

Bram Bulte, Leen Sevens, and Vincent Vandeghinste. 2018. Automating lexical simplification in dutch. Master's thesis, Centre for Computational Linguistics, KU Leuven, Belgium, June.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Wuwei Lan Yang Zhong Wei Xu Chao Jiang, Mounica Maddela. 2021. Neural crf model for sentence alignment in text simplification. Master's thesis, Department of Computer Science and Engineering The Ohio State University, August.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time?

Jordan Clive, Kris Cao, and Marek Rei. 2022. Control prefixes for parameter-efficient text generation.

Mischa Corsius, Vera Lange, Yvette Linders, Henk Pander Maat, Els van der Pool, Nina Sangers, Keun Sliedrecht, Wouter Sluis-Thiescheffer, and Charlotte Swart. 2023. Monitor begrijpelijkheid overheidsteksten.

James Cox and Keni Brayton. 2008. How to build the best questionnaires in the field of eduction. pages 2–13.

T. A. Cramwinckel. 2014. De belastingdienst als vertaler: van wettekst naar webtekst. een casestudy. *MBB*, (7-8):299–312.

Pieter Delobelle, Thomas Winters, and Nettina Berendt. 2020. Robbert: a dutch roberta-based language model. Master's thesis, Department of Computer Science, KU Leuven, Faculty of Electrical Engineering and Computer Science, TU Berlin, September.

Alice Delorme Benites and Caroline Lehr. 2021. Neural machine translation and language teaching – possible implications for the cefr. Master's thesis, Zürcher Hochschule für Angewandte Wissenschaften Institut für Übersetzen und Dolmetschen.

Ashwin Devaraj, William Sheffield, Byron C. Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. *Association for Computational Linguistics*, 1(60):7331–7345.

Thibault Sellam Dipanjan and Das Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. (58):7881–7892.

Marc Dols. 2018. Brieven aan burgers: Analyse en evaluatie. Master's thesis, Communicatie- en Informatiewetenschappen Specialisatie: Bedrijfscommunicatie en Digitale Media (BDM) Faculteit Geesteswetenschappen Universiteit van Tilburg, May.

Michael Eisele. 2019. On automatic summarization of dutch legal cases. Master's thesis, Hamburg Universität Facultät Für Mathematik, Informatik und Naturwissenschaften, October.

ELSA-Lab. 2022.

Björn Engelmann, Fabian Haak, Christin Katharina Kreutz, Narjes Nikzad Khasmakhi, and Philipp Schaer. 2023. Text simplification of scientific texts for non-expert readers.

Laura Faulkner. 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, I& Computers*, 3:379–383.

Hiske Feenstra, Karen Keune, Henk Pander Maat, Theo Eggen, and Ted Sanders. 2015. Geautomatiseerde beoordeling van schrijfvaardigheid. Master's thesis.

Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models.

Ellie Fossey, Carol Harvey, Fiona McDermott, and Larry Davidson. 2002. Understanding and evaluating qualitative research. *Australian and New Zealand Journal of Psychiatry*, 36:717–732.

Fritz, Morris, and Richler. 2012. Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141(1), 2–18.*

Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R. Lyu. 2019. Generating distractors for reading comprehension questions from real examinations. *Association for the Advancement of Artificial Intelligence*, pages 6423–6430.

Gebruiker-Centraal. 2022.

Dianne Gosens. 2008. Het effect van lexicale en syntactische wijzigingen op het begrip en de waardering van een autoverzekeringspolis. Master's thesis, Masteropleiding Communicatiestudies Faculteit Geesteswetenschappen Universiteit Utrecht, June.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. Master's thesis, Department of Computer Science, Cornell Tech Cornell University, New York, NY 10044, June.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Bart Haak. 2020. Information extraction from homicide-related dutch texts using bert. Master's thesis, Jheronimus Academy of Data Science, July.

Brian Harnish. 2023. Chatgpt alternatives you can try in 2023.

Melody A. Hertzog. 2008. Considerations in determining sample size for pilot studies.

Jason Holmes, Zhengliang Liu, Lian Zhang, Yuzhen Ding, Terence T. Sio, Lisa A. McGee, Jonathan B. Ashman, Xiang Li, Tianming Liu, Jiajian Shen, and Wei Liu. 2023. Evaluating large language models on a highly-specialized topic, radiation oncology physics.

Neslihan Iskender, Tim Polzehl, and Sebastian Moller. 2021. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. Master's thesis, Technische Universitat Berlin, Quality and Usability Lab, April.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, and Michael Ingrisch. 2022. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports.

J. Joshua. 2023. Data controls faq.

Suzanne Kleijn, Henk Pander Maat, and Ted Sanders. 2016. Effects of dependency length on the processing and understanding of texts. Master's thesis, Communicatie- en informatiewetenschappen Universiteit Utrecht.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid Google Research, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners.

Rogier Kraf, Leo Lentz, and Henk Pander Maat. 2011. Drie nederlandse instrumenten voor het automatisch voorspellen van begrijpelijkheid. *Tijdschrift voor Taalbeheersing*, (3):249–265.

Rogier Kraf and Henk Pander Maat. 2009. Leesbaarheidsonderzoek: oude problemen, nieuwe kansen. *Tijdschrift voor Taalbeheersing*, (2):97–123.

Yogesh Kumar, Komalpreet Kaur, and Sukhpreet Kaur. 2021. Study of automatic text summarization approaches in diferent languages. Master's thesis, Department of Computer Science Engineering, Chandigarh Group of Colleges, Landran, Mohali, India, January.

Naomi Langstraat. 2019. Creating a classroom-mt: Connecting simplification methods to language learner levels in monolingual machine translation. Master's thesis, Utrecht University, Faculty of Humanities, Bachelor of Science - Artificial Intelligence, July.

Alyssa Lees, Jeffrey Sorensen, and Ian Kivlichan. 2020. Jigsaw @ ami and haspeede2: Fine-tuning a pre-trained comment-domain bert model.

Leo Lentz. 2021. Wat zijn tekstbegrijpelijkheids voorspellingen waard? een vergelijkend onderzoek. *Handboek Didactiek Nederlands*, (4).

Leo Lentz and Sanne Elling. 2003. De voorspellende kracht van het ccc-model. *Tijdschrift voor Taalbeheersing*, (3):221–235.

Leo Lentz, Louise Nell, and Henk Pander Maat. 2017. Begrijpelijkheid van pensioencommunicatie: effecten van wetgeving, geletterdheid en revisies.

Leo Lentz and Henk Pander Maat. 2011. Een leesbare bijsluiter. *Tijdschrift voor Taalbeheersing*, (2):128–151.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Master's thesis, Facebook AI, Oktober.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. Master's thesis, Stanford University, January.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

van de Nick Luijtgaarden, Daniël Prijs, Marijn Schraagen, and Floris Bex. 2022. Abstractive summarization of dutch court verdicts using sequence-to-sequence models. Master's thesis, Utrecht University, The Netherlands, December.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

Qing Lyu, Josh Tan, Michael E. Zapadka, Janardhana Ponnatapura, Chuang Niu, Kyle J. Myers, Ge Wang, and Christopher T. Whitlow. 2023. Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: Promising results, limitations, and potential.

Ritch Macefield. 2009. How to specify the participant group size for usability studies: A practitioner's guide. *Journal of usability studies*, 5:34–35.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023. What makes chain-of-thought prompting effective? a counterfactual study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1448–1535, Singapore. Association for Computational Linguistics.

Paulo R. A. Margarido, Thiago A. S. Pardo, Gabriel M. Antonio, Vinícius B. Fuentes, Rachel Aires, Sandra M. Aluísio, and Renata P. M. Fortes. 2008. Automatic summarization for text simplification: Evaluating text understanding by poor readers. Master's thesis, HAN university of applied sciences.

Louis Martin, Angela Fan, Éric de la Clergerie, and Antoine Bordes Benoît Sagot. 2021. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. Master's thesis, Facebook AI Research, Paris, France, April.

Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of gpt-3 and advanced neural language models. Master's thesis, Center on

Terrorism, Extremism, and Counterterrorism Middlebury Institute of International Studies, September.

Rolf Molich. 2010. A critique of "how to specify the participant group size for usability studies: A practitioner's guide". *Journal of usability studies*, 5:124–128.

Elisa Nguyen, Daphne Theodorakopoulos, Shreyasi Pathak, Jeroen Geerdink, Onno Vijlbrief andMaurice van Keulen, and Christin Seifert. 2021. A hybrid text classification and language generation model for automated summarization of dutch breast cancer radiology reports. Master's thesis, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands, May.

Jeroen Offerijns, Suzan Verberne, and Tessa Verhoef. 2020. Better distractions: Transformer-based distractor generation and multiple choice question filtering. Master's thesis, Leiden Institute of Advanced Computer Science, Leiden University, October.

OpenAI. 2023. Gpt-4 technical report. Technical report.

Gavora P. 2012. Text comprehension and text readability. *Faculty of Humanistic studies, Tomas Bata University, Czech Republic*, pages 9–10.

Gustavo H. Paetzold and Lucia Specia. 2017. A survey on lexical simplification. Master's thesis, The University of Sheffield Western Bank Sheffield United Kingdom, November.

Kendeou Panayiota, Krista R. Muis, and Sandra Fulton. 2011. Reader and text factors in reading comprehension processes.

Henk Pander Maat and Nick Dekker. 2016. Tekstgenres analyseren op lexicale complexiteit met t-scan. *Tijdschrift voor Taalbeheersing*, 38(3):263–304.

Henk Pander Maat and Sanne Ditewig. 2017. Hoe worden onderwijsteksten vereenvoudigd, en helpt dat? *Handboek Didactiek Nederlands*, (39):245–263.

Henk Pander Maat, Leo Lentz, and Raynor D. K. 2015. How to test mandatory text templates: The european patient information leaflet. Technical report, PLOS one.

Henk Pander Maat and Thea van der Geest. 2021. Monitor begrijpelijkheid overheidsteksten.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *Association for Computational Linguistics*, 1(40):311–318.

Wouter Peer. 2023.

Daniël Prijs. 2022. On automatic summarization of dutch legal cases. Master's thesis, Utrecht University Graduate School of Natural Sciences Business Informatics, July.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Jan Renkema. 2011. Een kwalitatief onderzoek naar de begrijpelijkheid van digitale informatie van de belastingdienst.

Jan Renkema. 2013. Een kwalitatief verkennend onderzoek naar de kwaliteit van antwoordbrieven van de belastingdienst.

M. Heerkens K.R. Leuverink I. Veringa Renkema J. 2012. Minder belasting door meer begrip. *Universiteit van Tilburg Faculteit Geesteswetenschappen*, (4).

Rijksoverheid. 2023. Terugblik demo donderdag: Lees simpel app versimpelt overheidsinformatie. *Rijksprogramma voor Duurzaam Digitale Informatiehuishouding*.

Samuel Ronnqvist, Jenna Kanerva, and Tapio Salakoski Filip Ginter. 2020. Is multilingual bert fluent in language generation? Master's thesis, TurkuNLP, Department of Future Technologies University of Turku, Finland.

Alroobaea Roobaea and Pam J. Mayhew. 2014. How many participants are really enough for usability studies? Technical report, IEEE.

Alessandra Rossetti. 2019. Simplifying, reading, and machine translating health content: An empirical investigation of usability. Master's thesis, School of Applied Language and Intercultural Studies Dublin City University, April.

Muhammad Salman, Armin Haller, and Sergio J. Rodríguez Méndez. 2023. Syntactic complexity identification, measurement, and reduction through controlled syntactic simplification.

Ted Sanders and Carel Jansen. 2011. Begrijpelijke taal – fundamenten en toepassingen van effectieve communicatie. *Tijschrift voor Taalbeheersing*, (33):201–207.

Victor Sanh and Thomas Wolf Lysandre Debut, Julien Chaumond. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Juliën Schoonbrood. 2013. Onderzoek naar de invloed van financiële geletterdheid op pensioenkennis en de vind- en begripsprestaties van de startbrief. Master's thesis, Communicatie- en informatiewetenschappen Universiteit Utrecht, November.

Matthew Shardlow. 2014. A survey of automated text simplification.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable.

A. F. Siddiqi. 2014. An observatory note on tests for normality assumptions. *Journal of modelling in management*, (3):290–305.

Tialda Sikkema, Leo Lentz, Henk Pander Maat, and Nadja Jungmann. 2017. De taakgerichtheid van de aanmaning en de dagvaarding in incassozaken. *Tijdschrift voor Taalbeheersing*, 39(3):273–295.

Lynn Snyder, Caccamise Donna, and Wise Barbara. 2005. The assessment of reading comprehension. *Journal of modelling in management*.

S. Soberón and Winfried Stute. 2017. Assessing skewness, kurtosis and normality in linear mixed models. *Journal of Multivariate Analysis*, pages 123–140.

Sanja Stajner. 2021. Automatic text simplification for social good: Progress and challenges. pages 2637–2652.

Sanja Stajner, Ruslan Mitkov, and Horacio Saggion. 2014. One step closer to automatic evaluation of text simplification systems. pages 1–10.

S. Suha and Aqil M. Azmi. 2021. Automated text simplification: A survey. 54(2):36.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. Master's thesis, Department of Computer Science, The Hebrew University of Jerusalem, October.

Bowen Tan and Virapat Kieuvongngam. 2020. Automatic text summarization of covid-19 medical research articles using bert and gpt-2. Master's thesis, Laboratory of Molecular Genetics Rockefeller University New York, June.

(VNG) Vereniging Nederlandse Gemeenten. 2021. Regionaal verbeteren brieven omgevingswet. *Vereniging van Nederlandse Gemeenten (VNG)*.

Dr. S. Vijayarani, Ms. J. Ilamathi, and Ms. Nithya. 2020. Preprocessing techniques for text mining - an overview.

de Wietse Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert mode. Master's thesis, CLCG, University of Groningen, The Netherlands, December.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. Master's thesis, University of California, Santa Barbara, September.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study.

Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, and De Rosal Ignatius Moses Setiadi. 2022. Review of automatic text summarization techniques methods. 34(4):1029–1046.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. Master's thesis, School of Engineering and Applied Sciences Harvard University Cambridge, MA, USA, July.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. (7):401–415.

Divakar Yadav, Jalpa Desai, and Arun Kumar Yadav. 2022. Automatic text summarization methods: A comprehensive review. Master's thesis.

Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, and Chris Tanner. 2022. What gpt knows about who is who. Master's thesis, Institute for Applied Computational Science Harvard University, May.

Julio Christian Young and Makoto Shishido. 2023. Evaluation of the potential usage of chatgpt for providing easier reading materials for efl students.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. Towards better chain-of-thought prompting strategies: A survey.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2020. Retrospective reader for machine reading comprehension. Master's thesis, Department of Computer Science and Engineering, Shanghai Jiao Tong University.

7. Appendices

A. Letters used for randomized controlled trial

The letters used for the randomized controlled trial can be found on this github page: https://anonymous.4open.science/r/COLING-24-93E6/.
The letters were in Dutch and are translated here for comprehension of this research.

B. Questions prompt-engineering ChatGPT

The prompts for ChatGPT were in Dutch and are translated here for comprehension of this research. For all these prompts "de volgende tekst" refers here to the text of the letter.

Chat one: CEFR levels

- Kun je het niveau van de volgende tekst bepalen volgens de CEFR-classificaties?
- Kun je de volgende tekst vereenvoudigen naar CEFR-niveau A1?
- Kun je de volgende tekst tekst vereenvoudigen naar CEFR-niveau A2?
- Kun je de volgende tekst vereenvoudigen naar CEFR-niveau B1?
- Kun je de volgende tekst vereenvoudigen naar CEFR-niveau B2?
- Kun je de volgende tekst herschrijven naar CEFR-niveau A1?
- Kun je de volgende tekst herschrijven naar CEFR-niveau A2?
- Kun je de volgende tekst herschrijven naar CEFR-niveau B1?
- Kun je de volgende tekst herschrijven naar CEFR-niveau B2?
- Kun je de volgende tekst herschrijven naar CEFR-niveau C1?
- Kun je de volgende tekst herschrijven naar CEFR-niveau C2?

Translations:

- Can you determine the level of this text according to the CEFR classifications?
- Can you simplify this text to CEFR level A1?
- Can you simplify this text to CEFR level A2?
- Can you simplify this text to CEFR level B1?
- Can you simplify this text to CEFR level B2?
- Can you rewrite this text to CEFR level A1?
- Can you rewrite this text to CEFR level A2?
- Can you rewrite this text to CEFR level B1?
- Can you rewrite this text to CEFR level B2?
- Can you rewrite this text to CEFR level C1?
- Can you rewrite this text to CEFR level C2?

Chat two: Additional input

- Kunt u deze tekst vereenvoudigen met behulp van deze definitielijst waarbij in de eerste kolom het moeilijke woord staat en in de tweede kolom de eenvoudige definitie?
- Kunt u deze definitielijst gebruiken om deze tekst te vereenvoudigen?
- Kunt u deze definitielijst gebruiken om deze tekst te herschrijven?

Translations:

- Can you simplify this text with the use of this definition list having in the first column the difficult word and in the second column the simple definition?
- Can you use this definition list to simplify this text?
- Can you use this definition list to rewrite this text?

Chat three: Zero-shot vs Few-shot

1. Kun je de volgende tekst begrijpbaarder schrijven?
2. Kun je de volgende tekst versimpelen?
3. Kun je de volgende tekst opschrijven in bullet points?
4. Kun je deze bullet points in een eenvoudige tekst opschrijven?
5. Kun je hiervan een makkelijke tekst schrijven?

Translations:

1. Can you make the following text more comprehensible?
2. Can you simplify the following text?
3. Can you write the following text in bullet points?
4. Can you write a simple text based on these bullet points?
5. Can you write an easy text from this?

Chat four: Final version

1. Kun je de volgende tekst opschrijven in bullet points?
2. Kun je deze bullet points in een makkelijke tekst schrijven?

Translations:

1. Can you rewrite this text into bullet points?
2. Can you write these bullet points into an easy text?

C. Introduction and scenario description randomized controlled trial

The introduction and scenario descriptions were in Dutch but are translated for the comprehension of this research.

Introductie onderzoek

Beste lezer, Wat fijn dat je meedoet aan het lezersonderzoek van mijn thesis. Je krijgt zo drie teksten te zien over het thema zorg. Om de privacy van de gemeenten en de geadresseerden te bewaken, zijn de teksten geanonimiseerd en gesitueerd in de denkbeeldige gemeente Zilverdam. Ik wil je vragen om de brieven één voor één te lezen en daarbij te zeggen wat je denkt. Het is belangrijk om aan te geven als je iets niet begrijpt of onduidelijk vindt. Ik wil je vragen om deze stukken te markeren. Daarnaast worden er vragen gesteld door mij over de inhoud van de brieven tijdens het onderzoek en naderhand over de toon van brief. Deze vragen geven inzicht in hoe makkelijk:

- Je begrijpt wat er staat;
- Je begrijpt wat er gedaan moet worden;
- Je de toon gepast vindt.

Scenario script thema zorg:

Brief 1: WMO voorzieningen

Jouw tante Janny woont samen met haar man in de gemeente Zilverdam. Ze zijn beide met pensioen. Janny heeft steeds meer moeite met haar evenwicht. Ze loopt nu met een stok. Traplopen vindt ze erg lastig. De slaapkamer is boven en daarom wil Janny een traplift. Jij bent Janny's mantelzorger. Ze vraagt jou of je wilt kijken of wat er geregeld kan worden bij de gemeente.

Brief 2: Regels PGB

Janny en jij hebben een gesprek gehad met iemand van de gemeente. De traplift die Janny wil, zit niet in het aanbod van de gemeente. De gemeente zegt dat ze moet kijken naar een pgb. Lees de tekst om te kijken of een pgb iets voor Janny is.

Brief 3: Besluit PGB

Janny kon niet langer wachten en heeft alvast een traplift besteld. Met de gemeente maakte ze ondertussen een plan en deed de aanvraag voor een pgb. Lees de tekst om uit te leggen wat er is besloten.

Research Introduction

Dear reader, thank you for participating in the reader survey for my thesis research. You will now be presented with three texts on the topic of healthcare. To protect the privacy of municipalities and recipients, the texts have been anonymized and are situated in the imaginary municipality of Zilverdam. I kindly request you to read each of the letters one by one and share your thoughts as you do so. It's important to indicate if there is anything you do not understand or find unclear. Please mark these sections. Additionally, I will ask questions during and after the research about the content of the letters and the tone used in them.
These questions will provide insight into how easily you:

- Understand the content;
- Comprehend what needs to be done;
- Find the tone appropriate.

Scenario Script: Theme Care

Letter 1: WMO Facilities

Your aunt Janny lives with her husband in the municipality of Zilverdam. They are both retired. Janny is experiencing increasing balance issues and now uses a cane. Climbing stairs is challenging for her. The bedroom is upstairs, so Janny wants a stairlift. You are Janny's caregiver, and she has asked you to see if anything can be arranged with the municipality.

Letter 2: PGB Regulations

You and Janny had a conversation with someone from the municipality. The stairlift Janny wants is not part of the municipality's offerings. The municipality suggests she explore a Personal Budget (PGB). Please read the text to determine if a PGB is suitable for Janny.

Letter 3: PGB Decision

Janny couldn't wait any longer and has already ordered a stairlift. In the meantime, she worked with the municipality to create a plan and applied for a PGB. Please read the text to understand what decision has been made.

D. Questionnaires for randomized controlled trial

The questions and answers were in Dutch but are translated for the comprehension of this research.

Vraag	Juiste antwoord
Wat moet je doen als je hulp nodig hebt om zelfstandig te kunnen blijven wonen?	1. Melden hulpvraag
	2. Formulier in vullen
Heb je een DigiD nodig voor het invullen van het formulier?	Ja
Wat kun je doen als je geen DigiD hebt?	Telefonisch contact opnemen
Hoe meld je een hulpvraag?	Via de knop "Verzoek voor Wmo-voorziening"
Hoe heet het meldingsformulier?	Sociale dienstverlening
Binnen hoeveel tijd wordt de aanvraag beoordeeld?	8 weken
Waar wordt ernaar gekeken bij het beoordelen van de nodige zorg?	1. Hulp buren/familie
	2. Algemene voorzieningen
	3. Maatwerkvoorzieningen
Zijn maatwerkvoorzieningen persoonlijk?	Ja
Wat is een voorbeeld van een algemene voorziening?	Maaltijdservice
	Maatjesproject
Via wat kun je maatwerkvoorzieningen krijgen?	Zorg in natura (ZIN)
	Persoonsgebonden budget (PGB)
Wat houdt zorg in natura in?	De gemeente regelt alles
Hoe kun je opzoeken welke zorg de gemeente inkoopt?	Via de website Sociale kaart Zilverdam
Is de eigen bijdrage voor ZIN en PGB hetzelfde?	Ja
Wat houdt een persoonsgebonden budget in?	Zelf verantwoordelijk om de zorg te regelen
	(inkopen zorg, administratie, opstellen zorgovereenkomst)
Voor welke zorg en ondersteuning betaal je een eigen bijdrage?	Zie opsomming
Hoe hoog is de maximale eigen bijdrage?	19 Euro
Waarvan is de maximale eigen bijdrage afhankelijk?	leeftijd
	wel/geen partner
Wie bepaalt de hoogte van de eigen bijdrage?	CAK
Tot wanneer betaal je de eigen bijdrage?	Zolang je ondersteuning nodig heeft / tot de kostprijs bereikt is
Wat is het abonnementstarief?	De eigen bijdrage Wmo
Klopt het dat de eigen bijdrage af hangt van de hoeveelheid zorg?	Nee
Klopt het dat je tijdens een vakantie geen eigen bijdrage betaalt?	Nee
Waar vindt je meer informatie over de hoogte van de eigen bijdrage?	Website CAK
Aan wie betaal je de eigen bijdrage?	CAK
Hoevaak betaal je de eigen bijdrage?	Elke maand
Hoe betaal je de eigen bijdrage?	Automatische incasso / acceptgiro
Wat kun je doen als je meer informatie wilt over betalen?	Website CAK bezoeken / CAK bellen op 0800 1925
Wat moet je doen als je indicatie afloopt?	Contact opnemen zorgaanbieder
Wanneer moet je contact opnemen met de zorgaanbieder als je indicatie afloopt?	2 maanden voor het aflopen
Moet je voor het aflopen van de indicatie het aanvraagformulier invullen	Nee
Wat moet je doen voor vragen over sociale dienstverlening?	Bellen met 0900 1234
Waarvoor is de casemanager?	Overige vragen (als je er via andere manieren niet uitkomt)
time:	
Hoe duidelijk vind je de brief in het algemeen? (1-10)	
Vind je dat er overbodige informatie in de brief staat? Zo ja, waar?	
Wat vind je van de toon van de tekst? (bijvoorbeeld streng of vriendelijk)	
Hoe erg gepast vind je de toon van de brief in het algemeen? (1-10)	
Wat zou de schrijver van de tekst als eerste moeten veranderen?	

Figure 3: Questions with answers for letter one: WMO voorzieningen (Dutch).

Question	Correct answer
What should you do if you need help to continue living independently?	1. Report your request for help
	2. Fill out a form
Do you need a DigiD to fill out the form?	Yes
What can you do if you don't have a DigiD?	Contact by phone
How do you report a request for help?	Through the "Request for Wmo-provision" button
What is the name of the reporting form?	Social services
Within how much time is the request assessed?	8 weeks
What is considered when assessing the necessary care?	1. Help from neighbors/family
	2. General provisions
	3. Customized provisions
Are customized provisions personal?	Yes
What is an example of a general provision?	Meal service
	Buddy project
How can you receive customized provisions?	Nature of care (ZIN)
	Personal budget (PGB)
What does nature of care (ZIN) entail?	The municipality handles everything
How can you find out which care the municipality procures?	Through the website Social Map Zilverdam
Is the own contribution for ZIN and PGB the same?	Yes
What does a personal budget (PGB) entail?	Personally responsible for arranging care
	(purchasing care, administration, establishing care agreement)
For which care and support do you pay an own contribution?	See list
What is the maximum own contribution amount?	19 Euros
What is the maximum own contribution amount dependent on?	age
	marital status
Who determines the amount of the own contribution?	CAK
Until when do you pay the own contribution?	As long as you require support / until the cost threshold is reached
What is the subscription fee?	The Wmo own contribution
Is it true that the own contribution depends on the amount of care?	No
Is it true that you do not pay an own contribution during a vacation?	No
Where can you find more information about the amount of the own contribution?	CAK website
To whom do you pay the own contribution?	CAK
How often do you pay the own contribution?	Every month
How do you pay the own contribution?	Automatic debit / payment slip
What can you do if you want more information about payments?	Visit the CAK website / Call CAK at 0800 1925
What should you do when your indication expires?	Contact the care provider
When should you contact the care provider when your indication expires?	2 months before it expires
Do you need to fill out the application form before the indication expires?	No
What should you do for questions about social services?	Call 0900 1234
What is the casemanager for?	Other questions (if you cannot resolve them through other means)
time:	
How clear do you find the letter overall? (1-10)	
Do you think there is any unnecessary information in the letter? If yes, where?	
What do you think of the tone of the text? (e.g., strict or friendly)	
How appropriate do you find the tone of the letter overall? (1-10)	
What should the writer of the text change first?	

Figure 4: Questions with answers for letter one: WMO Facilities (translated).

Vraag	Juiste antwoord
Waarvoor wordt een pgb gebruikt?	Het regelen van ondersteuning of hulp
Klopt het dat bij een pgb je zelf de zorgverlener kiest?	Ja
Klopt het dat bij een pgb je zelf bepaalt wanneer je hulp krijgt?	Ja
Klopt het dat bij een pgb je zelf bepaalt hoe je hulp krijgt?	Ja
Heeft de gemeente heeft met alle (zorg)aanbieders een contract afgesloten?	Nee
Kun je een pgb gebruiken voor zorgaanbieders waarmee de gemeente geen contract mee h	Ja
Wat koop je met een pgb zelf in?	Jeugdzorg/ondersteuning/hulpmiddelen
Wat betekend de regie voeren in deze brief?	Zie opsomming
Wanneer mag je geen pgb gebruiken?	Zie opsomming
Wat moet er in het uitvoeringsplan staan?	Zie opsomming
Is een zorgovereenkomst hetzelfde als een uitvoeringsplan?	Nee
Wat moet er in de zorgovereenkomst staan?	Afspraken met de zorgverlener
Wie houdt in de gaten wanneer het pgb stopt?	Uzelf/de zorgvrager
Wat moet je doen als het pgb stopt en je nog zorg nodig hebt?	Nieuw gesprek bij het wijkteam aanvragen
Hoever voor het stoppen van het pgb moet je dit (gesprek wijkteam) aanvragen?	8 weken voor de einddatum van het pgb
Waarop controleert het wijkteam?	1. Onjuist gebruik pgb's
	2. Fraude
Hoe vraag je een pgb aan?	Volgen van het stappenplan/brochure/folder
Wat voor vragen kun je stellen aan de Sociale Verzekeringsbank?	Zie opsomming
Hoe en aan wie moet je vragen stellen over de inhoud van de zorgovereenkomst?	Gemeente, telefonisch op 14033
Waarover kan de contactpersoon bij het wijkteam meer informatie geven?	pgb-bedragen, toekenning pgb, rekeninstrument, hulpvraag
Hoe vraag je een pgb aan voor meerdere gezinsleden?	Via 1 plan
Wat zijn eisen van het plan voor een pgb voor meerdere gezinsleden?	1. Zorgbehoefte alle gezinsleden (die hulp nodig hebben) overzichtelijk
	2. Verband zorgbehoefte alle gezinsleden (die hulp nodig hebben) duidelijk
Wat is een alternatief van een pgb?	Zorg in natura (zin)
Wat is het verschil tussen een pgb en zin?	Pgb heeft andere zorgaanbieders dan zin
Wat is de eigenbijdrage voor een pgb?	19 Euro
Wat kun je doen als je hulp nodig hebt en je weet niet bij wie je moet zijn?	1. Bellen gemeente 14033
	2. Mailen gemeente info@zilverdam.nl
Waar moet je naartoe als je persoonlijk iemand wilt spreken?	Informatiewinkel
Hoe vind je een informatiewinkel in de buurt?	Indebuurt033
Met wat voor vragen helpt de gemeente u niet?	Specifieke vragen over de gezondheid
time:	
Hoe duidelijk vind je de brief in het algemeen? (1-10)	
Vind je dat er overbodige informatie in de brief staat? Zo ja, waar?	
Wat vind je van de toon van de tekst? (bijvoorbeeld streng of vriendelijk)	
Hoe erg gepast vind je de toon van de brief in het algemeen? (1-10)	
Wat zou de schrijver van de tekst als eerste moeten veranderen?	

Figure 5: Questions with answers for letter two: Regels PGB (Dutch).

Question	Correct answer
What is a personal budget (PGB) used for?	Managing support or assistance
Is it true that with a PGB, you can choose your own care provider?	Yes
Is it true that with a PGB, you decide when you receive assistance?	Yes
Is it true that with a PGB, you decide how you receive assistance?	Yes
Has the municipality entered into contracts with all (care) providers?	No
Can you use a PGB for care providers with whom the municipality has no contract?	Yes
What do you purchase with a PGB?	Youth care/support/aids
What does "taking control" mean in this letter?	See list
When are you not allowed to use a PGB?	See list
What should be included in the implementation plan?	See list
Is a care agreement the same as an implementation plan?	No
What should be included in the care agreement?	Agreements with the care provider
Who monitors when the PGB stops?	Yourself/the care recipient
What should you do if the PGB stops, and you still need care?	Request a new conversation with the neighborhood team
How long before the PGB expiration date should you request this (neighborhood team conversation)?	8 weeks before the PGB's end date
What does the neighborhood team check?	1. Improper use of PGBs
	2. Fraud
How do you apply for a PGB?	Follow the step-by-step guide/brochure/folder
What kind of questions can you ask the Social Insurance Bank?	See list
How and to whom should you ask questions about the content of the care agreement?	Municipality, by phone at 14033
What additional information can the contact person at the neighborhood team provide?	PGB amounts, PGB allocation, calculation tool, assistance request
How do you apply for a PGB for multiple family members?	Through one plan
What are the plan requirements for a PGB for multiple family members?	1. Clear overview of the care needs of all family members (needing assistance)
	2. Clear connection between the care needs of all family members (needing assistance)
What is an alternative to a PGB?	Care in kind (ZIN)
What is the difference between a PGB and ZIN?	PGB has different care providers than ZIN
What is the own contribution for a PGB?	19 Euros
What can you do if you need assistance and don't know who to contact?	1. Call the municipality at 14033
	2. Email the municipality at info@zilverdam.nl
Where should you go if you want to speak with someone in person?	Information desk
How can you find an information desk in the area?	Indebuurt033
What kind of questions does the municipality not help with?	Specific health-related questions
time:	
How clear do you find the letter overall? (1-10)	
Do you think there is any unnecessary information in the letter? If yes, where?	
What do you think of the tone of the text? (e.g., strict or friendly)	
How appropriate do you find the tone of the letter overall? (1-10)	
What should the writer of the text change first?	

Figure 6: Questions with answers for letter two: PGB Regulations (translated).

Vraag	Juiste antwoord
Wat is er op 13-09-2022 gebeurd?	Mevrouw (Brons) gemeld bij de gemeente met een hulpvraag
Wat is er op 18-09-2022 gebeurd?	Huisbezoek bij mevrouw (Brons)
Wat is de hulpvraag?	Een traplift
Wat is er besloten?	1. Mevrouw (Brons) krijgt een persoonsgebonden budgt (pgb)
	2. Mevrouw (Brons) krijgt een onderhoudsbedrag voor de traplift
Wat is een voorbeeld van een wijziging volgens de tekst?	Verzuizen/samenwonen/afname/toename beperking
Wat moet je doen bij een verandering van de persoonlijke situatie?	Doorgeven aan de gemeente
Hoe moet je een verandering doorgeven aan de gemeente?	Bellen naar 012-3456789 en vragen naar team Zorg
Wanneer moet je bezwaar maken?	Als je het niet eens bent met het besluit
Wat moet er in het bezwaar staan?	Zie opsomming
Tot wanneer kun je bezwaar indienen?	Tot 6 weken na de dag waarop het besluit is verzonden
Hoe maak je bezwaar?	1. Brief sturen naar adres van de gemeente
	2. Via www.zilverdam.nl -> loketten
Kun je bezwaar maken zonder een DigiD?	Ja
Wat is clientenondersteuning?	Meer hulp via een maatschappelijk werker
Hoeveel kost de clientenondersteuning?	Niks/Gratis
Hoe kun je clientenondersteuning aanvragen?	1. Bellen naar 012-3456789
	2. Mailen naar info@indebuurt.nl
	3. Binnenlopen informatiewinkel in de buurt
Wat moet je doen bij andere vragen/opmerkingen?	Contact opnemen Mevrouw Oudklomp
Kun je mevrouw Oudklomp op vrijdag bellen?	Nee
time:	
Hoe duidelijk vind je de brief in het algemeen? (1-10)	
Vind je dat er overbodige informatie in de brief staat? Zo ja, waar?	
Wat vind je van de toon van de tekst? (bijvoorbeeld streng of vriendelijk)	
Hoe erg gepast vind je de toon van de brief in het algemeen? (1-10)	
Wat zou de schrijver van de tekst als eerste moeten veranderen?	

Figure 7: Questions with answers for letter three: Besluit PGB (Dutch).

Question	Correct answer
What happened on 13-09-2022?	Mrs. (Brons) reported to the municipality with a request for assistance
What happened on 18-09-2022?	Home visit to Mrs. (Brons)
What is the request for assistance?	A stairlift
What was decided?	1. Mrs. (Brons) will receive a personal budget (PGB)
	2. Mrs. (Brons) will receive a maintenance amount for the stairlift
What is an example of a change according to the text?	Moving/cohabiting/decrease/increase in limitations
What should you do in case of a change in your personal situation?	Report it to the municipality
How should you report a change to the municipality?	Call 012-3456789 and ask for the Care team
When should you file an objection?	If you disagree with the decision
What should be included in the objection?	See list
Until when can you file an objection?	Up to 6 weeks after the day the decision was sent
How do you file an objection?	1. Send a letter to the municipality's address
	2. Via www.zilverdam.nl -> service counters
Can you file an objection without a DigiD?	Yes
What is client support?	Additional assistance through a social worker
How much does client support cost?	Nothing/Free
How can you request client support?	1. Call 012-3456789
	2. Email info@indebuurt.nl
	3. Visit an information desk in the neighborhood
What should you do for other questions/comments?	Contact Mrs. Oudklomp
Can you call Mrs. Oudklomp on Fridays?	No
time:	
How clear do you find the letter overall? (1-10)	
Do you think there is any unnecessary information in the letter? If yes, where?	
What do you think of the tone of the text? (e.g., strict or friendly)	
How appropriate do you find the tone of the letter overall? (1-10)	
What should the writer of the text change first?	

Figure 8: Questions with answers for letter three: PGB Decision (translated).

E. Results automatic evaluation metrics

Finance

Betalen_in_delen			ChatGPT	Naive	BERT	Original
LiNT			38	31	23	30
Rouge_1	rouge_1	recal	0.58839	0.94204	0.85232	-
	rouge_2	recal	0.34656	0.9018	0.73573	-
	rouge_L	recal	0.44591	0.94204	0.85232	-
BLEURT			0.60454	0.76511	0.37346	-
BLEU	blue_1	score	0.35816	0.87371	0.79755	-
		precision	0.8	0.9234	0.83544	-
	blue_2	score	0.27154	0.83336	0.71883	-
		precision	0.45985	0.84009	0.67865	-
	blue_3	score	0.21218	0.79659	0.64936	-
		precision	0.28938	0.76923	0.55508	-
	blue_4	score	0.16575	0.76533	0.59179	-
		precision	0.17647	0.71734	0.46921	-

Gemeentelijke_belastingen			ChatGPT	Naive	BERT	Original
LiNT			36	42	27	40
Rouge_1	rouge_1	recal	0.57904	0.89091	0.76554	-
	rouge_2	recal	0.30203	0.81092	0.60113	-
	rouge_L	recal	0.37684	0.88112	0.75424	-
BLEURT			0.45893	0.68226	0.3884	-
BLEU	blue_1	score	0.32691	0.87615	0.74517	-
		precision	0.8445	0.88595	0.76879	-
	blue_2	score	0.23403	0.8301	0.65409	-
		precision	0.4328	0.79526	0.59233	-
	blue_3	score	0.17002	0.78757	0.57542	-
		precision	0.23181	0.71688	0.45946	-
	blue_4	score	0.13084	0.74721	0.5059	-
		precision	0.15405	0.64525	0.3547	-

Kwijtschelding			ChatGPT	Naive	BERT	Original
LiNT			47	42	26	38
Rouge_1	rouge_1	recal	0.7171	0.90206	0.79377	-
	rouge_2	recal	0.50076	0.83204	0.62419	-
	rouge_L	recal	0.60817	0.90206	0.78859	-
BLEURT			0.74787	0.73536	0.35939	-
BLEU	blue_1	score	0.5406	0.87316	0.76936	-
		precision	0.8764	0.90576	0.80688	-
	blue_2	score	0.44505	0.83157	0.67622	-
		precision	0.59398	0.82152	0.62334	-
	blue_3	score	0.37568	0.79089	0.59755	-
		precision	0.43396	0.74211	0.48936	-
	blue_4	score	0.32404	0.75189	0.53231	-
		precision	0.33712	0.67018	0.39467	-

Care

Wmo_voorzieningen			ChatGPT	Naive	BERT	Original
LiNT			47	45	33	43
Rouge_1	rouge_1	recal	0.61753	0.90183	0.80489	-
	rouge_2	recal	0.39362	0.83122	0.65976	-
	rouge_L	recal	0.43692	0.89995	0.80019	-
BLEURT			0.62513	0.68564	0.52028	-
BLEU	blue_1	score	0.38767	0.86475	0.77323	-
		precision	0.85267	0.90684	0.81007	-
	blue_2	score	0.3055	0.82143	0.69577	-
		precision	0.52952	0.81827	0.65589	-
	blue_3	score	0.24258	0.78087	0.62825	-
		precision	0.33641	0.74	0.53663	-
	blue_4	score	0.19843	0.74165	0.56777	-
		precision	0.23889	0.66635	0.43905	-

Regels_pgb			ChatGPT	Naive	BERT	Original
LiNT			43	44	32	47
Rouge_1	rouge_1	recal	0.75641	0.90377	0.75617	-
	rouge_2	recal	0.51204	0.83886	0.57032	-
	rouge_L	recal	0.59615	0.90223	0.75	-
BLEURT			0.66788	0.77866	0.29067	-
BLEU	blue_1	score	0.67914	0.85129	0.7146	-
		precision	0.78803	0.90236	0.75748	-
	blue_2	score	0.55015	0.80455	0.61266	-
		precision	0.51712	0.80599	0.55678	-
	blue_3	score	0.45461	0.75998	0.5259	-
		precision	0.36021	0.7188	0.41074	-
	blue_4	score	0.38566	0.71817	0.45419	-
		precision	0.2732	0.64241	0.31013	-

Besluit_pgb			ChatGPT	Naive	BERT	Original
LiNT			48	43	30	40
Rouge_1	rouge_1	recal	0.60591	0.90746	0.80238	-
	rouge_2	recal	0.42469	0.84347	0.65672	-
	rouge_L	recal	0.52463	0.90547	0.79841	-
BLEURT			0.66899	0.63879	0.3776	-
BLEU	blue_1	score	0.3972	0.87275	0.77832	-
		precision	0.8255	0.92229	0.81377	-
	blue_2	score	0.33269	0.834	0.70161	-
		precision	0.57912	0.84221	0.66126	-
	blue_3	score	0.28377	0.798	0.62895	-
		precision	0.42905	0.77207	0.52846	-
	blue_4	score	0.24584	0.76326	0.56211	-
		precision	0.3322	0.70576	0.41955	-

Figure 9: Automatic evaluation metric results of the simplification models.

F. Table with descriptive statistics of variables from the randomized controlled trial

variable	mean	sd	median	min	max	range	skew	kurtosis	se
group	4.5	2.307367258	4.5	1	8	7	0	-1.286697163	0.271925839
age	41.93055556	17.57104403	43	18	83	65	0.371331742	-1.072921957	2.070767398
gender	0.5	0.503508815	0.5	0	1	1	0	-2.027584877	0.059339083
language	0.902777778	0.298339169	1	0	1	1	-2.66263155	5.161875509	0.035159608
edu	2.513888889	0.787097638	3	1	3	2	-1.155913305	-0.405255457	0.092760346
reading_work	3.708333333	2.497533995	4	0	8	8	-0.056788253	-1.28511434	0.294337204
reading_spare	2.152777778	1.61122578	2	0	12	12	3.174404923	16.90846477	0.189884779
disability	0.180555556	0.38734884	0	0	1	1	1.626480752	0.655114473	0.045649499
letters	13.51388889	10.268224	10	2	50	48	1.437533288	1.521199667	1.210121803
grade_clarity	6.736111111	1.861341751	7	2	10	8	-0.61367629	-0.469347613	0.219361229
grade_tone	6.916666667	1.535954078	7	3	10	7	-0.414277684	-0.172643999	0.181013924
l1_g	0.5	0.503508815	0.5	0	1	1	0	-2.027584877	0.059339083
l2_g	0.5	0.503508815	0.5	0	1	1	0	-2.027584877	0.059339083
l3_g	0.5	0.503508815	0.5	0	1	1	0	-2.027584877	0.059339083
l1_action_good	82.06018519	17.83793556	83.33333333	41.66666667	100	58.33333333	-0.636727324	-0.820435435	2.102220866
l1_understanding_good	78.62103175	15.64798818	82.14285714	35.71428571	96.42857143	60.71428572	-0.696801138	-0.452457676	1.844133093
l1_tone	7.013888889	1.605387545	7	2	10	8	-0.465337643	0.10820614	0.189196737
l1_u_grade	5.972222222	1.887484519	6.75	2	8	6	-0.605014758	-0.974974131	0.222442184
t1	16.77777778	6.426665887	17	6	35	29	0.642435057	0.325082181	0.757389838
l2_action_good	3.930555556	2.844943074	2.5	1	8	7	0.221212221	-1.720206341	0.335279757
l2_understanding_good	12.08333333	5.343813061	13.5	1	18	17	-0.420209405	-1.23082262	0.629774409
l2_tone	6.652777778	1.548827109	7	1	9	8	-1.033864164	1.592069942	0.182531025
l2_u_grade	5.854166667	1.372356614	6	2	8	6	-0.53682276	0.296717762	0.161733778
t2	14.94444444	4.515665655	15	6	30	24	0.167020914	0.465562666	0.532176301
l3_action_good	3.5	2.455232986	4	1	7	6	0.160466908	-1.707572903	0.289351982
l3_understanding_good	8.763888889	3.151092167	10	1	12	11	-0.531208236	-0.965990445	0.371359773
l3_tone	6.923611111	1.66923218	7	1	10	9	-0.849591298	1.093499008	0.196720899
l3_u_grade	6.909722222	1.655818569	7	1	10	9	-0.93570467	1.211396159	0.19514009
t3	8.819444444	4.193645054	8	3	20	17	0.958874797	0.306674783	0.494225809
t_totaal	40.54166667	11.34595231	42.5	22	70	48	0.295444261	-0.442937443	1.337133303
average_g_tone	6.863425926	1.44819957	7	1.333333333	9.333333333	8	-0.808477719	1.465542263	0.170671956
average_g_understanding	6.24537037	1.310536398	6.5	1.666666667	8.333333333	6.666666666	-0.833036412	0.709394273	0.154448196
t_action_good	85.87655644	8.915723494	9.413333335	67.45888889	100	32.54111111	-0.284391643	-0.72255892	1.05072809
t_understanding_good	79.95641992	8.514992439	8.080876007	57.87545788	93.95604396	36.08058608	-0.60070304	0.157339839	1.003501483

Table 1: Descriptive statistics of the variables from the randomized controlled trial.

G. Results (Generalised) Linear Models analyses

	(g)lm_letter_1		(g)lm_letter_2		(g)lm_letter_3		(g)lm_totaal	
	l1_action_good	l1_understaning_good	l2_action_good	l2_understanding_good	l3_action_good	l3_understanding_good	t_action_good	t_understanding_good
l1_g	**9.26** ***	**10.94** ***	-	-	-	-	**4.96** ***	**7.36** ***
l2_g	-	-	**6.12** ***	**11.80** ***	-	-	**3.82** ***	**4.86** ***
l3_g	-	-	-	-	**5.58** ***	**7.81** ***	**2.46** *	**3.40** **
age	**-2.96** **	-0.71	**-2.17** *	-1.34	-1.19	-0.51	**-3.77** ***	-1.50
gender	-0.46	0.74	-0.11	0.43	-0.04	-0.31	-0.33	0.48
language	-0.1	0.70	**3.60** ***	-0.53	0.36	1.11	**2.06** *	0.69
edu	-0.25	1.30	1.09	-0.34	-0.06	**-1.80** .	0.45	-0.35
reading_work	-0.73	0.41	-0.13	-0.15	-0.79	0.69	-0.97	0.13
reading_spare	**2.03** *	**2.20** *	-0.60	-0.51	1.02	-0.43	1.60	0.74
disability	-0.86	0.41	0.41	-0.03	**2.00** *	1.00	0.90	0.44
letters	0.13	0.51	1.61	**2.37** *	0.79	-0.20	1.26	1.30
grade_clarity	**2.50** *	**1.98** .	1.40	**2.53** *	1.49	0.48	**3.23** ***	**2.73** **
grade_tone	**-2.34** *	-1.62	-0.15	-0.62	-0.15	0.59	-1.63	-1.14
N	72	72	72	72	72	72	226	226
AIC (glm)	573.35	536.86	578.90	526.83	577.33	520.66	493.25	468.91
Multiple R-squared (lm)	0.63	0.71	0.51	0.74	0.42	0.58	0.53	0.63
Adjusted R-squared (lm)	0.56	0.65	0.42	0.69	0.32	0.50	0.43	0.55
F-statistic (lm)	9.15	13.19	5.70	15.66	4.00	7.43	5.13	7.81

Figure 10: Results of both the Generalised Linear Models analyses and Linear Models with significant values in **bold**. The codes for significance are: **.** p<0.10, * p<0.05, ** p<0.01 and ***p<0.001. *l1_g,l2_g, l3_g* present the dummy variables for the three letters respectively. The model descriptives are defined per model on the bottom lines of the table.

H. Results MANOVA analyses

	manova_letter_1	manova_letter_2	manova_letter_3	manova_totaal
l1_g	**0.75** ***	-	-	**0.53** ***
l2_g	-	**0.76** ***	-	**0.38** ***
l3_g	-	-	**0.68** ***	**0.28** **
age	**0.32** ***	**0.26** **	**0.31** ***	**0.40** ***
gender	0.03	0.08	0.05	0.09
language	0.11	**0.27** **	0.09	**0.20** *
edu	0.10	**0.21** *	0.16	**0.16** .
reading_work	0.12	0.03	0.10	0.05
reading_spare	0.11	0.07	0.09	0.09
disability	0.08	0.02	0.11	0.03
letters	0.04	0.09	0.02	0.05
grade_clarity	0.04	**0.17** .	0.08	**0.16** .
grade_tone	**0.21** *	**0.17** .	0.09	**0.22** *
N	72	72	72	226

Figure 11: Results MANOVA analyses: Pillai's trace values with significant values in **bold**. The codes for significance are: **.** p<0.10, * p<0.05, ** p<0.01 and ***p<0.001. *l1_g,l2_g, l3_g* present the dummy variables for the three letters respectively.

Legal Science and Computer Science: A Preliminary Discussion on How to Represent the "Penumbra" Cone with AI

Angela Condello[1], Giorgio Maria Di Nunzio[2]
[1] Department of Law, University of Messina, Italy
[2] Department of Information Engineering, University of Padua, Italy
[1]Piazza Pugliatti, 1 - 98122 Messina - Italy
[2]Via Gradenigo, 6/a - 35131 Padova - Italy
angela.condello@unime.it, giorgiomaria.dinunzio@unipd.it

Abstract

Legal science encounters significant challenges with the widespread integration of AI software across various legal operations. The distinction between signs, senses, and references from a linguistic point of view, as drawn by Gottlob Frege at the end of the 19th century, underscores the complexity of legal language, especially in multilingual contexts like the European Union. In this paper, we describe the problems of legal terminology, examining the "penumbra" problem through Herbert Hart's legal theory of meaning. We also analyze the feasibility of training automatic systems to handle conflicts between different interpretations of legal norms, particularly in multilingual legal systems. By examining the transformative impact of Artificial Intelligence on traditional legal practices, this research contributes to the theoretical discussion about the exploration of innovative methodologies for simplifying complex terminologies without compromising meaning.

Keywords: Legal terminology, Linguistic Sign, Terminology

1. Background

In this paper, we explore the multifaceted challenges facing legal science considering the widespread adoption of AI software across various legal operations, such as verification, drafting, risk analysis, and prediction,[1] with specific reference to the potential confusion between signs, senses, and their reference. Such distinction, as it is widely known, was drawn by Gottlob Frege in a renowned paper published in 1892 (Frege, 1892). Frege thereby defines it in such a kind that to the sign there corresponds a definite sense and to that - in turn - a definite reference, while to a given reference (an object) there does not belong only one single sign. From this perspective, the same sense (e.g., equality) can have different expressions in different languages and realms, and even in the same language.

Within our interdisciplinary context, between legal philosophy and computer engineering, we aim at narrowing down onto a pivotal issue: the evolving dynamics of legal terminology due to the pervasive and ever-increasing use of AI software by legal professionals, including lawyers, judges, and notaries (Rissland et al., 2003), with specific reference to the potential confusion between signs, senses and references caused by such use of AI software for legal professionals. How can we prevent the blurring of this fundamental differentiation in philosophy of language, a differentiation that is extremely delicate in legal science?

Central to our investigation is Herbert Hart's theoretical framework (Hart and Leslie, 2012), which posits that legal concepts, mediated through the terms that indicate them, exhibit a dual nature. While, in theory, they possess a core of settled meaning, they are also surrounded by a "penumbra" of debatable cases, known as "hard cases" (Dworkin, 1975), wherein the application of words is neither evidently applicable nor categorically ruled out. As (Rissland et al., 2003) explain, legal rules derive their dynamic nature in part through the dynamic, open-textured nature of the terms used in the rules. As new situations arise, interpretation of the meaning of these terms changes as well. Such background is complexified in realms like the European Union, where translation of legal concepts is per se a very problematic issue both for lawyers and linguists.

Against this background, the integration of AI software in legal practice raises critical questions that we want to explore; in particular, whether it is conceivable to anticipate the potential emergence of hard cases and, subsequently, prepare legal software to navigate the intricate core-penumbra problem inherent in legal meaning. In addition, the increasing sophistication of these technologies and their availability have generated two divergent narratives about their potential implications, as described by (Whalen, 2022). These narratives alternately express excitement

[1] https://joinup.ec.europa.eu/collection/justice-law-and-security/solution/leos-open-source-software-editing-legislation/discussion/smart-leos-which-new-functionalities-should-be-implemented-next-and-what-can-be-learnt-corrigenda

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 179–184

about legal technology's potential to make the law more efficient and improve access to justice, or concern about the ways in which it may exacerbate existing biases or otherwise systematically harm justice.

Our research extends beyond the theoretical questions and addresses practical considerations tied to the intersection of AI and legal semantics. In this context, one main issue arises: can automatic systems be trained to foresee the contours of hard cases and adapt to the subtle distinctions of legal meaning? Can we measure the uncertainty of legal concepts and argumentation to handle the conflicts between different interpretations of norms (da Costa Pereira et al. 2017)? Can we foresee, by working with interdisciplinary methodology, the potential confusion between words and the concepts they refer to, especially in multilingual legal systems? Could it be too risky to use AI systems also in multilingual realms?

This challenge involves understanding how AI systems can effectively discern the relevance of contextual complications as well as societal changes, a task that is very important in the context of Hart's legal theory of meaning.

As we examine the implications of AI software on legal terminology, our analysis recognizes the transformative impact on traditional legal practices. The diffusion of AI technologies introduces a paradigm shift, necessitating a reevaluation of established legal methodologies. We explore the potential repercussions of this shift on the interpretation of legal documents and the inherent stability (or instability) of legal concepts. The balance between settled meanings and the penumbra of hard cases becomes increasingly important in a field where AI contributes to legal decision-making processes. By examining the core-penumbra problem through the lens of Hart's legal theory of meaning, we shed light on the challenges and opportunities posed by the integration of AI in legal science. Through this interdisciplinary analysis, we contribute to the ongoing discourse on the evolving nature of legal semantics in an era marked by the influence of augmented intelligence.

2. State of the Art: Penumbra and Simplification of Legal Texts

In this section, we present an overview of the state-of-the-art of the recent research papers that have been dealing with the issue of penumbra and text simplification in legal texts. It is important to highlight also a recent comprehensive systematic review in legal natural language processing (Quevedo et al., 2023) that complements this overview from an NLP point of view. In our analysis, we have searched Google Scholar to create a sort of systematic review of the topic with two queries: 1) "legal text" and "penumbra", 2) "legal text" and "text simplification"; then, we have kept only research papers that were published in the last two years in conferences or journals related to computer science/engineering or to interdisciplinary fields. We also kept the most recent articles that have been made available through the arXiv platform.

In (Stathis et al., 2024), the authors introduce Intelligent Contracts (iContracts), a new field blending AI and law, facing challenges like data quality. The focus is on Proactive Control Data (PCD) to enhance iContracts, a novel area in research. The extent to which Proactive Data impacts an Intelligent Contract depends on their quantitative identification and qualitative assessment, as well as the use of relevant technologies that integrate the risk assessment data when a contract is generated. Results include successful PCD generation, significant impact on contract drafting, and methods for assessing PCD quality.

The work presented by (Jiang et al., 2024) discusses leveraging large language models (LLMs) to enhance legal education for non-experts by employing storytelling. Since Law is, by nature, a sensitive domain, and computational tools must be designed responsibly, it is critical to design comprehension tools in ways that do not oversimplify or overgeneralize the nuances of legal jargon. In this context, the authors introduce a new dataset called LEGALSTORIES, comprising complex legal doctrines explained through stories and multiple-choice questions generated by LLMs. The idea is that storytelling aids in relating legal concepts to personal experiences and exhibits higher retention rates among non-native speakers.

In (Engel and McAdams, 2024), the authors test Large Language Models (LLM), ChatGPT in particular, to generate evidence on the ordinary meaning of statutory terms taking into account that many terms qualify as penumbral, and the legislative context often has some influence. The authors emphasize the importance of considering a distribution of replies rather than solely relying on the "best" reply identified by ChatGPT. Using Chat 3.5, the setting of these experiments defines prompts and refine them given contextual factors and historical periods. These experiments represent the first attempt to use GPT for empirical data on statutory term meanings, indicating potential for improving legal interpretation despite the need for caution.

The study presented in (Dixit et al., 2024) explores the effectiveness of extractive text summarization for condensing legal documents while retaining crucial aspects. In particular, the proposed approach of decreasing any lexical or

syntactic intricacy related to the text without modifying the substance of the text is carried out. It is the pre-processing stage that finally results in the selection of a useful sentence. This work evaluates different classification models using the ROUGE scores. Extractive summarization selects relevant content chunks, ensuring well-structured summaries with all legal elements intact. These methods are favored in legal documents for their preservation of original content. The study advocates for comprehensive legal summaries covering all aspects.

In (Westermann et al., 2023), the author addresses the challenge laypeople face in understanding legal opportunities and remedies due to difficulty in assessing legal issues from factual descriptions. Understanding which legal opportunities or remedies are available to laypeople requires an analysis of which legal issues are raised by these facts, which may be difficult for laypeople to assess. This gap can cause laypeople to miss out on benefits or be unable to resolve their disputes. This research proposes an automated approach to analyze layperson-provided descriptions and map them to relevant legal issues, enhancing access to justice. The findings offer insights for legal professionals and developers to bridge the gap between layperson language and legal issues, potentially improving access to justice through legal decision support systems.

The study presented by (Kiliroor et al., 2023) addresses the challenge of understanding lengthy and complex legal documents, highlighting the importance of accessibility for impartiality. It proposes a text simplification method tailored to the legal domain, aiming to make legal text more comprehensible. The model identifies complex words and substitutes them with simpler alternatives using a word embedding model and sentiment analysis model. Trained on a dataset combining Indian Legal Documents Corpus, the approach successfully detects and replaces complex words with simpler ones, maintaining the original sentiment. This method has the potential to enhance accessibility to legal texts, saving time for individuals navigating legal documents while promoting impartiality.

In (Billi et al., 2023), the authors promote the integration of Large Language Models into rule-based legal systems to enhance accessibility, usability, and transparency, aligning with democratic principles in legal technology. This paper introduces a methodology to translate rule-based system explanations into natural language, enabling clearer and faster interactions for all users. Additionally, it empowers laypeople to perform complex legal tasks independently through a chain of prompts, facilitating autonomous legal comparisons. This approach aims to democratize legal technology, making it more inclusive and comprehensible for users, while also promoting transparency and stakeholder involvement in the legal decision-making process.

3. Representing “Penumbra” in Machine Learning

In this section, we briefly list some possibilities to represent the concept of penumbra from a machine learning point of view and we try to clarify how such representation becomes more complex in multi-linguistic contexts such as the European Union’s courts (ECJ, ECHR). Particularly, in the context of natural language processing (NLP) and decision-making algorithms, the penumbra can be linked to areas of uncertainty or ambiguity where the model's predictions may not be unequivocal.

3.1 Uncertainty in Model Predictions:

In machine learning models, especially those based on probabilistic frameworks like Bayesian models, predictions are often associated with a degree of uncertainty (Neil et al., 2019). The model may provide a probability distribution over possible outcomes rather than a definitive answer. This uncertainty may reflect the penumbral aspect, where certain instances may fall into a gray area, making it challenging for the model to make a clear-cut decision.

3.2 Boundary Cases:

Much like legal penumbra involving hard cases, machine learning models may struggle with boundary cases (Atkinson and Bench-Capon, 2019). These are instances that lie on the edge of the decision boundary, where small changes in input features can lead to different predictions. These boundary cases represent situations where the model's confidence is lower, and decisions may be less straightforward.

3.3 Context Sensitivity

In legal terms, the interpretation of a term may vary based on the context in which it is used. Similarly, machine learning models, specifically NLP models, often rely on context to make accurate predictions (Sosa Andrés, 2023). The model's understanding of certain terms or features may exhibit variability based on the surrounding context, introducing a level of interpretation flexibility analogous to the legal penumbra.

3.4 Language Sensitivity

Language is also a big part of the analysis and interpretation of legal terminology (Kalinina and Kudryashova, 2022). The linguistic nature of the term, the specific characteristic of a legal concept, the discrepancies between the state legal systems, the socio-cultural content of legal terms in different languages are only a few examples of the issues concerned with language. Therefore, the combination of language and legal

knowledge, as well as culture understanding, is necessary in understanding the content and translating it into another language functionally and in accordance with the target group.

3.5 Interpretable Machine Learning:

Interpretable machine learning models aim to provide transparency into how decisions are made (Farayola et al., 2023). Despite efforts to achieve interpretability, there may still be instances where the model's reasoning is not entirely clear. This lack of clarity aligns with the penumbral nature, where certain cases may defy straightforward interpretation.

In essence, the concept of penumbra in legal science, with its shades of interpretation ambiguity, can find in machine learning models dealing with uncertainty, boundary cases, context sensitivity, and adaptability.

4. Conclusions and Future Perspectives

In this paper, we started a discussion about research efforts that can explore the intersection of machine learning and legal theory to develop novel approaches for representing the penumbra in legal texts. Drawing upon theoretical frameworks such as Hart's legal theory of meaning, Machine Learning researchers can develop computational models that capture the dynamic nature of legal concepts and their surrounding penumbra. By integrating legal theory into machine-learning algorithms, researchers can create more sophisticated representations of legal ambiguity and uncertainty, facilitating more accurate and contextually appropriate legal interpretations.

We highlighted some ideas of possible research into representing the penumbra concept in legal texts through machine learning which may hold significant promise for advancing the understanding and application of legal semantics in AI-driven legal systems. One important aspect of this investigation lies in refining machine learning models to effectively capture the semantic differences of legal ambiguity inherent in the penumbra. This entails developing algorithms capable of identifying the boundaries of uncertainty within legal texts in order to distinguishing between settled meanings and hard cases, as proposed by Hart. By incorporating probabilistic frameworks such as Bayesian models, researchers could explore how uncertainty in model predictions reflects the penumbral aspect of legal interpretation, providing a more nuanced understanding of complex legal concepts.

Furthermore, research could focus on enhancing machine learning models' sensitivity to contextual variations in legal texts. Just as legal interpretations may vary depending on the context in which terms are used, NLP models must be trained to recognize and adapt to fine-grained contexts within legal documents. Techniques such as contextual embeddings and attention mechanisms can help capture the subtle shifts in meaning that occur within different legal contexts, thereby improving the models' ability to navigate the penumbra of legal interpretation.

Additionally, investigating the impact of language sensitivity on machine learning representations of legal texts is essential, especially in multilingual legal systems like those found in the European Union. Understanding how linguistic differences influence legal interpretation can inform the development of more robust machine learning models capable of handling diverse linguistic and cultural nuances. More specifically, we aim to link the problem of the penumbra in other interdisciplinary areas, such as Digital Humanities. On one hand, the description of the uncertainty of concepts can be used to store and index automatically non-catalogued and unprocessed material, which has to be, not only preserved, but also described, shared and made accessible (Grbac, 2021). On the other hand, the methodology to represent uncertainty can be included in machine translation systems where we have difficulties in the translation processes required in a multilingual and multicultural environment such as that of international cooperation (Vezzani et al., 2022). Finally, the same methodology can be included in other experimental research that deals with semantic phenomena involved in the process of determining a conceptual expression such as synonymy, polysemy, and elliptical segments (Pulizzotto et al., 2018).

In conclusion, research into representing the penumbra concept in legal texts through machine learning offers a rich and multifaceted landscape of opportunities. By refining machine learning models' ability to capture legal ambiguity, sensitivity to contextual variations, and interpretability, researchers can develop more robust and trustworthy AI-driven legal systems. Moreover, integrating legal theory into machine-learning algorithms and exploring innovative learning techniques can further advance our understanding of legal semantics and pave the way for more effective and equitable legal decision-making processes.

5. Bibliographical References

Atkinson, Katie, and Bench-Capon, Trevor (2019). Reasoning with Legal Cases: Analogy or Rule Application? In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, 12–21. ICAIL '19. New York, NY, USA: Association for Computing

Machinery. https://doi.org/10.1145/3322640.3326695.
Billi, Marco, Parenti, Alessandro, Pisano, Giuseppe, and Sanchi, Marco (2023). Large Language Models and Explainable Law: A Hybrid Methodology. arXiv.Org. November 20, 2023. https://arxiv.org/abs/2311.11811v1.
da Costa Pereira, Célia, Tettamanzi, Andrea G. B., Liao, Beishui, Malerba, Alessandra, Rotolo, Antonino, and van der Torre, Leendert (2017). Combining Fuzzy Logic and Formal Argumentation for Legal Interpretation. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, 49–58. ICAIL '17. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3086512.3086532.
Dixit, Utkarsh, Gupta, Sonam, Yadav, Arun Kumar, and Yadav, Divakar (2024). Analyzing the Impact of Extractive Summarization Techniques on Legal Text. In *Proceedings of Data Analytics and Management*, edited by Abhishek Swaroop, Zdzislaw Polkowski, Sérgio Duarte Correia, and Bal Virdee, 585–602. Lecture Notes in Networks and Systems. Singapore: Springer Nature. https://doi.org/10.1007/978-981-99-6544-1_44.
Dworkin, Ronald (1975). Hard Cases. *Harvard Law Review*, vol. 88, n. 6, pp. 1057-1109 https://doi.org/10.2307/1340249
Engel, Christoph and McAdams, Richard H. (2024) Asking GPT for the Ordinary Meaning of Statutory Terms. *MPI Collective Goods Discussion Paper*, No. 2024/5, U of Chicago, Public Law Working Paper No. 848, http://dx.doi.org/10.2139/ssrn.4718347
Farayola, Michael Mayowa, Irina Tal, Regina Connolly, Takfarinas Saber, and Malika Bendechache (2023). Ethics and Trustworthiness of AI for Predicting the Risk of Recidivism: A Systematic Literature Review. *Information* 14 (8): 426. https://doi.org/10.3390/info14080426.
Frege, Gottlob (1892). On Sense and Reference ["Über Sinn und Bedeutung"]. Zeitschrift für Philosophie und philosophische Kritik, vol. 100, pp. 25–50.
Grbac, Deborah (2021). The United Nations Depository Libraries System as an "open community": The ongoing evolution from a knowledge base to a knowledge network. Umanistica Digitale 199–216. https://doi.org/10.6092/issn.2532-8816/13676
Hart, Herbert. L. A. and Green, Leslie (2012). *The Concept of Law*. Third Edition, Third Edition. Clarendon Law Series. Oxford, New York: Oxford University Press.
Jiang, Hang, Zhang, Xiajie, Mahari, Robert, Kessler, Daniel, Ma, Eric, August, Tal, Li, Irene, Pentland, Alex, Kim, Yoon, Kabbara, Jad, Roy, Deb (2024). Leveraging Large Language Models for Learning Complex Legal Concepts through Storytelling. arXiv. https://doi.org/10.48550/arXiv.2402.17019.
Kalinina, Marina G., and Kudryashova, Sofya V. (2022). French, Spanish And German Terminology In Legal Discourse –Problematic Aspects In Translation. In *European Proceedings of Social and Behavioural Sciences* State and Law in the Context of Modern Challenges). https://doi.org/10.15405/epsbs.2022.01.45.
Kiliroor, Cinu C., Sagar, Som and Sundara Didde, Swani (2023). Sentiment-Based Simplification of Legal Text. In *Proceedings of the 4th International Conference on Communication, Devices and Computing*, edited by Dilip Kumar Sarkar, Pradip Kumar Sadhu, Sunandan Bhunia, Jagannath Samanta, and Suman Paul, 463–75. Lecture Notes in Electrical Engineering. Singapore: Springer Nature. https://doi.org/10.1007/978-981-99-2710-4_38.
Neil, Martin, Fenton, Norman, Lagnado, David, and Gill, Richard David (2019). Modelling Competing Legal Arguments Using Bayesian Model Comparison and Averaging. *Artificial Intelligence and Law* 27 (4): 403–30. https://doi.org/10.1007/s10506-019-09250-3.
Pulizzotto, Davide, Chartier, Jean-François, Lareau, Francis, Meunier, Jean-Guy, Chartrand, Louis, 2018. Conceptual Analysis in a computer-assisted framework: mind in Peirce. Umanistica Digitale https://doi.org/10.6092/issn.2532-8816/7305
Quevedo, Ernesto, Cerny, Tomas, Rodriguez, Alejandro, Rivas, Pablo, Yero, Jorge, Sooksatra, Korn, Zhakubayev, Alibek, and Taibi, Davide (2023). Legal Natural Language Processing from 2015-2022: A Comprehensive Systematic Mapping Study of Advances and Applications. *IEEE Access*, 1–1. https://doi.org/10.1109/ACCESS.2023.3339946.
Rissland, Edwina L., Ashley, Kevin D., and Loui, Ronald P. (2003). AI and Law: A Fruitful Synergy. *Artificial Intelligence*, AI and Law, 150 (1): 1–15. https://doi.org/10.1016/S0004-3702(03)00122-X.
Sosa Andrés, Maximiliano (2023). Legal Uncertainty and Its Consequences: A Natural Language Processing Approach. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-495521.
Stathis, Georgios, Biagioni, Giulia, de Graaf, Klaas Andries, Trantas, Athanasios, and van den Herik, Jaap (2024). The Value of Proactive Data for Intelligent Contracts. In *Intelligent Sustainable Systems*, pp. 107–25. Lecture Notes in Networks and Systems. Singapore: Springer Nature. https://doi.org/10.1007/978-981-99-7569-3_10.
Vezzani, Federica, Di Nunzio, Giorgio Maria, Silecchia, Sara (2022). La fraseologia dei trattati internazionali di disarmo: la risorsa terminologica DITTO. Umanistica Digitale 91–

117. https://doi.org/10.6092/issn.2532-8816/14796
Westermann, Hannes, Meeùs, Sébastien, Godet, Mia, Troussel, Aurore, Tan, Jinzhe, Savelka, Jaromir, and Benyekhlef, Karim (2023). Bridging the Gap: Mapping Layperson Narratives to Legal Issues with Language Models. In *Proceedings of the 6th Workshop on Automated Semantic Analysis of Information in Legal Text*, CEUR Workshop Proceedings. Braga, Portugal: CEUR. https://ceur-ws.org/Vol-3441/#paper5.
Whalen, Ryan (2022). Defining Legal Technology and Its Implications. *International Journal of Law and Information Technology* 30 (1): 47–67. https://doi.org/10.1093/ijlit/eaac005.

Simpler becomes Harder: Do LLMs Exhibit a Coherent Behavior on Simplified Corpora?

Miriam Anschütz, Edoardo Mosca, Georg Groh
Technical University of Munich
School of Computation, Information and Technology
{miriam.anschuetz, edoardo.mosca}@tum.de, grohg@cit.tum.de

Abstract

Text simplification seeks to improve readability while retaining the original content and meaning. Our study investigates whether pre-trained classifiers also maintain such coherence by comparing their predictions on both original and simplified inputs. We conduct experiments using 11 pre-trained models, including BERT and OpenAI's GPT 3.5, across six datasets spanning three languages. Additionally, we conduct a detailed analysis of the correlation between prediction change rates and simplification types/strengths. Our findings reveal alarming inconsistencies across all languages and models. If not promptly addressed, simplified inputs can be easily exploited to craft zero-iteration model-agnostic adversarial attacks with success rates of up to 50%.

Keywords: text simplification, model robustness, model consistency

1. Introduction

Automatic text simplification (ATS) is a popular natural language processing task that creates texts in plain language, preserving the original message of the source text. Plain or simplified language is a version of the English language with reduced text complexity and uses only well-known vocabulary. This simplified version aims to increase accessibility and, thus, gives people with learning impairments or reading difficulties access to information on the internet (Martin et al., 2022). While simplifications must alter some text features to reduce its overall complexity, they should still preserve the original source's content. Indeed, content coherency between the original source and simplified output is a core element of text simplification, spanning across various aspects (sentiment, emotion, topic, etc.) (Saggion and Hirst, 2017). For instance, if a strong sentiment or emotion, e.g., anger about something, is conveyed in the original text, this emotion should also be perceivable in the simplified version as well.

In line with this thought, this paper investigates whether models also exhibit this coherent behavior and assesses pre-trained classifiers and recent large language models (LLM) like GPT3.5 on original and simplified texts. For this, we exploit text simplification corpora across different languages, let the models classify content-related features such as the addressed topic, emotion, or sentiment, and analyze potential variations in these labels.

Our results show that models change their predictions for up to 50% of the samples, depending on the language and task used. Figure 1 shows an example of an incoherent model behavior on a manually created sample. The simplified version replaces "conference" with "science meeting". Apart from that, both versions convey the same message and sentiment. Nevertheless, a pre-trained emotion classifier assigns different labels to the two samples. By using human-created or human-aligned simplification corpora, we can ensure that our benchmark samples are natural and that humans consider them meaning-preserving and valid alterations.

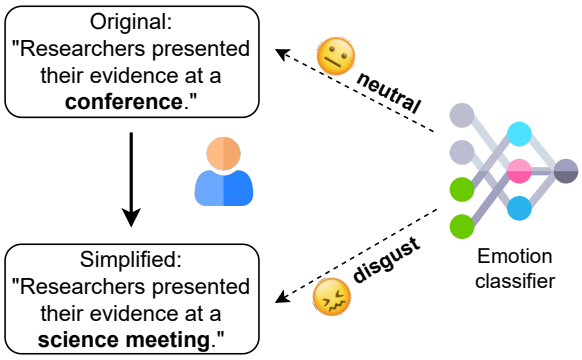


Figure 1: Manually created sentence pair. The simplified version simplifies the word "conference" but preserves the meaning and neutral sentiment of the original sentence. However, a pre-trained emotion classifier behaves incoherent and predicts a different label for the simplified sentence.

Therefore, our contribution can be summarized as follows:

- We compile a strict selection of human-created or human-aligned simplification datasets as model consistency benchmarks. This ensures the naturalness and correctness of our benchmark samples.
- We test pre-trained classifiers for their pre-

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 185–195

diction consistency on normal and simplified language, covering multiple tasks across different languages to give extensive insight into the models' shortcomings.

- Results show concerning discrepancies between the model behavior on normal and simplified inputs. If not addressed, these discrepancies can be easily exploited to produce zero-iteration model-agnostic adversarial attacks with a success rate of up to 50%.

2. Background and related work

Previous work by Elazar et al. (2021) defined model consistency as follows: Given two equivalent paraphrases, a consistent model creates non-contradictory predictions for both of them. In their study, they probed language models with cloze-phrases and evaluated whether the predictions for the original and a paraphrase prompt were similar. The models produced contradicting predictions for 39%-51% of the samples, depending on the model. We extend the consistency definition by Elazar et al. (2021) to classification tasks and deem a model consistent if it assigns the same label to both versions.

Our work is inspired by model robustness checks with adversarial attacks. To create an adversarial attack, a pre-trained classification model f, text samples x_i, and model predictions $f(x_i) = y_i$ are given. Then, the attacker tries to find adversarial samples x' that fool the model to change its prediction compared to the original sample, hence $f(x'_i) = y'_i \neq y_i$. The changes to the original sample x should be minimal and not change its ground truth. In addition, humans should perceive the alteration as valid and natural (Qi et al., 2021). If a model changes its prediction for the sample x', it is considered sensitive to the specific adversarial attack.

Approaches to adversarial attacks can be classified based on their knowledge of the target model. In a *white-box* scenario, the attacker has full access to the model and can optimize perturbations using its output and gradients to find adversarial samples. *Black-box* attacks, on the other hand, only have access to predictions without further model information (Yoo et al., 2020). *Blind or zero-iteration* attacks lack any model feedback and can only apply one perturbation step to deceive the model. This study adopts a zero-iteration setting to assess model sensitivity to text simplification perturbations.

Similar to our objective, Van et al. (2021) examined how NLI models change their predictions when the samples are pre-processed by an automatic simplification model and observed a performance drop of up to 50%. We extend this research to further tasks and languages. Another study investigated whether models are robust to text-style transfer attacks. Qi et al. (2021) utilized a pre-trained style transfer model to transfer common datasets for sentiment or topic prediction, e.g., into Twitter, bible, or poetry language style. In most cases, at least one style adaption yielded the model to alter its prediction. However, a human survey found that many transfers altered the ground truth of samples, indicating that changing predictions was appropriate behavior. To avoid unnatural paraphrases that change the samples' ground truth, we rely on human-supervised datasets for model coherence checks instead of generating adversarial samples with pre-trained simplification models.

3. Methodology

This paper examines whether models maintain a consistent behavior when dealing with normal and simplified texts. For this, we investigate whether and to what extent human-created simplification datasets can lead to output changes in pre-trained models published to the Hugging Face model hub (Wolf et al., 2020). We assume that although the altered text style is perceptible to humans, it is still a natural and meaning-preserving paraphrase of the original. Simplifications are targeted toward people with lower reading understanding capabilities. As such, the simplifications should reduce the complexity of the text but still preserve the topic or sentiment conveyed in the original texts. We expect simplification corpora, used to train ATS systems, to reflect this aspect.

In our experiments, we test various classification tasks covering topic, emotion, fake news/toxicity, and sentiment prediction. For all of them, we expect the original and simplified text to have the same content-related features and, thus, get the same labels from the classifiers. The specific tasks vary across languages, depending on the availability of pre-trained models for the respective language. In the following, we introduce the models under test, our selected datasets, and their pre-processing.

3.1. Models

Table 1 shows the pre-trained classification models we selected from the Hugging Face model hub (Wolf et al., 2020). Not all classification tasks are suited for our experiments as some text features in the simplified versions, for example, the complexity, are altered on purpose. Therefore, we selected only content-related prediction tasks like topic or sentiment classification. Furthermore, we picked

Prediction task	Model	#Classes	Domain	Benchmark
	English			
Topic	bert-agnews (Lee, 2023)	4	news	94%*
Sentiment	bert-base-multilingual-uncased-sentiment (NLPTown, 2020)	5	online	67%*
Emotion	emotion-english-distilroberta-base (Hartmann, 2022)	7	diverse	62%*
Fake news	roberta-fake-news-classification (Benyamina, 2022)	2	news	100%*
	German			
Topic	bert-base-german-cased-gnad10 (Lai-King, 2021)	9	news	68%*
Sentiment	german-news-sentiment-bert (Lüdke et al., 2021)	3	news	96%*
Toxicity	distilbert-base-german-cased-toxic-comments (Buschmeier, 2022)	2	social media	78%*
	Italian			
Topic	it5-topic-classification-tag-it (Papucci, 2022)	10	online	57%
Sentiment	feel-it-italian-sentiment (Bianchi et al., 2021)	2	twitter	84%*
Emotion	feel-it-italian-emotion (Bianchi et al., 2021)	4	Twitter	73%*

Table 1: Pre-trained classifiers used to perform different classification tasks. The classifiers vary in their number of classes and training data domain.
* performance copied from the Hugging Face model page

models with varying numbers of classes to experiment with different task difficulties and tried to reflect the datasets' domains to avoid misclassifications due to domain mismatch. However, we could not always find models with matching domains and, thus, preferred a mismatching domain over skipping the task in the respective language.

As expected, the English language has the highest availability of pre-trained models, and we found models with matching domains for all four tasks. For German, we found in-domain classifiers to predict the sentiment and topic but not for fake news or emotion prediction. To counteract the low availability of suited models, we included a toxicity classifier that detects toxic comments on social media data. Finally, for Italian, we could only find an in-domain topic classifier and had to include an out-of-domain emotion and sentiment classifier.

As shown in Table 1, the models show different performances on relevant benchmark datasets. For most of the model evaluations, we used the test set accuracies reported in the respective model pages on the hub. Only for the Italian topic classifier, no score was reported. Therefore, we evaluated the model on the TAG-it test set (Cimino et al., 2020) ourselves.

3.2. Datasets

A simplification corpus contains texts in standard language aligned with their simplified version. These texts can be sentences or paragraphs, and the simplifications span from replacing single words to completely rewriting the text.

While there exists a collection of simplification corpora across many languages (Ryan et al., 2023), we had specific requirements for our benchmark datasets. First, we selected only datasets where humans created the texts or alignments to avoid label changes due to misalignments. Second, a paragraph can address different topics and have multiple sentiments, resulting in ambiguous classifications. Hence, we only selected sentence-level datasets where the simplifications had a large enough overlap with the original version. In addition, we restricted ourselves to corpora with multiple levels of simplifications to examine whether the strength of the simplification had an impact on the prediction change rates. Moreover, the availability of pre-trained classification models limited the language diversity. Finally, we compiled a benchmark dataset for investigations in English, German, and Italian.

Table 2 shows the datasets we selected for your study. For English, we used Newsela (Xu et al.,

Dataset	Domain	#Simplification levels	#Samples
	English		
Newsela EN (Xu et al., 2015)	news	4	61k
	German		
TextComplexityDE (Naderi et al., 2019)	wikipedia	2*	249
DEplain (Stodden et al., 2023)	online	2	1.846
	Italian		
Simpitiki (Tonelli et al., 2016)	wikipedia	1	1.163
AdminIT (Miliani et al., 2022)	wikipedia, government	2*	736
Terence/Teacher (Brunato et al., 2015)	online, literature	2*	1.146

Table 2: Simplification datasets used to retrieve adversarial data with their number of simplification levels and covered domains. Datasets with the number of simplification levels marked with a * differ from their original version.

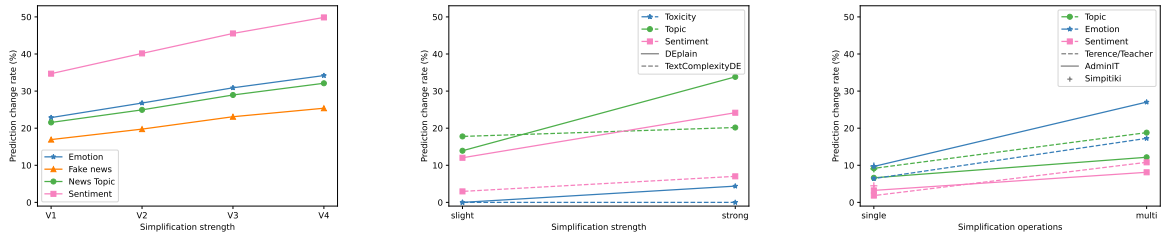
2015), the gold standard for English simplification (Martin et al., 2022): This dataset was created by language experts and consists of sentences from news articles that were simplified into four different levels, where *V1* is the mildest and *V4* the strongest simplification.

For German, we selected multiple datasets. The TextComplexityDE dataset by Naderi et al. (2019) consists of sentences from Wikipedia and their simplified versions. The simplifications were obtained from native speakers and are annotated by their simplification strength. Initially, the dataset has three simplification levels. However, we discarded the sample annotated with "could not be simplified", resulting in the simplification strengths *slightly simplified* and *strongly simplified*. The second corpus, DEplain (Stodden et al., 2023), is compiled from online articles with document- and sentence-level alignments. We picked the test split of the sentence-level dataset. The samples in the dataset were aligned manually and were annotated by their CEFR language level (Council of Europe, 2001). The original samples are at level B2 or C2 and simplified into A1 or A2. To match the simplification levels of the TextComplexityDE data, samples with an original language level of B2 are considered *slight simplifications*, while samples with original level C2 are *strong simplifications*.

We investigate three different simplification corpora for Italian. Simpitiki (Tonelli et al., 2016) uses the edit history of Italian Wikipedia to select edits with the annotation "simplification". The authors manually labeled the samples by their simplification operation and removed non-simplified versions. The second corpus, AdminIT (Miliani et al., 2022), contains a subset of the Simpitiki data and sentences from Italian municipality homepages. The samples are categorized into three simplification strategies: samples with the label *OP* show only a single simplification operation like sentence split or lexical substitution, while other samples were either manually rewritten (label *RS*) or manually aligned based on simplified documents (label *RD*). We grouped the later categories together, yielding two levels of simplifications present in the corpus: single simplification operation and more complex rewritings. The final dataset consists of two subcorpora, the Terence and the Teacher corpus (Brunato et al., 2015). Terence is created from books for children that were simplified manually. In contrast, the Teacher corpus contains original/simplified texts from educational homepages for teachers. Both datasets were manually annotated by their simplification operations. We divided the total number of annotations per text by the number of sentences and grouped the samples into two simplification levels, one with one or fewer simplification operations per sentence and one with multiple.

4. Evaluation

For each sample in the text simplification corpora, we obtained each classifier's prediction for the original and the simplified text and compared their predictions. For our consistency analysis, we do not evaluate whether any of the classifiers predicts a wrong label. We expect the classifiers to label all samples the same, whether they are the original or simplified versions. If the predictions deviate, we consider the respective classifier inconsistent with these samples. We then counted the number of samples with deviating predictions and compared the counts to the full dataset size to obtain the prediction change rate (PCR) for each classifier. Figure 2 shows the PCRs for all classifiers across different languages. We observe rates around 20% on average and up to 50% for English Newsela. These change rates are high, especially considering that the simplified samples are created without any knowledge of the models and can thus be con-



(a) English (b) German (c) Italian

Figure 2: Prediction change rates across different languages and tasks, sorted by the simplification strength of the samples. All classifiers show more deviating predictions the stronger the simplification strength. Overall, the English models are least coherent.

sidered model-agnostic adversarial samples. They are used for all classifiers simultaneously. Other works that limited the number of model-specific changes to the sample achieved only prediction change rates below 10% (Yoo et al., 2020, Fig. 2).

A common trend across all languages is that the model is more likely to change its prediction the stronger the simplification is—i.e., the more simplification operations are performed. Comparing the topic classifiers among all languages (green curves in Figure 2), the English classifier has four classes with an accuracy of 94% and is sensitive to simplification in more than 30% of the time, while the Italian topic classifier has ten classes with an accuracy of only 57% that only shows deviating predictions for 10% of the samples. This indicates the number of labels and the classifier's performance seems to have little impact on the prediction change rate. Similarly, the Italian models have a strong domain mismatch (Wikipedia data vs. Twitter-trained models), while the English data and most of the models are from the news domain. Nevertheless, the English models are more easily affected by simplified inputs. Overall, the human-created or human-aligned samples in the different simplification corpora evoke an alarming amount of prediction changes.

In the following sections, we discuss further experiments to investigate factors influencing classifiers behavior (sections 4.1 - 4.4) as well as examine whether s.o.t.a. LLMs like OpenAI's GPT models (OpenAI, 2023) are also sensitive to simplifications (section 4.5).

4.1. Edit distances

As stated before, the prediction change rates increase with a higher level of simplification. An obvious assumption would be that this is due to increasing differences between the original and simplified samples, especially since higher-level simplifications sometimes remove parts of the original information. To verify this assumption, we calculated the Levenshtein distances, normalized by the original sample's lengths, between the original and simplified versions using the Python Levenshtein[1] package. In Figure 3, these ratios are correlated with the number of classifiers that changed their prediction for the respective sample. Some samples have a normalized distance larger than 1, e.g., when an explanation is added in the simplified version. Among all languages, the German samples have the highest ratios. For all languages, the samples with no prediction change have the smallest normalized distances. However, the average distance for samples with two or more classifiers with prediction changes stays the same or even decreases. Overall, we observe a Spearman correlation of 0.34 for English, 0.35 for German, and 0.34 for Italian between the normalized Levenshtein distance and the number of classifiers with changing predictions[2]. Therefore, only a weak correlation exists between the normalized distance and coherency of the models. This indicates that the simplification operation and the choice of vocabulary to simplify the samples are more relevant to the classifiers than the pure number of edit operations or parts removed from the original sentence.

4.2. Reducing task complexities

Fine-grained emotion and sentiment prediction tasks can be difficult and may induce some ambiguity for the models. The English emotion and sentiment classifiers have seven and five classes, respectively. To control for these potential ambiguities, we reduced the number of classes and, thus, the task complexities. For the emotion task, we summarized all negative emotions, like anger, disgust, fear, sadness, and surprise, into a negative

[1] https://rapidfuzz.github.io/Levenshtein/levenshtein.html#distance

[2] p-values are 0.0 (en), $8.5\text{e}{-65}$ (de), and $1.4\text{e}{-85}$ (it); calculated with SciPy: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

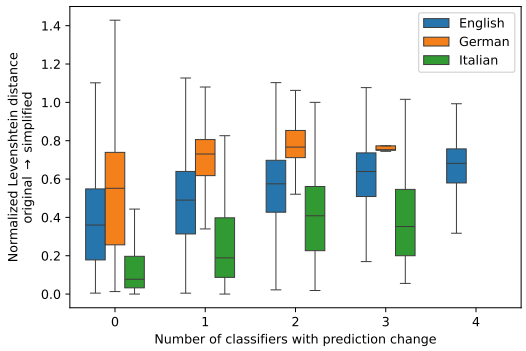


Figure 3: Number of classifiers with changing predictions per sample and their Levenshtein distances between the original and simplified sentences. The distances were normalized by the sample's lengths.

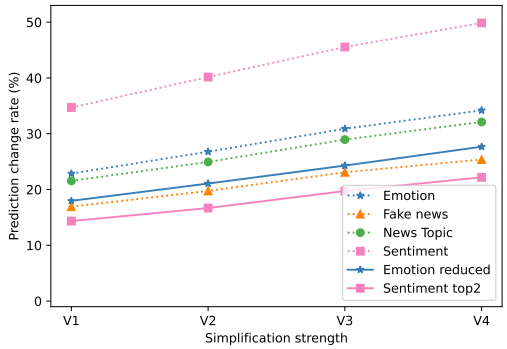


Figure 4: Predictions change rates for tasks with reduced number of classes. The reduced tasks have a better performance but are still susceptible to simplifications.

class. The only positive class, joy, and the neutral class were kept, resulting in a three-class classification task with which we can detect emotion flips. For the sentiment task, we calculated the top-2 accuracy. That is, a prediction was only considered as a deviating prediction if the difference was at least two steps (e.g., strongly negative vs. negative is no prediction change, but strongly negative to neutral is). Figure 4 shows how the classifier predictions change for these reduced tasks compared to the original tasks. The prediction change rate drops by five percentage points for the emotion tasks and more than 20 percentage points for the sentiment task. The rate is still higher with increased simplification strength. We conclude that the difficulty and ambiguity of the classification task can influence the prediction change rate. Nevertheless, even with the reduced tasks, the classifiers still perform inconsistently with simplified versions.

4.3. Simplification operations

We further investigated which simplification operations especially tempt the classifiers to change their predictions. For this, we utilized the Italian Simpliti corpus (Tonelli et al., 2016). The samples in this corpus are annotated by the operation performed to obtain the simplified version. These operations can be on the word level, such as deletion or replacement of single words, or on the phrase level, for example, splitting a sentence into two or transforming the verbal voice. Samples with word-level changes are closer to their original than phrase-level ones. Therefore, we expected the word-level operations to result in lower change rates. However, Figure 5 shows that they are on par or lead to even more prediction changes than the phrase-level simplifications. Word substitution is a combination of word deletion and insertion and, as such, has the highest PCR of all word-level perturbations. As such, replacing a word with its synonym has been used as word-level adversarial attacks before (Chiang and Lee, 2023). On the phrase level, merging two sentences does not affect the sentiment and topic classifiers. Yet, the topic classifier is sensitive to the information order, exhibiting the highest PCR of almost 20

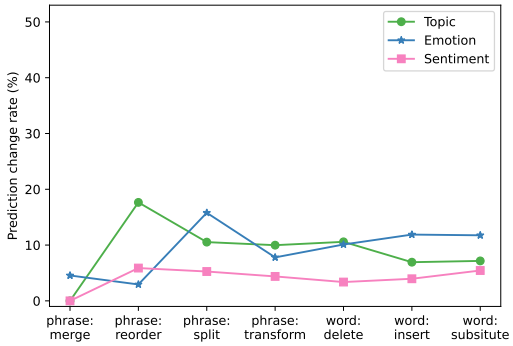


Figure 5: Prediction change rate of different simplification operations as annotated in the Italian Simpitiki corpus (Tonelli et al., 2016).

4.4. Masking named entities

Named entities (NE) can strongly impact the sentiment or topic of a phrase as, sometimes, they only occur in a particular context. In simplified language, NEs are sometimes generalized or removed. To test which influence named entities have on the models' prediction consistency, we compared the predictions when masking the named entities.

We utilized Spacy (Honnibal et al., 2020) to detect named entities in our original and simplified phrases. We searched for tokens with a tag in this list: ['EVENT', 'GPE', 'LANGUAGE', 'LAW',

'LOC', 'NORP', 'ORG', 'PERSON', 'PRODUCT', 'WORK_OF_ART']. If such a token was found, we masked it by replacing it with the placeholder "NAME". With this, the NE in the aligned pairs are the same and do not impact the classification outcome.

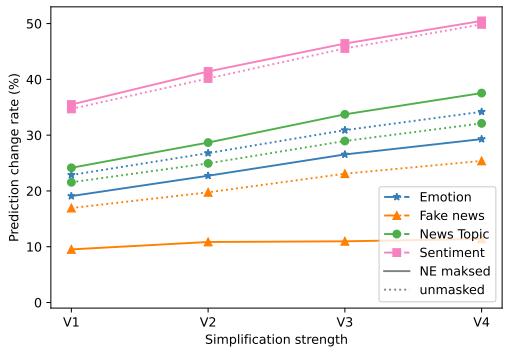


Figure 6: Comparison of prediction change rates of masked (solid) and unmasked (dashed) named entities on the English Newsela corpus. It depends on the task if NE masking increases or decreases the PCR.

In Figure 6, we compare the performances of the classifiers on the named-entity-masked and unmasked data. For the fake news and emotion classification tasks, masking the named entities reduces the prediction change rates. In contrast, for the topic prediction task, the masking even increases the change rates, probably because the sentences become more ambiguous. Therefore, it depends on the classifier and its task whether the possibly changing NEs in the simplifications impact its consistency.

4.5. ChatGPT

OpenAI's GPT models (OpenAI, 2023) are among the NLP models displaying the strongest capabilities in terms of generalization and performance. Hence, we also examined whether these models are sensitive to our simplifications. We used the OpenAI API to query the `gpt-3.5-turbo` model and predicted samples from the Newsela dataset in a one-shot manner. We prompted each sample individually to avoid biases due to previously seen non-simplified versions. We probed the same tasks with the same labels as the English models from the Hugging Face model hub described above. In addition, we asked the model to return a dictionary with all predictions simultaneously and set the temperature parameter to zero. The full prompt can be found in Figure 7. Due to the long processing time of an API request and financial limitations, we restricted ourselves to English and only classified the first 1000 samples per simplification level.

```
{"role": "system", "content": "You are
    an assistant designed to label news
    texts. Users will paste in a string
    of text and you will respond with
    labels you've extracted from the
    text as a JSON object. The topic
    must be one of world, sports,
    buisiness or sci/tech. The
    sentiment is on a scale from 1 to 5
    stars. Fake news can be true or
    false. Emotion can be one of anger,
    disgust, fear, joy, neutral,
    sadness or surprise"},
{"role": "system", "name":
    "example_user", "content": "Predict
    the sentence: Even a big first-day
    jump in shares of Google (GOOG)
    couldn't quiet debate over whether
    the Internet search engine's
    contentious auction was a hit or a
    flop."},
{"role": "system", "name":
    "example_assistant", "content": "{
      'topic': 'sci/tech',
      'sentiment': '2 stars',
      'fake_news': False,
      'emotion': 'sadness'}"},
{"role":"user", "content": f"Predict
    the sentence: {sentence}"}
```

Figure 7: OpenAI system description for our one-shot classification approach.

Figure 8 shows the prediction change rates on this subset. We compare different simplification levels and tasks for the ChatGPT model and our Hugging Face classifiers. The classification change rates for the fake news detection task decrease significantly compared to the model by Benyamina (2022). The PCRs for the other models decrease slightly, and the emotion classification is almost on par with the Hugging Face classifier in the strong simplifications. As we have seen before with the task-specific models, the changes increase with stronger simplification levels. Therefore, even ChatGPT is not robust to text simplification and makes incoherent predictions.

5. Discussion

This paper shows that even s.o.t.a models like GPT3.5 are sensitive towards text simplification as a special form of text style transfer and produce incoherent predictions for texts and their simplified versions. We observe this behavior on human-generated or human-aligned text simplification datasets. That means that our samples were validated by humans before and, thus, should be natural, grammatically correct, and meaning-

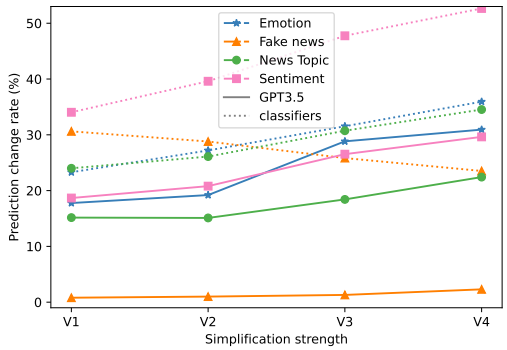


Figure 8: Comparison of Hugging Face classifiers with OpenAI's GPT 3.5. Especially for the emotion and the sentiment task, GPT is only slightly more robust than the smaller, task-specific models.

preserving. However, previous work by Devaraj et al. (2022) has shown that even human-curated simplification datasets can contain factual errors. While pre-processing and filtering those datasets seems promising for text simplification training (Ma et al., 2022), we worked with the original dataset versions. Therefore, it is possible that some of the labels change due to factual errors in the datasets. Unfortunately, such investigations only exist for English corpora. Assuming that the corpora in other languages are of higher factuality, this could explain why the English classifiers showed the highest prediction change rates even though English is usually the best-resourced language with top performance.

We tried to select well-documented classifiers with high accuracy rates. Still, the classifiers show varying performance on relevant benchmark datasets and, thus, potentially produce misclassifications. While our experiments show that the language with the strongest domain mismatches and weakest classifiers, Italian, has the lowest prediction change rates, we can not guarantee that the models do not misclassify some of the samples. Nevertheless, for our investigation, the actual label is not important as we are only interested in whether the labels for the original and simplified versions change. Even if the initial classification is wrong, consistent models should still produce the same label for all simplifications.

We observe alarmingly high prediction change rates across all languages and even OpenAI's GPT-3.5 model. This suggests that the pre-training data of these models lack samples in simplified language and, thus, that there is still only little information available in plain language. Increasing internet accessibility and, hence, increasing the amount of data in simplified language is the most promising approach to improving plain language understanding in language models. If such incoherence is not addressed, our findings empirically show that simplification can easily be exploited as a zero-iteration model-agnostic adversarial attack. Attackers would only need to simplify any input text to achieve success rates of up to 50% with little effort.

6. Conclusion

In this paper, we have investigated how coherent models perform on simplified texts. We have shown that different classifiers across multiple languages struggle with plain language samples and exhibit incoherent behavior. Such incoherency seems to affect also s.o.t.a LLMs like OpenAI's GPT3.5, which are not robust to simplifications from our benchmark datasets. We exploited human-created or human-aligned text simplification corpora to ensure natural and meaning-preserving samples. In this setting, we have observed prediction change rates of more than 40%, indicating a severe lack of plain language understanding in pre-trained language models.

In future studies, we aim to expand our experiments to include more languages and settings involving automatically generated simplifications. Additionally, we believe that improving model coherence on simplified inputs can be achieved through human-preference tuning techniques—such as RLHF and DPO—and we encourage researchers to explore this direction further.

Ethical considerations

We investigated whether model predictions change between original and simplified versions of the same text. Our findings can be valuable for identifying models' shortcomings and improving their robustness. However, our results can potentially be misused as adversarial attacks, and as such, they can threaten applications based on large language models. However, in our approach, we only re-use existing corpora and do not craft a new adversarial threat.

Especially for people with reading difficulties, the availability of information in plain language is crucial. Our experiments demonstrate that pre-trained language models are sensitive to plain language and that simplified samples are still underrepresented in pre-training corpora. This can cause severe problems when these people use LLM applications such as ChatGPT. Hence, we ask content creators and data scientists to remember the need for plain language and provide resources accordingly.

Data/Code availability statement

Our work utilizes existing text simplification corpora. These corpora are publicly available except for the English Newsela corpus (Xu et al., 2015). In addition, we publish our experiment code and links to these corpora at https://github.com/MiriUll/LLM-consistency-simplification.

7. Bibliographical References

Cheng-Han Chiang and Hung-yi Lee. 2023. Are synonym substitution attacks really synonym substitution attacks? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1853–1878, Toronto, Canada. Association for Computational Linguistics.

Andrea Cimino, Felice Dell'Orletta, and Malvina Nissim. 2020. Tag-it@ evalita 2020: Overview of the topic, age, and gender prediction task for italian. *Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*.

Council for Cultural Co-operation. Education Committee. Modern Languages Division Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

Salijona Dyrmishi, Salah Ghamizi, and Maxime Cordy. 2023. How do humans perceive adversarial text? a reality check on the validity and naturalness of word-based adversarial attacks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8822–8836, Toronto, Canada. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Lin, Jiajie Zou, and Nai Ding. 2021. Using adversarial attacks to reveal the statistical bias in machine reading comprehension models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 333–342, Online. Association for Computational Linguistics.

Yuan Ma, Sandaru Seneviratne, and Elena Daskalaki. 2022. Improving text simplification with factuality error detection. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 173–178, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

OpenAI. 2023. Gpt-4 technical report.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4580, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

Horacio Saggion and Graeme Hirst. 2017. *Automatic text simplification*, volume 32. Springer.

Hoang Van, Zheng Tang, and Mihai Surdeanu. 2021. How may I help you? using neural text simplification to improve downstream NLP tasks. In *Findings of the Association for Computational*

Linguistics: EMNLP 2021, pages 4074–4080, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dhruv Verma, Yash Kumar Lal, Shreyashee Sinha, Benjamin Van Durme, and Adam Poliak. 2023. Evaluating paraphrastic robustness in textual entailment models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 880–892, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jin Yong Yoo, John Morris, Eli Lifland, and Yanjun Qi. 2020. Searching for a search method: Benchmarking search algorithms for generating NLP adversarial examples. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 323–332, Online. Association for Computational Linguistics.

8. Language Resource References

Benyamina, Hamza. 2022. *RoBERTa Fake news classification*. PID https://huggingface.co/hamzab/roberta-fake-news-classification.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2021. "FEEL-IT: Emotion and Sentiment Classification for the Italian Language". In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first Italian corpus for text simplification. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 31–41, Denver, Colorado, USA. Association for Computational Linguistics.

Konstantin Buschmeier. 2022. *German Toxic Comment Classification*. PID https://huggingface.co/ml6team/distilbert-base-german-cased-toxic-comments.

Hartmann, Jochen. 2022. *Emotion English DistilRoBERTa-base*. PID https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Mathieu Lai-King. 2021. *German BERT for News Classification*. PID https://huggingface.co/Mathking/bert-base-german-cased-gnad10.

Lee, Jiaqi. 2023. *BERT AGnews*. PID https://huggingface.co/JiaqiLee/bert-agnews.

Simon Lüdke and Josephine Grau and Martin Drawitsch. 2021. *German sentiment BERT finetuned on news data*. PID https://huggingface.co/mdraw/german-news-sentiment-bert.

Martina Miliani, Serena Auriemma, Fernando Alva-Manchego, and Alessandro Lenci. 2022. Neural readability pairwise ranking for sentences in Italian administrative language. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 849–866, Online only. Association for Computational Linguistics.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective assessment of text complexity: A dataset for german language.

NLPTown. 2020. *bert-base-multilingual-uncased-sentiment*. PID https://doi.org/10.57967/hf/1515.

Michele Papucci. 2022. *it5-topic-classification-tag-it*. PID https://huggingface.co/mpapucci/it5-topic-classification-tag-it.

Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. Simpitiki: a simplification corpus for italian. In *CLiC-it/EVALITA*, pages 4333–4338.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Pre-Gamus

Reducing Complexity of Scientific Literature as a Support against Misinformation

Nico Colic[1], Jin-Dong Kim[2], Fabio Rinaldi[1]
[1]Dalle Molle Institute for Artificial Intelligence, Lugano, Switzerland
[2]Database Center for Life Science, ROIS-DS, Chiba, Japan
[1]{fabio.rinaldi, nicola.colic}@idsia.ch
[2]jdkim@dbcls.rois.ac.jp

Abstract

Scientific literature encodes a wealth of knowledge relevant to various users. However, the complexity of scientific jargon makes it inaccessible to all but domain specialists. It would be helpful for different types of people to be able to get at least a gist of a paper. Biomedical practitioners often find it difficult to keep up with the information load; but even lay people would benefit from scientific information, for example to dispel medical misconceptions. Besides, in many countries, familiarity with English is limited, let alone scientific English, even among professionals. All this points to the need for simplified access to the scientific literature. We thus present an application aimed at solving this problem, which is capable of summarising scientific text in a way that is tailored to specific types of users, and in their native language. For this objective, we used an LLM that our system queries using user-selected parameters. We conducted an informal evaluation of this prototype using a questionnaire in 3 different languages.

Keywords: LLM, text summarisation, text simplification

1. Introduction

Today's age of information abundance presents the challenge of navigating complex scientific literature, particularly biomedical. Even professional practitioners struggle to keep up with most recent literature. For example, a clinical doctor might be confronted with an unusual disease, and might need to get information which is only available in the literature, yet might not have the time and disposition to read a scientific paper which might or might not answer that particular information need (Cohen and Hersh, 2005). More so, the gap between scientific publications and lay peoples' understanding hinders dissemination of biomedical insights, fuelling misinformation and making informed decision-making difficult (Kandula et al., 2010). The COVID-19 pandemic, in particular, has underscored this vital role of accurate biomedical information in public health (Bin Naeem and Kamel Boulos, 2021). However, traditional scientific literature often presents dense, technical content, inaccessible to non-experts. This disparity, coupled with the rapid pace of scientific research, exacerbates the challenge of making the knowledge of biomedical literature usable by experts and lay people. This challenge is particularly pronounced among language groups with limited proficiency in English, as the majority of medical research is published in English (Frayne et al., 1996). This problem is not just limited to lay people, but to biomedical professionals, as well.

Addressing this challenge, we present an application that facilitates the understanding of biomedical texts, generating concise, easily comprehensible summaries and simplifications in the users' chosen language. This application, thus, is faced with three different tasks:

- Text simplification
- Text summarisation
- Machine translation

We begin with an overview of current research in these fields (section 2); explain the implementation of our application (sections 3 and 4) and our first evaluation (section 5). We close with a brief discussion of the results and future work (section 6).

2. Background

Text simplification is the process of making a text easier to understand by rewriting it in simpler language, while retaining the original meaning and key information. This often involves replacing complex words (lexical simplification) and structure (grammatical simplification) with simpler alternatives, rephrasing sentences to be more concise, and breaking down complex ideas into more manageable portions (Al-Thanyyan and Azmi, 2021).

Text summarisation, on the other hand, refers to the task of condensing a longer piece of text into a

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 196–201

shorter version while preserving its essential information. Traditionally, this process involves identifying the most important sentences or paragraphs and presenting them in a cohesive and concise manner, thereby providing a condensed overview of the original content (El-Kassas et al., 2021).

The rise of Large Language Models (LLMs) marks a paradigm shift in natural language processing, revolutionizing the way text generated and manipulated (Min et al., 2023). LLMs, such as OpenAI's GPT series and Google's BERT, have demonstrated unprecedented capabilities in capturing and generating human-like text across various domains, including biomedical literature. Essentially, these models are pre-trained on vast amounts of text data, learning to predict the next word in a sequence based on the context provided by the preceding words. This pre-training process allows LLMs to capture complex linguistic patterns and semantic relationships within the data. In particular, they have also proven to outperform previous approaches in the above tasks by far (Van Veen et al., 2023; Al-Thanyyan and Azmi, 2021; Kocmi and Federmann, 2023).

In the context of biomedical text summarisation, simplification, and translation tasks, leveraging LLMs offers several advantages. Rather than treating these tasks as independent processes, performing them simultaneously in one integrated workflow makes intuitive sense. LLMs possess the capability to understand complex biomedical texts, extract salient information, paraphrase content into simpler language, and translate it into multiple languages in a unified manner.

Mainstream use of LLMs such as ChatGPT, however, seems to be mostly business-focused (Wenxue Zou and Tang, 2023), and we presume that especially among more elderly professionals, adoption is hindered by the somewhat more modern chat-based interaction (Sarcar et al., 2023).

One danger of using LLMs, however, is their well-known tendency to *hallucinate*, that is, to invent facts (Zhang et al., 2023). For our application, this is particularly detrimental, as they may pose a health risk and can be a source of the very misinformation we're trying to combat with our application.

3. Implementation

3.1. Design Decisions

By integrating the tasks of text simplification, summarisation and translation into a single workflow, we harness the full potential of LLMs to streamline the process and hope to produce more coherent and linguistically accurate outputs. Because of this, we generally use the term *summary* when referring to the model's output in this paper, and mean it to also imply simplification and translation. As it turns out, however, some "summaries" can be longer than the original text, as the simplification aspect causes additional sentences to be included that explain complicated concepts.

We designed our application with a simple, easy-to-use interface to ease adoption by elderly professionals, in particular. In our application, we have decided to use *personas* to allow users to select their preferred level of text simplification. These personas cater to different user groups, allowing individuals to choose a simplification level that aligns with their comprehension needs and background knowledge. The personas available for selection include:

- *Teenagers*, who need simpler language and lexical simplification.
- *Adult Laypeople*, who need explanation of medical concepts.
- *Professional Clinician*, who mostly need summarisation.

The application is thus implemented as a web service with a simple-to-use interface that allows our users to obtain simplified summaries of biomedical texts in various languages. It is available here[1]; and its code can be found there[2]. The application is composed of 3 sections:

- Text selection
- Parameter selection
- Output

The application is written in `python` using `streamlit`, which facilitates the development of web applications.

3.2. Text Selection

The text selection section allows the user to either enter their own text, select from 10 pre-selected demonstration papers, or to enter a PubMed ID to automatically fetch the corresponding abstract. In the latter case, the application downloads the abstract for the given PubMed ID from the PubMed repository using the Entrez library, and displays it to the user in case the text needs editing.

3.3. Parameter Selection

The parameter selection allows the user to enter all the necessary information to generate the query that is sent to the ChatGPT API. Here, the **maximum number of tokens** indicates only a hard

[1] `pre-gamos.streamlit.app/`
[2] `github.com/Aequivinius/pre-gamos.ai`

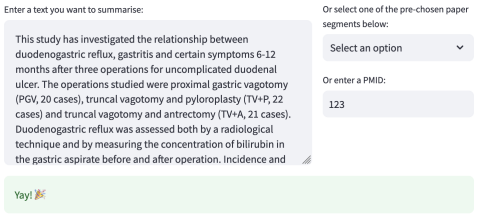


Figure 1: The user submitted a PubMed ID to download the corresponding abstract. It is displayed in the input field to the left, where they can make changes to the abstract, or enter a new text.

cut-off of the response in order to keeping costs incurred by using the API in check. However, it does not affect the length of the summary, as the model does not have a mechanism to control its output length. However, the choosing between different **personas** allows the user to direct the linguistic complexity of the summary and degree of simplification. Currently, the application supports *teenager*, *adult layperson* and *professional clinician* as possible personas. **Temperature** indicates to the model how much determinism is required, with lower temperature making its responses more deterministic (Ouyang et al., 2023). Finally, the user can select the target **language** of the response.

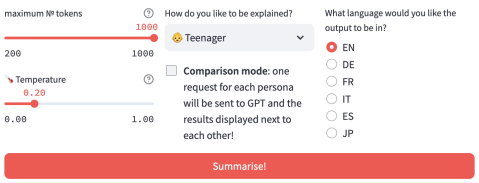


Figure 2: All values are left to their defaults. A simplification of the text selected above will be produced, suitable for a teenager in English, with moderate non-determinism.

There is an additional **Comparison mode** checkbox, which will generate simplifications for all personas simultaneously and display them side-by-side. This feature was added to allow for easier evaluation.

3.4. Output

Once the response from the model has been received, it is presented to the user. In addition, the Flesch readability score (Farr et al., 1951) is computed, and download buttons for different export formats displayed.

Since ChatGPT (3.5) only takes into account 5000 tokens for generating its responses (Floridi and Chiriatti, 2020), longer texts submitted through

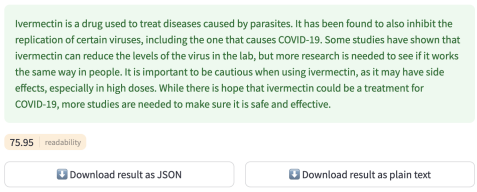


Figure 3: Resulting summary displayed along with its reading score, and different export options.

our applications are chunked and submitted individually for simplification.

If the comparison mode is activated, this section will also display a pair-wise comparison of the responses generated for each persona.

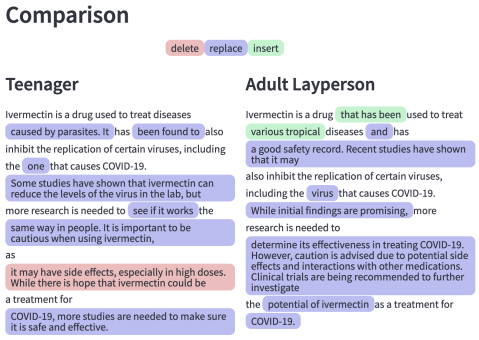


Figure 4: Outputs generated for teenagers and adult layperson displayed side-by-side, with differences highlightened.

For both the reading score and comparison, the text is naively tokenised by splitting on the empty space. In order to tokenise Japanese, where this approach is not viable at all, the dedicated tokeniser `fugashi` is employed (McCann, 2020).

The requests send to the model and its responses are cached; so for repeat queries, the application serves previous responses instantaneously.

4. PA-LLM

As an auxiliary tool, we developed a similar application using the same technology called PA-LLM[3]. However, here we allow the user to select *which* LLM the request is sent to, allowing them to compare the quality of the different models. Currently, only *GPT* and *BARD* are supported; but more APIs can be easily added.

PA-LLM also allows the user to upload the summaries and simplifications obtained through it to PubAnnotation, a repository for storing and displaying annotations of PubMed articles (Kim and

[3] `pa-llm.streamlit.app/`

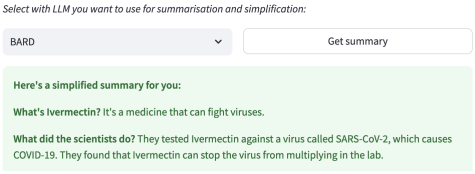


Figure 5: Showing part of the PA-LLM application, where the user has selected BARD to obtain their response.

Wang, 2012). The text simplifications are added as paragraph-level annotations. This allows the user to save their results, and superimpose different simplifications and more traditional annotations such as named entities.

While this is only an adjunct to the project, we realised that the explosive development of and progress in LLMs necessitates the need for researchers to be able to easily compare them.

5. Evaluation

Evaluating text summarisation automatically is notoriously difficult (Bhandari et al., 2020), and to evaluate it manually costly (Steinberger and Ježek, 2009). While we are preparing a formal evaluation, in the scope of this short-term project we can only present the results of an informal evaluation.

For this, we used in-depth questionnaires with 5 participants from 3 different language groups (1 for English, 3 for Spanish, 1 for Japanese). These questionnaires presented 9 summaries to the participants, generated for 3 different input texts and for 3 different personas, and asked participants to rate the summaries according to how appropriate they were for each of the personas, and how coherent (logical order of facts) and consistent (lack of contradiction).

For the input texts, we selected a balanced paper about the use of ivermectin during the Covid pandemic, a retracted paper about hydroxychloroquine, and finally a paper univocally endorsing the use of masks. We picked these different types of publications in order to see if it affects the model's performance.

The summaries were shown to the participants without any information about the persona they had been generated for; and participants were asked to rate their response on a scale of 1 to 5, with 5 being the highest.

In a second part, participants were shown a pair of summaries (as generated by the comparison mode described in 3.3), and were asked to indicate specificities that made one summary more appropriate for one persona than the other.

For English, the summaries generated for professional clinicians were always rated maximally appropriate for them; while for adult laypeople and teenagers they were only rated 4.3 and 4 out of 5, respectively.

For Spanish, summaries aimed at professional clinicians were rated only 4 out of 5 in average for appropriateness, somewhat higher than the 2.7 and 3.8 for adult laypersons and teenagers. In fact, the former was rated much more appropriate for professional clinicians (3.88 out of 5). For the Spanish-speaking participants of the questionnaire, however, we note difference in response profiles, with one participant having a very high variance of 2.2 across the questions; and the other two participants having a low variance of 0.3, but different averages of 2.7 and 4.7. In fact, the k-alpha score for the responses was -0.103, which indicates that there was a slight, but systematic disagreement between participants.

For Japanese, conversely, the summaries generated for professional clinicians were deemed most inappropriate, with only 3.7 out of 5, as opposed to the 4.3 rating both adult layperson and teenager texts received.

For coherence and consistency, all responses were rated 5 in all languages; with the exception for the responses from one Spanish participant who consistently rated responses either 2 or 3, for all questions.

We also computed Flesch reading ease scores, which gives a measure of how easily understandable a piece of text is. It is computed based on the average sentence length and the average number of syllables per word in the text; and higher scores point to simpler language (Farr et al., 1951).

The readability scores vary across the languages, but for all clearly show a difference for the generated summaries depending on the target persona. For English, the readability scores averaged to 75.9, 45.6 and 24.9 for teenagers, adult laypeople and professional clinicians across 10 sample abstracts. While the actual averages differ across languages (for German, for example, they are 40.5, 30.0 and 19.3 for teenagers, laypeople and professional clinicians, respectively), in all languages did summaries for teenagers result in the highest readability scores, and those for clinicians in the lowest.

We also asked participants to point out specific words or structures that made one summary more suitable for a teenager or for a professional clinician. For all five participants, they noted simpler terms for the former, and more accurate and more complicated terms for the latter. For Japanese, however, the participant noted that some terms were incorrectly translated.

For English, interestingly, our participant pointed out that the summary contained an explanation about a drug mentioned in the original text. The explanation itself, however, was not part of the original

text.

The questionnaires used for our evaluation and results can be found here[4].

6. Discussion and Conclusion

The evaluation above clearly shows that a more rigid evaluation is necessary; but already gives some first insights.

Firstly, for the evaluation of appropriateness, the same summaries received vastly different ratings from different survey participants. This shows that future evaluations need to include examples of appropriate summaries so that participants can calibrate their ratings.

Secondly, it shows that our approach is promising. The summaries were generally deemed most appropriate for the target personas they were generated for, and also their readability scores seem to support this.

Thirdly, a more careful evaluation of language differences is needed. While all texts for all languages were rated highly in terms of coherence and consistency, we noted some irregularities for some languages. For Japanese, it was the mistranslation of terms; for English, it was the hallucination of background information.

This last point deserves special attention, because it shows that even for summarisation tasks, the model does use knowledge not provided in the input text to generate its response. While in our particular case this was, in fact, helpful to make the text more easily understood, it can be a source of misinformation and put at risk the trustworthiness of applications such as ours. This is indeed in line with similar research (Zaretsky et al., 2024), where LLMs introduced misinformation on a similar simplification task.

7. Acknowledgements

The work described in this paper has been partially funded by the grant “Platform for an Epidemic-Related Guard Against Misinformation that is Understandable and grounded in Science” (PREGAMUS, Hasler Foundation) to Fabio Rinaldi, by the grant “Brisk.AI” (RPG2120, Leading House for the Latin American Region - University of St. Gallen) to Fabio Rinaldi, and by a “Strategic Research Projects” grant from ROIS (Research Organization of Information and Systems) to Jin-Dong Kim.

Special thanks to Oscar Lithgow Serrano (IDSIA) for contributing to the design of the experiments described in this paper, and to Yalbi Balderas Martinez (INER, Mexico) for contributing to the evaluation in Spanish.

[4] drive.google.com/drive/folders/12sBQDW_h59BWq-6dXgLZ1l6g0nHiQdwg

8. Bibliographical References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Manik Bhandari, Pranav Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. *arXiv preprint arXiv:2010.07100*.

Salman Bin Naeem and Maged N Kamel Boulos. 2021. Covid-19 misinformation online and health literacy: a brief overview. *International journal of environmental research and public health*, 18(15):8091.

Aaron M Cohen and William R Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

James N Farr, James J Jenkins, and Donald G Paterson. 1951. Simplification of flesch reading ease formula. *Journal of applied psychology*, 35(5):333.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Susan M Frayne, Risa B Burns, Eric J Hardt, Amy K Rosen, and Mark A Moskowitz. 1996. The exclusion of non-english-speaking persons from research. *Journal of general internal medicine*, 11:39–43.

Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. In *AMIA annual symposium proceedings*, volume 2010, page 366. American Medical Informatics Association.

Jin-Dong Kim and Yue Wang. 2012. Pubannotation-a persistent and sharable corpus and annotation repository. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 202–205.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Paul McCann. 2020. fugashi, a tool for tokenizing japanese in python. *arXiv preprint arXiv:2010.06858*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Shuyin Ouyang, Jie M Zhang, Mark Harman, and Meng Wang. 2023. Llm is like a box of chocolates: the non-determinism of chatgpt in code generation. *arXiv preprint arXiv:2308.02828*.

Sayan Sarcar, Cosmin Munteanu, Jaisie Sin, Christina Wei, and Sergio Sayago. 2023. Designing conversational user interfaces for older adults. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages 1–5.

Josef Steinberger and Karel Ježek. 2009. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. 2023. Clinical text summarization: Adapting large language models can outperform human experts. *Research Square*.

Yunkang Yang Wenxue Zou, Jinxu Li and Lu Tang. 2023. Exploring the early adoption of open ai among laypeople and technical professionals: An analysis of twitter conversations on #chatgpt and #gpt3. *International Journal of Human–Computer Interaction*, 0(0):1–12.

Jonah Zaretsky, Jeong Min Kim, Samuel Baskharoun, Yunan Zhao, Jonathan Austrian, Yindalon Aphinyanaphongs, Ravi Gupta, Saul B. Blecker, and Jonah Feldman. 2024. Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format. *JAMA Network Open*, 7(3):e240357–e240357.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Author Index