LREC-COLING 2024

3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI 2024)

Workshop Proceedings

Editors
Rodrigo Wilkens, Rémi Cardon, Amalia Todirascu and Núria Gala

20 May, 2024
Torino, Italia

Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI 2024)

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

Message from the General Chair

Recent studies show that the number of children and adults facing difficulties in reading and understanding written texts is steadily growing. Reading challenges can show up early on and may include reading accuracy, speed, or comprehension to the extent that the impairment interferes with academic achievement or activities of daily life. Various technologies (text customization, text simplification, text-to-speech devices, and screening for readers through games and web applications, to name a few) have been developed to help poor readers to get better access to information as well as to support reading development. Among those technologies, text adaptations are a powerful way to leverage document accessibility by using NLP techniques.

The 3rd Workshop on Tools and Resources for REAding DIfficulties (READI), collocated with the International 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), aims at presenting current state-of-the-art techniques and achievements for text adaptations together with existing reading aids and resources for lifelong learning. The materials can be addressed to children struggling with difficulties in learning to read, to the community of teachers, speech-language pathologists and parents seeking solutions, but also to professionals involved with adults struggling with reading (illiterates, aphasic readers, low vision readers, etc.).

18 propositions have been submitted at this third edition, from which 9 were accepted, i.e., a 50% acceptance rate. This acceptance rate is lower than in the first and second editions, which had an acceptance rate of 66% and 71%, because the third edition takes place on a half-day, unlike the previous ones, which were full-day events. The accepted papers come from 41 different authors from 11 different countries (United Kingdom 6, Spain 4, Belgium 4, Germany 4, France 3, Japan 3, United States 3, Switzerland 2, Iceland 1, Italy 1, and Russian Federation 1). READI also features one invited speaker, Giulia Venturi from the Istituto di Linguistica Computazionale Antonio Zampolli (CNR, Pisa, Italy).

We are thankful to the authors who submitted their work to this workshop, to our Program Committee members , the reviewers and the additional reviewers who did a thorough job evaluating submissions, to Giulia Venturi who kindly accepted to be our invited speaker, and to the LREC committee for including this workshop into their program.

Organizing Committee

Organizers

Rémi Cardon, Université catholique de Louvain, Belgium
Thomas François, Université catholique de Louvain, Belgium
Núria Gala, Aix Marseille Université, France
Amalia Todirascu, Université de Strasbourg, France
Rodrigo Wilkens, Université catholique de Louvain, Belgium

Program Committee:

Delphine Bernhard, Lilpa, Université de Strasbourg, France
Dominique Brunato, Institute of Computational Linguistics "A. Zampolli" (ILC-CNR), Pisa, Italy
Rémi Cardon, CENTAL, ILC, Université Catholique de Louvain, Belgium
Mireia Farrus, Universitat de Barcelona, Spain
Thomas François, UCLouvain, CENTAL, Belgium
Núria Gala, LPL-CNRS, Aix Marseille Université, France
Lingyun Gao, UCLouvain, Belgium
Itziar Gonzalez-Dios, HiTZ Basque Center for Language Technologies - Ixa, University of the Basque Country UPV/EHU, Spain
Arne Jonsson, Linköping University, Sweden
Euan McGill, Universitat Pompeu Fabra, Spain
Didier Schwab, Univ. Grenoble Alpes, France
Matthew Shardlow, Manchester Metropolitan University, United Kingdom
Kim Cheng Sheang, Universitat Pompeu Fabra, Spain
Anaïs Tack, KU Leuven; imec; UCLouvain, Belgium
Amalia Todirascu, University of Strasbourg, France
Vincent Vandeghinste, Instituut voor de Nederlandse Taal, Netherlands
Giulia Venturi, Institute of Computational Linguistics "Antonio Zampolli" (ILC-CNR), Italy
Elena Volodina, University of Gothenburg, Sweden
Rodrigo Wilkens, CENTAL, ILC, Université Catholique de Louvain, Belgium

Invited Speaker:

Giulia Venturi, ILC-CNR, Pisa, Italy

Table of Contents

Whorshop Program

Monday, May 20, 2024

9:10–10:00 Invited speaker

9:10–10:00 *Putting the right book into the hands of the right reader" in the era of Large Language Models*
Giulia Venturi

10:00–10:30 Oral Section presentation

10:00–10:30 *Evaluating Document Simplification: On the Importance of Separately Assessing Simplicity and Meaning Preservation*
Liam Cripwell, Joël Legrand and Claire Gardent

10:30–11:00 coffee break

11:00–12:00 Poster Section presentation

11:00–12:00 *Malmon: A Crowd-Sourcing Platform for Simple Language*

Helgi Björn Hjartarson and Steinunn Rut Fririksdóttir

11:00–12:00 *Automatic Generation and Evaluation of Reading Comprehension Test Items with Large Language Models*
Andreas Säuberli and Simon Clematide

11:00–12:00 *An Extensible Massively Multilingual Lexical Simplification Pipeline Dataset using the MultiLS Framework*
Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Md Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Marcos Zampieri and Horacio Saggion

11:00–12:00 *SIERA: An Evaluation Metric for Text Simplification using the Ranking Model and Data Augmentation by Edit Operations*
Hikaru Yamanaka and Takenobu Tokunaga

11:00–12:00 *Transfer Learning for Russian Legal Text Simplification*

Mark Athugodage, Olga Mitrofanove and Vadim Gudkov

Monday, May 20, 2024 (continued)

Evaluating Document Simplification: On the Importance of Separately Assessing Simplicity and Meaning Preservation

Liam Cripwell[†], Joël Legrand[†,‡], Claire Gardent[†]
[†]LORIA, CNRS, Inria, Université de Lorraine, Nancy, France
[‡]Centrale Supélec, Metz, France
{liam.cripwell, joel.legrand, claire.gardent}@loria.fr

Abstract

Text simplification intends to make a text easier to read while preserving its core meaning. Intuitively and as shown in previous works, these two dimensions (simplification and meaning preservation) are often-times inversely correlated. An overly conservative text will fail to simplify sufficiently, whereas extreme simplification will degrade meaning preservation. Yet, popular evaluation metrics either aggregate meaning preservation and simplification into a single score (SARI, LENS), or target meaning preservation alone (BERTScore, QuestEval). Moreover, these metrics usually require a set of references and most previous work has only focused on sentence-level simplification. In this paper, we focus on the evaluation of document-level text simplification and compare existing models using distinct metrics for meaning preservation and simplification. We leverage existing metrics from similar tasks and introduce a reference-less metric variant for simplicity, showing that models are mostly biased towards either simplification or meaning preservation, seldom performing well on both dimensions. Making use of the fact that the metrics we use are all reference-less, we also investigate the performance of existing models when applied to unseen data (where reference simplifications are unavailable).

Keywords: simplification, evaluation, out-of-domain

1. Introduction

Text simplification is the task of rewriting a text such that it is easier read and understood by a wider audience, while still conveying the same central meaning. This generally involves transformations such as lexical substitution (Paetzold and Specia, 2017; North et al., 2023) or structural modifications (sentence splitting) according to the text syntax (Narayan et al., 2017) or discourse structure (Niklaus et al., 2019; Cripwell et al., 2021). Although the main motivation is to promote accessibility (Williams et al., 2003; Kajiwara et al., 2013), it can also be a useful preprocessing step for downstream NLP systems (Miwa et al., 2010; Mishra et al., 2014; Štajner and Popovic, 2016; Niklaus et al., 2016).

While early simplification work has focused on individual sentence inputs (Nisioi et al., 2017; Martin et al., 2020; Cripwell et al., 2022; Yanamoto et al., 2022), recent progress has been made on document-level simplification (Sun et al., 2021; Cripwell et al., 2023b,a). However, several challenges stand in the way of further progress on simplification tasks, including the limited ability to transparently perform automatic evaluation and most popular metrics' requirement of multiple references.

Recent investigation into the quality of sentence-level test data and system outputs has found many instances of factual incoherence not previously detected during data collection or evaluation (Devaraj et al., 2022). This raises questions of how faithful simplifications are to their inputs and whether or not these concerns also apply to the document-level task. Although attempts to automatically evaluate semantic faithfulness in sentence simplification have seen limited success (Devaraj et al., 2022), summarization literature contains a lot of work that could be transferable to document simplification (Laban et al., 2022; Fabbri et al., 2022).

Despite their ability to generate highly fluent texts, the commonly used end-to-end neural systems rely heavily on the quality of data they are trained on. In text simplification, training data is scarce, with most existing corpora being compiled via automatic alignment methods. These are known to contain a lot of noise and imbalanced distributions of possible transformation types (Sulem et al., 2018; Jiang et al., 2020). As a result, end-to-end systems are very conservative in the amount of editing they perform, often making little to no changes to the input (Alva-Manchego et al., 2017). With some works observing an inverse correlation between meaning preservation/faithfulness and simplicity (Schwarzer and Kauchak, 2018; Vu et al., 2018), this raises the question of whether those models sufficiently simplify the input text (since some amount of degradation is a requirement for performing simplification).

Evaluation poses additional challenges, with the suitability of popular automatic metrics remaining unclear (Alva-Manchego et al., 2021; Scialom et al., 2021b; Cripwell et al., 2023c). As most automatic metrics require multiple, high-quality references, studies are usually restricted to a small pool of im-

perfect datasets that include reference simplifications, making it difficult to gauge how well systems actually perform on real-world out-of-domain data. Furthermore, most metrics produce a single score that aims to quantify overall quality, despite the fact that quality aspects are often highly correlated or definitionally at odds with each other (Schwarzer and Kauchak, 2018; Vu et al., 2018). As such, results are often difficult to interpret, making it unclear where models succeed and fail.

In this work, we compare various document-level simplification models in terms of meaning preservation and simplicity, with specific focus on English-language data. Departing from single-value, reference-based scores such as SARI or BERTScore, we exploit distinct, reference-less metrics for these two dimensions. For meaning preservation, we rely on existing reference-less metrics such as SummaC (Laban et al., 2022), QAFactEval (Fabbri et al., 2022), entity matching and BLEU (Papineni et al., 2002) with respect to the input document.

For simplicity, we introduce a variation of the SLE metric proposed in (Cripwell et al., 2023c), which we refer to as ϵSLE. It is able to estimate how close a simplification is to the target reading level without relying on any references. We also report the Flesch-Kincaid grade level (**FKGL**) (Kincaid et al., 1975), a simple document readability metric which is based on a regression model that considers the average length of sentences and syllable count of words in the document.

To assess how well existing models perform on each of these two dimensions, we apply these metrics to the output of four document level simplification models using both in and out of domain test data. We find that none of these four models ranks first on both dimensions, confirming the tension between meaning preservation and simplification. Models with high meaning preservation scores tend to be conservative and under-simplify. Conversely, models that simplify more tend to under-perform in terms of meaning preservation. We further show that for a given model, the trade-off may invert when evaluating on an out-of-domain test set.

2. Related Work

Document Simplification. Document simplification work began by iteratively applying sentence simplification methods over documents (Woodsend and Lapata, 2011; Alva-Manchego et al., 2019), which was quickly found to be insufficient for certain operations, often leading to damaged discourse coherence (Siddharthan, 2003; Alva-Manchego et al., 2019). Some works then began reducing the problem scope, focusing on specific subtasks of document simplification, including sentence deletion (Zhong et al., 2020; Zhang et al., 2022), insertion (Srikanth and Li, 2021), and reordering (Lin et al., 2021).

Sun et al. (2020) used a sentence-level model with additional encoders to embed tokens from the preceding and following sentences, which they attend to during generation. However, this proved incapable of outperforming a sequence-to-sequence baseline (Sun et al., 2021). Cripwell et al. (2023b) achieved state-of-the-art performance by first using high-level document context to generate a document plan and then using this plan to guide a sentence simplification model downstream. Later, Cripwell et al. (2023a) iterated on this framework by exploring the importance of context within the simplification component and proposing several alternate downstream models that lead to further performance increases.

Faithfulness in Simplification. The goal of text simplification is not only to make a text easier to read, but also to ensure the same information is conveyed. Until recently, explicit evaluation of the faithfulness of simplification outputs has been somewhat overlooked. In general, semantic adequacy with the original complex text is only manually considered during human evaluation, with automatic metrics mostly focusing on semantic similarity to reference simplifications (which are assumed to be sufficiently faithful). Even during human evaluation, the typical criterion for faithfulness is rather relaxed, demanding only that the text continues to generally convey the core meaning.

A recent manual investigation into common faithfulness errors in both system outputs and test data found many issues undetected by common evaluation metrics (Devaraj et al., 2022). However, this analysis was limited to sentence-level simplification and many of the issues uncovered do not extend to the document-level case — a limitation which the authors acknowledge. For instance, content that appears to be wrongly inserted or deleted when considering a pair of aligned sentences in isolation could easily have been moved to or from other sentences in the same document. They also attempted to train a model to automatically evaluate faithfulness, to limited success.

Outside of explicit evaluation, some sentence simplification works have considered faithfulness within their training processes. Guo et al. (2018) train a multi-task simplification model with entailment as an auxiliary task. Nakamachi et al. (2020) integrate the semantic similarity between an input and generated output within the reward function of their reinforcement learning (RL) framework for simplification, while Laban et al. (2021) include an inaccuracy guardrail that rejects generated sequences that contain named entities not present in the input.

Ma et al. (2022) attempt to improve performance by down-scaling the training loss of examples with similar entity mismatches. However, these works either do not explicitly evaluate the faithfulness of their system outputs or find that they do not actually prevent the final model from generating unfaithful simplifications.

On the related task of summarization, there has been much more work on this front (Maynez et al., 2020; Pagnoni et al., 2021). The evaluation of semantic faithfulness in summarization is broadly split into either entailment-based (Falke et al., 2019; Kryscinski et al., 2020; Koto et al., 2022) or question answering (QA)-based methods (Wang et al., 2020; Durmus et al., 2020; Scialom et al., 2021a), with comprehensive benchmarks being established for each (Laban et al., 2022; Fabbri et al., 2022).

Simplicity Evaluation. The most popular evaluation metrics (e.g. SARI, BERTScore) used in simplification generally require multiple high-quality references to perform as intended (Xu et al., 2016; Zhang et al., 2019). This poses problems for practitioners seeking to apply simplification models to novel data, as it is impossible to gauge performance without going through the difficult and expensive process of manually creating references — a problem that is exacerbated in the document-level case.

Recent investigations into the validity of these metrics also raise concerns over whether they do in fact measure simplicity itself and not correlated attributes like semantic similarity to references (Scialom et al., 2021b; Cripwell et al., 2023c). However, a reference-less sentence simplicity metric (showing high correlations with human judgements) has also been recently proposed, which could allow for meaningful evaluation of out-of-domain performance (Cripwell et al., 2023c). Despite this, the efficacy of existing evaluation metrics when applied at the document level remains unexplored.

3. Experimental Setup

Our global aim is to perform a more thorough investigation into the performance of existing document simplification systems, with particular focus on providing more interpretable results that differentiate between faithfulness and simplicity. We also investigate the out-of-domain performance of existing systems and reconsider how this should be evaluated given a lack of diverse references.

3.1. Data

We primarily rely on the Newsela (Xu et al., 2015) corpus, which is often considered the gold-standard document simplification dataset. It consists of 1,130 English news articles that have been manually rewritten by professional editors at five different discrete reading levels (0-4) of increasing simplicity.[1] The main drawback of using Newsela is that it requires a licence to use in research, meaning that it is not necessarily made available to all practitioners. This makes it somewhat more difficult to compare and reproduce results, but unfortunately nothing comes close in terms of quality.

As we intend to focus on reference-less evaluation, we can also consider model performance on out-of-domain data for which we have no reference simplifications. For this, we use standard English Wikipedia (EW) articles from Wiki-auto (Jiang et al., 2020). Although EW corpora with automatically aligned reference simplifications from simple English Wikipedia (SEW) exist, they are known to contain a lot of noise, being of particularly poor quality when considered at the document level (Xu et al., 2015; Cripwell et al., 2023b). To assess performance on longer documents, we only consider those that contain at least 10 sentences and 3 paragraphs. To diversify the domain of articles, we annotate each with a semantic type according to their WikiData (Vrandečić and Krötzsch, 2014) entry. We select 19 of the most common types, group them into 5 broad categories and sample articles equally from each to obtain a final test set of 1000 documents (further details are given in Appendix A).

3.2. Simplification Systems

We consider several document simplification systems (at or near state-of-the-art) from existing works, which have all been trained on Newsela.

PG$_{Dyn}$ (Plan-Guided Simplification with Dynamic Context) is a pipeline system that first generates a document simplification plan using high-level context, then conditions a sentence simplification model on said plan (Cripwell et al., 2023b). The plan consists of a sequence of simplification operations (split, delete, copy or rephrase), with one for each sentence in the input document.

From Cripwell et al. (2023a) we include three additional systems: (i) **LED$_{para}$** — a paragraph-level Longformer (Beltagy et al., 2020) model which is the best performing end-to-end system; (ii) $\hat{O} \rightarrow$ **LED$_{para}$**, which uses the same Longformer model, but is conditioned on a plan from the same planner as PG$_{Dyn}$; and (iii) $\hat{O} \rightarrow$ **ConBART** — a modification of the BART (Lewis et al., 2020) architecture that attends to a high-level document context during decoding, while also conditioning on a plan.

Table 1 provides a summary of the model attributes.

System	Description
PG$_{Dyn}$	- Sentence-level text input - Plan-guided
LED$_{para}$	- Paragraph-level text input - No plan-guidance - Longformer-based end-to-end model
$\hat{O} \rightarrow$ LED$_{para}$	- Paragraph-level text input - Plan-guided - Longformer-based simplification component
$\hat{O} \rightarrow$ ConBART	- Sentence-level text input - Plan-guided - Simplification model with cross-attention over high-level representation of document sentences

Table 1: Descriptions of the different document simplification systems we consider.

As these Newsela-trained models have all been prefixed with target reading-level control tokens during training, we must also specify this during inference. For in-domain evaluation, we consider the performance of the various models on each of the four target simplification levels present in Newsela. On the out-of-domain Wikipedia data, we set the target reading-level to 3 (on a scale of 0-4) for all models. Ideally, this will result in substantial editing during simplification while limiting the over-deletion of content.

3.3. Evaluating Faithfulness

We consider two existing reference-less metrics for evaluating faithfulness: **SummaC** (an NLI entailment-based metric) (Laban et al., 2022) and **QAFactEval** (a QA-based metric) (Fabbri et al., 2022). Both are from the summarization literature and should therefore be considered with a level of caution when being applied to simplification. For example, as summarization outputs are generally much shorter than their inputs, it is likely that these metrics will skew in favour of very short and concise simplifications (i.e. precision) even when too much information has been removed. In response, we also use variations of each that focus more on recall.

SummaC (Summary Consistency) (Laban et al., 2022) first works by using an out-of-the-box NLI model[2] to compute an NLI entailment matrix

over a document. This is an $M \times N$ matrix of entailment scores between each of the M input sentences and N output sentences. This is transformed into a histogram form of each column and a convolutional layer is used to convert the histograms into a single score for each output sentence, which are then averaged. As such, this metric is naturally more precision-oriented and therefore could favour shorter, lexically conservative simplifications. In response, we also compute a recall-oriented version, whereby scores are calculated for each input sentence (i.e. high scores will require generating a simple document that retains as much source information as possible).

QAFactEval (Fabbri et al., 2022) is a state-of-the-art QA-based metric that consists of several components within a pipeline. In order they are: answer selection → question generation → question answering → overlap evaluation → question filtering. Questions and correct answers are first generated given a summary, then answers are predicted given the input document as context. For each of these, an answer overlap score is computed using the LERC metric (Chen et al., 2020), which estimates the semantic similarity between the true and predicted answers. The final result is the average of these answer overlap scores for the questions remaining after a question filtering phase (those that are considered answerable).

If an overly short simplification leads to only a few questions being generated it is possible that this could achieve high scores. Further, the process of simplification itself (lexical subtitution in particular) might challenge this metric as the QA model must be able to accurately recognize the semantic similarity between substituted phrases in order to gauge the validity of an answer. As with SummaC, we compute both precision- and recall-oriented versions of this metric. In the recall case we generate questions from the source document instead of the output.

Entity Matching. Another heuristic for assessing the semantic faithfulness of generated text is to consider the similarity between entities present in the input vs. output — sometimes referred to as entity-based semantic adequacy (**ESA**) (Wiseman et al., 2017; Laban et al., 2021; Faille et al., 2021; Ma et al., 2022). We extract named entities from input documents using the spaCy library[3] and compute the precision, recall, and F1 with respect to those found in the generated simplifications.

Conservativity. Given the nature of semantic faithfulness being tied to the input, high scores

for these metrics can be obtained by overly conservative models. So, to better contextualize results, we also include the average lengths of outputs (no. of tokens and sentences) as well as the **BLEU** (Papineni et al., 2002) with respect to the input (BLEU$_C$). Generally, simplifications are slightly shorter than their inputs and often contain more sentences (a result of splitting). This BLEU$_C$ score will give a further indication of the amount of editing that has been performed and therefore flag whether a system has potentially achieved high faithfulness scores as a result of over-conservativity.

3.4. Evaluating Simplicity

Most popular evaluation metrics for simplification have well documented limitations, such as their reliance on high-quality references. Furthermore, their efficacy has not been fully explored for the document-level task. Given this and the fact that the scope of our study covers performance on out-of-domain data, for which there are no references, we instead rely on reference-less alternatives that are known to correlate well with pure simplicity (Cripwell et al., 2023c).

FKGL. We report the Flesch-Kincaid grade level (FKGL) (Kincaid et al., 1975) — a simple document readability metric with a long history of usage. It is based on a regression model that considers the average length of sentences and syllable count of words in the document. However, FKGL gauges simplicity in absolute terms, assuming a simpler output is universally more valuable. Because of this, it is not ideal for evaluating simplicity for specific target groups (e.g. the different reading grade levels supported by Newsela).

ϵ**SLE**$_{doc}$. Given that most document simplification systems target a specific reading level during generation, it would be more useful to evaluate the divergence from this target level of simplicity, rather than measuring raw simplicity alone. To this end, we modify the **SLE** sentence level simplicity metric proposed in (Cripwell et al., 2023c) to obtain a simplicity metric for documents which we dub ϵSLE$_{doc}$.

SLE is trained to predict a sentence's simplicity level following a leveling scheme similar to Newsela. We adapt this to the document level by computing the prediction for a document Y as the mean of its sentences' **SLE** scores:

$$\text{SLE}_{doc}(Y) = \frac{1}{|Y|} \sum_{i=1}^{|Y|} \text{SLE}(y_i) \qquad (1)$$

where y_i is the ith sentence of document Y. We further adapt this to our task by deriving the simplicity level error (ϵ**SLE**$_{doc}$) of a system as the mean

absolute error (MAE) between the predicted and target document reading levels.

$$\epsilon\text{SLE}_{doc} = \frac{1}{N} \sum_{i=1}^{N} \left| \text{SLE}_{doc}(\hat{Y}_i) - l_i \right| \qquad (2)$$

where l_i is a target simplicity level. ϵSLE is able to estimate how close a simplification is to the target reading level without relying on any references, allowing it to avoid the limitations and rigidity of most other popular evaluation metrics. Although SLE was initially proposed for sentence-level evaluation, it was also trained with document-level labels and to optimize document-level accuracy. As such, we believe SLE$_{doc}$ should work well as a document-level metric. Although individual sentences within a document might have diverse simplicity levels, in aggregate they should converge to the global document level, following the central limit theorem (SLE distributions per reading level are shown in Appendix B).

4. Results and Discussion

4.1. Newsela Performance

Faithfulness and simplicity results on the Newsela test set are shown in Tables 2 and 3, respectively.

References. We observe that the references achieve much better FKGL and ϵSLE$_{doc}$ than any system indicating that simplification models simplify less than Newsela editors. Similarly, all models have higher meaning preservation scores than the references which shows that they are more conservative (since they were hand written by professionals, we assume that references are sufficiently faithful to the input). This suggests that there is indeed a trade-off between faithfulness and simplicity and more specifically, that models with high meaning preservation scores under-simplify with respect to their target simplification level.

In summation, there are still improvements that can be made to reduce conservativity and improve simplification in current document simplification systems.

End-to-End vs Planning. We see a similar trend when comparing end-to-end (LED$_{para}$) and plan-guided models (PG$_{Dyn}$, $\hat{O} \rightarrow$ LED$_{para}$, $\hat{O} \rightarrow$ ConBART).

The end-to-end model is more meaning preserving than the plan-guided models but simplifies less. Specifically, while the end-to-end model (LED$_{para}$) achieves the highest scores across all three faithfulness metrics, it also has the highest BLEU$_C$, produces outputs that are much longer than the references or any other system and achieves the

worst simplicity performance, both in terms of absolute (FKGL) and relative (ϵSLE$_{doc}$) criteria.

In contrast, the plan-guided models achieve faithfulness results not too far from LED$_{para}$ while still generating outputs much closer to the references in terms of length and BLEU$_C$.

Together these results suggest that plan-guidance allows models to avoid conservativity and make necessary edits to achieve high simplicity, although at the cost of some reduced faithfulness to the input.

Local vs Global Context. The simplification components of the plan-guided models each consider document context differently. While PG$_{Dyn}$ has no notion of document context, $\hat{O} \rightarrow$ LED$_{para}$ considers the local, token-level context of the surrounding paragraph, and $\hat{O} \rightarrow$ ConBART considers a high-level representation of more global context (SBERT encodings of 26 surrounding sentences).

The results indicate that the more local paragraph context leads to slight improvement in terms of faithfulness, but a reduction in simplicity performance. $\hat{O} \rightarrow$ ConBART achieves the best overall simplicity (FKGL) as well as ϵSLE$_{doc}$. Interestingly, both $\hat{O} \rightarrow$ ConBART and $\hat{O} \rightarrow$ LED$_{para}$ are much better than the other systems at simplifying to the highest level of simplicity (level 4 in Table 2), mirroring the human evaluation observations of Cripwell et al. (2023a) where plan-guided, context-aware systems appeared particularly strong in cases where major editing is required.

4.2. Out-of-Domain Performance

Out-of-domain performance is assessed by testing the Newsela-trained models on EW data. Results are shown in Tables 4 and 5. The difference in performances between in- and out-of-domain data with the same target reading level is shown in Appendix D.

End-to-End vs Planning. The end-to-end, Longformer model (LED$_{para}$) produces much shorter output documents than the plan-guided models — the opposite of what is seen for Newsela. As EW articles have longer paragraphs on average, this could be a result of over-fitting (i.e. being biased towards Newsela paragraph length observed during training and therefore generating overly short simplifications when applied to the longer EW texts. This could also be a result of over-deletion due to a lack of plan-guidance, as the other paragraph-level model ($\hat{O} \rightarrow$ LED$_{para}$) does not share this behaviour, potentially suggesting that planning also helps models better adapt to unseen domains.

On the other hand, $\hat{O} \rightarrow$ ConBART achieves the lowest faithfulness scores out of all dedicated systems, particularly on QAFactEval. As this model attends over a wider document context, it is possible that this increase in model variance could have led to some overfitting on the Newsela data. The ConBART network achitecture also contains additional layers that were not pretrained before finetuning on the Newsela dataset, further pointing towards potential overfitting. However, it is still close to PG$_{Dyn}$ on SummaC and ESA, while also achieving the best simplicity scores, which could mean the lower faithfulness scores are a result of the trade-off with simplicity. Without reference simplifications, it seems difficult to draw strong conclusions before examining human evaluation results.

Sentences vs Paragraphs. In terms of simplicity, the sentence-level models (PG$_{Dyn}$ and $\hat{O} \rightarrow$ ConBART) achieve much lower FKGL and ϵSLE$_{doc}$ than the two paragraph-level models. However, like on Newsela, they are markedly outperformed by the paragraph models on faithfulness metrics, particularly in terms of precision. While paragraph models produced longer outputs on in-domain data, they now produce shorter texts than sentence-level models, particularly in terms of the number of sentences. This could indicate potential conservativity with respect to sentence splitting, or an over-deletion of sentences.

5. Human Evaluation

To confirm system performance on the out-of-domain data, we also conduct a human evaluation. Due to the difficulty of comparing full documents, we follow existing document simplification work in evaluating at the paragraph-level (Cripwell et al., 2023a). We present annotators with a complex paragraph and an extract from a generated simplification corresponding to that paragraph. Evaluators are then asked to judge whether the generated text is fluent, consistent with, and simpler than the input.

We randomly sample 250 paragraphs from the test set that contain between 3-6 sentences. We consider the outputs from all tested systems and ask annotators to rate them on each dimension. We pose each as a binary (yes/no) question in order to avoid the inter-annotator subjectivity that is inherent when using a Likert scale. The proportion of positive ratings is used as the final score. Further details are given in Appendix C.

5.1. Human Evaluation Results

Table 6 shows the results of the human evaluation.

Despite achieving the best fluency, the end-to-end model (LED$_{para}$) underperforms on both meaning preservation and simplicity compared to the

System	SummaC ↑			QAFactEval ↑			ESA ↑			Length		BLEU$_C$
	P	R	F1	P	R	F1	P	R	F1	Tokens	Sents	
Input	-	-	-	-	-	-	-	-	-	866.9	38.6	-
Reference	0.61	0.47	0.53	3.86	3.02	3.39	0.59	0.47	0.52	671.5	42.6	44.6
PG$_{Dyn}$	0.65	0.47	0.55	3.95	3.10	3.47	**0.61**	0.48	0.53	667.2	42.6	47.6
LED$_{para}$	**0.66**	**0.52**	**0.58**	**4.00**	**3.29**	**3.61**	0.60	**0.51**	**0.55**	712.9	44.9	51.5
$\hat{O} \to$ LED$_{para}$	0.65	0.50	0.57	3.98	3.16	3.52	0.60	0.49	0.54	683.1	42.8	49.1
$\hat{O} \to$ ConBART	0.65	0.48	0.56	3.95	3.11	3.48	0.60	0.48	0.53	671.6	43.0	47.5

Table 2: **In-Domain Evaluation.** Faithfulness results for systems evaluated on the Newsela test set.

System	FKGL ↓	ϵ**SLE**$_{doc}$↓				
		1	2	3	4	**Total**
Reference	4.93	0.22 (1.12)	0.21 (1.97)	0.24 (3.11)	0.22 (3.84)	0.23
PG$_{Dyn}$	4.98	0.30 (1.24)	**0.22** (2.02)	0.22 (3.07)	0.32 (3.69)	0.26
LED$_{para}$	5.15	0.29 (1.06)	0.24 (1.92)	0.24 (2.97)	0.34 (3.67)	0.28
$\hat{O} \to$ LED$_{para}$	5.09	**0.26** (1.13)	0.24 (1.87)	0.23 (3.02)	0.30 (3.72)	0.26
$\hat{O} \to$ ConBART	**4.96**	0.28 (1.23)	**0.22** (1.98)	**0.21** (3.06)	**0.29** (3.73)	**0.25**

Table 3: **In-Domain Evaluation.** Simplicity results for systems evaluated on the Newsela test set. Columns 1-4 shows the results on the test sets for each level of simplicity, 4 being the level for highest degree of simplification. Numbers in parentheses are the raw SLE averages for each level.

System	SummaC ↑			QAFactEval ↑			ESA ↑			Length		BLEU$_C$
	P	R	F1	P	R	F1	P	R	F1	Tokens	Sents	
PG$_{Dyn}$	0.70	0.38	0.50	3.28	2.18	2.62	0.58	0.34	0.43	614.5	40.6	31.4
LED$_{para}$	**0.76**	0.39	0.51	**3.78**	2.11	2.71	**0.64**	0.35	0.45	513.7	32.5	27.4
$\hat{O} \to$ LED$_{para}$	0.73	**0.41**	**0.53**	3.61	**2.28**	**2.79**	0.62	**0.37**	**0.47**	601.5	37.0	32.0
$\hat{O} \to$ ConBART	0.68	0.38	0.49	3.10	2.06	2.48	0.57	0.33	0.42	598.4	40.5	29.5

Table 4: **OoD Evaluation.** Faithfulness and Conservativity results on the out-of-domain Wikipedia test set.

System	FKGL ↓	ϵ**SLE**$_{doc}$ ↓
Input	10.07	- (0.89)
PG$_{Dyn}$	4.72	**0.21** (2.92)
LED$_{para}$	4.92	0.29 (2.78)
$\hat{O} \to$ LED$_{para}$	5.02	0.31 (2.76)
$\hat{O} \to$ ConBART	**4.58**	**0.21** (3.00)

Table 5: **OoD Evaluation.** Simplicity results on the out-of-domain Wikipedia test set. Numbers in parentheses are the raw SLE$_{doc}$ averages (0-4).

best balance between the two.

plan-guided systems. This corroborates the automatic results in suggesting that planning can help systems to adapt better to unseen domains. The best overall results are achieved by PG$_{Dyn}$, but this can largely be attributed to its very high simplicity ratings as it falls below $\hat{O} \to$ ConBART in terms of meaning preservation. Although this once again points towards a trade-off between these two dimensions, $\hat{O} \to$ ConBART manages to achieve the

In contrast to what is observed via the automatic faithfulness metrics, sentence-level systems also appear to outperform paragraph-level ones. This could be a result of the paragraph models having a wider text window in which to make potential mistakes/hallucinations, whereas the sentence-level systems are more constrained. Further, the EW paragraphs are longer on average than the Newsela ones used to train these models, which could result in them failing to maintain all information when extending to longer input sizes (this is alluded to by the drop in the number of sentences in paragraph-level model outputs when moving to the EW domain, Table 4). In fact, many of the cases where the end-to-end model achieves lower faithfulness scores are the result of the model fully deleting the input paragraphs.

System	Flu	Faith	Simp	**Mean**
PG$_{\text{Dyn}}$	0.898	0.732	**0.820**	**0.817**
LED$_{\text{para}}$	**0.932**	0.632*	0.664*	0.743*
$\hat{O} \rightarrow$ LED$_{\text{para}}$	0.890	0.684	0.760	0.778*
$\hat{O} \rightarrow$ ConBART	0.890	**0.760**	0.764	0.805

Table 6: Human evaluation results on Wikipedia. Ratings significantly different from the highest rated system on each attribute are denoted with $*$ ($p < 0.05$). Significance was determined with a Student's t-test.

6. Conclusion

In this work, we conducted an investigation into the simplicity and the semantic adequacy of outputs from state-of-the-art document simplification systems. By leveraging recent advancements in automatic faithfulness evaluation for summarization and the reference-less evaluation of simplification, we were also able to carry out an analysis of simplification performance on out-of-domain data.

Separately assessing the models' ability to preserve meaning and simplify allowed for a detailed analysis of how these two dimensions vary across models and between evaluation settings (in- vs out-of-domain evaluation).

While a state-of-the-art end-to-end model appears to achieve the best in-domain faithfulness results, it is also much more conservative than plan-guided systems, generating outputs with low simplicity. Plan-guided systems also appear better at adapting to unseen domains, but we continue to observe a general trade-off between faithfulness and simplicity. Consideration of this trade-off using only automatic metrics is challenging for out-of-domain settings as it is unclear what exactly constitutes a sufficient level of faithfulness without having references to use as a baseline.

Human evaluation results indicate that plan-guided, sentence-level simplification systems produce outputs with the highest meaning preservation when switching domains — a phenomenon not captured by the automatic faithfulness metrics. This highlights the need for further exploration into automatic methods of faithfulness evaluation for simplification systems. We hope our work motivates future investigations into more thorough simplification evaluation strategies and the development of training methods and architectures that can allow simplification systems to effectively adapt to unseen domains, rather than further optimizing performance on the most popular datasets.

7. Limitations

Paragraph-Level Human Evaluation Following previous document simplification studies, our human evaluation was performed using only paragraph-level extracts from simplified documents, rather than the entire documents themselves. This was done to limit the complexity of each human evaluation task as full-document annotation would likely be challenging for many workers. Because of this, it is possible that certain long-distance discourse phenomena are not properly considered during the evaluation. For example, important information may be excluded from a specific output paragraph, but may actually be present in a different part of the document. However, given the iterative nature of most systems tested, such cases should be uncommon. This shift in granularity also makes it difficult to compare automatic and human evaluation results as we cannot directly compute correlations between them.

English Only The datasets and systems we investigate are applicable only to English. It is possible that many of the insights from the study could equally apply in the case of other languages; however, independent analyses would need to be carried out to confirm this. Additionally, many of the evaluation metrics used (e.g. both simplification metrics – FKGL and SLE) are built specifically with English text in mind and therefore would not easily be adaptable to equivalent evaluations of simplification in other languages.

8. Acknowledgements

9. Bibliographical References

Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. 2017. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Confer-*

ence on Natural Language Processing (Volume 1: Long Papers), pages 295–305, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2019. Cross-sentence transformations in text simplification. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 181–184, Florence, Italy. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, pages 1–29.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2021. Discourse-based sentence splitting. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 261–273, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2022. Controllable sentence simplification via operation classification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2091–2103, Seattle, United States. Association for Computational Linguistics.

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.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023c. Simplicity level estimate (SLE): A learned reference-less metric for sentence simplification.

In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12053–12059, Singapore. Association for Computational Linguistics.

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.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Juliette Faille, Albert Gatt, and Claire Gardent. 2021. Entity-based semantic adequacy for data-to-text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1530–1540, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic multi-level multi-task learning for sentence simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 462–476, Santa Fe, New Mexico, USA. 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.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73, Kaohsiung, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

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. Technical report, Naval Technical Training Command Millington TN Research Branch.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. Ffci: A framework for interpretable automatic evaluation of summarization. *Journal of Artificial Intelligence Research*, 73:1553–1607.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul 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 (Volume 1: Long Papers)*, pages 6365–6378, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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.

Zhe Lin, Yitao Cai, and Xiaojun Wan. 2021. Towards document-level paraphrase generation with sentence rewriting and reordering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1033–1044, Punta Cana, Dominican Republic. 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, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Kshitij Mishra, Ankush Soni, Rahul Sharma, and Dipti Sharma. 2014. Exploring the effects of sentence simplification on Hindi to English machine translation system. In *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, pages 21–29, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. 2010. Entity-focused sentence simplification for relation extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 788–796, Beijing, China. Coling 2010 Organizing Committee.

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.

Shashi Narayan, Claire Gardent, Shay B. Cohen, and Anastasia Shimorina. 2017. Split and

rephrase. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 606–616, Copenhagen, Denmark. Association for Computational Linguistics.

Christina Niklaus, Bernhard Bermeitinger, Siegfried Handschuh, and André Freitas. 2016. A sentence simplification system for improving relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 170–174, Osaka, Japan. The COLING 2016 Organizing Committee.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2019. Transforming complex sentences into a semantic hierarchy. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3415–3427, Florence, Italy. 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.

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

G.H. Paetzold and L. Specia. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593. © 2017 AI Access Foundation, Inc. This is an author produced version of a paper subsequently published in Journal of Artificial Intelligence Research. Uploaded in accordance with the publisher's self-archiving policy.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. 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.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Max Schwarzer and David Kauchak. 2018. Human evaluation for text simplification: The simplicity-adequacy tradeoff. In *SoCal NLP Symposium*.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021a. Questeval: Summarization asks for fact-based evaluation.

Thomas Scialom, Louis Martin, Jacopo Staiano, Éric Villemonte de la Clergerie, and Benoît Sagot. 2021b. Rethinking automatic evaluation in sentence simplification.

Advaith Siddharthan. 2003. Preserving discourse structure when simplifying text. In *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003) at EACL 2003*, Budapest, Hungary. Association for Computational Linguistics.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online. Association for Computational Linguistics.

Sanja Štajner and Maja Popovic. 2016. Can text simplification help machine translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 162–173, Melbourne, Australia. Association for Computational Linguistics.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Renliang Sun, Zhe Lin, and Xiaojun Wan. 2020. On the helpfulness of document context to sentence

simplification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1411–1423, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Sentence simplification with memory-augmented neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 79–85, New Orleans, Louisiana. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Sandra Williams, Ehud Reiter, and Liesl Osman. 2003. Experiments with discourse-level choices and readability. In *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003) at EACL 2003*, Budapest, Hungary. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

K. Woodsend and Mirella Lapata. 2011. Wikisimple: Automatic simplification of wikipedia articles. In *AAAI*.

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.

Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. Controllable text simplification with

deep reinforcement learning. 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 2: Short Papers)*, pages 398–404, Online only. Association for Computational Linguistics.

Bohan Zhang, Prafulla Kumar Choubey, and Ruihong Huang. 2022. Predicting sentence deletions for text simplification using a functional discourse structure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 255–261, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9709–9716.

A.   WikiData Article Annotation

We selected Wikipedia articles to cover a range of diverse categories (shown in Table 7). However, we did not obverse any major performance differences between categories, apart from slightly lower scores for articles from more specialized categories (e.g. Science and Industry).

B.   SLE In-Group Distributions

Figure 1 shows the distribution of SLE scores predicted for reference sentences belonging to each original Newsela reading level group. We can see that although the mean is approximately equal to the reading level, there is substantial diversity within each group.

C.   Human Evaluation Details

Human judgements were obtained via the Amazon Mechanical Turk crowdsourcing platform. Annotators were sourced from majority English speaking countries (AU, CA, GB, IE, NZ, US) and were paid $0.18 USD per evaluation. According to preliminary tests, under this scheme participants earn approximately $16.2 USD per hour — which is higher than the minimum hourly wage of all countries. The form and instructions presented to human evaluators is shown in Figure 2.

Category	Sub-Category	Count
Biographical	Human	500
	Musical Group	250
	Fictional Human	250
Location	City	250
	Village	250
	Commune of France	250
	City in the United States	250
Media	Film	250
	Video Game	250
	Literary Work	250
	Television Series	250
Science	Taxon	250
	Class of Disease	250
	Chemical Compound	250
	Class of Anatomical Entity	250
Industry	Business	250
	Profession	250
	Organization	250
	Automobile Model	250

Table 7: Distribution of Wikipedia article categories.
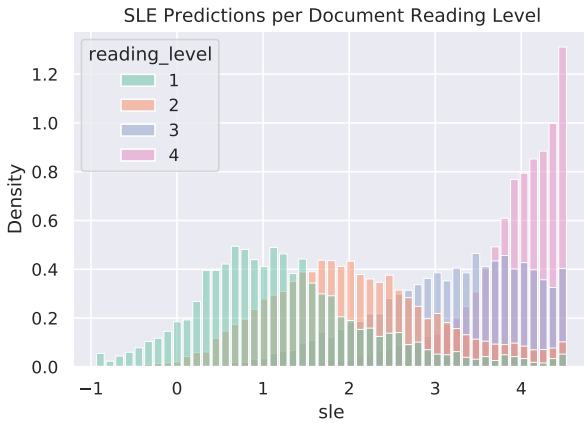
Figure 1: Distribution of SLE scores for reference sentences within each Newsela reading level group.

D. Extra Evaluation Results

Table 8 shows the relative change in automatic evaluation results when moving from in- to out-of-domain data (using the same target reading level of 3).

Figure 2: Submission form presented to annotators during the human evaluation.

System	SummaC ↑		QAFactEval ↑		ESA ↑		BLEU$_C$	FKGL ↓	ϵSLE$_{doc}$ ↓
	P	R	P	R	P	R			
PG$_{Dyn}$	0.04	-0.11	-0.66	-0.89	-0.02	-0.13	-14.09	-0.11	0.17 (-0.25)
LED$_{para}$	0.09	-0.13	-0.19	0.05	-0.15	-0.09	-21.0	-0.09	0.22 (-0.28)
$\hat{O} \rightarrow$ LED$_{para}$	0.07	-0.1	-0.33	-0.84	0.03	-0.11	-14.55	0.06	0.24 (-0.32)
$\hat{O} \rightarrow$ ConBART	0.02	-0.11	-0.86	-1.01	-0.03	-0.15	-16.04	-0.27	0.09 (-0.14)

Table 8: Difference in results for target-level 3 when moving from the in-domain Newsela to the out-of-domain Wikipedia test set.

Malmon: A Crowd-Sourcing Platform for Simple Language

Helgi Björn Hjartarson, Steinunn Rut Friðriksdóttir

University of Iceland

Sæmundargata 2, Reykjavík

hbh42@hi.is, srf2@hi.is

Abstract

This paper presents a crowd-sourcing platform designed to address the need for parallel corpora in the field of Automatic Text Simplification (ATS). ATS aims to automatically reduce the linguistic complexity of text to aid individuals with reading difficulties, such as those with cognitive disorders, dyslexia, children, and non-native speakers. ATS does not only facilitate improved reading comprehension among these groups but can also enhance the preprocessing stage for various NLP tasks through summarization, contextual simplification, and paraphrasing. Our work introduces a language independent, openly accessible platform that crowdsources training data for ATS models, potentially benefiting low-resource languages where parallel data is scarce. The platform can efficiently aid in the collection of parallel corpora by providing a user-friendly data-collection environment. Furthermore, using human crowd-workers for the data collection process offers a potential resource for linguistic research on text simplification practices. The paper discusses the platform's architecture, built with modern web technologies, and its user-friendly interface designed to encourage widespread participation. Through gamification and a robust admin panel, the platform incentivizes high-quality data collection and engagement from crowdworkers.

Keywords: automatic text simplification, crowd-sourcing platform, gamification

1. Introduction

Automatic text simplification (ATS) is a Natural Language Processing (NLP) task where the linguistic complexity of text is reduced in order to facilitate reading comprehension without losing its original information. This is particularly helpful for readers with low literacy, for instance due to cognitive disorders or dyslexia, children and non-native speakers learning a new language. Text simplification has also been shown to improve results when used at the preprocessing stage for other NLP tasks. The ATS process varies in nature; for instance, it can involve summarizing the text to remove any redundant information, simplifying the context of the text, or paraphrasing it so that key points are emphasized. Usually, ATS involves two steps, lexical simplification and syntactic simplification, where the former focuses on reducing complexity by replacing complex words with simpler synonyms and the latter reduces grammatical complexity, such as by removing or simplifying subordinate clauses that may be difficult for readers to comprehend.

One of the challenges of ATS is identifying the complexity of a given text and deciding the best way to reduce it. In order to train models that can perform this task automatically, it is fundamental to have access to extensive parallel corpora in which complex sentences are paired with their simplified versions. Various ATS corpora exist for high-resource languages like English (see Al-Thanyyan and Azmi (2021) for an overview) but far fewer for lower-resource languages. Recent methodologies in ATS are largely data-driven where simplification rules are inducted from the data. It is therefore

crucial to create simple, ready-to-use tools that researchers can use for their ATS data collection. In our work, we introduce Malmon, a crowdsourcing platform that can be used to collect training data for text simplification models. The platform is language independent, openly accessible[1] and easily adaptable for researchers wanting to collect their own data.

2. Collecting ATS Data

As creating an ATS corpus from scratch can be prohibitively expensive, many attempts at automating the data collection have been made. In their paper, Holmer and Rennes (2023) describe the creation of a pseudo-parallel ATS corpus for Swedish. They then fine-tune a BART model for sentence simplification on their data with promising results. Ormaechea and Tsourakis (2023) use a combination of automatic methods and manual annotation to align Wikipedia articles in French to their counterparts in the simplified version Vikidia. Similarly, Dmitrieva and Konovalova (2023) use Sentence Transformers to measure the similarity between Finnish news articles and their simplified counterparts to create sentence pairs which are then manually reviewed.

These semi-automatic methods can potentially speed up the data collection process and consequently reduce the cost required. However, they

---

[1]The source code for the platform is available on `https://github.com/polarparsnip/malmon`. We encourage other researchers to use and adapt this code to their needs.

are not without their setbacks. Holmer and Rennes (2023) mention some problems relating to the automating process, for instance that sentences with named entities often get aligned with sentences containing completely different entities. Another approach is to have expert annotators manually simplify sentences or excerpts of text. This approach was used for the Newsela corpus, whose 2016-01-29.1 version consists of 1,911 news articles and up to 5 simplified versions written by trained professionals (Paetzold and Specia, 2017). Similarly, the Alector corpus, intended to research the effectiveness of simplifying text for dyslexic children, was constructed by a group of experts who manually simplified 79 literary and scientific texts commonly used in French schools (Gala et al., 2020).

A similar approach, and the one we advocate here, is to collect ATS data using crowdsourcing. Katsuta and Yamamoto (2018) crowdsourced a parallel corpus for Japanese. Crowdworkers were asked to limit themselves to a core vocabulary of 2000 words so that the resulting simplifications was at an everyday conversational level. Appendix A shows the proposed guidelines for crowdworkers using our platform to simplify Icelandic text. While our guidelines do not include a core vocabulary similar to that of Katsuta and Yamamoto (2018), we encourage our users to avoid rare or complex words. We note that this frame of reference can be changed at will so that it better suits the needs of other researchers using the platform. We also note that our platform can be used both for crowdsourcing simplified versions of text examples directly (whether the data collection process is to be open to the public or conducted by expert crowdworkers) and for manually reviewing sentence alignments created with automatic methods. If the former is chosen, the resulting data could also be used in linguistic research on the way people simplify or paraphrase complex text, similarly to what was presented in Amancio and Specia (2014).

3. The Platform

3.1. Motivation

As previously mentioned, ATS can greatly benefit individuals with low literacy levels. Azab et al. (2015) used ATS methods to design a browser extension to help students learning English as a second language by annotating and substituting difficult words with simpler synonyms. A similar platform for people with aphasia was designed by Devlin and Unthank (2006), which presents users who have difficulty understanding or remembering a particular word with another word that has the same meaning but is more common or easier to understand. Javourey-Drevet et al. (2022) designed

an iPad application that presented French children with original and simplified versions of informative and narrative texts. Their results indicate that the simplified texts benefited poor readers and children with weaker cognitive skills, increasing their reading fluency and text comprehension.

ATS methods can also facilitate comprehension of particularly difficult or domain-specific text. This has been prominent within the medical field. Kushalnagar et al. (2018) used ATS methods to simplify information about breast cancer in order to improve comprehension for Deaf people who use American Sign Language. Phatak (2023) proposed several ATS methods to simplify complex biomedical literature in English for the general public and Cardon and Grabar (2020) did the same for French. Similarly, Truică et al. (2023) presented SimpLex, a software that uses ATS methods to simplify medical text in English for the general public. ATS systems have also been used as a preprocessing task for other NLP systems to improve their results. In their paper, Van et al. (2021) show that augmenting data with ATS to provide additional information during training significantly improves performances of various text classification and relation extraction models.

However, to be able to create such systems, it is fundamental that sufficient parallel data exist. TS-ANNO, introduced by Stodden and Kallmeyer (2022), is a crowd-sourcing platform that can be used for a variety of tasks related to ATS. Our platform, however, focuses solely on generating parallel complex-simple sentence pairs. It offers a simple, easy to use way of collecting the data via crowdsourcing. As discussed in Section 3.3, the users of our platform are presented with three options only, to simplify, verify or download the resulting data. This straight-forward navigation leaves little room for confusion as to what is expected of the users, particularly crowdworkers that might not have previous experience with work in NLP. Our platform may prove especially useful for lower-resource languages where the number of expert annotators might be scarce and priority must be placed on straight-forward solutions aimed at the general public. Researchers interested in using the platform can access the source code and modify it freely.

3.2. Technical Information

The platform, which we call Malmon, is built as a full-stack website utilizing a SQL database set up with PostgreSQL to store all sentences and user data. The back-end web server is built in Javascript utilizing Express.js to handle http connections and the front-end of the website is made using Next.js, which is a React-based Javascript framework. When a user is logged in, the server

checks if they are an admin or a general user and redirects them to the appropriate section of the site. The server makes sure that general users can't access any of the admin areas and that a logged in admin has access to all necessary admin functionalities.

User registration and login on the site are straightforward. Users are required to enter a username, e-mail, and password when they register an account. Since user accounts are tied to each individual user's progress, it is important to be able to recover an account in the event that a user loses their password. Tying e-mail addresses to user accounts could also possibly help to distinguish between different users if the need arises, for instance to detect outliers that may be the result of system spamming.

The language of the platform can be chosen with one environment variable when setting it up for hosting, with current supported languages being: Icelandic, English, Norwegian, Danish, Swedish, Faroese, and Italian[2]. Additionally, adjusting existing language settings or adding more supported languages is a straightforward process that only involves modifying one file.

3.3. Functionality

When not logged in, site visitors are presented with a simple website with a front page detailing how the platform works. In the footer, they have the option to log in or to register a new account. In the navigation menu, they again have two options: to log in and to get data. The latter option is the only functionality available to users when they are not logged in, apart from actions such as creating a new account or signing in. This option allows visitors to the platform to fetch the current state of the resulting dataset as either a JSON or CSV file, with the files containing complex-simplified sentence pairs. This means that the dataset being collected at each given time is open to everyone who wishes to use it for model training or other similar purposes.

Once logged in, users still see the option to download the dataset but are also presented with options in the navigation menu that are only visible to logged in users. They now see options for navigating to their *account* section and a *score-table* section, both of which will be covered in more detail in Section 3.4. They are also presented with the option to go to an FAQ page detailing the guidelines for submitting sentences and the options to go to the *simplify* (see Figure 1) and the *verify* sections (see Figure 2). These two last sections are the

---

[2]Note that the proposed guidelines are only available in Icelandic and English as of this publication. The other languages have been translated using ChatGPT and thus require further review.

main areas in which users contribute to the dataset being collected on the platform.

Once users navigate to the *simplify* section, they are given a random sentence from the database and a fast and simple CAPTCHA-like task to verify that they are a human and not a bot. After completing the task, they are presented with an input text field in which they can enter a simplified version of the sentence they were given. This CAPTCHA-like task makes sure sentences are being submitted by humans and prevents bots and/or other spam methods from being able to submit sentences. Once the user feels like they have entered a good enough simplified version of the sentence they received, they can click submit and the sentence is then saved in the database.

When a simplified sentence submitted by a user is saved in the database, it is marked as unverified and is therefore not yet part of the dataset which can be downloaded. To be included in the dataset, a submitted sentence must first be verified by a separate user on the platform which is the purpose of the *verify* section. Once users navigate to that section, they are again given a sentence along with a simplified version of that sentence submitted by another user. After completing the same CAPTCHA-like task, they are presented with two buttons, a "confirm" button and a "reject" button. If a user feels like the simplification submitted by another user is a good representation of the original sentence, they can approve it by pressing the "confirm" button. If they feel like the simplified version is not a good representation of the original sentence, they can reject it by pressing the "reject" button. If a simplified sentence is confirmed by the user, it is marked as *verified* and is now part of the collected dataset which people can download. If it is rejected, it is taken out of circulation and will no longer appear to users.

These two sections form the data collection portion of the platform and are the main ways users interact with the website.

3.4. Gamification

Crowdsourcing is a data collection process whereby content is obtained by having a group of people use their leisure time to make their contributions at a minimal cost. As crowdsourcing is generally performed by non-expert volunteers, there needs to be some incentive for participation, as what is considered interesting from a scientific perspective may not be enjoyable for the general public. One way to achieve this is through gamification, which incorporates video game elements to improve user experience in a non-game service which in turn can enhance user engagement (see for instance Deterding et al., 2011; Quecke and Mariani, 2021). Competitive game elements, such

Figure 1: The simplification page of the platform. Users are presented with a complex sentence and are asked to write a simplified version of the sentence.

as points and immediate performance feedback, have been found to positively affect crowdworker motivation and, consequently, participation (Yang et al., 2021).

In our site's navigation menu, all users can access a *score-table* that details which users have contributed the most in terms of submitted simplified sentences and the amount of submitted sentence verifications. Users are ranked based on the lower of the two aforementioned attributes, so if a user has, for example, submitted 33 simplified sentences and verified 22 sentences, they will be ranked based on the number 22. This guarantees that users can't focus exclusively on one method in order to receive a good score, instead providing incentive to contribute to both areas in order to boost their ranking on the scoreboard.

In the *account* section, users can view their information which includes their username, as well as how many sentences they have submitted and how many sentence verifications they have completed. Also contained in the *account* section is a digital pet tied to their account that grows according to their sentence submissions and sentence verifications, based on the same system as the scoreboard. When a user has only just created an account and not yet taken any action, the digital pet appears as an egg. As they contribute to the platform, their pet evolves into higher stages similar to creatures in franchises like Pokémon or Digimon.

These features encourage and reward users for their contributions to the platform and can act as a basis for other reward systems which could then be integrated with them. For instance, the fully evolved pet could be accompanied by a lottery ticket in the form of a QR code where a diligent user gains the

chance to win a real-world price. Adding an image to the last stage of the pet is straightforward and only involves adding two environment variables when the platform is set up in hosting. Since users have to confirm they are human before submitting sentences, it will be difficult to try to cheat the system to gain whatever rewards are in place.

3.5. Admin Functionality

One of the key focus points for the platform was to have extensive and user friendly admin functionality. When an admin is logged in, they have instant access to the admin dashboard. This dashboard allows an admin to access the editor areas for sentences, simplified sentences, and users.

In the *sentences* area, an admin can view all the saved sentences from the database. The sentences are displayed 10 at a time with the option to move forward to the next 10 sentences. Each sentence is displayed individually with an option to update that specific sentence or delete it, so if an admin notices a sentence containing errors or one that should not be there, they have the option to react accordingly. There is also a form on the page where an admin can register a new sentence and add it to the list of complex sentences.

In the *simplified sentences* area, an admin can view all the simplified sentences that have been submitted, 10 at a time. The simplified sentences are displayed individually with information on whether the sentence has been confirmed or rejected by another user. They are also accompanied by options for an admin to either delete the sentence or delete a user rejection. An admin may delete a user rejection when they feel like a simplified sentence was unjustly rejected, and so by
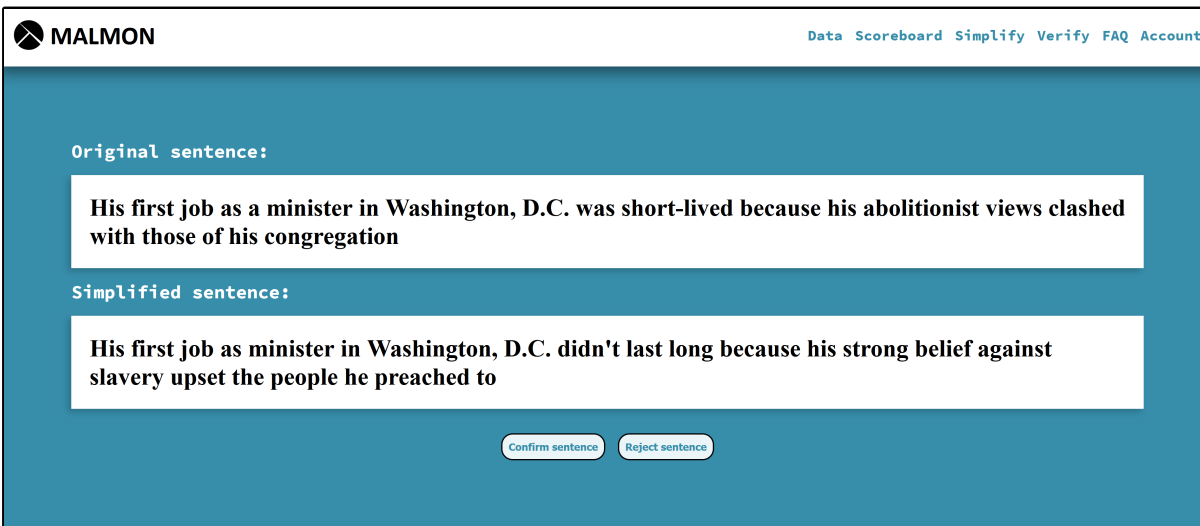
Figure 2: The verification page of the platform. Users are presented with a complex and simple sentence pair and are asked to verify the quality of the simplified sentence.

deleting the rejection it re-enters circulation and awaits confirmation by a user.

In the *user* area, an admin can view a list of all registered users on the website, again 10 at a time. The username and registration date of each user is displayed, as well as the number of simplified sentences the user has submitted and the number of verifications the user has completed. For each listed user, an admin is given the option to delete that particular user.

Then, in the *upload* area, an admin can upload a CSV file containing a list of complex sentences they wish to add to the database of the platform. These sentences will then be added to the collection of complex sentences on the platform that users are presented with.

In addition to these functionalites, an admin can also access all normal user pages and interact with the page as a user would.

4. Conclusions

We present Malmon, a data collection platform intended for crowdsourcing complex-simple sentence pairs that can be used to train automatic text simplification or ATS systems. The source code for the platform is available on Github and can be freely adapted to the needs of individual researchers. We have discussed potential use of such data, particularly in aiding people with low literacy levels, whether due to reading comprehension or cognitive disabilities, second language learners, or children. ATS can also benefit the general public when used to simplify complex, domain-specific texts, such as in the medical field, or as a preprocessing step to increase the performance of other NLP systems.

The platform combines data collection and data verification and brings it all together in combination with a simple reward system. Users can freely and easily submit simplified sentences and verify sentences from other users. Contributing to the platform evenly in both submissions and verifications increases a user's score on the scoreboard and in their account. Each user additionally receives a digital pet that grows in accordance with their score. The reward system on the site can be used by itself but it can also easily be built upon or combined with other reward systems to further incentivize user participation in the crowdsourcing process. One example of this would be a lottery-based system where users can participate in the lottery by completing the evolution of their digital pet, which can only be done by participating on the site. This could for example be done by adding a one-time-use QR code adjacent to the final stage of the digital pet.

Even though it is easy to implement other ideas with the existing reward system framework, future improvements could include additional admin functionality such as a menu for choosing reward system options and combinations. This would allow anyone to choose their preferred method of crowdsourcing without interacting with the technical side of the platform. Other possible additions to the platform worth mentioning include increasing the digital pet functionality, allowing more interaction between users, and possibly expanding the platform to allow collection of other types of data.

We hope that this platform can benefit other researchers interested in ATS, particularly those working with low-resource languages.

19

6. Bibliographical References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated Text Simplification: A Survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Marcelo Amancio and Lucia Specia. 2014. An analysis of crowdsourced text simplifications. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 123–130, Gothenburg, Sweden. Association for Computational Linguistics.

Mahmoud Azab, Chris Hokamp, and Rada Mihalcea. 2015. Using Word Semantics to Assist English as a Second Language Learners. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 116–120.

Rémi Cardon and Natalia Grabar. 2020. French Biomedical Text Simplification: When Small and Precise Helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716. International Committee on Computational Linguistics.

Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. 2011. Gamification. Using Game-Design Elements in Non-Gaming Contexts. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 2425–2428.

Siobhan Devlin and Gary Unthank. 2006. Helping Aphasic People Process Online Information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 225–226.

Anna Dmitrieva and Aleksandra Konovalova. 2023. Creating a Parallel Finnish–Easy Finnish Dataset from News Articles. In *1st Workshop on Open Community-Driven Machine Translation*, page 21.

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.

Daniel Holmer and Evelina Rennes. 2023. Constructing Pseudo-parallel Swedish Sentence Corpora for Automatic Text Simplification. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 113–123.

Ludivine Javourey-Drevet, Stéphane Dufau, Thomas François, Núria Gala, Jacques Ginestié, and Johannes C Ziegler. 2022. Simplification of Literary and Scientific Texts to Improve Reading Fluency and Comprehension in Beginning Readers of French. *Applied Psycholinguistics*, 43(2):485–512.

Akihiro Katsuta and Kazuhide Yamamoto. 2018. Crowdsourced corpus of sentence simplification with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Poorna Kushalnagar, Scott Smith, Melinda Hopper, Claire Ryan, Micah Rinkevich, and Raja Kushalnagar. 2018. Making Cancer Health Text on the Internet Easier to Read for Deaf People who Use American Sign Language. *Journal of Cancer Education*, 33:134–140.

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 edition of the Swiss Text Analytics Conference*, pages 30–40. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2017. Lexical Simplification with Neural Ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 34–40.

Atharva Phatak. 2023. *Medical Text Simplification: Bridging the Gap Between Medical Research and Public Understanding*. Ph.D. thesis, Lakehead University.

Anna Quecke and Ilaria Mariani. 2021. How to Design Taskification in Video Games. A Framework for Purposeful Game-Based Crowdsourcing. In *CEUR Workshop Proceedings*, volume 2934, pages 1–10. CEUR-WS.

Regina Stodden and Laura Kallmeyer. 2022. 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*, pages 145–155, Dublin, Ireland. Association for Computational Linguistics.

Ciprian-Octavian Truică, Andrei-Ionuţ Stan, and Elena-Simona Apostol. 2023. SimpLex: A Lexical Text Simplification Architecture. *Neural Computing and Applications*, 35(8):6265–6280.

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*.

Congcong Yang, Hua Jonathan Ye, and Yuanyue Feng. 2021. Using Gamification Elements for Competitive Crowdsourcing: Exploring the Underlying Mechanism. *Behaviour & Information Technology*, 40(9):837–854.

A. Proposed Guidelines

The following guidelines, translated to English, are aimed at crowdworkers using the platform to simplify Icelandic text. The guidelines can be changed freely by researchers using the platform so that the better suit the needs of their languages. We also include examples in English for clarity purposes.

Your task is to simplify the proposed sentences in such a way that the resulting text is better suited for readers with language difficulties (such as people that have dyslexia or aphasia), L2 speakers and/or children. When simplifying the sentences, please keep the following in mind:

- The simplified sentence should only contain common, everyday vocabulary. Please avoid specialized or uncommon words as much as possible unless the sentence explicitly explains the meaning of such words. If you are not sure whether or not the word you are using is uncommon, please refer to the following website: `https://ordtidni.arnastofnun.is/`. At the bottom of the page, you will find a frequency list for words in their base form as well as for their conjugations. You can also search for a specific word using the search bar above. If the base form of a given word has a frequency below 30.000, it should probably be avoided.

- Drop unnecessary information. The simplified sentences should maintain the meaning of the original sentences but non-important information can be omitted.
  `Example:` Snæfell er hæsta staka fjall landsins, 1833 m yfir sjó. −> Snæfell er hæsta fjall Íslands. Það er 1833 metra hátt.

  - `An example in English:` Mount Everest, is Earth's highest mountain above sea level, located in the Mahalangur Himal sub-range of the Himalayas. −> The tallest mountain in the world is Mount Everest. It is located in the Himalayas.

- Avoid unnecessary verbosity.
  `Example:` Samkvæmt ráðleggingum stofnunarinnar er mælt með því að börn hreyfi sig a.m.k. 60 mínútur á dag. −> Stofnunin mælir með því að börn hreyfi sig a.m.k. 60 mínútur á dag.

  - `An example in English:` According to the guidelines of the institution, it is recommended that children exercise for at least 60 minutes per day. −> The instutution recommends that children exercise for at least 60 minutes per day.

- Simplify sentences so that they contain as few subordinate clauses as possible. If the original sentence contains such clauses, the simplified version should rather contain multiple sentences, separated by a period.
  `Example:` Hérna er fjallið *sem mér þótti svo vænt um*. −> Hérna er fjallið. Mér þótti vænt um það.

  - `An example in English:` Watching Star Wars, which has lots of special effects, is my favorite thing to do. −> I love watching Star Wars. It has lots of special effects.

- Avoid unusual word order and stylization. Simplified sentences should preferably be in the active voice and the indicative mood.
  `Example:` Gagnrýnin *sem fram hefur komið* á fullan rétt á sér. −> Gagnrýnin *sem hefur komið fram* á fullan rétt á sér.

  - `An example in English:` Across the river and through the woods go Ella and Larry. −> Ella and Larry go across the river and through the woods.

Automatic Generation and Evaluation of Reading Comprehension Test Items with Large Language Models

Andreas Säuberli, Simon Clematide
Department of Computational Linguistics, University of Zurich
{andreas, simon.clematide}@cl.uzh.ch

Abstract

Reading comprehension tests are used in a variety of applications, reaching from education to assessing the comprehensibility of simplified texts. However, creating such tests manually and ensuring their quality is difficult and time-consuming. In this paper, we explore how large language models (LLMs) can be used to generate and evaluate multiple-choice reading comprehension items. To this end, we compiled a dataset of German reading comprehension items and developed a new protocol for human and automatic evaluation, including a metric we call *text informativity*, which is based on guessability and answerability. We then used this protocol and the dataset to evaluate the quality of items generated by Llama 2 and GPT-4. Our results suggest that both models are capable of generating items of acceptable quality in a zero-shot setting, but GPT-4 clearly outperforms Llama 2. We also show that LLMs can be used for automatic evaluation by eliciting item reponses from them. In this scenario, evaluation results with GPT-4 were the most similar to human annotators. Overall, zero-shot generation with LLMs is a promising approach for generating and evaluating reading comprehension test items, in particular for languages without large amounts of available data.

Keywords: reading comprehension, automatic item generation, question generation, evaluation, large language models

1. Introduction

Assessing reading comprehension is not only a crucial part of language testing in an educational context, it is also useful in many scenarios related to evaluation in natural language processing (NLP) – for example, when evaluating the comprehensibility of automatically simplified texts (Alonzo et al., 2021; Leroy et al., 2022; Säuberli et al., 2024), benchmarking the natural language understanding capabilities of large language models (LLMs) (Lai et al., 2017; Bandarkar et al., 2023), or determining factual consistency in text summarization (Wang et al., 2020; Manakul et al., 2023). Multiple-choice tests are the most common way of assessing human reading comprehension because administering and grading them is simple. However, designing good multiple-choice reading comprehension (MCRC) items which actually test comprehension (as opposed to other things like the test taker's world knowledge or the readability of the item itself) is notoriously difficult (Jones, 2020; Jeon and Yamashita, 2020). Given the recent advancements in the zero-shot capabilities of LLMs (Wei et al., 2021; Ouyang et al., 2022), automatically generating MCRC items appears to be a promising option.

Evaluating MCRC items poses an additional challenge. While test developers in language assessment rely on extensive expert reviews and large pilot studies to determine the quality of test items (Green, 2020; Gierl et al., 2021), these evaluation methods are not practicable for fast-paced and iterative development of NLP models. In NLP research,
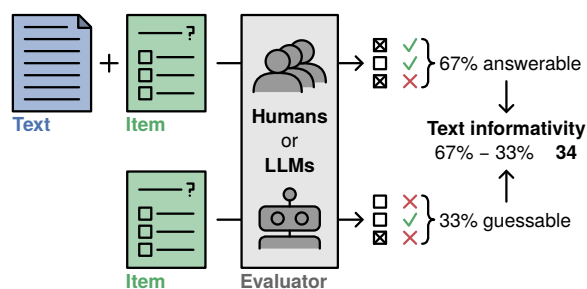
Figure 1: Our evaluation protocol measures the answerability and guessability of MCRC items by letting high-performing humans or LLMs respond to them with and without seeing the text. The text informativity metric is the difference between answerability and guessability and denotes to what degree the text informs the item responses.

there is still no consensus on evaluation methodologies and a lack of valid metrics for automatic evaluation (Circi et al., 2023; Mulla and Gharpure, 2023).

In this paper, we address both the generation and the evaluation of MCRC items in the German language. We propose a new evaluation metric called **text informativity** combining answerability and guessability and use it for human and automatic evaluation. Our main contributions can be summarized as follows:

1. We compile a dataset of German MCRC items from online language courses.

2. We present a protocol for human evaluation

of MCRC items and use it to evaluate items generated by two state-of-the-art LLMs.

3. We demonstrate that the same protocol can also be used for automatic evaluation by replacing the human annotators with LLMs.

2. Background and Related Work

2.1. Automatic Item Generation

Automatic item generation (AIG) has been of interest in educational and psychological assessment for several decades (Haladyna, 2013). Until now, rule-based approaches based on manually written templates have been used in these fields (Lai and Gierl, 2013; Circi et al., 2023). Recent NLP research introduced neural approaches and especially pre-trained transformer models to generate comprehension questions (Yuan et al., 2017; Du et al., 2017; Zhou et al., 2017; Gao et al., 2019; Lopez et al., 2020; Berger et al., 2022; Rathod et al., 2022; Ghanem et al., 2022; Uto et al., 2023; Fung et al., 2023), multiple-choice distractors (Maurya and Desarkar, 2020; Shuai et al., 2021; Xie et al., 2022), or entire MCRC items based on a text in an end-to-end fashion (Jia et al., 2020; Dijkstra et al., 2022). Several works have reported promising results using zero-shot or few-shot prompting of LLMs (Attali et al., 2022; Raina and Gales, 2022; Kalpakchi and Boye, 2023). While most previous research has focused on the English language, our work is the first to evaluate LLMs for zero-shot generation of German reading comprehension items.

2.2. Evaluation of Generated Items

Most NLP works on question generation and AIG report reference-based similarity metrics borrowed from machine translation or text summarization, such as BLEU, ROUGE, and METEOR (Amidei et al., 2018; Circi et al., 2023; Mulla and Gharpure, 2023). These metrics are unsuitable for generating MCRC items because similarity does not imply high quality for this task. Human evaluation is mostly done by asking experts or crowd workers to rate generated items in terms of fluency, relevance, difficulty, and other categories (e.g. Jia et al., 2020; Gao et al., 2019; Ghanem et al., 2022; Uto et al., 2023). Attali et al. (2022) is a notable exception, conducting both expert reviews and a large-scale pilot study to evaluate LLM-generated test items.

Several studies have examined the possibility of using question answering (QA) models to evaluate generated items instead of human test takers. Most commonly, this is done by letting a QA model respond to the items and equating a high response accuracy to good answerability (Yuan et al., 2017; Klein and Nabi, 2019; Shuai et al., 2021; Rathod

et al., 2022; Raina and Gales, 2022; Uto et al., 2023). In addition to answerability, Berzak et al. (2020), Liusie et al. (2023), and Raina et al. (2023) also measured guessability by benchmarking the model's ability to answer the items without seeing the text. Finally, Lalor et al. (2019) and Byrd and Srivastava (2022) used large ensembles of models responding to human-written items and applied item response theory to determine psychometric measures such as difficulty and discrimination. Our evaluation protocol builds on these ideas and additionally leverages the recent advances in the natural language understanding capabilities to simplify and improve automatic evaluation.

3. Evaluation Protocol

We propose a protocol for evaluating MCRC test items, including a new metric we call **text informativity** for evaluating an item's capability of measuring reading comprehension. It involves measuring the response accuracy of high-performing test takers when they have access to the text (**answerability**) and comparing it to their response accuracy when guessing the correct answer without seeing the text (**guessability**). To obtain these accuracies from human test takers, we first show them the items without the corresponding text and ask them to guess the correct answers. We then reveal the text and let them answer the same items again. Text informativity is then calculated as the difference between answerability and guessability. Intuitively, this metric represents to what degree the information extracted from the text helps the test takers to answer the test items. Since reading comprehension is essentially the ability to extract meaningful information from a text, a high text informativity indicates that the item actually measures the comprehension of the given text.

To apply this protocol for automatic evaluation, we replace human test takers with LLMs and we design prompts to elicit item responses twice for each item; once the text is included in the prompt, and once the model is instructed to guess the correct answers based on world knowledge. The assumption (which we are going to test) is that the LLMs are comparable to highly proficient human readers in terms of world knowledge and comparable reading comprehension capabilities.

Figure 1 illustrates the evaluation protocol. In the experiments described below, we will apply it to human-written and automatically generated items and compare the results to subjective ratings of item quality.

4. Experimental Setup

4.1. Data

We compiled a dataset of German texts and MCRC items from free online language courses[1] offered by *Deutsche Welle* (DW), a broadcast company based in Germany. The target users for these courses are non-native speakers. We included the lessons from the *Top-Thema* course[2], which consists of news articles which were summarized and simplified to match the B1 level in the Common European Framework of Reference for Languages (CEFR). The average text length is 327 tokens (*spaCy* tokenization). Each text comes with several types of exercises, including three MCRC items. Almost all of these have three answer options, and in 66% of the items, the user is allowed to select multiple answer options as correct. For simplicity, we will treat all items as if multiple correct answer options were possible.

We randomly selected 50 texts and all corresponding MCRC items as a test set for the experiment. For the human evaluation, we only used a subset of ten texts to reduce the workload for the annotators.

Scripts for scraping and preprocessing the dataset are available on GitHub[3]. The dataset itself is currently not licensed for redistribution. We hope to publish the dataset for research purposes in the near future to enable more reproducible research.

4.2. Models

We selected two state-of-the-art instruction-tuned LLMs for generating MCRC items and as evaluators for the automatic evaluation:

1. Llama 2 Chat (70B parameters; `meta-llama/Llama-2-70b-chat-hf` on Hugging Face) (Touvron et al., 2023)

2. GPT-4 (unknown model size; snapshot `gpt-4-0613`) (OpenAI, 2023)

4.3. Zero-Shot Item Generation

For each of the 50 texts, we prompted the two LLMs to generate three MCRC items with three answer options each, including which answer options were correct (refer to Appendix A for the full prompts). We used a sampling temperature of 0 (i.e., greedy decoding) for both models.

Since Llama 2 is an English-centric model, it sometimes switched to English. We detected these

---

cases using a language detection library and regenerated outputs with a temperature of 0.5 until at least 80% of the output was identified as German. In cases where Llama 2 generated more than three items, we only kept the first three.

Appendix B contains examples of human-written and generated items for one of the texts.

4.4. Human Evaluation

We recruited six annotators for the human evaluation. All were university students or recent graduates and native German speakers. Considering that the texts and items in our dataset are targeted at CEFR level B1, it is safe to assume that the annotators can respond correctly to answerable items. The annotators took part on a voluntary basis and did not receive monetary compensation. The total workload was between 30 minutes and two hours per person.

We collected three types of annotations: (1) item responses without seeing the text, (2) item responses while seeing the text, and (3) item quality ratings.

Every annotator annotated all ten texts. For each text, the annotation involved two stages. In the **guessing stage**, the three items from *one* generator (human, Llama 2, or GPT-4) were presented, and the annotator was asked to guess for each answer option whether it is correct or incorrect. The reason for only showing the items from a single generator is that the items from different generators would often contain very similar questions, but with different answer options (see Appendix B). This meant that the answer to an item from one generator were sometimes guessable based on the set of answer options from another generator. In the **comprehension stage**, the text and the items from *all* generators were shown. Annotators were asked to respond to the items again and additionally rate the quality of each item on a scale from 1 (unusable) to 5 (perfect). The following criteria for quality were listed, but annotators were free in how they weighted the criteria:

- The item refers to the content of the text.

- The item is comprehensible and grammatically correct.

- The item is unambiguously answerable.

- The item is answerable without additional world knowledge.

- The item is only answerable after reading the text (not through world knowledge alone).

We randomized the order of the texts, items, and answer options for each annotator. Screenshots of the evaluation interface are shown in Appendix C.

---

[1] https://learngerman.dw.com
[2] https://learngerman.dw.com/de/top-thema/s-55861562
[3] https://github.com/saeub/dwlg

4.5. Automatic (LLM-Based) Evaluation

We used zero-shot prompting to elicit item responses from Llama 2 and GPT-4 in two settings. In the **guessing setting**, each prompt contained the text, the stem of a single item, and a single answer option, and the models were instructed to respond with a binary (true/false) label. In the **comprehension setting**, the prompt did not include the text (refer to Appendix A for the full prompts). Both settings used a sampling temperature of 0.

Note that this procedure is different from the human evaluation in that only a single answer option is shown at a time. The main reason for this is to simplify parsing the LLM output. Particularly with Llama 2, showing all answer options and prompting the model to list all correct answer options in a consistent way was not feasible.

While GPT-4 consistently produced responses in the requested format, Llama 2 frequently responded with wordy disclaimers (e.g., "Without seeing the text, it is difficult to say …"). To bypass this behavior for Llama 2, we compared the predicted probabilities (i.e., softmaxed output scores) for the first generated token to determine which label was more likely.

Llama 2 showed a strong bias towards positive responses. We therefore considered the response to be positive if $P(\mathit{true})/(P(\mathit{true})+P(\mathit{false})) \geq \tau$. We optimized the threshold τ to maximize response accuracy in each setting separately on an additional 50 texts from the same dataset. The resulting thresholds were $\tau_{\text{with text}} = 0.9952$ and $\tau_{\text{without text}} = 0.9849$. No such optimization was done for GPT-4.

The code for the automatic evaluation is available on GitHub[4].

5. Results

5.1. Text Informativity

Figure 2 shows the guessability and answerability estimates for the items of the three generators (human, Llama 2, and GPT-4) according to the three evaluators (humans, Llama 2, and GPT-4). For easier comparability, the text informativity metrics are also reported in Table 1.

The three evaluators agree on several observations: human-written items have the lowest guessability, items generated by GPT-4 have the highest answerability, and items generated by Llama 2 have the lowest text informativity.

Overall, GPT-4 as an evaluator outperformed humans in terms of response accuracy both when guessing and when seeing the text. However, since
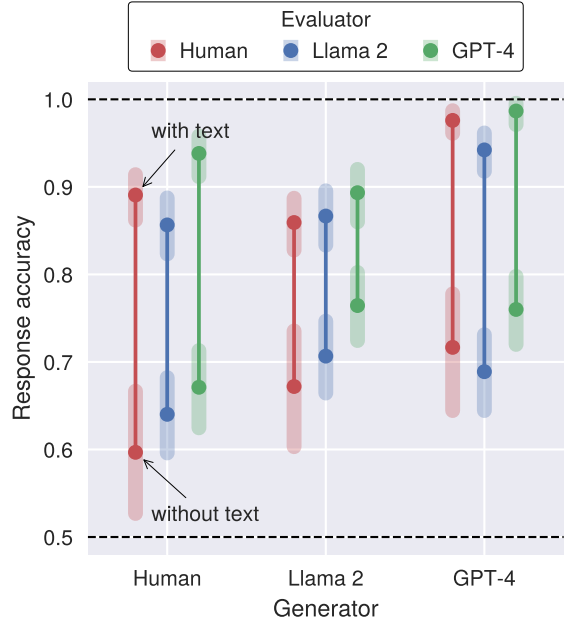
Figure 2: Mean human and LLM response accuracies on human-written and LLM-generated items. The distance between the two points corresponds to text informativity. Accuracies are on the level of answer options, therefore random guessing is at 0.5. For human evaluators, means are based on 10 texts and around 185 responses without text and around 546 responses with text. For LLM evaluators, means are based on 50 texts and around 451 responses in both settings. Error bars are bootstrapped 95% confidence intervals.

	Evaluator		
	Human	Llama 2	GPT-4
Human	**0.294** [0.220, 0.367]	0.216 [0.161, 0.272]	**0.267** [0.219, 0.316]
Llama 2	0.187 [0.115, 0.262]	0.160 [0.109, 0.213]	0.129 [0.082, 0.178]
GPT-4	0.259 [0.193, 0.328]	**0.253** [0.204, 0.302]	0.227 [0.187, 0.269]

Table 1: Text informativity (↑) for all combinations of generators and evaluators. The best text informativity estimates per evaluator are marked in bold. Numbers in brackets are bootstrapped 95% confidence intervals.

text informativity is the difference between the accuracies in both settings, this difference in performance has little effect on text informativity, as evidenced by the similar values in Table 1 between human and GPT-4 evaluators. Llama 2 appears to be less reliable in this respect.

---

[4] https://github.com/saeub/item-evaluation
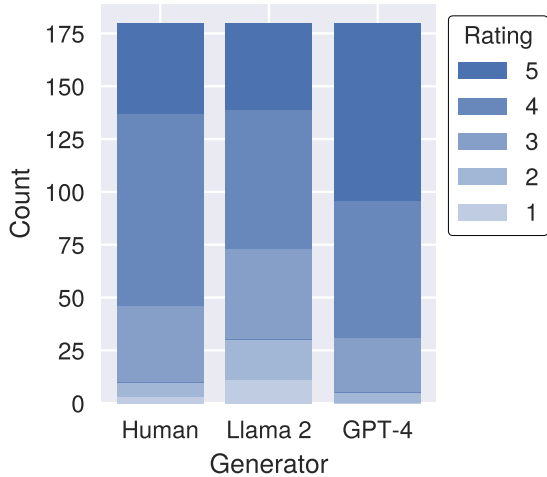
(a) Rating distributions for different item generators. On average, items generated by GPT-4 received the highest ratings, Llama 2 the lowest.
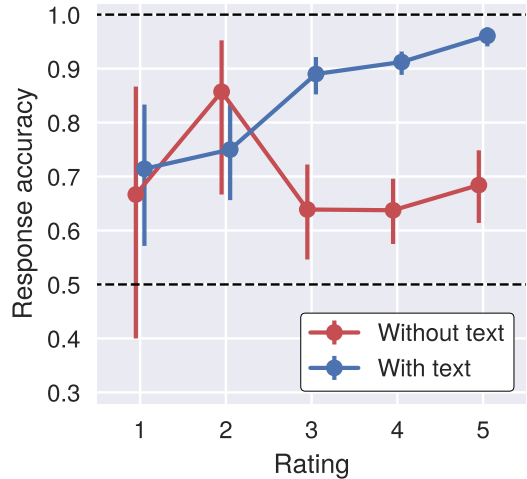
(b) Mean human response accuracies with and without text grouped by item rating (irrespective of generator). Items rated higher tend to have better answerability. Error bars are bootstrapped 95% confidence intervals.

Figure 3: Distributions of human quality ratings and their relation to human response accuracy. A rating of 1 means *unusable*, 5 means *perfect*.

5.2. Quality Ratings

The distribution of human quality ratings is shown in Figure 3a. Across all generators, more than half of the ratings were *good* (4) or *perfect* (5), indicating that most items were of acceptable quality. Items generated by Llama 2 were rated the worst on average, with 11/180 *unusable* (1) ratings. Surprisingly, GPT-4 received considerably more *perfect* ratings (84/180) than human and Llama 2 items.

Comparing the ratings to the response accuracies in Figure 3b reveals that highly rated items tended to have higher answerability, while there is no clear relationship between ratings and guessability. This suggests that the annotators prioritized answerability over guessability in their ratings and explains the higher ratings for items generated by GPT-4, which tended to be both highly answerable and easily guessable (see Figure 2).

5.3. Inter-Annotator Agreement

To quantify how human-like the responses by the LLM evaluators are, we measured the agreement between the two models and the group of human annotators. To achieve this, we calculated pairwise inter-annotator agreements (IAAs) using Cohen's κ between the binary responses from the LLM and each of the humans, for both Llama 2 and GPT-4. We then compared the mean of this pairwise model-human IAA to the mean human-human IAAs. If the model-human IAA is similar to the human-human IAAs, this indicates that the model's response be-

| | Mean IAA with (other) humans | |
Evaluator	without text	with text
Human 1	0.185	0.712
Human 2	0.015	0.679
Human 3	0.000	0.677
Human 4	0.400	0.669
Human 5	0.000	0.634
Human 6	0.216	0.729
Humans 1–6 (average)	0.136	0.683
Llama 2	0.051	0.651
GPT-4	**0.185**	**0.724**

Table 2: Mean Cohen's κ between responses by evaluators and (other) humans with and without text across items (irrespective of generator). The IAAs in the setting with text are based on 93 binary responses (10 texts, 30 items). The values in the setting without text are less reliable because each human only annotated a third of all items in this setting. The mean pairwise agreement between GPT-4 and humans (0.724) is larger than the average agreement between the six humans (0.683).

havior is similar to that of the human annotators.

The results in Table 2 show that GPT-4 provided the most human-like responses and even exceeded the average human-human IAA in both settings. IAA between Llama 2 and humans was lower, but still within the range of human-human IAAs.

5.4. Qualitative Analysis

To provide a tangible explanation for why some items are more guessable or less answerable than others, we conducted a qualitative analysis of generated and human-written items that were either guessed correctly without the text or answered incorrectly with the text by a majority of human annotators. We describe the most common phenomena here and refer to Appendix D for specific examples.

The main reason why items are highly guessable is that they ask about real-world concepts or events that are widely known even without reading the text. This is especially common in our dataset because the texts are news articles about current events. Items that are difficult to guess tend to involve questions about the text itself rather than the events described in it. Examples of such questions are "What is the text about?" or "What does the text say about ...?" where all answer options may be plausible, but not all are true given the text. For most texts, there is at least one question of this type among the human-written items in our dataset, while GPT-4 and Llama 2 tend to generate fewer of these questions.

The explanations for items not being perfectly answerable are more diverse. We found three common features of unanswerable items, listed here in descending order of frequency:

1. **Wrong label:** The item has an incorrect *true*/*false* label for some answer options. This occurred most frequently with Llama 2, and especially when none of the generated answer options are correct, but the model still produced the *true* label for one of them.

2. **Unclear answer options:** The item is phrased in a way that leaves room for interpretation. In particular, some answer options paraphrase information from a text such that not all annotators may agree that they still bear the same meaning.

3. **Insufficient evidence:** The text does not provide the necessary evidence to decide conclusively whether an answer option is correct. In many of these cases, answering correctly requires additional world knowledge.

6. Discussion

6.1. LLMs for Item Generation

One of the aims of this paper was to evaluate LLMs for zero-shot generation of MCRC items in German. Given the lack of data, zero- and few-shot learning are the most promising techniques for this language, and our results strongly suggest that state-of-the-art instruction-tuned LLMs are capable of generating items of acceptable quality. In particular, items generated by GPT-4 are close to the human-written items in our dataset in terms of text informativity. Llama 2 also produced noteworthy results, considering that only 0.17% of the pre-training data is German (Touvron et al., 2023), this is still an impressive result. Using more multilingual or German-centric LLMs could further improve this performance.

A common problem with both models was that they produced easily guessable items, as our evaluations showed. Guessability can be measured in a straightforward manner with human or LLM annotators, and this feedback could be used to improve AIG performance in future work, e.g., through reinforcement learning from human or model feedback (Ouyang et al., 2022; Bai et al., 2022). Previously, Yuan et al. (2017) and Klein and Nabi (2019) have used similar approaches to improve the answerability of generated items.

6.2. Evaluation Protocol

We presented a simple method and a metric for evaluating reading comprehension items. There are several advantages to our method in comparison to previous work. Compared to ratings, this method is more objective and human-centric. It is also more meaningful and interpretable than similarity-based metrics like BLEU, and it does not rely on references. Another advantage is that the same protocol can be used for human and automatic evaluation.

One of the most important limitations is that text informativity only considers two aspects of item quality, i.e., answerability and guessability. Although these are some of the most difficult and critical criteria to meet, there are other aspects that can lead to low item quality (Jones, 2020). For example, our approach cannot detect items where the correct answer options use the same wording as in the text, meaning that no comprehension is required for a correct response. Some of these cases can easily be detected, e.g., using string matching. Characteristics such as grammaticality and difficulty would also have to be addressed separately. We leave these for future work.

Another limitation is that the protocol relies on highly proficient test takers, while the test items in our dataset are targeted at language learners. This is by design, as the goal is to measure the items' answerability given that the text was fully understood, but it still means that the response behavior in the human evaluation is not representative of the target user group.

6.3. LLMs for Item Evaluation

The evaluation protocol we presented measures guessability and answerability by item responses from human annotators with and without showing them the text. By replacing the humans with LLMs, we are making two assumptions:

1. The LLMs have similar world knowledge to humans, resulting in similar guessability estimates.

2. The LLMs have similar reading comprehension abilities to humans, resulting in similar answerability estimates.

Based on the results presented in Section 5.1, using GPT-4 leads to an over-estimation of both guessability and answerability. In contrast to previous work focusing only on answerability (Yuan et al., 2017; Klein and Nabi, 2019; Shuai et al., 2021; Rathod et al., 2022; Raina and Gales, 2022; Uto et al., 2023), using text informativity as a metric normalizes this difference to some degree. The high IAA between GPT-4 and human annotators also suggest that using GPT-4 as an evaluator is a viable option. In contrast, results from Llama 2 were less consistent with humans, both at the dataset level and the response level. Moreover, Llama 2 only yielded usable results after optimizing the classification threshold on additional data as described in Section 4.5 (meaning that the responses were not technically zero-shot in this case). However, depending on the use case, it may still be a good open-source option for evaluation.

Compared to previous work (Berzak et al., 2020; Liusie et al., 2023; Raina et al., 2023), using LLMs for estimating answerability and guessability has several advantages: since we use zero-shot generation, no training is required. This is particularly convenient for languages such as German, where no large MCRC datasets exist. Zero-shot generation also prevents overfitting on dataset-specific features that would go unnoticed by human test takers (compare Berzak et al. (2020), where a fine-tuned RoBERTa classifier consistently outperformed human test takers at guessing the correct answer).

A limitation of our approach is that a single LLM is unable to capture human label variation. On the one hand, this means that we cannot model how strongly different human annotators will agree on their responses to a specific item, which can be useful for evaluation (Plank, 2022). On the other hand, it means that evaluating the quality of a single item is not feasible, which is why we only reported text informativity at the level of an entire dataset. Possible solutions to this problem include using multiple models (Lalor et al., 2019; Byrd and Srivastava, 2022) or prompt variation (Portillo Wightman et al., 2023) to determine uncertainty.

7. Conclusion and Future Work

The overarching goal of this paper was to explore the potential of LLMs for generating and evaluating MCRC items. To this end, we introduced a new evaluation protocol and metric, text informativity, and demonstrated its applicability for both human and automatic evaluation. We used this protocol to evaluate two state-of-the-art LLMs for zero-shot item generation based on a dataset of German texts and MCRC items from online language courses. Our results show that both GPT-4 and Llama 2 are capable of generating items of acceptable quality, but GPT-4 clearly outperforms in terms of text informativity and human quality ratings. We also found that using GPT-4 for automatic evaluation is a viable option, while Llama 2 is less reliable.

These insights have significant implications: they show that zero-shot learning can make automatic item generation and evaluation feasible in languages where MCRC resources are scarce. Our evaluation protocol also addresses the lack of automatic evaluation metrics for the task. In a more general sense, using LLMs to generate reading comprehension items *and* to predict how humans will respond to these items is a promising approach – not only for language assessment in education, but also for comprehensibility evaluation in text simplification and readability assessment.

Future work could focus on improving item generation, e.g., by using text informativity as a reward for reinforcement learning, or improving item evaluation, e.g., by making LLM responses more human-like and reflective of individual variability and uncertainty.

8. Acknowledgements

9. Bibliographical References

Oliver Alonzo, Jessica Trussell, Becca Dingman, and Matt Huenerfauth. 2021. Comparison of methods for evaluating complexity of simplified texts among deaf and hard-of-hearing adults at different literacy levels. In *Proceedings of the*

2021 CHI Conference on Human Factors in Computing Systems, CHI '21. ACM.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*. Association for Computational Linguistics.

Yigal Attali, Andrew Runge, Geoffrey T. LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A. von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants.

Gonzalo Berger, Tatiana Rischewski, Luis Chiruzzo, and Aiala Rosá. 2022. Generation of English question answer exercises from texts using transformers based models. *2022 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pages 1–5.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Matthew Byrd and Shashank Srivastava. 2022. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

Ruhan Circi, Juanita Hicks, and Emmanuel Sikali. 2023. Automatic item generation: foundations and machine learning-based approaches for assessments. *Frontiers in Education*, 8.

Ramon Dijkstra, Zülküf Genç, Subhradeep Kayal, and J. Kamps. 2022. Reading comprehension quiz generation using generative pre-trained transformers. In *iTextbooks@AIED*.

X. Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

Yin-Chun Fung, Lap-Kei Lee, and Kwok Tai Chui. 2023. An automatic question generator for Chinese comprehension. *Inventions*.

Yifan Gao, Lidong Bing, Wang Chen, Michael Lyu, and Irwin King. 2019. Difficulty controllable generation of reading comprehension questions. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, IJCAI-2019. International Joint Conferences on Artificial Intelligence Organization.

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer McIntosh von der Ohe, and Alona Fyshe. 2022. Question generation for reading comprehension assessment by modeling how and what to ask.

Mark J. Gierl, Hollis Lai, and Vasily Tanygin. 2021. *Advanced Methods in Automatic Item Generation*. Routledge.

Rita Green. 2020. Pilot testing: Why and how we trial. In *The Routledge Handbook of Second Language Acquisition and Language Testing*, chapter 11, pages 115–124. Routledge.

Thomas M. Haladyna. 2013. Automatic item generation: A historical perspective. In Mark J. Gierl and Thomas M. Haladyna, editors, *Automatic Item Generation: Theory and Practice*, chapter 2, pages 13–25. Routledge, New York.

Eun Hee Jeon and Junko Yamashita. 2020. Measuring L2 reading. In *The Routledge Handbook of Second Language Acquisition and Language Testing*, chapter 25, pages 265–274. Routledge.

Xin Jia, Wenjie Zhou, Xu Sun, and Yunfang Wu. 2020. EQG-RACE: Examination-type question generation.

Glyn Jones. 2020. Designing multiple-choice test items. In *The Routledge Handbook of Second Language Acquisition and Language Testing*, chapter 9, pages 90–101. Routledge.

Dmytro Kalpakchi and Johan Boye. 2023. Quasi: a synthetic question-answering dataset in Swedish using GPT-3 and zero-shot learning. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 477–491, Tórshavn, Faroe Islands. University of Tartu Library.

Tassilo Klein and Moin Nabi. 2019. Learning to answer by learning to ask: Getting the best of GPT-2 and BERT worlds.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations.

Hollis Lai and Mark J. Gierl. 2013. Generating items under the assessment engineering framework. In Mark J. Gierl and Thomas M. Haladyna, editors, *Automatic Item Generation: Theory and Practice*, chapter 6, pages 77–101. Routledge, New York.

John P. Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. 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)*. Association for Computational Linguistics.

Gondy Leroy, David Kauchak, Diane Haeger, and Douglas Spegman. 2022. Evaluation of an online text simplification editor using manual and automated metrics for perceived and actual text difficulty. *JAMIA Open*, 5(2).

Adian Liusie, Vatsal Raina, and Mark Gales. 2023. "World knowledge" in multiple choice reading comprehension. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*. Association for Computational Linguistics.

Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. 2020. Simplifying paragraph-level question generation via transformer language models.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. MQAG: Multiple-choice question answering and generation for assessing information consistency in summarization. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 39–53, Nusa Dua, Bali. Association for Computational Linguistics.

Kaushal Kumar Maurya and Maunendra Sankar Desarkar. 2020. Learning to distract: A hierarchical multi-decoder network for automated generation of long distractors for multiple-choice questions for reading comprehension. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: A review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, pages 1–32.

OpenAI. 2023. GPT-4 technical report.

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.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gwenyth Portillo Wightman, Alexandra Delucia, and Mark Dredze. 2023. Strength in numbers: Estimating confidence of large language models by prompt agreement. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 326–362, Toronto, Canada. Association for Computational Linguistics.

Vatsal Raina and Mark Gales. 2022. Multiple-choice question generation: Towards an automated assessment framework.

Vatsal Raina, Adian Liusie, and Mark Gales. 2023. Analyzing multiple-choice reading and listening comprehension tests.

Manav Rathod, Tony Tu, and Katherine Stasaski. 2022. Educational multi-question generation for reading comprehension. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*.

Andreas Säuberli, Franz Holzknecht, Patrick Haller, Silvana Deilen, Laura Schiffl, Silvia Hansen-Schirra, and Sarah Ebling. 2024. Digital compre-

hensibility assessment of simplified texts among persons with intellectual disabilities.

Pengju Shuai, Zixi Wei, Sishun Liu, Xiaofei Xu, and Li Li. 2021. Topic enhanced multi-head co-attention: Generating distractors for reading comprehension. *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2023. Difficulty-controllable neural question generation for reading comprehension using item response theory. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners.

Jiayuan Xie, Ningxin Peng, Yi Cai, Tao Wang, and Qingbao Huang. 2022. Diverse distractor generation for constructing high-quality multiple choice questions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:280–291.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study.

A. Prompts

The instructions used for generating items and responses for the automatic evaluation are specified in Tables 3 and 4. For both models, the instructions were provided as the first user message, and no system instructions were specified.

German	English
Text: $[T]$	Text: $[T]$
Schreibe 3 Multiple-Choice-Verständnisfragen zum Text oben, in deutscher Sprache. Jede Frage soll 3 Antwortmöglichkeiten haben. Schreibe hinter jede Antwort in Klammern, ob sie richtig oder falsch ist. Zwischen 0 und 3 Antworten können richtig sein. Die falschen Antworten sollten plausibel sein, wenn man den Text nicht gelesen hat.	Write 3 multiple-choice comprehension questions about the text above, in German language. Each question should have 3 answer options. After each answer, write whether it is correct or incorrect in parentheses. Between 0 and 3 answers can be correct. The incorrect answers should be plausible, not having read the text.

Table 3: The German prompt template for item generation and a translation into English. In the text T, headings and paragraphs were separated by a newline character.

	German	English
With text	Text: $[T]$ Frage: $[q]$ Antwort: $[a]$ Gemäß dem Text oben, ist diese Antwort richtig (R) oder falsch (F)? Gib nur den Buchstaben R oder F an.	Text: $[T]$ Question: $[q]$ Answer: $[a]$ Based on the text above, is this answer correct (C) or incorrect (I)? Indicate only the letter C or I.
Without text	Die folgende Frage und Antwort stammen aus einer Multiple-Choice-Verständnisaufgabe zu einem unbekannten Text. Frage: $[q]$ Antwort: $[a]$ Ohne den Text zu kennen, nur basierend auf Allgemeinwissen, ist es plausibler, dass die Antwort richtig (R) oder falsch (F) ist? Gib nur den Buchstaben R oder F an.	The following question and answer are from a multiple-choice comprehension task about an unknown text. Question: $[q]$ Answer: $[a]$ Without knowing the text, only based on general knowledge, is this answer more likely to be correct (C) or incorrect (I)? Indicate only the letter C or I.

Table 4: The German prompt templates for item evaluation and a translation into English. In the text T, headings and paragraphs were separated by a newline character.

B. Examples of Human-Written and Generated Items for the Same Text

The following sections show all human-written and generated items for one of the texts in the test set. The check marks (✓) and crosses (✗) indicate whether the answer option is correct or incorrect (according to the author/generator).

The corresponding lesson on the DW website (including the German text) can be found at `https://learngerman.dw.com/de/l-46996604`. The text is about Yemen's national football team, who had qualified for the Asia Cup in the United Arab Emirates, but faced challenges preparing for the championship due to political tensions.

B.1. Human-Written Items

German	English
Der Text handelt vor allem von ...	The text is mainly about ...
✗ Fußballfans im Jemen und wie sie versuchen, die Vereine in ihren Orten zu unterstützen.	✗ Football fans in Yemen and how they try to support the clubs in their towns.
✗ einer großen nationalen Sportveranstaltung, die im Jemen stattfinden sollte.	✗ a major national sports event that was supposed to take place in Yemen.
✓ den Vorbereitungen der jemenitischen Nationalmannschaft auf eine wichtige Meisterschaft.	✓ the preparations of the Yemeni national team for an important championship.
Was sagt Bashir Sinan?	What does Bashir Sinan say?
✓ Wenn Fußball gespielt wird, vergessen die Jemeniten den Krieg in ihrem Land.	✓ When playing football, Yemenis forget about the war in their country.
✓ Das jemenitische Fußballteam leidet unter den politischen Problemen im Land.	✓ The Yemeni football team suffers from the political problems in the country.
✗ Für die Fußballer ist das nächste Ziel, die Asienmeisterschaften zu gewinnen.	✗ The next goal for the football players is to win the Asia Cup.
Was ist richtig?	What is correct?
✓ Viele Jemeniten fliehen vor dem Bürgerkrieg in ihrem Land.	✓ Many Yemenis are fleeing the civil war in their country.
✓ Die jemenitische Mannschaft will, dass es Friedensgespräche gibt.	✓ The Yemeni team wants peace talks to take place.
✗ Katar und Saudi-Arabien kämpfen im Krieg gemeinsam gegen den Jemen.	✗ Qatar and Saudi Arabia are fighting together in the war against Yemen.

B.2. Items Generated by Llama 2

German	English
Warum dürfen keine Fußballspiele mehr in Jemen stattfinden? ✓ Weil das Land zu unsicher ist ✗ Weil es ein Bürgerkrieg gibt ✗ Weil die Bevölkerung nicht interessiert ist	Why are no more football games allowed in Yemen? ✓ Because the country is too unsafe ✗ Because there is a civil war ✗ Because the population is not interested
Wie haben die jemenitischen Fußballspieler sich für die Asienmeisterschaften qualifiziert? ✓ Sie haben in verschiedenen arabischen Ländern trainiert ✗ Sie haben ein wildcard [*sic*] erhalten ✗ Sie haben sich nicht qualifiziert	How did the Yemeni football players qualify for the Asia Cup? ✓ They trained in various Arab countries ✗ They received a wildcard ✗ They did not qualify
Was ist die Botschaft der jemenitischen Fußballmannschaft an die Bevölkerung und die Politik? ✓ Lassen wir die Gewehre schweigen und einen vernünftigen Dialog und Frieden an ihre Stelle treten ✗ Lassen wir die Gewalt weitergehen und uns nichts ausreden ✗ Lassen wir uns auf die politischen Spannungen einigen und die Fußballmeisterschaft boykottieren	What is the message of the Yemeni football team to the population and the politicians? ✓ Let's silence the guns and have a reasonable dialogue and peace take their place ✗ Let's let the violence continue and not talk ourselves out of anything ✗ Let's agree on the political tensions and boycott the football championship

B.3. Items Generated by GPT-4

German	English
Warum dürfen im Jemen keine Fußballspiele mehr stattfinden? ✓ Wegen des Bürgerkriegs und der damit verbundenen Unsicherheit. ✗ Weil die Fußballstadien zerstört wurden. ✗ Weil die jemenitischen Fußballspieler alle das Land verlassen haben.	Why are no more football games allowed in Yemen? ✓ Because of the civil war and the associated insecurity. ✗ Because the football stadiums were destroyed. ✗ Because all Yemeni football players have left the country.
Wo fand das Training für die Asienmeisterschaften statt? ✓ In verschiedenen arabischen Ländern, wie Saudi-Arabien und Katar. ✓ In Malaysia. ✗ Im Jemen.	Where did the training for the Asia Cup take place? ✓ In various Arab countries, such as Saudi Arabia and Qatar. ✓ In Malaysia. ✗ In Yemen.
Was ist die Botschaft der jemenitischen Fußballmannschaft an die Bevölkerung und die Politik? ✗ Sie sollten mehr Geld in den Fußball investieren. ✓ Sie sollten die Gewehre schweigen lassen und einen vernünftigen Dialog und Frieden an ihre Stelle treten. ✗ Sie sollten die Fußballspieler besser unterstützen.	What is the message of the Yemeni football team to the population and the politicians? ✗ They should invest more money in football. ✓ They should silence the guns and have a reasonable dialogue and peace take their place. ✗ They should support the football players better.

C. User Interface for Human Evaluation

Figure 4: Screenshot of the user interface for the human evaluation, without text.

Figure 5: Screenshot of the user interface for the human evaluation, with text and quality ratings.

D. Examples of Guessable and Unanswerable Items

D.1. Guessable Items

	German	English
Item (human-written)	Die Saturnalien ... ✗ wurden an den längsten Tagen im Jahr gefeiert. ✓ waren ein Fest, bei dem unfreie Menschen mit den Herrschern die Rollen wechselten. ✓ fanden unter anderem in der römischen Stadt Köln statt.	The Saturnalia ... ✗ were celebrated on the longest days of the year. ✓ were a festival where unfree people switched roles with the rulers. ✓ took place in the Roman city of Cologne, among other places.
Item (generated by GPT-4)	Was bedeutet der Name „Karneval" aus dem Lateinischen übersetzt? ✗ „Fest der Freude" ✓ „Fleisch, leb wohl" ✗ „Tanz der Narren"	What does the name "Carnival" mean when translated from Latin? ✗ "Festival of joy" ✓ „Flesh, farewell" ✗ „Dance of the fools"

D.2. Unanswerable Items

D.2.1. Wrong label

	German	English
Text excerpt	[...] *Die Musikwissenschaftlerin Marina Schwarz meint dazu: „Das ist Teil der immer noch patriarchalischen Gesellschaft, in der wir leben." Offenbar finden auch viele Frauen, die in dieser Gesellschaft aufgewachsen sind, solche Texte normal.* [...]	[...] *Musicologist Marina Schwarz says: "This is part of the still patriarchal society in which we live." Apparently, many women who have grown up in this society also find such lyrics normal.* [...]
Item (generated by Llama 2)	Was ist laut Text Marina Schwarz' Meinung zu sexistischen Texten im Schlager? ✓ Sie findet sie inakzeptabel. [should be ✗] ✗ Sie findet sie normal, weil es Teil der patriarchalischen Gesellschaft ist. ✗ Sie findet sie nicht sexistisch, sondern nur humorvoll.	According to the text, what is Marina Schwarz' opinion on sexist lyrics in Schlager? ✓ She finds them unacceptable. [should be ✗] ✗ She finds them normal, because it is part of the patriarchal society. ✗ She does not find them sexist, just humorous.

D.2.2. Unclear answer options

	German	English
Text excerpt	[...] *Für viele Deutsche zählt beim Kiosk eher die Atmosphäre – besonders in der warmen Jahreszeit.* [...]	[...] *For many Germans, the atmosphere is more important at the kiosk – especially in the warm season.* [...]
Item (human-written)	Viele Menschen kaufen Alkohol am Kiosk, weil ... ✓ er dort billiger ist als in Bars und Kneipen. ✓ sie die schöne Stimmung vor Ort mögen. [unclear if ✓ or ✗] ✓ sie auf dem Weg zu einer Party etwas trinken möchten.	Many people buy alcohol at the kiosk because ... ✓ it is cheaper there than in bars and pubs. ✓ they like the nice atmosphere on site. [unclear if ✓ or ✗] ✓ they want to drink something on the way to a party.

D.2.3. Insufficient evidence

	German	English
Text excerpt	[...] *Besonders im Rheinland sind die Straßen voll mit kostümierten Menschen, die tanzen, singen und feiern* [...]	[...] *Especially in the Rhineland, the streets are full of people in costumes who dance, sing and celebrate* [...]
Item (generated by Llama 2)	Wo finden die meisten Karnevalsumzüge und -feiern statt? ✓ In Köln ✗ In Rom ✗ In Berlin	Where do most carnival parades and celebrations take place? ✓ In Cologne ✗ In Rome ✗ In Berlin

An Extensible Massively Multilingual Lexical Simplification Pipeline Dataset using the MultiLS Framework

Matthew Shardlow[1], Fernando Alva-Manchego[2], Riza Batista-Navarro[3], Stefan Bott[4], Saul Calderon Ramirez[5], Rémi Cardon[6], Thomas François[6], Akio Hayakawa[4], Andrea Horbach[7,8], Anna Hülsing[7], Yusuke Ide[9], Joseph Marvin Imperial[10,14], Adam Nohejl[9], Kai North[11], Laura Occhipinti[12], Nelson Peréz Rojas[5], Nishat Raihan[11], Tharindu Ranasinghe[13], Martin Solis Salazar[5], Marcos Zampieri[11], Horacio Saggion[4]

[1]Manchester Metropolitan University [2]Cardiff University [3]University of Manchester
[4]Universitat Pompeu Fabra [5]Tecnológico de Costa Rica [6]UCLouvain [7]University of Hildesheim
[8]CATALPA, FernUniversität in Hagen [9]NARA Institute of Science and Technology
[10]National University Philippines [11]George Mason University
[12]University of Bologna [13]Aston University [14]University of Bath
m.shardlow@mmu.ac.uk

Abstract

We present preliminary findings on the MultiLS dataset, developed in support of the 2024 Multilingual Lexical Simplification Pipeline (MLSP) Shared Task. This dataset currently comprises of 300 instances of lexical complexity prediction and lexical simplification across 10 languages. In this paper, we (1) describe the annotation protocol in support of the contribution of future datasets and (2) present summary statistics on the existing data that we have gathered. Multilingual lexical simplification can be used to support low-ability readers to engage with otherwise difficult texts in their native, often low-resourced, languages.

Keywords: lexical simplification, lexical complexity prediction, MultiLS

1. Introduction

The lexical simplification pipeline is a family of systems concerned with the task of automatically identifying and replacing complex vocabulary with simpler alternatives (North et al., 2023b). The lexical simplification pipeline provides a more targeted approach to simplification than automated text simplification (Al-Thanyyan and Azmi, 2021; Alva-Manchego et al., 2020) which directly rewrites entire sentences. The two core operations included in the lexical simplification pipeline are (1) lexical complexity prediction and (2) the replacement of complex words with simple synonyms. Other varied operations exist in the text simplification ecosystem (Cardon and Bibal, 2023) which may be handled within lexical simplification depending on the specific implementation of the pipeline.

The task of Lexical Complexity Prediction (LCP) (Shardlow et al., 2020, 2022; North et al., 2023b), a form of Complex Word Identification (CWI) (Shardlow, 2013), involves assigning continuous values in the range 0-1 to given tokens in context, representing the difficulty that an intended reader population may associate with that target word. LCP was previously explored through a shared task (Shardlow et al., 2021) at SemEval 2021.

The second task, often referred to just as lexical simplification (Saggion et al., 2022) involves generating simple substitutions for target words in context. This task has been explored for single words and multi-word expressions, and is related to the identification of simple paraphrases (Pavlick and Callison-Burch, 2016; Maddela et al., 2021).

In addition to these two tasks, lexical simplification pipeline systems often take into account word sense disambiguation (Saggion et al., 2016), independent substitution generation / selection (Qiang et al., 2020) and grammaticality filtering (Gooding and Kochmar, 2019) steps — which are not explicitly explored in our dataset.

We identify two shortcomings of current work on the lexical simplification pipeline as follows:

1. Current datasets only explore one pipeline operation, but no dataset exists with multiple operations on the same target words in context. This means that systems that are trained on one task are unsuitable for the other. Systems trained using multiple datasets may experience 'genre drift', where the text type across datasets differs.

2. The existing data is overwhelmingly in the English language. Whereas some recent efforts exist to provide open source data in languages other than English, there is no guarantee that these datasets are created using similar protocols.

38

We introduce the MultiLS dataset[1] to address these two issues, based on the MultiLS framework (North et al., 2024). MultiLS is a new dataset that unites the related tasks of LCP and lexical simplification. Each instance in MultiLS contains a single target within an authentic context, which has been annotated for both the difficulty of the target (0-1) and relevant simplifying substitutions for the target. MultiLS is available in 10 languages and each language has the same amount of data, providing equality in provision between language sources.

2. Related Work

Current systems adopting the lexical simplification pipeline make use of transformer technology as described in detail in a recent survey by North et al. (2023b). In this section we particularly focus on the multilingual resources available for (1) Full-pipeline lexical simplification systems (2) LCP datasets and (3) Lexical simplification datasets.

Whilst several recent works exist implementing the lexical simplification pipeline in English making use of transformer-based technology (Qiang et al., 2021a; Baez and Saggion, 2023), there have also been significant efforts in Spanish to implement lexical simplification systems both for European Spanish (Alarcon et al., 2021; Stajner et al., 2023) and for Latin American variants such as Ecuadorian Spanish (Ortiz-Zambrano et al., 2023). The full pipeline has also been implemented in Swedish (Graichen and Jonsson, 2023), French (Rolin et al., 2021) and Chinese (Qiang et al., 2021b), making use of language-specific monolingual transformer based models. The lexical simplification pipeline is typically implemented as a monolingual task. However, there are also efforts to implement multilingual systems for simplification (Sheang and Saggion, 2023; Liu et al., 2023), which rely on multilingual language models trained on the TSAR-2022 shared task data for English, Spanish and Portuguese (Štajner et al., 2022).

An LCP dataset comprises of target words in context with a continuous value representing the difficulty of that target. LCP datasets were released for *English* through previous shared tasks (Yimam et al., 2018; Shardlow et al., 2021). There are 70K instances of LCP judgements available for English across these three shared task datasets, with additional data released through these efforts for *Spanish* and *German*. Recent research addressed the prediction of lexical difficulty for foreign language readers of French (Tack, 2021). Additionally, other research for French has implemented LCP annotations in the medical context

(Sheang et al., 2022; Koptient and Grabar, 2022). For *Japanese*, the recent JaLeCoN dataset (Ide et al., 2023) provides 10K LCP annotations for news text and 8K LCP annotations for governmental texts. Published work to develop LCP annotations for languages other than English has also taken place for *Russian* (Abramov and Ivanov, 2022; Abramov et al., 2023), *Turkish* (Ilgen and Biemann, 2023), *Chinese* (Yang et al., 2023) and *Malay* (Omar et al., 2022).

Lexical simplification datasets comprise of a context, with a marked target word and a list of potential simplifying substitutions for that target word. The TSAR-2022 shared task data (Štajner et al., 2022) provides instances of lexical simplifications for *English*, *Spanish* (Ferrés and Saggion, 2022) and *Portuguese* (North et al., 2022, 2023a). Additionally, for Spanish, the EASIER Corpus (Alarcon et al., 2023) provides further simplification data. We also identified suitable simplification resources for *French* (Billami et al., 2018), *Japanese* (Kajiwara and Yamamoto, 2015; Kodaira et al., 2016), *Chinese* (Qiang et al., 2021b) and an additional resource for Portuguese (Hartmann et al., 2018).

3. MultiLS Dataset

We introduce the MultiLS dataset and describe the Trial data, comprising of 30 instances per language for 10 available languages. MultiLS provides LCP and lexical simplification annotations on common targets and contexts for each available language, significantly extending the availability of multilingual lexical simplification pipeline data. The full MultiLS dataset including a further 5,600 test instances across all 10 languages will be released as part of the 2024 MLSP shared task (Shardlow et al., 2024).

3.1. Annotation Protocol

We gathered an international team of 21 researchers representing 14 institutions and based across 8 countries. Each researcher was tasked with coordinating the annotations for one or more of the languages in our dataset. To guide the varied teams of dataset providers, we produced a comprehensive set of annotation guidelines. The key points from these guidelines are described below. The full guidelines are available with the dataset to encourage future contributions of additional languages.

3.1.1. Data Preparation

Dataset providers selected appropriate instances in their target language, focusing on contexts or single words (i.e., not multi-word expressions).

---

[1]https://github.com/MLSP2024/MLSP_Data

The definition of *word* may change from one language to another, especially when handling languages with non-Alphabetic scripts. The words were selected in each language to ensure sufficient difficulty to warrant lexical complexity annotation, and particularly to ensure that annotators will be able to find some simpler substitutions for the word in context. We provided a sample list of 200 words in English with the aim of encouraging common words across languages. However, due to language-specific constraints, not all providers used this list to make their selection. Whilst this was not enforced, there are some common targets across language pairs which can be used for future investigations.

Once the words had been selected, dataset providers identified 200 contexts in their target language, where each context contained one of the target words. Data providers were also free to select 200 contexts and then choose appropriate target words within those contexts. The contexts were selected from a readily available source in each language, specifically one that is related to an educational setting and released under a license that allows further redistribution of the text. For each context, an additional 2 words were selected for annotation. The requirement to select 200 contexts, with 3 words per context gave rise to 600 instances in total per language. An example is given below in English, with the selected target words highlighted in bold text:

Folly is set in **great dignity**.

Note that the highlighted words: 'Folly', 'great' and 'dignity' all bear semantic content. The remaining words (the copula 'is' and the preposition 'in') are short words that do not have much influence on the overall meaning of the text. Particularly, it would be hard to find substitutions for these.

3.1.2. Annotator Selection

We requested that data providers solicited a minimum of 10 annotations per instance. Data providers were instructed to select annotators according to a 'Target group', which was also recorded as metadata indicating that the annotations received were reflective of the needs of the target group. For each annotator, the following additional elements of metadata were collected: (1) The number of years the annotator has spent in education; (2) Whether or not they are a native speaker of the language that is being annotated; (3) Age; (4) Typical number of hours they spend reading per week; (5) First Language; and (6) Number of languages they speak.

Dataset providers were able to either choose the same annotators to perform both lexical simplification and LCP, or to choose different groups for

each task. For example if the target group was language learners, the data provider may have chosen to ask the learners to provide LCP annotations and their teachers to provide lexical simplification annotations.

3.1.3. Data Annotation for Lexical Complexity

Annotations for lexical complexity were performed using a 5-point Likert scale, with the following points translated into each language:

1. Very Easy - Words that are very familiar to you

2. Easy - Words that are mostly familiar to you

3. Neutral - Words that are neither difficult nor easy to you

4. Difficult - Words whose meaning is unclear, but that you may be able to infer from context

5. Very Difficult - Words that you have never seen before, or whose meaning is very unclear

Each instance was presented to the annotator with the full context and the annotators were asked to provide an independent judgement for each of the three highlighted words per context. Dataset providers additionally performed manual quality control on the resulting annotations, such as checking that the annotators had used the full range of annotations and that the complexity judgements were in line with those of other annotators'. The 1-5 annotations were converted to 0-1 following the Complex 2.0 format (Shardlow et al., 2022).

We did not typically enforce annotator agreements, but instead relied on manual evaluation of the outputs of annotators by the dataset providers. All provided data was quality checked and adjusted to ensure consistency where needed.

3.1.4. Data Annotation for Lexical Simplification

For each target word, annotators provided a minimum of 1 and a maximum of 3 words that could be used to simplify the target in the given context. The substitutions were selected to ensure (a) that the meaning of the original word and the overall context was preserved, and (b) that the substitution was easier to understand than the original target. For some of the target words, it was not easy to find appropriate simplifications in the contexts that they are presented in. For instance, a word may already be sufficiently simple, or despite being complex there may be no simpler alternatives. In these cases, the annotators were instructed to write the original word, or to leave the field blank

and indicate that the original word is the simplest word that could fit in this context.

Data providers performed quality control through manual verification of the submissions of each annotator by checking (a) the suitability of the substitutions within the context, and (2) the frequency with which annotators were unable to find a simplification.

For some languages, a substitution may cause issues regarding the agreement with surrounding words (e.g. a masculine noun replaced by a feminine substitution will require to revise the gender of its related adjectives or determiners). We decided to treat morphological adaptation as a separate task that is left aside. Annotators were informed that they may propose substitutions that do not strictly fit in the grammatical context regarding gender/number agreement.

3.2. MultiLS Trial Data

Presently, we have released 30 instances per language for the 10 languages in Table 1. We report the aggregated metadata for each language, as well as summary statistics on the overall dataset.

4. Discussion

In this work we have presented a data annotation effort for LCP and lexical simplification which is intended to be extensible to a wide variety of languages. Currently, 8 out of the 10 languages represented are Indo-European, with 5 Romance Languages (French, Italian, Spanish, Portuguese and Catalan), 2 West-Germanic languages (English and German) and Sinhala which is of the Indo-Aryan family. Additionally we have Filipino and Japanese which are of the Austronesian and Japonic families respectively. Eight of the languages make use of the Latin script, with Sinhala and Japanese being exceptions to this. The Latin script languages are alphabetic, whereas Sinhala is an abugida language (characters represent a combination of vowel and consonants) and Japanese script features kanji (logographic characters loaned from Chinese) and kana (syllabic characters). The available languages are a result of the collaborative team that we were able to gather. We hope to extend the language families, scripts and script types represented in future iterations of the dataset.

The target groups and text genres represented in the language subsets of our dataset are varied as shown in the second and third column of Table 1. This reflects the availability of target texts in each language as well as the available pools of annotators that we were able to access. We expect that this will result in some variations be-

tween datasets reflecting the interests of the target groups. We have exposed this information to help those working with our dataset to adjust systems according to each target group. We will also make summary metadata regarding our annotators for each subset available alongside the dataset.

The average complexity in our dataset varies across subsets. The average complexity of each subset is below 0.5 (0 = Very Easy, 1 = Very Difficult). All datasets contain examples of complex language (> 0.5), but the low average complexities represents the fact that the majority of identified tokens were assigned easier complexity values (< 0.5) by annotators. This is representative of Zipfian language distributions, where most frequent words are familiar, with few rare complex words.

The context length also varies between subsets of our data. Japanese has a particularly short context length as each kanji character is a logographic unit, leading to fewer characters per sentence. Considering the other languages, the texts selected for Filipino have a typically short context length (64.066) whereas those selected for Catalan have a generally long context length (239.533). We also note significant variations in the number of unique substitutions per language with an average of 3.967 substitutions per instance for Filipino and 15.8 for Japanese. Each language is unique and the variations arise from the target groups, text genres, annotator pool and language specific factors. We deliberately present the MultiLS dataset as a composite of sub-language datasets to allow and encourage the development of language-specific and multilingual simplification interventions and technologies.

We initially aimed for a high degree of uniformity in our dataset across language subsets. However, to prioritise the inclusion of more languages we chose to relax the inclusion criteria to incorporate existing efforts to annotate LCP/LS for interesting and diverse text types and genres. Additionally, our approach gave significant agency to the native-speaking dataset providers in each language-setting to make linguistically appropriate decisions for their bespoke context. The result is a dataset with lower intra-lingual conformity, but ultimately a larger, more diverse and easily extensible dataset.

Recent efforts in lexical simplification within English have focussed on personalised approaches to (a) complexity detection (Gooding and Tragut, 2022) and (b) simplification (Sukiman et al., 2024), which seeks to model the individual reader, as opposed to building a single model for all readers. Our proposed dataset only provides a single aggregated judgement per instance, meaning that it is not useful for personalised lexical simplification pipelines in its current form. The authors will ex-

Language	Target Group	Text Genre	Mean Complexity	Mean Context Length	Mean # Unique Subs
Catalan	Varied	News	0.487 (0.125)	239.533 (70.128)	14.167 (3.354)
English	University Students	Wikibooks	0.200 (0.201)	111 (36.992)	6.167 (1.859)
Filipino	University Staff	Educational Books	0.171 (0.126)	64.066 (22.137)	3.967 (1.098)
French	Language Learners	Varied	0.371 (0.229)	129.1 (45.564)	10.067 (3.463)
German	High-School Students	Wiki / Literary	0.413 (0.191)	195.733 (59.604)	8.067 (2.791)
Italian	Native Speakers	Wikibooks/Wikiquote	0.248 (0.168)	168.4 (67.614)	7.800 (2.952)
Japanese	Language Learners	Varied	0.259 (0.173)	37.8 (7.303)	15.800 (4.634)
Portuguese	MTurk Workers	Varied	0.273 (0.165)	165.9 (74.062)	5.367 (1.217)
Sinhala	University Staff	News / Religious	0.243 (0.214)	163.4 (52.554)	4.333 (0.606)
Spanish	Varied	Educational Books	0.449 (0.233)	178.7 (48.075)	10.867 (3.785)

Table 1: Dataset metadata and statistics for the MultiLS trial data organised alphabetically by the English name of the language. All values given as mean average with standard deviation in brackets. Context length is reported as character length for cross-lingual comparison.

plore the use of the unaggregated annotator-level complexity predictions to better understand how we can use this data for personalised judgements.

5. Conclusion

We present the MultiLS dataset, comprising of LCP and lexical simplification data for 10 languages. MultiLS is an extensible framework and is open to contributions of additional languages and to additional data for the existing languages (North et al., 2024). MultiLS will allow future researchers to develop truly multilingual lexical simplification pipeline systems. We include one instance per language in Table 2 in the Appendix.

6. Acknowledgments

7. Bibliographical References

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Comput. Surv.*, 54(2).

Rodrigo Alarcon, Lourdes Moreno, and Paloma Martínez. 2023. Easier corpus: A lexical simplification resource for people with cognitive impairments. *Plos one*, 18(4):e0283622.

Rodrigo Alarcon, Lourdes Moreno, and Paloma Martínez. 2021. Lexical simplification system to improve web accessibility. *IEEE Access*, 9:58755–58767.

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

Anthony Baez and Horacio Saggion. 2023. LSLlama: Fine-tuned LLaMA for lexical simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 102–108.

Mokhtar B. Billami, Thomas François, and Núria Gala. 2018. ReSyf: a French lexicon with ranked synonyms. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2570–2581, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rémi Cardon and Adrien Bibal. 2023. On operations in automatic text simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 116–

130, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Daniel Ferrés and Horacio Saggion. 2022. ALEXSIS: A dataset for lexical simplification in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3582–3594, Marseille, France. European Language Resources Association.

Sian Gooding and Ekaterina Kochmar. 2019. Recursive context-aware lexical simplification. 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 4853–4863, Hong Kong, China. 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.

Emil Graichen and Arne Jonsson. 2023. Context-aware Swedish lexical simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 11–20, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Nathan S Hartmann, Gustavo H Paetzold, and Sandra M Aluísio. 2018. SIMPLEX-PB: A lexical simplification database and benchmark for Portuguese. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pages 272–283. Springer.

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.

Tomoyuki Kajiwara and Kazuhide Yamamoto. 2015. Evaluation Dataset and System for Japanese Lexical Simplification. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 35–40, Beijing, China. Association for Computational Linguistics.

Tomonori Kodaira, Tomoyuki Kajiwara, and Mamoru Komachi. 2016. Controlled and Balanced Dataset for Japanese Lexical Simplification. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 1–7.

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.

Kang Liu, Jipeng Qiang, Yun Li, Yunhao Yuan, Yi Zhu, and Kaixun Hua. 2023. Multilingual lexical simplification via paraphrase generation. *arXiv preprint arXiv:2307.15286*.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553.

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, Toronto, Canada. Association for Computational Linguistics.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023b. Deep Learning Approaches to Lexical Simplification: A Survey.

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

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.

Jenny A Ortiz-Zambrano, César Espin-Riofrio, and Arturo Montejo-Raéz. 2023. LegalEc: A new corpus for complex word identification research in law studies in Ecuatorian Spanish. *Procesamiento del Lenguaje Natural*, 71:247–259.

Ellie Pavlick and Chris Callison-Burch. 2016. Simple PPDB: A Paraphrase Database for Simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 143–148, Berlin, Germany. Association for Computational Linguistics.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, Yang Shi, and Xindong Wu. 2021a. LSBert: Lexical simplification based on BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3064–3076.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34-05, pages 8649–8656.

Jipeng Qiang, Xinyu Lu, Yun Li, Yunhao Yuan, and Xindong Wu. 2021b. Chinese lexical simplification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1819–1828.

Eva Rolin, Quentin Langlois, Patrick Watrin, and Thomas François. 2021. FrenLyS: A Tool for the Automatic Simplification of French General Language Texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3 September, 2021*, pages 1196–1205. INCOMA Ltd.

Horacio Saggion, Stefan Bott, and Luz Rello. 2016. Simplifying words in context. experiments with two lexical resources in spanish. *Computer Speech and Language*, 35:200–218.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Matthew Shardlow. 2013. 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, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

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

Kim Cheng Sheang, Anaïs Koptient, Natalia Grabar, and Horacio Saggion. 2022. Identification of complex words and passages in medical documents in French. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 116–125, Avignon, France. ATALA.

Kim Cheng Sheang and Horacio Saggion. 2023. Multilingual controllable transformer-based lexical simplification. *Procesamiento del Lenguaje Natural*, 71:109–123.

Sanja Stajner, Daniel Ibanez, and Horacio Saggion. 2023. LeSS: A computationally-light lexical simplifier for Spanish. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1132–1142, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Safura Adeela Sukiman, Nor Azura Husin, Hazlina Hamdan, and Masrah Azrifah Azmi Murad. 2024. A Hybrid Personalized Text Simplification Framework Leveraging the Deep Learning-based Transformer Model for Dyslexic Students. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 34(1):299–313.

Anaïs Tack. 2021. *Mark my words! On the automated prediction of lexical difficulty for foreign language readers*. Ph.D. thesis, UCL-Université Catholique de Louvain.

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.

8. Language Resource 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.

Aleksei V Abramov, Vladimir V Ivanov, and Valery D Solovyev. 2023. Lexical complexity evaluation based on context for russian language. *Computación y Sistemas*, 27(1):127–139.

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.

Salehah Omar, Juhaida Abu Bakar, Maslinda Mohd Nadzir, Nor Hazlyna Harun, and Nooraini Yusoff. 2022. Malay lexical simplification model for non-native speaker. In *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–6.

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.

Cheng-Zen Yang, Jin-Jian Li, and Shu-Chang Lin. 2023. Lexical complexity prediction using word embeddings. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 279–287, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

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.

Appendix

Language	Target	Context	Complexity	Substitutions
Catalan	successius	Els manifestants han denunciat que en els darrers anys "els successius governs d'aquí i d'allà han passat del menyspreu al descessius governs d'aquí i d'allà han passat del menyspreu al descessius al manament de l'escola pública, començant a Catalunya per la pròpia LEC" i han alertat que s'està "deteriorant greument la qualitat de l'educació pública i les condicions laborals dels treballadors".	0.45	diferents, successius, contigus, differents, diversos, continus, consecutius, seguit, seguits, ultims, tots, tot el conjunt, succesius, anteriors, previs
English	external	(If your robot has an external ROM chip, then that is the one that is pulled and replaced.	0.05	outer, outside, exterior, external
Filipino	agaw	Akin na 'yan! biglang agaw ni Karlo sa laruan ni Lara.	0.075	agaw, kuha, kinuwa, nakaw
French	entretien	Cette gratuité n'est que partielle puisque une partie des impôts fonciers payés par les propriétaires est consacrée à l'entretien des réseaux de distribution.	0.675	réparation, maintien, nettoyage, gestion, entretien, tenue, restauration, réfection, revue, maintenance, soins
German	Grausen	Das Grausen überwältigte alle seine Sinne, er stürzte verworren aus dem Zimmer durch die öden widerhallenden Gemächer und Säulengänge hinab.	0.6	Angst, Furcht, Schreck, Panik, Gruseln, Grauen, Entsetzen
Italian	perduta	Aveva l'aria di una perduta nobiltà: la miseria gli si leggeva tutta nel volto e nella camicia sbrindellata sul petto.	0.08	persa, andata, antica, finita, inesistente, passata, smarrita"
Japanese	掲載した	ドラマに関する感想を募集し、週ごとにピックアップして回答も掲載した。	0.32	載せた、書いた、紹介した、発表した、公開した、知らせた、紙面に載せた、記載し情報を伝えた
Portuguese	estradas	as equipes contratadas pelo departamento de estradas de rodagem do paraná (der/pr) estão fazendo a primeira camada de asfalto no trecho da rodovia que passa por baixo da trincheira da jacob macanhan e também nos acessos laterais para a avenida camilo di Iellis e para os bairros da região de pinhais	0.08	ruas, caminhos, vias, rodovias
Sinhala	වාර්තාවක්	අසල තිබිලා ලුහාඬයෝ ලෝඛාවන්ම ලේඛකයාව වැඩ විදා තරුණයෝ එතනදී සමහ සමකියා නම මූ ලෙස්ටර් සමහර සදීධ පැවිදිය පුවිදියේ ඌබ මිනි ජීවනයක් ලැබිගේ වෙ නම අඟහර පෙරෙ දුන ලේ.	0.67	ලුහුරුගාස, සාත්තුවක්, ලුදය, සමුදුරුස්කාය, ලුදරුක
Spanish	notifique	Notifique a su Banco o institución financiera la pérdida o robo de sus tarjetas, chequeras o si sospecha que alguien está utilizando sus cuentas sin su permiso.	0.45	diga, avise, informe, comunique, avisa, alerten, comuniquen, expliquen, indique

Table 2: One example per language in the dataset

SIERA: An Evaluation Metric for Text Simplification using the Ranking Model and Data Augmentation by Edit Operations

Hikaru Yamanaka and Takenobu Tokunaga

School of Computing, Tokyo Institute of Technology

Tokyo Meguro Ôokayama 2-12-1, Japan

hyamanak@lycorp.co.jp, take@c.titech.ac.jp

Abstract

Automatic evaluation metrics are indispensable for text simplification (TS) research. The past TS research adopts three evaluation aspects: fluency, meaning preservation and simplicity. However, there is little consensus on a metric to measure simplicity, a unique aspect of TS compared with other text generation tasks. In addition, many of the existing metrics require reference simplified texts for evaluation. Thus, the cost of collecting reference texts is also an issue. This study proposes a new automatic evaluation metric, SIERA, for sentence simplification. SIERA employs a ranking model for the order relation of simplicity, which is trained by pairs of the original and simplified sentences. It does not require reference sentences for either training or evaluation. The sentence pairs for training are further augmented by the proposed method that utilizes edit operations to generate intermediate sentences with the simplicity between the original and simplified sentences. Using three evaluation datasets for text simplification, we compare SIERA with other metrics by calculating the correlations between metric values and human ratings. The results showed SIERA's superiority over other metrics with a reservation that the quality of evaluation sentences is consistent with that of the training data.

1. Introduction

Text simplification (TS) rewrites texts into simple and understandable ones while retaining their original meaning (Alva-Manchego et al., 2021). TS is expected to be an assistive technology for readers like children, non-native speakers and people with reading difficulties (Gooding, 2022). Recent TS models can generate fluent sentences by leveraging neural machine translation techniques, transforming a complicated sentence to its simplified counterpart within the same language (Al-Thanyyan and Azmi, 2021).

The performance of TS systems has been evaluated in terms of the following three aspects (Martin et al., 2018; Alva-Manchego et al., 2020, 2021).

- Fluency: Is the simplified text natural and free from grammatical errors?

- Meaning preservation: Does the simplified text retain the core meaning of the original?

- Simplicity: Is the simplified text easier to understand than the original?

Fluency and meaning preservation are common evaluation aspects in text generation tasks in general, and several automatic evaluation metrics have been proposed (Sai et al., 2022). In particular, BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020) are popular metrics for evaluating fluency and meaning preservation of texts. In contrast, simplicity is a unique evaluation aspect of TS and indispensable for TS research.

Automatic evaluation metrics are classified into two categories: reference-based and reference-free metrics. Reference-based metrics utilize reference texts for calculating evaluation scores for the target text, while reference-free metrics do not. Evaluation metrics for the text generation tasks are often reference-based. However, collecting manually written references for evaluation is expensive and time-consuming. In addition, it is inappropriate to regard the reference texts as the only correct output since there can be multiple acceptable simplified texts. Against this backdrop, we develop a reference-free metric for simplicity in this study.

To evaluate simplicity in TS, several automatic evaluation metrics have been proposed, including both reference-free (Kincaid et al., 1975; Sulem et al., 2018b) and reference-based (Papineni et al., 2002; Xu et al., 2016; Zhang et al., 2020) methods. However, it has been reported that these existing metrics are inappropriate for evaluating simplicity because of low correlation with manual evaluation (Alva-Manchego et al., 2020, 2021; Scialom et al., 2021). The evaluation metric of simplicity in TS research is still an open problem.

In this study, we limit the scope of TS to a sentence and propose a novel reference-free metric for evaluation of sentence simplicity, which we call **SIERA** (SImplification metric based on Edit operation through learning to RAnk). SIERA requires only parallel corpora of original and simplified sentences for training the evaluation model. The references are not necessary for calculating the evaluation scores. Following the framework of previous reference-free trainable evaluation met-

rics for other than text simplification (Wu et al., 2020; Maeda et al., 2022), the training procedure of SIERA consists of two parts: (1) learning-to-rank for determining the order relation of simplicity in training parallel corpora and (2) data augmentation to increase the number of training sentence pairs using edit operations between the original and simplified sentence pairs.

We summarize our contribution as follows.

- We propose a novel reference-free automatic evaluation metrics SIERA for sentence simplification, which can be trained only by a parallel corpus of original and simplified sentences.

- We develop a data augmentation method for extending the parallel corpus by considering edit operations between the original and simplified sentences.

- We demonstrate the superiority of SIERA to other automatic evaluation metrics for TS on three different evaluation datasets.

2. Related Work

Reference-based metrics
Reference-based metrics need reference sentences written by humans to evaluate simplified sentences. SARI (Xu et al., 2016), BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020) are common metrics in TS. SARI is a widely used metric for evaluating simplicity, which calculates the percentage of correctly added, kept, and deleted n-grams among the input, output and reference sentences. However, because SARI was initially proposed to evaluate lexical simplification, it is less suitable for evaluating simplified sentences with multiple rewriting operations (Alva-Manchego et al., 2020).

BLEU calculates a similarity score based on n-gram matching between output and reference sentences. Despite its simple computation and interpretability, BLEU is not recommended as a simplicity metric for sentences with splitting operations (Sulem et al., 2018a).

In contrast to BLEU and SARI, which rely on surface features like n-grams, BERTScore utilizes the BERT (Devlin et al., 2019) embeddings to compute sentence similarities considering contextual meaning. BERTScore aligns tokens of the output and reference sentences to calculate the cosine similarity between the aligned token embeddings. Alva-Manchego et al. (2021) reported that BERTScore is superior to BLEU and SARI in simplicity evaluation.

Recently, learnable reference-based automatic evaluation metrics have also been proposed. Maddela et al. (2023) developed LENS, which employs an adaptive ranking loss to weight reference

sentences based on their similarity to the sentence to evaluate in terms of edit operation. The LENS metric correlates more with human evaluation than the previous reference-based metrics, such as SARI and BERTScore.

However, the reference-based methods have a drawback; collecting manually written references is expensive and time-consuming. Also, Alva-Manchego et al. (2021) pointed out that a high similarity to the reference does not necessarily indicate high simplicity since there can be acceptable sentences other than references, and manually written references have diverse levels of simplicity against the original sentences.

Reference-free metrics
Reference-free metrics evaluate sentences without references. SAMSA (Sulem et al., 2018b) calculates whether the semantic structure between the input and output sentences is maintained after sentence splitting. However, SAMSA focuses on simplification by sentence splitting; its evaluation performance is poor for simplification with multiple rewriting operations (Alva-Manchego et al., 2021). FKGL (Kincaid et al., 1975) is another reference-free metric, which calculates from the average number of words and syllables. FKGL was initially proposed as a readability metric for grade levels in the United States, but it is often used to evaluate simplicity in TS. Tanprasert and Kauchak (2021) showed that FKGL is less robust against superficial edit operations, claiming that it is inappropriate for simplicity evaluation.

Vajjala and Meurers (2016) proposed the first pairwise ranking model to predict the readability of sentences. They created a classical ranking model that takes into account lexical and syntactic features to predict the readability of sentences and proposed to use it as an automatic evaluation metric of TS. Lee and Vajjala (2022) proposed a Neural Pairwise Ranking Model (NPRM) to predict sentence readability, which is a pairwise ranking model based on neural dense layers and BERT embedding. NPRM has not yet been investigated to see if it can be used for simplicity evaluation. We propose SIERA by extending the NPRM architecture.

More recently, Cripwell et al. (2023) proposed a learnable reference-free metric Simplicity Level Estimate (SLE) that calculates the absolute simplicity score of a single sentence. The goodness of simplification from the original to simplified sentences is calculated by the difference in their SLE scores. Unlike SLE, SIERA directly evaluates simplification for a given pair of original and simplified sentences.

Edit operations
Edit operations are often utilized in the simplification models. Alva-Manchego et al. (2017) pro-

posed a sequence transformation model that predicts edit operation tags such as deletion, replacement, and addition. Dong et al. (2019) extended this model to EditNTS, which performs edit operation prediction and adaptation in parallel, leveraging data that is automatically assigned token-by-token edit operations using dynamic programming. In recent years, a sentence simplification model has been proposed, which incrementally adds edit operations to improve a simplicity metric through unsupervised learning (Dehghan et al., 2022).

There is also a trend toward developing a typology of edit operations. Cardon et al. (2022) manually annotated a TS corpus with edit operations according to their thoughtful typology, suggesting the importance of TS evaluation in terms of fine-grained edit operation units. Yamaguchi et al. (2023) proposed a taxonomy of edit operations at the surface and content levels for understanding TS systems. Heineman et al. (2023) also organized 21 categories of edit operations for TS evaluation. Recently, there have been attempts to automatically generate these typologies of edit operations using LLM (Cardon and Bibal, 2023).

3. Resources

3.1. Training data

We use Newsela (Xu et al., 2015) as the training data for SIERA. Newsela is built upon data from 1,130 English news articles manually rewritten by professional editors in four levels of plain language to match the grade level of children, i.e. in principle, the original article has four variants corresponding to each simplicity level (1– 4), with 4 being the most simple level. The Newsela data is composed of parallel data aligned by sentences using Jaccard similarity. The number of total sentence pairs is 141,582.

3.2. Evaluation data

The evaluation data set for TS evaluation metrics consists of pairs of original and simplified text and manually assigned evaluation ratings for each pair regarding fluency, meaning preservation, and simplicity. The evaluation metrics are evaluated by measuring the correlation between these manual ratings and the evaluation scores obtained from the evaluation metrics. This study uses three sets of evaluation data, which are English corpora.

Simplicity-DA
Simplicity-DA (Alva-Manchego et al., 2021) is a data set consisting of original sentences from TurkCorpus (Xu et al., 2016) and corresponding simplified sentences created by six automatic sim-

plification models[1]. Each model generated 100 simplified sentences, resulting in 600 sentence pairs. The manual ratings are assigned as continuous values ranging from 0 to 100. Fifteen ratings are collected for each sentence pair.

Human-Likert
Human-Likert (Scialom et al., 2021) also uses TurkCorpus as the original sentences, but unlike Simplicity-DA, the simplified sentences are written manually, comprising 100 sentence pairs in total. Thirty human ratings ranging from 0 to 100 are collected for each sentence pair.

SimpDA$_{2022}$
SimpDA$_{2022}$ (Maddela et al., 2023) uses source texts extracted from Wikipedia from 22/10/2022 to 24/11/2022, to evaluate long and complex sentences. These source sentences are simplified by two humans and four recent TS models[2], resulting in a total of 360 sentence pairs. Three manual ratings ranging from 0 to 100 were assigned to each sentence pair.

4. SIERA ranking model

We propose SIERA by extending NPRM (Lee and Vajjala, 2022). This section describes the outline of NPRM and its possible improvement. Then, we propose a SIERA ranking model based on the NPRM architecture.

4.1. NPRM

Training NPRM uses only parallel data consisting of original sentences and their simplified sentences during training. Let n be the total number of the original sentences, s_i be the i-th original sentence, and s_i' be the corresponding simplified sentence. The instances for training data are made by concatenating s_i and s_i' by separating a SEP token in both orders as in (1). The arrow over p_i indicates the order of the original and simplified sentences in the pair, i.e. the source of the arrow indicates the original sentence.

$$\overrightarrow{p_i} = \textit{concat}(s_i; \text{SEP}; s_i')$$
$$\overleftarrow{p_i} = \textit{concat}(s_i'; \text{SEP}; s_i) \tag{1}$$

We use notation p_i for denoting both $\overrightarrow{p_i}$ and $\overleftarrow{p_i}$. The expected correct label y_i for p_i is either row vector $[0, 1]$ for $\overrightarrow{p_i}$ or $[1, 0]$ for $\overleftarrow{p_i}$. The element value 1 indicates the simplified sentence position in a pair.

---

[1]ACCESS (Martin et al., 2020), DMASS-DCSS(Zhao et al., 2018), Dress-Ls (Zhang and Lapata, 2017), Hybrid (Narayan and Gardent, 2014), PBMT-R (Wubben et al., 2012) and SBMT-SARI (Xu et al., 2016).

[2]GPT-3.5 (Ouyang et al., 2022) w/ zero and few-shot, Muss (Martin et al., 2022) and T5-3B (Raffel et al., 2019).

NPRM calculate the output o_i through BERT (Devlin et al., 2019) and a fully-connected feed-forward neural network (FFNN) as in (2), where BERT($\cdot$) denotes an output vector corresponding to the CLS token of BERT[3]. The output o_i is a two-dimensional column vector, where each vector element represents the probability that the sentence corresponding to that element is a simplified sentence.

$$o_i = softmax(\text{FFNN}(\text{BERT}(p_i))) \qquad (2)$$

The created training data p_i and corresponding labels y_i are fed into the model and trained with a loss function (3),

$$L = -\sum_{i=1}^{n} y_i \cdot \log(o_i), \qquad (3)$$

where y_i and o_i represents both $\overrightarrow{y_i}$ and $\overleftarrow{y_i}$, and both $\overrightarrow{o_i}$ and $\overleftarrow{o_i}$ respectively and correspondingly. $\log(o_i)$ denotes the element-wise application of the logarithmic function.

Inference   Given an original sentence s and its simplified sentence s', NPRM calculates a readability score as in (6).

$$\overrightarrow{p} = concat(s; \text{SEP}; s'), \qquad (4)$$
$$\overrightarrow{o} = softmax(\text{FFNN}(\text{BERT}(\overrightarrow{p}))), \qquad (5)$$
$$readability\ score = [0, 1] \cdot \overrightarrow{o}. \qquad (6)$$

We can utlize this readability score to measure the simplicity of s' against s.

4.2.   Improvement of NPRM

Comparing the training phase ((1) and (2)) and the inference phase ((4), (5) and (6)) of NPRM, we find that NPRM utilizes both forwardly and backwardly-ordered pairs ($\overrightarrow{p_i}$ and $\overleftarrow{p_i}$) for training, but utilizes only the forwardly-ordered pairs for inference. We suspect that the inference phase of NPRM does not fully utilize the learned result.

We propose to utilize the backwardly-ordered sentence pair ($\overleftarrow{p}$) in addition to the forwardly-ordered sentence pair ($\overrightarrow{p}$) also in the inference phase to calculate the score as in (9). Equation (7) and (8) are the counterpart of (4) and (5), respectively.

$$\overleftarrow{p} = concat(s'; \text{SEP}; s), \qquad (7)$$
$$\overleftarrow{o} = softmax(\text{FFNN}(\text{BERT}(\overleftarrow{p}))), \qquad (8)$$
$$simplicity\ score = \frac{1}{2}([0, 1] \cdot \overrightarrow{o} + [1, 0] \cdot \overleftarrow{o}). \qquad (9)$$

---

[3]Although NPRM has freedom in the choice of neural network architectures; we adopt BERT and FFNN following Lee and Vajjala (2022)'s experimental setting.

5.   Data Augmentation

This section describes a method to extend the parallel data for training the SIERA ranking model. We utilize edit operations for simplification to increase the original and simplified sentence pairs. Given a pair of an original sentence s and its simplified sentence s', the simplification can be represented by a set of edit operations that transform s to s'. Alva-Manchego et al. (2020) reported that applying more edit operations for simplification makes the resultant sentence simpler. Following their finding, we apply subsets of the edit operations that bridge between s and s' to create new sentences which are simpler than s but less simple than s'. We call them *intermediate sentences*. Suppose we create an intermediate sentence $\hat{s}$ from s by applying a subset of edit operations; we can create new sentence pairs $(s, \hat{s})$ and $(\hat{s}, s')$. Theoretically, if we can transform s to s' through N operations, we could create $2^N - 2$ intermediate sentences; thus we obtain $2(2^N - 2)$ new sentence pairs for training the SIERA ranking model.

Following Dong et al. (2019), we consider two levels of the edit operation: *token unit edit operation* (TE) and *span unit edit operation* (SE). TE is an edit operation applied to each token in the original sentence to transform it into a simplified sentence. There are three types of TE: ADD token, DELete token, and KEEP token. To extract TEs from given sentence pairs, we adopt the implementation by Dong et al. (2019)[4]. SEs are constructed by concatenating consecutive TEs except for KEEP. There are following three types of SEs. Figure 1 shows an example of extracted TEs and SEs.

- ADD-DEL span: A span in which one or more consecutive ADDs and DELs are combined in this order. This corresponds to lexical simplification and sentence splitting.

- DEL span: One or more consecutive DEL spans other than the ADD-DEL span. This corresponds to the deletion of unnecessary information.

- ADD span: One or more consecutive ADD spans other than the ADD-DEL span. This corresponds to the addition of necessary information.

We create intermediate sentences by applying a subset of the extracted SEs to the original sentence. However, an arbitrary subset of the extracted SEs does not always produce a valid sentence. For instance, among four SEs in Figure 1,

---

[4]https://github.com/YueDongCS/EditNTS/blob/master/label_edits.py

Original	According to Ledford , Northrop executives said they would build substantial parts of the bomber in Palmdale , creating about 1,500 jobs .
TEs	KEEP KEEP KEEP KEEP KEEP DEL KEEP KEEP KEEP KEEP ADD(most) DEL DEL KEEP KEEP KEEP ADD(parts) KEEP KEEP ADD(.) ADD(It) ADD(would) ADD(create) DEL DEL DEL KEEP KEEP KEEP
SEs	KEEP KEEP KEEP KEEP KEEP DEL KEEP KEEP KEEP KEEP ADD(most) DEL DEL KEEP KEEP KEEP ADD(parts) KEEP KEEP ADD(.) ADD(It) ADD(would) ADD(create) DEL DEL DEL KEEP KEEP KEEP
Simplified	According to Ledford, Northrop said they would build most of the bomber parts in Palmdale. It would create 1,500 jobs .

Figure 1: Example of extracted TEs and SEs. Each highlighted color represents ADD-DEL span , DEL span , and ADD span . In this example, four SEs are extracted in total.

we can apply the first DEL SE and the last ADD-DEL SE independently, but the second and third SEs must be applied simultaneously to rewrite "substantial parts of the bomber" to "most of the bomber parts". Applying only one of them produces invalid sentences. Although we can theoretically create $2^4 - 2 = 14$ intermediate sentences from this example, invalid sentences should be excluded from them.

To exclude irrelevant intermediate sentences, we discard intermediate sentences dissimilar from the original and simplified sentences in terms of BERTScore (Zhang et al., 2020). More concretely, we calculate BERTscore$_{f1}$ of an intermediate sentence with its original and simplified sentence each and average them. These average scores are further averaged across the entire generated intermediate sentences to determine a threshold. We discard the intermediate sentences that have lower average scores than the threshold. We randomly choose m sentences from the remaining intermediate sentences to augment the training sentence pairs.

6. Experiment

6.1. Experimental settings

Training data and models to compare
We use the Newsela dataset for training the SIERA ranking model. Newsela comprises original news articles and corresponding simplified variants over four simplification levels. We first train a baseline model (Base) using 16,084 sentence pairs of the original and its most simplified sentence in Newsela. Next, we extend the sentence pairs for the baseline model by our proposed augmentation method described in section 5, resulting in 38,120 sentence pairs in total. We adopt a single intermediate sentence for each original sentence pair, i.e. the hyperparameter $m = 1$. Theoretically, we should obtain three times the original number of sentence pairs, but we have fewer sentence pairs in reality due to the filtering process to exclude irrelevant intermediate sen-

tences. We call the SIERA model trained with this extended data +Silver. Furthermore, we extend the sentence pairs for the baseline using manually simplified sentences of the intermediate level of Newsela, resulting in 46,470 sentence pairs. The difference from the +Silver's training data is that the quality of intermediate sentences is guaranteed because they are written by professional editors. Therefore, we do not apply filtering in this data augmentation. Despite no filtering, the total number of sentence pairs is slightly fewer than three times that of the Base training data. This is because some Newsela articles do not have simplified sentences of intermediate levels. We call the SIERA model trained with this extended data +Gold.

We also consider the variants of these three models in the inference phase. The SIERA model uses both forwardly and backwardly-ordered sentence pairs in the inference phase (a two-way model). We consider the models that use only one of them in the inference phase and denote them by putting an arrow over the model name, i.e. $\rightarrow$ and $\leftarrow$ indicate using only forwardly or backwardly-ordered sentence pairs, respectively (a one-way model). Note that $\overrightarrow{\text{Base}}$ is equivalent to NPRM.

We also consider the existing reference-based (SARI, BLEU[5] , BERTScore[6]) and reference-free metrics (SAMSA, FKGL)[7].

Hyperparameters
We used the bert-base-uncased[8] model from Huggingface Transformers as a pre-training model and a ranking model was implemented using Pytorch Lightning[9]. We set the parameters of the FFNN

---

[5]Sacrebleu with max_ngram_order = 4 (https://github.com/mjpost/sacrebleu)

[6]The official implementation with roberta-large model (https://github.com/Tiiiger/bert_score)

[7]EASSE (Alva-Manchego et al., 2019)

[8]https://huggingface.co/bert-base-uncased

[9]https://www.pytorchlightning.ai

	Simplicity-DA				Human-Likert				SimpDA$_{2022}$			
	Pearson		Spearman		Pearson		Spearman		Pearson		Spearman	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Base	.029	.015	.016	.011	.549	.034	.541	.026	.394	.025	.387	.018
$\overrightarrow{\text{Base}}$ (NPRM)	.012	.030	.000	.013	.434	.057	.451	.042	.308	.047	.336	.021
$\overleftarrow{\text{Base}}$	.035	.031	.024	.023	.458	.052	.477	.031	.334	.040	.366	.036
+Silver	**.049**	.017	**.027**	.016	**.580**	.035	**.547**	.033	.366	.026	**.401**	.028
$\overrightarrow{\text{+Silver}}$	**.017**	.034	**.003**	.068	**.452**	.069	**.483**	.051	.255	.049	**.342**	.058
$\overleftarrow{\text{+Silver}}$	**.059**	.023	**.046**	.026	**.559***	.060	**.501**	.044	.337	.046	**.384**	.032
+Gold	.052	.017	.026	.013	.607	.022	.604	.017	.446	.027	.465	.025
$\overrightarrow{\text{+Gold}}$	.025	.020	.019	.020	.535	.038	.561	.027	.393	.039	.421	.026
$\overleftarrow{\text{+Gold}}$	.065	.016	.033	.011	.555	.047	.561	.028	.412	.033	.459	.027
SARI	.358	-	.326	-	.390	-	.373	-	-	-	-	-
BLEU	.507	-	.482	-	.349	-	.312	-	-	-	-	-
BERTScore$_p$	.628	-	.660	-	.417	-	.387	-	-	-	-	-
BERTScore$_r$	.505	-	.502	-	.374	-	.401	-	-	-	-	-
BERTScore$_{f1}$	.590	-	.579	-	.393	-	.393	-	-	-	-	-
SAMSA	.060	-	.068	-	−.374	-	−.319	-	−.083	-	−.122	-
FKGL	.117	-	.110	-	−.353	-	−.359	-	−.387	-	−.353	-

Table 1: Correlations of SIERA (top half) and other metrics (bottom half) with three evaluation datasets. The mean and standard deviation of ten runs with different seeds are shown for SIERA. The single calculation result is shown for other metrics since they have no seed. Because SimpDA$_{2022}$ has no reference, the results for the reference-based methods are not available. The bold values for the +Silver family indicate superiority over the corresponding Base value. The asterisk (*) denotes the significant difference at p < .05 of the two-sided permutation test.

Dataset	Simplicity-DA		Human-Likert		SimpDA$_{2022}$	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
$\Delta_{(\text{+Silver,Base})}$	.020	.011	.031	.006	−.028	.014
$\Delta_{(\text{+Gold,+Silver})}$	.003	−.001	.027	.057	.080	.064

Table 2: Difference of the correlation coefficient (mean) between the models

and the BERT last layer learnable. AdamW[10] was chosen as the optimization algorithm, with the number of epochs set to 10 and the batch size to 16. We used cross-entropy loss[11] as the loss function and the learning rate was set to 10^{-4}. Twenty percent of the training data was used for validation. We adopted early stopping based on the loss with the validation data and selected the checkpoint at the epoch with the lowest loss[12] We conduct the experiment with random seed values ten times, and report their average results.

Evaluation data
We use three data sets, Simplicity-DA, Human-Likert and SimpDA$_{2022}$, for evaluating the models. Correlations between the model prediction scores

---

[10]https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html
[11]https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html
[12]The codes are available at https://github.com/hyama1569/siera.

and the human ratings are calculated in two ways: Pearson's correlation coefficient and Spearman's rank correlation coefficient.

6.2. Results and discussion

The top half of Table 1 shows the results of the SIERA models with different settings. Comparing the two-way models (model name without an arrow) and the one-way models (those with an arrow), the two-way models are consistently superior to their one-way counterparts for Human-Likert and SimpDA$_{2022}$ but not for Simplicity-DA. Furthermore, the backwardly-ordered models are superior to the forwardly-ordered models for all datasets. We could not find an explanation for this asymmetry yet. This is an unfortunate result for NPRM, which employs the forwardly-ordered model.

+Silver outperforms Base except for the case of Pearson's coefficient with SimpDA$_{2022}$. Table 2 shows the difference between the mean values of the correlation coefficients in Table 1, where $\Delta_{(X,Y)}$

	Simplicity-DA				Human-Likert				SimpDA$_{2022}$			
	Pearson		Spearman		Pearson		Spearman		Pearson		Spearman	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Base	.029	.015	.016	.011	.549	.034	.541	.026	.394	.025	.387	.018
+Silver(50%)	.065	.016	.033	.011	.555	.047	.561	.028	.349	.026	.400	.026
+Silver	.049	.017	.027	.016	.580	.035	.547	.033	.366	.026	.401	.028
$\Delta_{(+Silver(50\%),Base)}$	.050		.017		.006		.020		−.045		.013	
$\Delta_{(+Silver,+Silver(50\%))}$	−.016		−.006		.025		−.014		.017		.001	
+Gold(50%)	.039	.018	.015	.017	.601	.032	.593	.032	.439	.028	.459	.019
+Gold	.052	.017	.026	.013	.607	.022	.604	.017	.446	.027	.465	.025
$\Delta_{(+Gold(50\%),Base)}$	.010		−.001		.052		.052		.045		.072	
$\Delta_{(+Gold,+Gold(50\%))}$	.028		.011		.006		.011		.007		.006	

Table 3: Correlations of SIERA using half of the augmented data

denotes a difference of X's value from Y's value. This result confirms the effectiveness of the proposed data augmentation method. As the augmented data used for training +Gold are made from manually written intermediate sentences by professionals, we can consider the results of +Gold as an upper bound regarding the data augmentation. Table 2 shows that the gains from Base to +Silver tend to be larger than that from +Silver to +Gold for Human-Likert and SimpDA$_{2022}$. This result suggests room for improvement in the quality of the intermediate sentences derived by applying edit operations. Although we employed the BERTScore-based filtering to exclude irrelevant intermediate sentences, this filtering is still limited. We discarded dissimilar intermediate sentences to their original and simplified sentences regarding BERTScore$_{f1}$. The similarity judgement was done against a threshold calculated by averaging the similarity of all generated intermediated sentences. The thresholds for +Silver and +Gold datasets are quite close, i.e. 0.554 and 0.547, respectively. Therefore, sentence similarity is not enough for filtering irrelevant sentences. We need to consider more effective methods for obtaining high-quality intermediate sentences.

We conducted a supplemental experiment using half of the augmented data. The result is shown in Table 3. The rows $\Delta_{(+Silver,+Silver(50\%))}$ and $\Delta_{(+Gold,+Gold(50\%))}$ indicate that the augmented data size reduction does not significantly impact the correlation with the human ratings. They also show that the gains from Base to +Gold(50%) are consistently larger than that from +Gold(50%) to +Gold for Human Likert and SimpDA$_{2022}$. However, this does not hold for +Silver. This difference suggests that the quality of the augmented pairs, i.e. the intermediate sentences, has more impact on the correlation than their size, supporting our claim on the importance of the intermediate sentence quality.
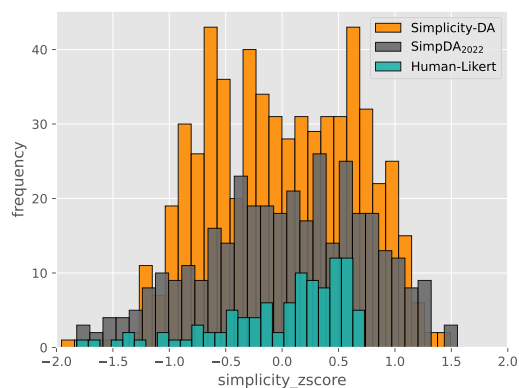
Figure 2: Distribution of simplicity scores of three datasets

The SIERA's correlation coefficients are far lower for Simplicity-DA than for the other two evaluation data sets. A possible explanation is a quality gap of simplified sentences between the training (Newsela) and evaluation data (Simplicity-DA). Simplified sentences in Newsela were written by human experts, while those in Simplicity-DA were generated by the six automatic simplification models, and four of them were rather old RNN-based models. SimpDA$_{2022}$ also includes two-thirds of its data automatically generated. However, they are all Transformer-based models, which can generate more fluent sentences than the RNN-based models used for Simplicidy-DA. We suspect that the difference between Simplicity-DA and SimpDA$_{2022}$ comes from the quality of simplified sentences to evaluate.

Figure 2 shows the distribution of human simplicity ratings of simplified sentences in the three evaluation datasets. Following Alva-Manchego et al. (2021), we normalized the ratings to z-scores ranging from −1 to 1. We notice that Simplicity-DA has a distinct lump on the left side, i.e. it has more

	Pearson		Spearman	
	mean	std	mean	std
Base	.236	.023	.264	.027
+Silver	.266	.033	.284	.031
+Gold	.281	.028	.288	.028
SARI	.121	-	.137	-
BLEU	.212	-	.243	-
BERTScore$_p$	.195	-	.203	-
BERTScore$_r$	.073	-	.157	-
BERTScore$_{f1}$	.114	-	.144	-
SAMSA	.038	-	$-.005$	-
FKGL	$-.041$	-	$-.062$	-

Table 4: Result for high-quality Simplicity-DA subset

low-rated sentences than the other two. Considering we trained the SIERA model using only human-written high-quality sentences, we suspect that the SIERA model could not learn to score low-quality simplified sentences. To confirm our hypothesis, we select high-quality simplified sentences from Simplicity-DA and calculate the correlation between their human ratings and the SIERA scores. We normalize the scores by transforming the human-rated fluency, meaning preservation and simplicity scores into z-scores, and select simplified sentences where all three scores are positive. The resultant high-quality subset contains 196 sentences, about one-third of the entire Simplicity-DA. Table 4 shows the result for the subset, which supports our hypothesis. Our above claims in Human-Likert and SimpDA$_{2022}$ also hold in the high-quality Simplicity-DA subset.

	Pearson		Spearman	
	mean	std	mean	std
Base	.623	.028	.639	.018
+Silver	.621	.025	.654	.023
SARI	.324	-	.293	-
BLEU	.573	-	.511	-
BERTScore$_p$	.602	-	.595	-
BERTScore$_r$	.507	-	.476	-
BERTScore$_{f1}$	.578	-	.535	-
SAMSA	.078	-	.096	-
FKGL	.076	-	.082	-

Table 5: Results of Simplicity-DA-trained SIERA

To further confirm our hypothesis, we trained a SIERA model using a part of the Simplicity-DA data. We randomly chose 80 simplified sentences from each simplification model of Simplicity-DA for training, resulting in 480 sentence pairs. The remaining 120 sentence pairs were held for testing. Although the model outputs could be a simplified sentence in a pair, their human rating might be very low because the models do not always work well. We define a simplified sentence in a pair based on the human simplicity rating of the model outputs. When the human rating of the model output is lower than the average rating of the entire training data, the original sentence is considered a simplified sentence. Since the number of training data was small, we increased the training data by paring system outputs from the same original sentence. The sentence with a higher human rating in each pair is considered a simplified sentence. This operation increased the training data size to 1,538 sentence pairs in total.

This training data creation refers to human ratings. This is not a normal way of training the SIERA model, which uses only pairs of original and simplified sentences without human ratings. The experiment only aims to confirm our hypothesis. Table 5 shows the result, reinforcing our hypothesis on the sentence quality gap between the training and test data.

The bottom half of Table 1, 4 and 5 shows the correlations of the other evaluation metrics. BERTScore$_p$ shows good performance for both Simplicity-DA and Human-Likert, which is consistent with the results of the previous study (Alva-Manchego et al., 2021). Comparing +Silver with the other evaluation metrics, we can see that +Silver consistently beats these metrics for the high-quality Simplicity-DA subset, Human-Likert and SimpDA$_{2022}$[13] Not to mention its high correlation, SIERA has a strong point that it does not require reference simplified sentences for evaluation. As we discussed, however, we need to be careful about the training data for the SIERA model, which should be consistent with the quality of the target sentences.

7. Conclusion and future work

We presented SIERA, a novel reference-free metric for evaluating sentence simplicity. SIERA adopts a pair-wise ranking model to predict the order relations of simplicity in the paired sentences. The model is trained by pairs of original and simplified sentences. Evaluating simplified sentences with SIERA requires only pairs of the original and simplified sentences, i.e. reference sentences are unnecessary. We also propose a data augmentation method by applying automatically extracted edit operations to the original sentence to generate intermediate sentences. The intermediate sentences are expected to have middle simplicity between the original and corresponding simplified sentences.

We evaluated SIERA using three evaluation data sets for text simplicity. The experimental results

---

[13]FKGL with Pearson's coefficient is the exception.

showed that as far as the quality of target sentences is consistent with that of the training data, SIERA correlates better with human ratings than other simplicity metrics, including reference-based metrics. SIERA does not require reference sentences but needs training. We must carefully choose the training data to maximize SIERA's potential.

Also, the augmented data by the proposed method contributed to improving SIERA's correlation with human ratings. To augment the training data, we automatically extracted edit operations from a pair of the original and simplified sentences and applied a subset of the operations to the original sentence to obtain intermediate sentences. However, we did not consider dependencies among operations in the application, which may cause irrelevant sentences, as we discussed in section 5. We applied the BERTScore-based filtering to exclude irrelevant sentences, but the experimental result suggested this filtering had a limitation. Improving the quality of intermediate sentences is one of the future research directions. Considering the syntactic structure of sentences might help to generate more relevant intermediate sentences.

As we have limited parallel corpora for text simplification, we could not verify the effectiveness of metrics employing a trainable model like SIERA for different domain texts from the training data. We found that the quality gap between the training and test data impacts the performance of the SIERA model. Likewise, the domain shift would have an impact as well. Parallel corpora for simplification have been built in several domains like administrative documents (Scarton et al., 2018), general medical documents (Devaraj et al., 2021), and radiology reports (Yang et al., 2023). However, they have not necessarily been assigned human ratings. They can be used for training models but not for the evaluation of metrics. Collecting parallel corpora for simplification in various domains that can be used both for training and evaluation is indispensable.

8. Bibliographical References

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Computing Surveys*, 54(2).

Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. 2017. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 295–305, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

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. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Rémi Cardon and Adrien Bibal. 2023. On operations in automatic text simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 116–130, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Rémi Cardon, Adrien Bibal, Rodrigo Wilkens, David Alfter, Magali Norré, Adeline Müller, Watrin Patrick, and Thomas François. 2022. Linguistic corpus annotation for automatic text simplification evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1842–1866, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Simplicity level estimate (SLE): A learned reference-less metric for sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12053–12059, Singapore. Association for Computational Linguistics.

Mohammad Dehghan, Dhruv Kumar, and Lukasz Golab. 2022. GRS: Combining generation and revision in unsupervised sentence simplification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 949–960, Dublin, Ireland. Association for Computational Linguistics.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

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.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

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.

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, Singapore. Association for Computational Linguistics.

J. Peter Kincaid, Robert P. Fishburne, R 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. In *Research Branch Report 8-75*.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland. 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.

Koki Maeda, Masahiro Kaneko, and Naoaki Okazaki. 2022. IMPARA: Impact-based metric for GEC using parallel data. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3578–3588, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

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.

Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric de La Clergerie, Antoine Bordes, and Benoît Sagot. 2018. Reference-less quality estimation of text simplification systems. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 29–38, Tilburg, the Netherlands. Association for Computational Linguistics.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 435–445, Baltimore, Maryland. 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 F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

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 on Associ-*

ation for Computational Linguistics, page 311–318, USA. Association for Computational Linguistics.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys*, 55(2).

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. 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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Thomas Scialom, Louis Martin, Jacopo Staiano, Éric Villemonte de la Clergerie, and Benoît Sagot. 2021. Rethinking automatic evaluation in sentence simplification. *CoRR*, abs/2104.07560.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018a. 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.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 685–696, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *CoRR*, abs/1603.06009.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised reference-free summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3612–3621, Online. Association for Computational Linguistics.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

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.

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, Dubrovnik, Croatia. Association for Computational Linguistics.

Ziyu Yang, Santhosh Cherian, and Slobodan Vucetic. 2023. Data augmentation for radiology report simplification. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1922–1932, Dubrovnik, Croatia. Association for Computational Linguistics.

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*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. 2018. Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173,

Brussels, Belgium. Association for Computational Linguistics.

9. Language Resource References

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.

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.

Thomas Scialom, Louis Martin, Jacopo Staiano, Éric Villemonte de la Clergerie, and Benoît Sagot. 2021. Rethinking automatic evaluation in sentence simplification. *CoRR*, abs/2104.07560.

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.

Transfer Learning for Russian Legal Text Simplification

Mark Athugodage, Olga Mitrofanova, Vadim Gudkov

HSE University, Saint-Petersburg State University, Saint-Petersburg State University

mathugodage@hse.ru, o.mitrofanova@spbu.ru, vvagudkov@gmail.com

Abstract

We present novel results in legal text simplification for Russian. We introduce the first dataset for such a task in Russian - a parallel corpus based on the data extracted from "Rossiyskaya Gazeta Legal Papers". In this study we discuss three approaches for text simplification which involve T5 and GPT model architectures. We evaluate the proposed models on a set of metrics: ROUGE, SARI and BERTScore. We also analysed the models' results on such readability indices as Flesch-Kinkaid Grade Level and Gunning Fog Index. And, finally, we performed human evaluation of simplified texts generated by T5 and GPT models; expertise was carried out by native speakers of Russian and Russian lawyers. In this research we compared T5 and GPT architectures for text simplification task and found out that GPT handles better when it is fine-tuned on dataset of coped texts. Our research makes a big step in improving Russian legal text readability and accessibility for common people.

Keywords: Text Simplification, Text Readability, Legal text, Russian, New corpus, T5, GPT

1. Introduction

Legal documents in almost all languages are considered to be long, complex and difficult to read for people without a domain specific expertise. The texts of laws, regulations, and various resolutions are written in a very specific formal style. Legal language implies abundance of professional terms, latinisms, references to other legal documents, at the same time, these texts are considered as unemotional and syntactically complicated. It is not uncommon when a single sentence in a legal document can be a page-long (Ramaswamy et al., 2023).

The complexity of legal documents, and especially laws, complicates the life of citizens without a domain specific expertise since there is the famous Latin maxima "Ignorantia legis non excusat" ("ignorance of the law does not excuse anyone"). The only choice a simple man has is to appeal to the lawyer who may elucidate a certain law or a group of laws, but not a complete set of laws in the country.

It's notable that the government acknowledges the existence of a problem dealing with the clarity of legal documents. We can mention that the Russian Parliament recommended lawmakers use simple sentences with "SVO" structure: Subject + Verb + Object. However, some evidence suggest that Russian court resolution complexity gets even higher and higher each year (Dmitrieva, 2017).

Another significant aspect of the text complexity issue is of sociolinguistic nature: we cannot make a legal text so simple that it would be comprehensible for all citizens. The first reason concerns disabled people not all of whom are able to read and properly understand legislative documents. Then, the second reason is that the Russian Federation is a multiethnic and multilingual country, and this admits that among the citizens there may be those who do not speak Russian perfectly. Thus, the relevance of the study is explained by the high needs of society in tools and techniques for simplifying legal documents.

Since this paper is devoted to Text Simplification, it is useful to say few words about this task. Text Simplification is a text-to-text generation task, likewise summarization, machine translation, paraphrasing and style transfer. This task is often confused with summarization because of similar nature. While text summarization is always considered to be an operation of text compression, text summarization can either "compress" a text, leave it as it was or even make it larger (Fenogenova and Sberbank, 2021).

The main goal of Text Simplification is to make it easier for reader to understand a text. It becomes necessary when people without a domain specific expertise try to learn a narrow-field text, for example, a medical text. Text Simplification aims to make a specific-domain text more clear for a broader audience (Van et al., 2020) .

In the given paper we present results of research aimed at the substantiation of the possibility formal simplification of legal documents based on neural network models. We focus our attention at the development of specialised parallel legal corpus which includes the data extracted from "Rossiyskaya Gazeta Legal Papers", fine-tuning of neural models from T5 and GPT families for the simplification of legal texts and evaluation for assessing the quality of simplification. The structure of the paper is as follows: in section.

2. Related work

2.1. Approaches to text simplification

Modern state-of-the-art approaches for text simplification include neural-based and rule-based. Text Simplification can be performed at the lexical level, syntactic level and by means of hybrid approaches. Lexical and syntactic simplification procedures should be considered as time-tested and in most cases imply using rule-based methods. The hybrid method of text simplification is the most recent and popular at present. Yet another alternative is provided by the back-translation method. We can distinguish it as a separate class of solutions since it may be rule-based or neural-based. Although many researchers refer to backed-translation as a text simplification method, it has more in common with paraphrasing. For Russian (Galeev et al., 2021) tried back-translation as solution for a complex text: they fine-tuned a BART model for machine-translation task and then compiled "double translation". Recent works on text simplification are focused on adaptation and fine-tuning of existing neural networks - mostly Transformers. Transformers nowadays have proved to be a very efficient model for a vast list of NLP tasks - text simplification is not an exception. LS-Bert(Garimella et al., 2022), a Transformer-based lexical simplification model, is a bright example of lexical simplification method. LSBert finds complex words and generates the substitutions, taking into account the context. LSBert should be considered as a facilitated approach since it omits certain NLP procedures, e.g. morphological transformation. Beyond the most famous text-to-text simplification based on Transformers, there is also an edit-based method. A good example of edit-based model is EditNTS, where for each token or n-gram there are four actions offered: ADD (add token) KEEP (do not change the token; leave it as it is) DELETE (delete token) STOP If, for example, there is a sentence "She gazed at me", then EditNTS would simplify it to "She watched me". To make such simplification, EditNTS would need the following actions: KEEP for "she", DELETE for "gazed", DELETE for "at", ADD for "watched", KEEP for "me", and STOP (Dong et al., 2019). There are some similar models: TST (an adaptation of GEC-ToR corrector) (Omelianchuk et al., 2021), FELIX](Mallinson et al., 2020) and LaserTagger (Malmi et al., 2019). Such models reproduce the idea of text editing, but focus not on grammatical and orthographic errors but on simplification of complex words and phrases.

2.2. Hybrid methods using Transformers

In general there are two approaches for hybrid text simplification using neural networks: sentence-level simplification (sentence by sentence) and document-level simplification (a whole document at once). Nowadays, the most popular approach for text simplification is Transformer-based model. Transformer architecture is based on self-attention. Self-attention (also known as intra-attention) is a mechanism relating different positions of a single sequence of tokens, which makes possible the computing a representation of the sequence and modelling global dependencies. There are 3 main types of types of Transformers:

- Encoder Transformers: BERT (Devlin et al., 2018), Longformer (Beltagy et al., 2020), XL-Net (Yang et al., 2019), Transformer-XL (Dai et al., 2019);

- Decoder Transformers: GPT (Radford et al., 2019), CTRL (Keskar et al., 2019);

- Encoder-Decoder Transformers: T5 (Raffel et al., 2020), BART(Lewis et al., 2019), LED (Beltagy et al., 2020), PEGASUS (Zhang et al., 2020).

The most relevant choice for text-to-text task is encoder-decoder Transformers. We choose T5 Transformers since it is a classical example of encoder-decoder. In future, it would be reasonable to try BART as well, but BART is similar to BERT in the encoder part, which implies language masking, but at this moment BART fine-tuning with or without masking is not included in the experimental design. Then, other options for text2text generation are generative models, i.e. decoder Transformers. We use the most renowned of them - GPT, since the other model, CTRL, isn't available for Russian yet. GPT-2 has achieved competitive performance on text summarization and simplification tasks. GPT-3 and GPT-4, as well as their modifications, are not open-source models, thus they are not available for fine-tuning. Most researchers use GPT-2 and their modifications for fine-tuning (for example, researchers used GPT-2 to fine-tune Indonesian summarizer (Khasanah and Hayaty, 2023)). Beyond GPT, there are also LLaMa and LLaMA 2, open-source LLMs from Meta AI - researchers often fine-tune these models for their specific tasks, including text simplification (Baez and Saggion, 2023).

With the emergence of large language models, NLP researchers and engineers started using prompt-engineering for many seq2seq tasks. So do they for text simplification task. People extensively use GPT-4 (Wu and Huang, 2023) with other LLMs being less popular. Although some researchers claim that transfer learning is "dead" (Pu et al.,

2023), experiments show that smaller models like BART are still perform not worse than LLMs (Sun et al., 2023). Evidence suggests that though many companies apply LLMs for their seq2seq tasks (including text simplification), smaller models are still in need, since there are some cases when one cannot train and deploy large models (Sharir et al., 2020; Chahal et al., 2020; Ahmed et al., 2023) .

3. Experimental dataset

3.1. Rationale for data selection

The crucial problem in fine-tuning seq2seq model is data availability. This problem is much more fatal for text simplification task, since there are no large datasets for this task - one can compare with a similar task, text summarization, for which there are dozen of datasets: XLSum, Newsela, CNN/DailyMail, etc. Some recent solutions are data annotation with LLMs (Gray et al., 2024). However, we find this method too risky for such a delicate field as law. Although modern LLMs are almost impeccable in performance, there is still place model hallucination as well as factual errors (Xu et al., 2024).

Since there was no dataset for Russian legal texts, we developed our own one[1]. We present dataset "Rossiyskaya Gazeta Legal Papers"[2] , which we made available on Kaggle. The dataset is based on legal papers and their simplified versions from "Rossiyskaya Gazeta" web newspaper. "Rossiyskaya Gazeta" is an official newspaper published by the Government of Russia. It's one of the widely available sources of legal documents for the citizens of Russia - the other one is a state-owned website `pravo.gov.ru`. Every important legal document (decisions of the High Court of Russia, Constitutional Court of Russia, orders of the President of Russia and the Government of Russia and federal laws) are published by these two sources.

In course of corpus development we selected documents accompanied by commentaries (i.e., a simplified version). The newspaper provides such commentary to what it sees as the most vital of public documents. These commentaries have legal status since they are provided by official publisher. They are aimed to serve as a simpler description for the legal document for people without a domain-specific expertise. In total our corpus has 2963 pairs of original documents and simplified ones.

Figure 1: Types of documents in the dataset
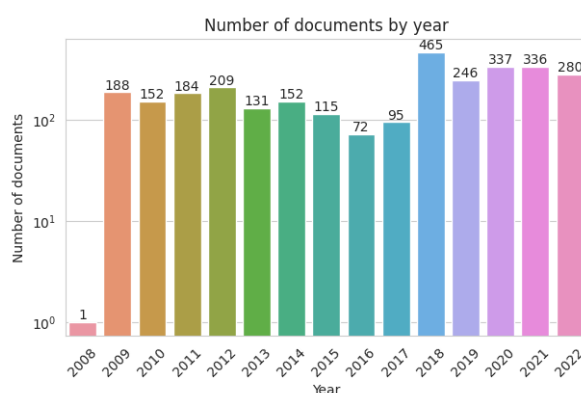
Figure 2: Distribution of the documents by year for the period from 2008 to 2022.

These documents are dated from December 2008 to November 2022. The distribution of the documents over years is shown at Figure 2.

3.2. Dataset filtration

Figure 4 clearly shows the difference in the amount of the legal documents and their simplified versions: the former are much larger than the latter. That proves the idea that the simplified version shouldn't be larger than the original text (with some exceptions), in this respect simplification is close to summarization.

When compiling the corpus, we encountered the problem of uneven distribution of documents by length, see Fig. 5 and Fig 6. E.g., the largest document of 2016 is over 100K tokens in size. In 2010, 2014, 2015 and 2019 there are documents of about 80k in size. These emissions are poorly consistent with the fact that the mean size of legal documents is about 1...2K tokens throughout the whole period. To make the dataset balanced as regards original text size - simplified text size ratio we manually fil-

---

[1] We used the following code `https://github.com/Athugodage/RuLawSimplification/tree/main/dataset%20creation%20code`

[2] `https://www.kaggle.com/datasets/athugodage/russian-legal-text-parallel-corpus`

61
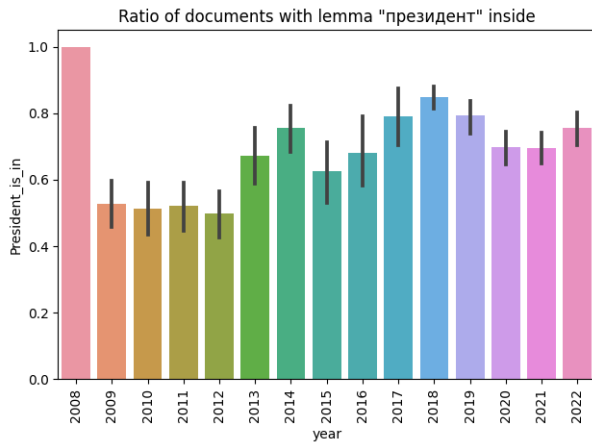
Figure 3: Ratio of documents by year where Russian lemma "president" appears
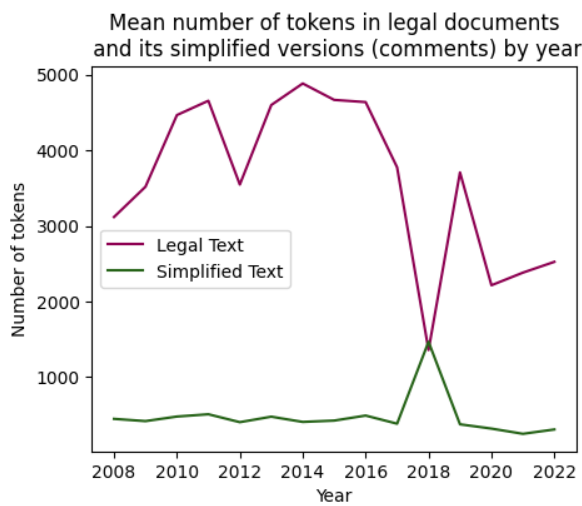
Figure 4: Mean number of tokens in legal documents set and in the set corresponding to its simplified versions (comments) by year

Figure 5: The distribution of document size (in number of tokens) by year

tered out documents of more than 40k tokens in size. We also deleted pairs of documents with the original text less than 400 tokens in size and the original text is smaller than its simplified version as the given data may lack linguistic features relevant for simplification procedure. We also believed that the original document shouldn't be smaller than its simplified version, so we deleted all pairs that fit this condition. In its final version the corpus prepared for fine tuning includes 2017 document pairs

3.3. Dataset pre-processing

For T5 model series we performed text alignment using the Natural Language Inference (NLI) model, based on RuBERT (Kuratov and Arkhipov, 2019). NLI allows us to see logical similarities between two texts. The standard model implies three-way

Figure 6: The distribution of simplified text size (in number of tokens) by year

inference[3]; it gives three probabilities (with values from 0 to 1): entailment (the fact that sentence B is a consequence of sentence A), neutral, contradiction (sentence B is a negation of sentence A). We used a two-way model (without neutral class) (Lin and Su, 2021). First, NLI checked that text A is a consequence of sentence B, and then vice versa. As a result, there were two entailment values - we summed them. In our case, values from 0.001 to 1.265 were obtained; sentences with mutual values less than 0.005 were determined to be slightly similar and deleted. Apart from the above described pre-processing we have also performed a custom pre-processing for GPT models. The process is described in sections below.

4. Experimental design: model selection and fine-tuning

Current T5 and GPT models for Russian do not fit text simplification task. T5 models for Russian can summarise, translate and paraphrase, but cannot simplify. Most GPT models for Russian (as well as for any other language) are intended for tasks like text generation, question answering and chatting, though some researchers tried to teach GPT simplify in Russian (Shatilov and Rey, 2021). The newest Open AI's ChatGPT-4, Yandex's YaGPT and Sber's Gigachat can simplify a text if a user asks it (however, there are still considerations on the quality of such simplification). This is why we decided to fine-tune our own models. In the following sections you can read about our fine-tuned models: *T5-RLS2000, GPT-simplifier-large-text, and GPT-simplifier25*. They are based on mainstream Russian models from Sber.

4.1. Larger T5 model (*T5-RLS2000*)

This model [4] is based on the Russian-language model T5 from Sber (Zmitrovich et al., 2023) on the entire aligned body of 2 thousand pairs of articles. The fine-tuning was conducted at a rate of 0.00002 on 3 epochs. More information is available in the model's card. This model cannot process multi-sentence texts - one may enter just one sentence in the input.

4.2. Larger GPT model (*GPT-simplifier-large-text*)

This model [5] is based on the Russian GPT3 model from Sber, which in turn is a trained GPT-2 model from OpenAI. The model is fine-tuned on a standard case of 2 thousand pairs of articles. The texts were submitted in full form without compression. The model is trained on 10 epochs with a learning rate of 0.00005. For faster and more efficient operation, gradient accumulation was used every 8 moves.

4.3. Smaller GPT model (*GPT-simplifier25*)

Working with the above mentioned models, we came to the conclusion that the main problem of legal text processing and simplification is the large size of the documents. This problem could be solved if we filter the document. The first sentence in every legal document is introductory *(Examples: "*Именем Российской Федерации*" -> "In the name of Russian Federation"; "*Принят Государственной Думой*" -> "Adopted by State Duma")*. The second phrase corresponds to the pattern like the following: *"*Конституционный Суд Российской Федерации в составе Председателя X, судей А, Б, В, Г, Д, Ж, руководствуясь статьей 100 Конституции Российской Федерации, пунктом 1 статьи 2 Гражданского Кодекса Российской Федерации [...]*" (in English: "Russian Constitutional Court, consisting of the Chairman X, judges A, B, C, D, E F, [made a decision] in accordance of article 100 of the Russian Constitution, paragraph 1 of the article 2 of the Russian Civil Code, [and so on... This listing can be page-long]")*. We skip these two sentences.

Also, with the help of regular expressions, sentences with too long references to other laws were removed. For example, it is common in Russian legal texts to give citation in brackets just in the middle of the sentence like this: *"*"(постановления от 30 октября 2003 года N 15-П, от 27 июня 2012 года N 15-П, от 18 июля 2013 года N 19-П и др.)*"*. We delete it.

This allowed us to examine a clear text without citation and unnecessary phrases. If the document was still too big (e.g. the document had more than 40 sentences), we left just last 35 sentences (removing all others). This action may seem controversial for some researchers, since one can claim that we let significant context be left aside. But that is not true, since the structure of Russian legal document itself is designed so that the most informative part is **always** left in the end of the document. The beginning of any document has somewhat a ritualistic nature. It is almost always filled

---

[3] https://huggingface.co/cointegrated/rubert-base-cased-nli-threeway
[4] https://huggingface.co/marcus2000/T5-RLS2000

[5] https://huggingface.co/marcus2000/GPT_simplifier_large_text

with some phrases like those mentioned above. In contrast, all important decisions, terms and regulations traditionally placed in the end. Thus, cutting text leaving just last 35 sentences did not affect completeness of the content. After that, fine-tuning of the same model of Sber was carried out. The new model[6] was named *GPT-simplifier25*, because it was trained on 25 epochs. Comparing these two GPT models may be scientifically interesting, since it shows whether text reduction is possible (in our case) and, if it is, whether the model which was fine-tuned on dataset with reduced texts has better results than the other one. We did this to check a hypothesis that data economy could positively impact on the result. The following document [7] is a good example of our claim: the document itself starts with some external information, then the first paragraph is the argumentation of the order; the real content starts from the second paragraph.

5. Automatic evaluation

Automatic metrics for simplification include primarily SARI and SAMSA (Grabar and Saggion, 2022). In addition, there is a number of metrics that are often used to evaluate simplification, but in fact they are common for any seq2seq task in NLP. For example, ROUGE is almost always mentioned in similar studies, but this metric was originally designed for summarization (Lin, 2004). There are some other rare metrics which were primarily designed for a specific contest, as with RuSimScore, which was introduced during RuSimpleSentEval (RSSE) in 2021 (Orzhenovskii, 2021). Having analysed different groups of metrics, we focused our attention on ROUGE, BERTScore and SARI. To evaluate each of the pre-trained models, we proposed our own approach. GPT models were evaluated on a set of 2500-characters-long excerpts (because these models cannot have a limited context) from original documents (from the test set). T5 model were evaluated on the test set of the aligned sentences: the algorithm checked to what extent the model simplifies each sentence. We see that on ROUGE metrics our models show bad results, comparing to summarization models. The best our model in this case is GPT-large. On another equally interesting BERTScore metric, the best results are obtained for our T5 model. Still, the most prominent metric for us remains SARI, since only this (of proposed ones) shows the real efficiency of the text simplification model. The best result (54.96) on SARI belongs to *T5-RLS2000*. This is an excellent result; for comparison, in 2021 the state-of-the-art

result in text simplification was 44.3(Omelianchuk et al., 2021). In some other works the SARI score is around 35 (Sun et al., 2021) . Further evaluation of the simplification abilities of the models was performed using readability indices. We examined a set of resulting simplified texts from our fine-tuned GPT models and selected Gunning Fog Index and Flesch-Kincaid Readability Index to evaluate them. We made our own script to evaluate these indices because standard versions of them that are available in open-source Python libraries, are more suitable for English. Our version of these formulas allow take into account the specificity of Russian text. Table 2 shows the results of checking simplified texts from the test sample on the Gunning Fog Index. The table shows the average number for 100 documents. The Gunning Fog Index gives a difficulty score for each text individually. The table below shows the complexity index of the original and the text simplified by a specific model.

The same table shows the results of checking the Flesch-Kinkaid readability index in the values of the training classes, i.e. how much you need to study (on average) to understand this or that text. As can be seen from the two tables with estimates of the readability indices, the small GT3 model copes with simplification much better than the large one (Blinova and Tarasov, 2022). The T5 models did not participate in the evaluations on the readability index, because these are simplification models based on proposals. However, we offer table 3 to show the readability estimates for other T5 models for summarization and paraphrasing. Such a comparison is also interesting because it clearly shows the fundamental difference between the task of simplification and summarization.

6. Human evaluation

For a more qualified evaluation we asked 20 respondents to access simplified texts generated with our models. The respondents were presented with four legal documents:

- Federal law dated 30.12.2021 № 454-FZ "About seed production"[8]

- Resolution of the Chief State Sanitary Doctor of the Russian Federation dated 02.07.2021 No. 17 "On Amendments to the Resolution of the Chief State Sanitary Doctor of the Russian Federation dated 03/18/2020 No. 7 "On ensuring the isolation regime in order to prevent the spread of COVID-2019"[9]

---

[6]https://huggingface.co/marcus2000/ GPT_simplifier25

[7]http://publication.pravo.gov.ru/ Document/View/0001202210170033

[8]http://publication.pravo.gov.ru/ Document/View/0001202112300119

[9]http://publication.pravo.gov.ru/ Document/View/0001202107060020

Table 1: Automatic evaluation using ROUGE, BERTScore, SARI.

| Metric | ROUGE | | | | BERTScore | | SARI |
Model	1	2	3	LSUM	P [a]	F1 [b]	SARI
T5-RLS2000	5	0.6	0.05	5	**0.65**	0.64	**54.96**
GPT s. 25	2.1	0	1.79	1.84	0.61	0.6	40.96
GPT s. large	7.16	1.25	6.7	6.85	0.61	0.6	39.9
rut5 base sum gazeta	9.25	**2.39**	**9.2**	**9.39**	0.6	0.6	35.5
ruT5 large	**10.2**	0.5	**9.2**	9.2	0.6	0.58	34.51
mbart ru sum gazeta	7	1.16	6.59	6.54	-	-	53.9
rut5 base paraphraser	3.3	0.22	2.47	2.44	0.53	0.53	35.62

[a] Mean Precision in BERTScore metric
[b] Mean F1 score in BERTScore metric

Table 2: Gunning Fog Readability and Flesch-Kincaid Grade Level Readability Indices evaluations on our models

| | Gunning Fox Index | | FKGL | |
Model	Original	Simplified	Original	Simplified
GPT simplifier 25	**59.5**	**41.8**	**26.58**	**18.23**
GPT simplifier large	59.5	54	26.58	25.48

- Federal law dated 03.04.2023 N 108-FZ "About making changes to Federal Law "On State Regulation of Production and Turnover of Ethyl Alcohol, Alcoholic and Alcohol-Containing Products and on Restriction Consumption (Drinking) of Alcoholic Products"[10]

- Federal law dated 21.11.2022 № 455-FZ "On amendments to Federal Law "On State Benefits to Citizens with Children"[11]

Each of the documents had five simplified versions (four of them generated with our T5 and GPT models [12], one being the commentary from Rossiyskaya Gazeta newspaper). Respondents were asked to rate each text with a score from 0 to 10. During the evaluation, respondents were recommended to give special priority to the following criteria:

- literacy,
- readability (easy to read, no complicated lexical items)
- conveys the basic principles of the document (the more specified, the better),
- authenticity of facts.

The assessment was conducted in the form of a survey in Google Forms. The results were evaluated in two groups of respondents – in a group of experts with a degree in law (or, at least, a law student), and in a group of experts without a legal education. Eventually, 20 people took part in the survey. Among them five respondents confirmed their qualification in legal sciences, 15 respondents

Figure 7: Average score of the human evaluation for each of the models. Orange columns (the left-side graph) represent the scores among lawyers; green columns - among people without a domain specific expertise

turned out to be non-specialists in law. On average, the assessment of legal experts is 1 point higher than experts without a degree in law. Judging from Fig. 7, it can be concluded that lawyers consider the model *GPT-simplifier25*, trained on abbreviated texts, as the most accurate. The rest consider the simplification model according to the proposals of *T5-RLS2000* to be the best. It should be noted that in both cases, the proposed models got higher ratings than the Rossiyskaya Gazeta commentary.

In general, the fact that the proposed neural network models coped with a lot of documents better than comments written by a living person can be

---

[10] http://publication.pravo.gov.ru/Document/View/0001202304030011
[11] http://publication.pravo.gov.ru/Document/View/0001202211210043
[12] Examples are given in Appendix A.

Table 3: Gunning Fog Readability and Flesch-Kincaid Grade Level Readability Indices evaluations on other models (for comparison)

Model	Gunning Fox Index		FKGL	
	Original	Simplified	Original	Simplified
rut5-base-sum-gazeta (summarization)	53.5	62.7	26.58	29.04
ruT5-large (summarization)	53.5	36.56	26.58	10.86
rut5-base-paraphraser (paraphrasing)	53.5	137.9	26.58	126

considered a success of experiments on fine-tuning simplification models for Russian legal texts.

7. Conclusion

In this article we discussed the problem of automatic simplification of Russian legal texts. The work presents three new fine-tuned neural network models: T5-based and GPT-based. In order to fine-tune the models we developed a new parallel corpus based on Russian legal documents and commentaries. This corpus contains a pair of an original legal text and its description, provided by Rossiyskaya Gazeta (a newspaper published by the Government of Russia). The discussed language models have significant differences since the size of models' datasets varied a lot. The models were evaluated with ROUGE, SARI and BERTScore. The generated texts were analysed as regards readability indexes Flesch-Kinkaid Grade Level and Gunning Fog Index. We asked 20 respondents to participate in human evaluation of the fine-tuned models.

The proposed solutions take a big step in expanding the availability and readability of legal documents for wide audience. With the help of the proposed models, it is possible to simplify professional legal texts so that they can be understood by almost everyone. However, at this stage, simplified texts may have some shortcomings, thus, verification of the simplified texts by experts or editors may be required. Our next challenge is to improve existing simplification technology so that the user could read generated texts immediately after the procedure. The future work deals with fine-tuning Longformer Encoder-Decoder and LongT5 for simplification task and with reduction of defects in generated texts.

8. Acknowledgments

9. Bibliographical References

References

Ishrat Ahmed, Yu Zhou, Nikhita Sharma, and Jordan Hosier. 2023. Text summarization for call center transcripts. In *Intelligent Systems Conference*, pages 542–551. Springer.

Anthony Baez and Horacio Saggion. 2023. Lsllama: Fine-tuned llama for lexical simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 102–108.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Olga Blinova and Nikita Tarasov. 2022. A hybrid model of complexity estimation: Evidence from russian legal texts. *Frontiers in Artificial Intelligence*, 5:248.

Dheeraj Chahal, Ravi Ojha, Manju Ramesh, and Rekha Singhal. 2020. Migrating large deep learning models to serverless architecture. In *2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 111–116. IEEE.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

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*.

Arina Dmitrieva. 2017. Art of legal writing: quantitive analysis of the resolutions of the constitutional court of russian federation. *Comparative Constitutional Review (Saint-Petersburg, Russia)*, 3:125–133.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. Editnts: An neural programmer-interpreter model for sentence simplification through explicit editing. *arXiv preprint arXiv:1906.08104*.

Alena Fenogenova and SberDevices Sberbank. 2021. Text simplification with autoregressive models. *Proc. Computational Linguistics and Intellectual Tech*, pages 1–8.

Farit Galeev, Marina Leushina, and Vladimir Ivanov. 2021. rubts: Russian sentence simplification using back-translation. *Proc. Computational Linguistics and Intellectual Tech*, pages 1–8.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nanda Kambhatla. 2022. Text simplification for legal domain:{I}nsights and challenges. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 296–304.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *Traitement Automatique des Langues Naturelles*, pages 453–463. ATALA.

Morgan A Gray, Jaromir Savelka, Wesley M Oliver, and Kevin D Ashley. 2024. Empirical legal analysis simplified: reducing complexity through automatic identification and evaluation of legally relevant factors. *Philosophical Transactions of the Royal Society A*, 382(2270):20230155.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Aini Nur Khasanah and Mardhiya Hayaty. 2023. Abstractive-based automatic text summarization on indonesian news using gpt-2. *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, 10(1):9–18.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

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. *arXiv preprint arXiv:1910.13461*.

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

Yi-Chung Lin and Keh-Yih Su. 2021. How fast can bert learn simple natural language inference? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 626–633.

Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. 2020. Felix: Flexible text editing through tagging and insertion. *arXiv preprint arXiv:2003.10687*.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. *arXiv preprint arXiv:1909.01187*.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text simplification by tagging. *arXiv preprint arXiv:2103.05070*.

Mikhail Orzhenovskii. 2021. Rusimscore: unsupervised scoring function for russian sentence simplification quality. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 524–532.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

Sankar Ramaswamy, R Sreelekshmi, and G Veena. 2023. Complexity analysis of legal documents. In *International Conference on Artificial Intelligence on Textile and Apparel*, pages 141–154. Springer.

Or Sharir, Barak Peleg, and Yoav Shoham. 2020. The cost of training nlp models: A concise overview. *arXiv preprint arXiv:2004.08900*.

AA Shatilov and AI Rey. 2021. Sentence simplification with rugpt3. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue*, pages 1–13.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. *arXiv preprint arXiv:2110.05071*.

Renliang Sun, Wei Xu, and Xiaojun Wan. 2023. Teaching the pre-trained model to generate simple texts for text simplification. *arXiv preprint arXiv:2305.12463*.

Hoang Van, David Kauchak, and Gondy Leroy. 2020. Automets: the autocomplete for medical text simplification. *arXiv preprint arXiv:2010.10573*.

Shih-Hung Wu and Hong-Yi Huang. 2023. A prompt engineering approach to scientific text simplification: Cyut at simpletext2023 task3.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. 2023. A family of pretrained transformer language models for russian.

10. Appendix A. Examples

Original text: Link to full text: https://rg.ru/2022/01/11/semenovodstvo-dok.html

Rossiyskaya Gazeta simplified version: Новая редакция закона о семеноводстве поможет в том числе побороть импортозависимость России от зарубежных семян. Закон об этом публикует "Российская газета". Закон о семеноводстве не касается дачников, использующих семена для собственных нужд. Но он важен для российского АПК. Документ, в частности, регламентирует основные процессы по производству, хранению, реализации, транспортировке и использованию семян растений, а также по их импорту и экспорту из России. В новой редакции определены базовые понятия ("семена "сорт "гибрид"и многие другие). Закон запрещает ввозить и использовать для посевов в России семена, содержащие ГМО. Для борьбы с фальсификатом будет создана федеральная государственная информационная система. Документ предполагает создание госреестра селекционных достижений, допущенных к использованию. Кроме того, в новой редакции предусмотрена необходимость оформления удостоверений о качестве семян сельхозрастений. Предыдущая версия закона была принята еще в 1997 году и нуждалась в обновлении.

Новая редакция долго и бурно обсуждалась в отраслевом сообществе. В результате в окончательную версию были внесены принципиальные изменения - исключены необходимость лицензирования и сертификации семян как избыточные меры. Лицензирование заменяется реестром производителей семян, сертификация - удостоверением качества. Закон должен создать условия для развития рынка семян в России и повышения их качества. Власти неоднократно обращали внимание на слабую обеспеченность российского АПК отечественными семенами. Семена некоторых культур полностью завозятся из-за рубежа. Новые нормы закона вступят в силу с 1 сентября 2023 года.

GPT-simplifier-large-text: В России с 2021 года вводится новый вид разрешенного использования семян - "сорт-сорт". В России он появился в конце прошлого года. По новому закону, если в одном регионе выращивается более трех тысяч сортов и гибридов сельскохозяйственных растений, то в каждом регионе будет действовать свой вид разрешенного использования семян.

GPT-simplifier25: В России впервые появился единый госреестр сельхозкультур с правами экспонирования до 2015 г., уточнили "РГ"во время обсуждения документа минсельхозпрода РФ Дмитрием Медведевым после заседания правительственной комиссии при Росреестре 25 ноября прошлого году о развитии отрасли сельского хозяйства... Подробнее об этом можно прочесть здесь. Сегодня же этот порядок продлен еще более чем наполовину, говорит замдиректора департамента агропродовольственных отношений Минсельхоза Михаил Блинкин ("РГ"). По словам эксперта ведомства Алексея Сидорова : это связано со снижениемелностью цен производителей сельскохозяйственно й продукции вследствие снижения ее доли импорта как сырья... Это происходит потому что производители меньше закупают импортной техники у фермеров перед началом сезона хранения урожая либо перепрофилируют свои машины под хранение влаги осенью этого периода без необходимости проведения дополнительных работ даже если урожай уже созрели полностью.. При снижении себестоимости выпускаемой аграрной культуры затраты производителя могут быть уменьшены только благодаря снижению затрат посредников между производителями тракторов зернохранилищ ввиду отсутствия перекрестников обработки зерна через один агрегатный аппарат."Это значит лишь то, -что новые правила распространяются исключительно непосредственно именно сейчас!

А раньше они были введены просто так – чтобы поддержать спрос населения сказал он.

T5 RLS2000: В России вводится режим самоизоляции. В России вводится режим самоизоляции. В России вводится генетическая экспертиза семян. В России вводится новый порядок лицензирования сельскохозяйстве нных предприятий. В частности, в законе о семеноводстве. В частности, в отношении сельскохозяйстве нных культур. В России вводится система агрострахования. В частности, в области семеноводства. В России будет создана единая государственная политика в области семеноводства. В частности, в области семено- водства. В частности, в России будет создано министерство. сельского хозяйства. В России будет создана государственная информационная система в области семеноводства. В частности, в России будет создан банк стандартных образцов семян. В частности, в России будет создано министерство сельского хозяйства. В частности, в отношении семян сельскохозяйстве нных культур. В частности, в области семеноводства. В частности, в ходе проведения лабораторно - биологических исследований. В частности, за счет бюджета регионов. В частности, в области семеноводства. В документе, в частности - о порядке проведения научных исследований. В России вводится система агрострахования. В зависимости от этапа производства семян сельскохозяйстве нных культур определяются категории агрокультур

Accessible Communication: a systematic review and comparative analysis of official English Easy-to-Understand (E2U) language guidelines

Andreea Deleanu, Constantin Orasan, Sabine Braun
University of Surrey (Guildford, United Kingdom)
m.deleanu@surrey.ac.uk, c.orasan@surrey.ac.uk, s.braun@surrey.ac.uk

Abstract

Easy-to-Understand (E2U) language varieties have been recognized by the United Nation's Convention on the Rights of Persons with Disabilities (2006) as a means to guarantee the fundamental right to Accessible Communication. Increased awareness has driven changes in European (European Commission, 2015, 2021; European Parliament, 2016) and International legislation (ODI, 2010), prompting public-sector and other institutions to offer domain-specific content into E2U language to prevent communicative exclusion of those facing cognitive barriers (COGA, 2017; Maaß, 2020; Perego, 2020). However, guidance on what it is that makes language actually 'easier to understand' is still fragmented and vague. For this reason, we carried out a systematic review of official guidelines for English Plain Language and Easy Language to identify the most effective lexical, syntactic and adaptation strategies that can reduce complexity in verbal discourse according to official bodies. This article will present the methods and preliminary results of the guidelines analysis.

Keywords: Accessibility, Easy-to-understand language variety, Accessible Communication, systematic review

1. Introduction

Accessibility as we conceive it today was first mentioned in the Universal Declaration of Human Rights (UDHR, 1948). The definition has since been extended to take people's individual (dis)abilities into account, with the European Standard EN 17161 (2019) defining accessibility as the "extent to which products, systems, services, environments and facilities can be used by people from a population with the widest range of user needs, characteristics and capabilities to achieve identified goals in identified contexts of use". As context of use also include the interaction between people, Accessible Communication has become a fundamental right in itself (UNCRPD, 2006). Accessible Communication includes "any form of communication that prevents communicative exclusion" (Perego, 2020) so that all users have equal opportunities (UNCRPD, 2006) regardless of their communicative resources, abilities or access to the mode or channel (Maaß, 2020). This entails that when users cannot or cannot completely access information in its original form (Greco, 2016), an alternative should be provided to overcome any potential barrier. Barriers range from sensory to cognitive, from language and culture to expert-knowledge, from motoric to individual skills (Maaß, 2020). As far as the cognitive barrier is concerned, it arises when a person cannot *make sense of* or *cannot fully understand* information because of its complexity. This in turn affects their experience and social and cultural participation. Complexity can be *intrinsic*, meaning that complex information is inaccessible because of the way it has been developed or presented by content creators. Complexity can also be *extrinsic*, however, when an individual's diminished cognitive abilities reduce the ease with which information is received, processed, stored, retrieved, and used (COGA, 2017). In order to address the cognitive barrier, Easy-to-Understand language varieties have been proposed as a means to overcome complexity of verbal written communication for a variety of users (UNCRPD, 2006).

Easy-to-understand (E2U) is an umbrella term encompassing a wide range of "functional language varieties of different national languages with reduced linguistic complexity, which aim to improve comprehensibility" (Hansen-Schirra & Maaß, 2020b) in verbal communication. E2U varieties aim at overcoming cognitive, linguistic (for non-native speakers), cultural and expert-knowledge barriers encountered by a wide pool of users, including migrants, functional illiterates, vulnerable age groups (Maaß, 2020) and people with diverse cognitive abilities[1]. These language varieties thus differ from standard language as they are user-oriented and their main function is to help *understand* and *use* information provided (Hansen-Schirra & Maaß, 2020a). *Plain* and *Easy Language* are two of the most used E2U varieties to facilitate access to information. While the use of E2U promises to overcome cognitive barriers and achieve seamless and accessible communication, several issues arise, undermining its success.

Firstly, the UNCRPD (2006) does not (yet) provide practical guidance on E2U principles nor specifies which conditions end-users have, leaving signatories to develop guidelines and best practices at company,

---

[1] We use 'people with diverse cognitive abilities' and 'cognitively diverse individuals' as umbrella terms to identify individuals with temporarily impaired cognitive abilities (due to fatigue, inattention, a learning difficulty, age and/or injury-related cognitive decline) and individuals with permanent impairments. Temporary and permanent impairments include, but are not limited to, the conditions identified by the American Psychiatric Association as 'mental disorders' (APA, 2013). Cognitively diverse audiences can possess varied degrees of cognitive resources in the areas of attention, executive functions, knowledge, language, literacy, memory, perception, behaviour and/or reasoning (Diamond, 2013; COGA, 2017).

national[2] or transnational[3] level according to their users and target languages. This in turn proves detrimental to the legal implementation of E2U, as lack of consistency weakens its status. Secondly, reception studies with end-users in the field of Accessible Communication are scarce and often rely on individual endeavours. All this results in a lack of an official E2U taxonomy and a growing pool of vague, context-specific or unreliable guidelines being created by academia and the public and private sectors. Needless to say, this means official and non-official guidelines proliferate based on intuition or individual expertise of both professional and amateur adaptors rather than based on evidence – albeit with some exceptions (Fajardo, et al., 2014). Adaptors, in turn, find themselves having to pick and choose from several recommendations, often in contrast with each other. What is worse, contrasting guidelines and inconsistent terminology to identify the variety *and* the user group, have supported the stigma and rejection of *Easy Language* (Hansen-Schirra & Maaß, 2020b), often considered an impoverished version of standard language (Bredel & Maaß, 2019; Maaß, 2020). Thirdly, Accessible Communication has so far mainly promoted the use of E2U in written domain-specific communication. As far as other formats are concerned, the cognitive barrier is yet to be fully addressed in spoken interactions, audiovisual and multimodal settings (Maaß & Hernández Garrido, 2020; Maaß, 2020; Perego, 2020), with a few exceptions[4]. This further excludes people with diverse cognitive abilities from a truly accessible communicative environment and constitutes a significant gap in Accessible Communication research.

This research is conducted within the framework of a project in Media Accessibility, with a focus on overcoming cognitive barriers in audiovisual formats for English-speaking audiences. The final goal of the project was to identify best practice and recommendations applicable to audiovisual content, and more specifically, to the adaptation of film narratives for cognitively diverse audiences. This has resulted in the creation of an audiovisual mode called 'Accessible Cues'. The mode relies on text on screen and an integrated additional narrator to explain and clarify complex elements of the film narrative. However, for these explanations to be effective, they need to be understandable, hence the need to use E2U varieties. To achieve this, we carried out a review and classified existing official English E2U guidelines to identify shared recommendations, discrepancies and grey areas. Such a review of existing guidelines and their subsequent analysis has, to our knowledge, never been attempted before. Although the focus is on English guidelines, we believe our approach to be applicable to other languages as well, albeit integrated by language-specific lexical and syntactic

recommendations. As inconsistency and vagueness abound in the analysed guidelines, it was also deemed essential to investigate current practice, to help identify patterns in E2U that could prove effective in reducing verbal complexity and thus enhance comprehension. The findings from the analysis of the guidelines have informed the analysis of two parallel corpora, namely a corpus of standard vs. adapted news articles by the *Guardian Weekly* (Onestopenglish, 2007) and the standard vs. adapted corpus developed in the in the FIRST project (Orasan, Evans and Mitkov, 2018). We conducted the corpus analysis to identify strategies used by professionals to adapt standard language texts into E2U and to identify further significant E2U strategies applicable to audiovisual formats (forthcoming). In this article, we focus on categorizing, analysing and contrasting E2U guidelines to identify adaptation patterns. This has been pursued by analysing 10 official *Plain* and *Easy Language* guidelines which provide guidance on how to create from scratch and/or adapt a standard language text into E2U.

Our contributions can be summarised as follows:

(1) we conduct a comprehensive alternative classification of 10 official E2U guidelines for the adaptation of English texts and provide an alternative methodology to classify E2U guidelines.

(2) we additionally conduct a qualitative analysis to identify strategies covered by existing guidelines, including shared, discrepant and incomplete (or "grey areas") recommendations.

Relevant background information will be reviewed in Section 2 by providing a brief overview of the verbal and non-verbal strategies used in *Plain* and *Easy Language*. This will be followed by Section 3 on the guidelines analysis which will focus on presenting the guidelines and methodology used. Section 4 will cover a discussion on the guidelines analysis results. Section 5 will provide conclusions and an overview on future work. Section 6 will conclude with a brief discussion on limitations.

2. Background information

2.1 Plain and Easy Language

Several E2U language varieties have been developed throughout the years to address text complexity. Among these, *Plain Language* (PL) and *Easy Language* (EL) are the most widely used and known varieties. PL is primarily used to facilitate expert-lay communication by empowering lay-users to make informed decisions about health, legal actions, rights and finances (Matveeva, et al., 2018; Hansen-Schirra & Maaß, 2020b). Its primary users include lay-recipients and functional illiterates who struggle with the expert-knowledge barrier posed by public

---

[2] See UNE 153101:2018 EX, Accessibility Standard on *Easy Language* (here called *easy to read*).
[3] See Lindholm & Vanhatalo, 2021 for a discussion on the application of E2U language varieties across the EU.

[4] See the EU project SELSI (*Spoken Easy Language for Social Inclusion)* on spoken *Easy Language*. See the EU project EASIT *(Easy Access for Social Inclusion Training*) on training materials for the adaptation of existing audiovisual access services.

administration, legal or governmental documents and rhetoric (Perego, 2020). There are also secondary users who have benefitted from PL, such as vulnerable age groups (IFLA, 2010; García Muñoz, 2012; Matveeva, et al., 2018; Bernabé Caro, 2020, Perego, 2020; PLAIN, 2011a); migrants (McGee, 2010; PLAIN, 2011a), people with reading difficulties (Maaß & Hernández Garrido, 2020) and people with disabilities who do not have access to EL texts (Maaß, 2020). While PL has been dominating the scene for the past 50 years (Mazur, 2000), EL has just started gaining momentum, driven by increased awareness of the importance of Accessible Communication (ODI, 2010; European Commission, 2015, 2021; European Parliament, 2016). EL is also known as *Easy-to-Read* (E2R; EtR), *Easy Reading* (ER) or *Easy English* (EE) (Maaß, 2020; Perego, 2020; Scope Australia, 2015; García Muñoz, 2012), further creating conceptual chaos, as previously discussed in the introduction. Although initially designed to meet the needs of people with learning difficulties (Hansen-Schirra, et al., 2020) with a focus on legibility[5] (IFLA, 2010), EL has become a means of inclusion for a wide pool of cognitively diverse users[6]. Primary users of EL also include sign-language users (Maaß, 2020), pre-lingually deaf (IFLA, 2010; Maaß, 2020) and deaf-blind people (IFLA, 2020; Rink, 2019). Secondary users belong to different age groups and rely on EL in expert-lay communication contexts, as it is the case for non-experts (Maaß & Hernández Garrido, 2020); non-native language speakers (Maaß, 2020; Saggion, et al., 2011); people with limited education and functional illiterates (IFLA, 2010; Maaß, 2020), especially when no PL version is available. Both language varieties rely on verbal strategies to make language more *accessible* and on non-verbal strategies to make meaning *easier* to *retrieve* and *perceive* (Perego, 2020).

2.2 Verbal and non-verbal E2U strategies

The adaptation or creation from scratch of E2U material is achieved through verbal and non-verbal strategies. These are applied according to the expected knowledge of target users, their literacy level, communication needs, the text type and text function (Bernabé Caro, 2020; Perego, 2020). Comprehension is improved at verbal level by manipulating language. Non-verbal strategies manipulate the overall text instead, by relying on visual aids (e.g., images, pictures, pictograms, ideograms, symbols and icons) to help users visualize and co-reference information (Tuset et al. 2011), and on textual and layout techniques (e.g., tables, headings, bullet points and lists) to provide more organized, and therefore linkable and clear information. Strategies used to manipulate information rely on two adaptation strategies, namely *simplification* and *easification*. This article will only discuss non-verbal strategies that directly affect language rather than strategies concerning legibility, page design and visual aids.

Simplification can be defined as "the process of transforming a text into an equivalent which is more understandable" (Saggion, et al., 2011). It does so by reducing linguistic complexity (WCAG 2.1, 2019) and it consists in the adaptation of the form and content of a text "to produce either a 'simplified version' or a 'simple account' of the original text" (Bhatia, 1983) to facilitate comprehension without distorting meaning. Input is here manipulated by resorting to lexical and syntactic transformations at sentence, paragraph and overall text level.

Easification, on the other hand, makes text more accessible not by adapting its content but by developing in the reader specific learning strategies. (Bhatia, 1983). This includes guiding readers, raising awareness of potential ambiguities and difficulties (van den Bos, et al., 2007), introducing the topic by giving an overview of it, highlighting causal links and relations, supporting an argument with evidence, examples and references through visual aids (e.g., boxes, images, flow charts, diagrams, etc.) and restructuring, reorganising or rearranging information in the text (Bernabé Caro, 2020).

Regardless of their benefits, both simplification and easification have their limitations. In fact, both methods are based on *assumptions* (albeit expertise-based) made by the adapter and elaborations and changes may not fully transfer original meaning, maintain grammatical correctness, nor help readers develop their own coping strategies (Saggion, 2018; Fajardo, et al., 2014). Co-creation and validation with end-users would therefore be preferable. However, this is often not feasable due to economic and time constraints. A possible solution could be identifying patters in E2U adaptation by exploring official recommendations and/or practice. This would then provide a more holistic approach to E2U adaptation.

3. Guidelines analysis

The cognitive barrier is yet to be addressed beyond written verbal communication. As a point in case, guidance on Easy-to-Understand (E2U) practice in multimodal settings, and more specifically, in the audiovisual realm, is scarce and, to date, no solution has been proposed to improve access to film narrative. For this reason, we conducted a guidelines and corpus analysis (forthcoming) to extract recommendations relevant for the development of a mode that can improve access to and enjoyment of film narratives, i.e., 'Accessible Cues'. This was pursued by first exploring and comparing several official *Plain* and *Easy Language* guidelines designed for domain-specific written communication, as no guidance has been provided yet for other formats.

---

[5] Legibility is the interaction between the reader and language-independent elements which both impact comprehension and limit expression. When accounting for legibility, the level of visual and cognitive stress encountered by readers is lowered by making information perceivable, distinguishable and adaptable, thus facilitating readability (Bernabé Caro & Orero, 2019; Bernabé Caro, 2020; Bernabé Caro & Cavallo, 2021).

[6] See footnote 1 for a definition.

These guidelines were catalogued, classified, compared and analysed to extract meaningful recommendations applicable to multimodal communication at content, lexical and syntactic level.

3.1 Resources

Ten guidelines were taken into consideration for this study[7]. They range from government-led initiatives to promote *Plain Language* (PL), to charity-led guidelines for the application of *Easy Language* (EL). These were selected based on a series of criteria, such as the fact that they were freely available online; recent (i.e., published after the 90s) and developed in the United Kingdom, United States and Australia by official bodies. These include governments, national, transnational or European Union user associations and charities. Guidelines focusing on EL have referred to this variety under different labels i.e., *easy words and pictures, easy read, simple words and pictures*, *Aphasia Friendly* and even *plain language*. To overcome this incoherence, we decided to use the umbrella term 'Easy Language' in this analysis to distinguish this language variety from PL. An overview of the guidelines can be found in Table 1.

Guidelines	Variety	Author	Year	Pages
Am I making myself clear? Guidelines for accessible writing	PL	Mencap (UK association)	2000	31p.
Toolkit for Making Written Material Clear and Effective (11 parts)	PL	McGee Consulting (for the US Department of Health and Human Services)	2010	Part 3: 24p. part 4: 96p.
Federal Plain Language Guidelines	PL	Plain Language Action and Information Network (PLAIN, i.e., US)	2011a	118p.
Government Digital Service style guide and guidance on content design	PL	Government Digital Service (GDS, i.e., for UK Government online services)	2022	21p.
Make it Simple	EL	International League of Societies for the Mentally Handi-capped (ILSMH, i.e., for the EU)	1998	21p.
Information for All	EL	Inclusion Europe (for the EU)	2010	40p.
Guidelines for easy-to-read materials	EL	International Federation of Library Associations and Institutions (IFLA, i.e., UK)	2010	31p.
Making written information easier to understand for people with learning disabilities	EL	Office for disability issues (ODI) and advocacy group Value People (for UK government)	2010	40p. Additional resources: 25p.
Clear Written Communications	EL	Scope (Australian charity)	2015	23p.
How to make information accessible	EL	Change (UK charity)	2016	25p.

Table 1: Overview of analysed official guidelines

3.2 Methodology

The guidelines and their additional documentation were manually analysed by the first author based on existing E2U theory (Maaß, 2020; Perego, 2020) and the guidelines' own principles, i.e., their inherent characteristics and their declared premises, intent

and recommendations. Following this review, we created a list of draft categories for each individual set of guidelines. These draft categories were later contrasted to identify macro and micro categories. Four macro categories were identified in order to classify the guidelines, based on their individual characteristics and the recommendations they provided. The ten guidelines were therefore classified and analysed according to the following macro categories: main characteristics, recommendations for practice, guidance on alternative formats and non-verbal aids. An overview of each category is presented in Figure 1.
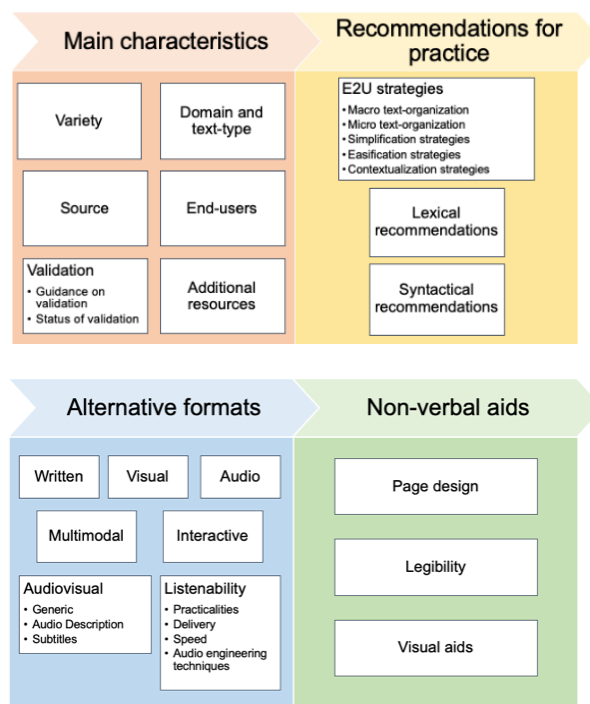
Figure 1: Framework used to categorize and analyse the guidelines

Main characteristics revolve around features such as the E2U language **variety** they discuss, **sources** used by guideline developers (e.g., other publications and guidelines, in-house or personal experience, common sense, empirical research with users with and without disabilities), intended **domain and text type** (e.g., medical, legal, written brochures, contracts, etc.,), expected **end-users**, **validation** (whether and how the guidelines have been checked with end-users) and additional resources provided (e.g., visual aids, samples, glossaries, checklists, lists of terms[8], external links, legal information, research results, etc.,).

Recommendations for **practice** were assigned to different categories to efficiently compare guidelines developed by different entities for heterogeneous end-users and domains. Individual categories were developed by color-coding similar recommendations across guidelines and establishing a hierarchy.

---

[7] The guidelines analysis data set can be accessed at https://bit.ly/m/Accessble-Cues

[8] For example, PLAIN (2011a) provides a list of over 200 words to be avoided. See PLAIN (2011b).

Categories identified are **E2U strategies**, **lexical** and **syntactic recommendations**. E2U strategies encompass information on **macro** and **micro text-organization** (e.g., style, purpose, appropriateness, accuracy, credibility, relevance of information, order, use of bullet points, linking words etc.,), **simplification** (i.e., elimination, reiteration and explanation of information) and **easification strategies** (e.g., introductions, summaries, visual aids, quiz formats etc.,), and lastly, **contextualization** strategies (i.e., at generic, narrative, spatial, temporal, terminological and inferential level). Some recommendations can be simultaneously ascribed to different categories, e.g., simplification *and* syntactic categories. These special recommendations have been first assigned to their generic 'E2U strategies' category, and, if additional information is provided, expanded in the lexical or syntactic categories accordingly.

Recommendations on **non-verbal aids** applicable to written content only will not be discussed in this article due to space restrictions. Yet, these make up a major component of the guidelines and cover suggestions on how to improve perceptibility at **page design** (e.g., paper, colour and printing recommendations) and **legibility**[9] level (i.e., font-size, font-type, layout). The use of **visual aids** is also recommended, to help facilitate visualization and co-reference of information (Tuset et al. 2011). **Alternative formats** will also not be discussed for the same reasons. However, these include recommendations on the use of creative visual and audio formats and the preference for multimodal and interactive interfaces to traditional written material, in opposition to actual practice. Audiovisual formats are also suggested, and recommendations provided, although brief and scarce, for Audio Description and subtitles. Generic recommendations on **listenability**, i.e., the ease with which information is perceived and understood (Perego & Blaž, 2018-2021) are also provided, stressing the need for more attention to audio and audiovisual formats. While all the analysed guidelines emphasise the importance of alternative formats, with audio and video at the fore, none provide explicit information. This could be due to the lack of expertise of guidelines issuers and multimodal versions proving more costly and time consuming, further highlighting the gap between theory, practice and users' best interests.

4. Discussion

We have briefly introduced the framework used to categorize official guideline recommendations for Easy-to-Understand language (E2U) in Section 3. In this section, we will briefly discuss the analysis outcomes of the following categories: main characteristics, E2U strategies, lexical and syntactic

recommendations (see Figure 1). Due to space restrictions, we have removed extensive examples and definitions for each of the discussed categories. However, relevant above-mentioned elements can be found in appendix. The section will conclude with a brief overview of which categories have been successfully and unsuccessfully addressed, in our opinion, to highlight those key areas which could benefit from future research.

4.1 Main characteristics

The first step in the analysis has been to identify the main characteristics of the analysed guidelines[10]. As far as domain and text-type are concerned, guidelines have been designed for healthcare, administration or government-related instructions, factsheets and newsletters, but also for non-traditional E2U communication means such as questionnaires and forms, fictional and non-fictional literature, news and commercial websites. While most guidelines focus on the provision of factual domain specific E2U information, suggestions have also been theorized to be applicable to fictional content as part of an enriching cultural community experience (IFLA, 2010; Scope, 2015) suggesting that there can be more to *Easy Language* than just provision of clear facts.

All guidelines claim to be based on in-house practice and expertise or research into reading behaviour and E2U reception studies. The extent of the validation and the way reception studies have been conducted were however not mentioned in any of the guidelines or the documentation they provided, suggesting that there might be no sound empirical basis.

4.2 E2U strategies

Macro strategies[11] suggested by guidelines revolve around **how** and **what information** should be provided. These range from using a conversational style and everyday spoken language to avoiding slangs, regional dialects and inappropriate language. As far as grammar is concerned, publications suggest abiding by grammatical rules and correct spelling (GDS, 2022; Scope, 2015) while ODI (2010) suggests traditional grammar does not apply and natural spoken language should be favoured instead in both written and oral communication, as the latter tends to occur in more informal and less rule-based environments. This could mean, for example, using Saxon Genitive[12] but not, surprisingly, using contractions for verbs, although this forms part of spoken everyday language.

Most guidelines stress the importance of age and culturally appropriate language, thus suggesting that content producers need to thoroughly know their audience (Mencap, 2000) to address their specific needs (McGee, 2010). This could mean explicitly saying who the material is for, what its purpose is, who

---

[9] See footnote 5 for a definition.
[10] See Table A in appendix.
[11] See Table B in appendix. Ticks represent elements the guidelines approve of, while crosses those which they reject. Blank rows indicate that no information has been provided.

[12] Singular and plural possessives associated with apostrophe to indicate possession. For example, *the boy's toy* to indicate the toy of the boy or *boys' toys* to indicate a range of toys designed for boys.

the people involved are and who to contact in case of need (PLAIN, 2011a).

Micro strategies encompassing text-organisation suggest grouping information on the same topic together, organizing information in a logical sequence and presenting exceptions and conditions after the main idea, unless brief. All of the analysed guidelines suggest that the **inverted pyramid approach**, i.e., organizing information from most important to secondary, is also the best way of facilitating retention of information. Additional recommendations regard the use of headings, content lists and bullet points to organize the structure of the text to increase its usability. Guidelines also suggest using topic sentences to introduce paragraphs or sections to help readers better navigate the document.

An interesting section regards **linking words**, with *Plain Language* guidelines providing a list of preferable words to be used to ensure coherence and to highlight pragmatic relations between paragraphs, sentences and words (McGee, 2010; PLAIN, 2011a). Linking words have been divided into **pointing words**, **echo links** and **connectives** to clearly state whether information is expanded, contrasted or changed[13]. Preferable connectives overlap between both publications, with PLAIN also providing a list of words and connectives to be avoided (PLAIN, 2011b). Although all *Easy Language* guidelines recommend presenting information in a chronological order using a clear logical structure, none mention coherence, cohesion or connectives to be used. This could be due to all *Easy Language* guidelines advising the use of short simple sentences and avoiding complex structures, i.e., connectives between words.

The next step has been identifying and categorizing **easification and simplification** strategies shared by the selected guidelines. An overview of their distribution is presented in Table 2. Ticks are used to identify strategies the guidelines approve of, while crosses identify those which the guidelines reject. Blank rows indicate that no information has been provided.

Source	Eliminate	Reiterate	Exemplify	Explain	Summarize	Introduce
Mencap (2000)	√			√		
McGee (2010)	√	√	√	√		
PLAIN (2011a)	√	√	√	√		
GDS (2022)		x	√	√	√	√
ILSMH (1998)	√	√	√	√	√	x
Inclusion Europe (2010)	√	√	√	√		
IFLA (2010)	√			√		
ODI (2010)	√	√		√	√	
Scope (2015)	√	√	√			√
Change (2016)	√	√	√	√	√	

Table 2: Overview of easification and simplification strategies in the analysed guidelines

Elimination consists in removing confusing and unnecessary content, introductions and comments, redundant words, fillers, prepositions and excess modifiers. **Reiteration** consists in repeating keywords and new concepts, their explanation and using consistent terminology to identify the same concept or important information throughout the text with next to no synonymity[14]. Reiteration is also applied at syntactic level, with a consistent use of structures to introduce semantically similar concepts and introducing sentences on the same topic with the same set of words (ODI, 2010). **Exemplification** is characterized by step-by-step instructions and use of familiar analogies introduced by cues such as *for example, such as, like* and *including* to help readers relate.* **Explanations** rely on the use of **definitions** within the text introduced by *meaning that, that is, that means*, analogies, comparisons, images, illustrated word banks or other easification tools such as boxes. Explanations also rely on **paraphrase** of code-specific terms, easification devices such as **glossaries** at the beginning or end of the document and **context clues**[15] for code-specific language to support or improve reading comprehension. An example of definition and reiteration is provided in Figure A in appendix. As far as easification devices are concerned, these include **summaries**[16], **introductions**[17], visual aids in the form of illustrations, symbols, diagrams, tables and graphs, captions (McGee, 2010), story and fact boxes (ILSMH, 1998; Mencap, 2000; Inclusion Europe, 2010; ODI, 2010; Scope, 2015), quiz and question formats (McGee, 2010) and even workbooks (ODI, 2010); as shown in Figure B in appendix.

All these easification and simplification strategies are to be used to provide context, explain complex relations (IFLA, 2010) or instructions (PLAIN, 2011a), spell out implications (McGee, 2010) and explain new or difficult concepts and terms as they are being used (Change, 2016) or shortly after (ODI, 2010). Overall, guidelines consistently suggest the use of elimination

---

[13] See Table C in appendix.

[14] For example: "You could donate **clothes** you no longer need to a charity shop. The **garments** you donate should be in good condition. The charity shop will not be able to sell **attire** that is badly worn" becomes "You could donate **clothes** you no longer need to a charity shop. The **clothes** you donate should be in good condition. The charity shop will not be able to sell **clothes** that are badly worn" (Change, 2016).

[15] These are definition, synonym, antonym (Gibbs, 2020), syntactic (Robinson, 1975) and semantic clues (Kusumarasdyati, 2001). They help readers understand unfamiliar words (Reed, et al., 2017; Nash & Snowling, 2006), draw inferences and develop expectations (Kusumarasdyati, 2001).

[16] Summaries describe what the content is about.

[17] Introductions are informative guided sections that present the topic, how to navigate the document and tell where resources, references and other versions of the document can be found.

to condense information, with explanations, examples and repetitions as additional strategies text producers can rely on to explicitate or clarify information. Easification devices are also mentioned as essential, as they help condense information and therefore reduce the size of the written document while also supporting comprehension.

Contextualization strategies include presenting the context or field of application and contextualizing information or narrative according to readers' abilities or expected world knowledge by presenting events **spatially** and **temporally**, **clarifying inferences**, using **terminology in context** or **adding context** to help retrieve knowledge or improve literacy. An overview of these strategies can be found in Table D in appendix.

Inferences have been found to pose a major difficulty in communicative exchanges, nevertheless, only some guidelines have confirmed the need to fill in coherence gaps (Bernabé Caro & Orero, 2021). This could be achieved by clearly stating the purpose of the document, assuming lack of background knowledge, or presenting key information only (McGee, 2010; Inclusion Europe, 2010; IFLA, 2010; PLAIN, 2011a). Additional suggestions regard spelling out implications as this helps readers identify personal implications, i.e., if the information provided is applicable to them and how it can be used (McGee, 2010: 56)[18]. However, only one example has been given, which does not help understand the extent to which implications need to be spelled out, suggesting that content creators are in charge of deciding *how much* is *too much* or *not enough* depending on their audience (Mencap, 2000).

The use of **terminology in context** implies the use of specific terms rather than the preference for short hypernyms, as these might only confuse readers about the field of application of the information, with only GDS (2022) stressing the importance of choosing specific words over short high-frequency words that could potentially be polysemic and therefore more ambiguous than low-frequency or technical terms, in contrast with traditional readability indices (for a discussion, see Fajardo, et al., 2014; Crowley, et al., 2008).

Providing **additional context** can help retrieve knowledge as it is the case for glosses (Inclusion Europe, 2010)[19], in line with suggestions by McGee (2010), claiming context needs to be given first,

followed by new information, definitions or explanations. As far as the **contextualization of narrative** is concerned, this mainly revolves around the length and type of information to be provided, with a focus on the functional and informative dimension of the text. This is achieved by avoiding lengthy descriptions that have a more aesthetic purpose, removing details audiences cannot relate to and removing elements that are not relevant for the comprehension of the plot and whose presence can prove confusing, overloading or misleading. For example, this could mean reducing setting descriptions, irrelevant characters or digressions but also contextualizing relevant elements based on the expected world knowledge and frames of reference possessed by audiences, to help them relate to an event[20] or story (IFLA, 2010)[21]. On the other hand, this does not mean that the language to be used in the adapted narrative should not be creative (Change, 2016) or that original E2U fiction should not be engaging and entertaining (IFLA, 2010). This once more highlights the creative freedom given to adaptors and, consequently, one of the reasons behind inconsistency in daily practice.

4.3 Lexical recommendations

Lexical recommendations are largely consistent across guidelines[22]. These include the suggestion to use clear familiar words and spoken everyday language characterized by high-frequency choices. Examples of high-frequency choices are 'not needed' for 'superfluous', 'tiring' for 'strenuous' and 'shared' for 'collaborative (Change, 2016). Yet, the extent to which high-frequency words are easier to understand has been criticised by GDS (2022) as high-frequency words tend to be polysemic and therefore the drawing of inferences can prove difficult due to the impossibility of disambiguating meaning. Additional suggestions are using conversational pronouns (*you, your, we, our*) to address the readers and clearly stating who "you" and "we" refer to. Other suggestions are the avoidance of abbreviations, acronyms, foreign words – unless in use or explained – and a ban on slang and regional words. An example of domestication can be found in the adapted text in Table E in appendix, where the French *Monsieur* is replaced by the familiar yet abbreviated '*Mr.*'. Recommendations also range from a ban on special characters to hyphens and large numbers in favour of digits, analogies, or euphemisms (*few, many, long time ago*). All guidelines stress the need for short words and sentences and some even provide some

---

[18] See Figure C in appendix.

[19] Only the following example has been provided: "Peter Smith spoke at the meeting" becomes "Peter Smith **is the president of a self-advocacy group**. Peter Smith spoke at the meeting". Peter's name has been associated with his profession, i.e., the gloss (Inclusion Europe, 2010).

[20] For example: "Your general practitioner might refer you to the hospital to have an **x-ray of your chest taken**" becomes "Your doctor might ask you to go to the hospital. At the hospital someone will take an x-ray of your chest. **An x-ray is like a photograph.** It allows the doctor to see

inside your body" Change (2016). In this case, readers are encouraged to relate medical procedures to their daily lives.

[21] See the adapted version of *The Count of Monte Cristo* (Dumas, 1997) by IFLA (2010) in Table E in appendix. In the adapted version, setting descriptions have been kept to a bare minimum, with a focus on actions and dialogues. Moreover, mentioned characters have been narrowed down to main ones.

[22] See Tables F, G and H in appendix for a sample of lexical recommendations. Ticks represent elements the guidelines approve of, while crosses those which they reject. Blank rows indicate that no information has been provided.

practical guidance in terms of maximum length. Unfortunately, the extent to which these suggestions are empirically valid has not been discussed in any of the above-mentioned guidelines. All guidelines stress the importance of using an active voice while the few recommendations given on adjectives, adverbs and compound nouns have been extracted from the examples and samples provided by guidelines themselves, rather than from prescriptive instructions. Based on IFLA's (2010) literary adaptation in table E in appendix, it can be hypothesized that adverbs of manner should be avoided while adjectives should be explicitated, removed or replaced with higher-frequency alternatives when of low-frequency. The example provided is "He was a young man of **between eighteen and twenty**, tall, slim, with **fine** dark eyes and **ebony-black** hair. His whole **demeanour** possessed the **calm** and **resolve** peculiar to men who have been accustomed **from childhood** to **wrestle with danger**" becoming "He was **at most** twenty years old. He was tall and slim, he had **beautiful** dark eyes and his hair was **black**. He **looked strong and steady**". In this example, the age number has not been transformed into digits, contrarily to most guidelines recommendations. Moreover, as shown by the words in bold, inferences to be drawn from the description of his personality have been explicitated, compound adjectives have been replaced by one-word synonyms and more familiar terms have been used.

A small number of ambiguous and inconsistent recommendations have been found, due to vague language being used to describe rules. As far as ambiguity is concerned, all guidelines insist on the use of concrete words against abstract words or abstractions. What this entails is however not specified as it seems to mean that abstract concepts such as love, ethics, justice etc., should not be mentioned in the guidelines themselves. This is however not the case, as Change (2016) suggests that texts about ideas, concepts and abstract themes (e.g., national identity, spirituality etc.,) can be translated through a more imaginative and creative use of pictures, thus relying on the visual channel to support meaning-making.

Vagueness regards the motto "avoid difficult words". All guidelines mentioned have yet to explain or quantify what makes a word *difficult*. Suggestions to answer this question range from circumlocutions, technical words and jargon, words ending in *–ion, –tion*, *–sion*, *–ance* and *–ment* (GDS, 2022) and nominalized verbs to be replaced with more familiar words or explanations, context-cues or even glossaries. Additional difficulties are posted by noun strings[23] and descriptive words that need to be replaced with prepositions and articles that clarify the

relation between words[24]. The extent to which these suggestions have undergone a reception study with end-users is however unclear.

While all guidelines concur on the ban on metaphoric and figurative language, two guidelines suggest that figures of speech and metaphors could be used *if* familiar and that symbolic language could be preserved in creative texts (ILSMH, 1998; IFLA, 2010). ODI (2010) also indicates that humour and jokes can be acceptable in its updated *Accessible Communication Formats* (Disability Unit & Cabinet Office, 2021) suggesting that a more informal approach might suit target audiences better, once more indicating that no consensus on user preferences has been found.

Traditional readability studies have suggested that a higher number of references, among which pronouns can be found, improves cohesion and thus supports text comprehension (Kintsch & Van Dijk, 1978; McNamara, et al., 2010). On the other hand, research has also found that ambiguous or inconsistent pronouns affect comprehension (Tavares, et al., 2015), that the number of referents negatively impacts on literal comprehension (Fajardo, et al., 2014), that low-skilled readers struggle with drawing inferences about pronominal antecedents (Oakhill & Yuill, 1986) and that the redundancy of references in simplified texts make the grammar more complex and unnatural (Meisel, 1980). Nevertheless, the use of pronouns is scarcely mentioned in the guidelines, suggesting that no consensus has been found in this case either. While some publications insist on the use of proper nouns (McGee, 2010; Scope, 2015), others suggest the use of pronouns *only* when they clearly refer to specific objects or people (Inclusion Europe, 2010; PLAIN, 2011a). Additionally, while some insist on the use of consistent, repetitive and reduced semantic nuance of words and phrases (Mencap, 2000; ODI, 2010; PLAIN, 2011a; Scope, 2015), others suggest in their examples, that when referencing a concept, personal pronouns, proper names or circumlocutions can all be used (Change, 2016). No consensus has been reached regarding the use of contractions, negations, modal verbs or tenses to be avoided, with Inclusion Europe (2010) using past tense and negations to write the guidelines and provide examples, while, at the same time, rejecting both in its recommendations, as shown in Table I in appendix.

4.4 Syntactic recommendations

Syntactic recommendations are also largely consistent across guidelines[25]. Recommendations range from presenting one idea per sentence to a ban on word splitting. They also include practical recommendations on sentence length and word

---

[23] These occur when three or more nouns follow in succession. For example, *Underground mine worker safety protection procedures development* is a noun string, as all nouns preceding 'development' act as its adjectives (PLAIN, 2011a).

[24] For example *National Highway Traffic Safety Administration's automobile seat belt interlock rule*, should

be explicitated into *The National Highway Traffic Safety Administration's interlock rule applies to automotive seat belts* (PLAIN, 2011a).

[25] See Table J in appendix.

order, with a preference for Subject-Verb-Object (SVO) simple sentences[26] and marked order being used to emphasize words. All guidelines recommend avoiding complex sentences, nevertheless examples and guidance prove insufficient, as no definition of 'complex' is given and examples mainly consist in adapted sentences taken out of context, with no step-by-step instructions. Additionally, guidelines insist on banning subordinates, regardless of this potentially disrupting meaning, as relations between sentences cannot be solely expressed by coordination. One inconsistency is provided in IFLA (2010), where original subordinates are replaced by relative clauses and coordinates in the adapted example[27], *de facto* increasing grammatical intricacy and thus text complexity (Halliday, 2008; To, 2017). This suggests that no agreement has been reached regarding the use of dependent clauses, regardless of them being banned in guidelines. Suggestions shared by all guidelines amount to avoiding subordinate clauses in general and exceptions and clauses indicating uncertain future[28] in particular; using simple sentences and resorting to *or*, *but*, *and*, commas and full stop to connect sentences. Nevertheless, McGee (2010) and PLAIN (2011a) have put forward a list of subordinate connectives[29] to support cohesion and coherence, suggesting that simple sentences or coordinates might not be enough to express pragmatic meaning.

4.5 "Grey areas"

As far as the main characteristics are concerned, future guidelines developed by official bodies should provide more explicit reference to how they were compiled, by whom and for what purpose, while also providing more extensive details on how the guidelines were validated or whether any end-users were consulted. This could help harmonize practice across official bodies and adaptors. Nevertheless, guidelines have been successfully explicit in their description of end-users, domain, text-types and additional resources adaptors can access. Macro and micro strategies have also been successfully addressed, with linking words being a major point of contention between guidelines. This inconsistency could be addressed by appraising end-users' comprehension and expectations in a reception study. The same is applicable to their ability to cope with and understand abstract concepts, figurative and metaphoric language. Difficult words should also be further defined to provide practical guidance, i.e., tools, that can help adaptors identify and evaluate them. Other lexical recommendation areas that could benefit from end-users' feedback involve references and pronouns, contractions (Saxon genitive and verb-related), negations, modal verbs and tenses. As far as other E2U strategies are concerned,

recommendations on contextualization have been explicit, although validation with sample populations would be preferable. Simplification and easification strategies have also been successfully addressed, although terminology and text organization of guidelines themselves could be streamlined. The systematic review could also benefit from additional official guidelines being categorized and an analysis of professional E2U practice, as this could shed light on the above-mentioned "grey areas" that have not been successfully addressed by the 10 guidelines we have analysed for this project.

5. Conclusions

The guidelines analysis has shown that different approaches to E2U communication can be taken for different users, depending to the content-creator's experience, purpose and preferences. As a result, no universal set of rules has been or can be identified. Although the analysis highlights inconsistencies and ambiguities of current approaches to E2U, it has also helped identify strategies that are shared across official guidelines. In addition, while the analysed guidelines tend to focus on informative text such as news, public information or domain-specific health or legal information, they mention various formats for achieving E2U, including *stories* to inform and entertain end-users (IFLA, 2010; Inclusion Europe, 2010; McGee, 2010; ODI, 2010; Scope, 2015. Audiovisual media content such as films and TV programs, has been identified as a further crucial area for Accessible Communication to thrive, beyond the realm of domain-specific interactions (IFLA, 2010; ODI, 2010; Inclusion Europe, 2010). As this research is conducted in the context of a project in Media Accessibility, we intend to address the gap in Accessible Communication by applying the best identified E2U strategies to an audiovisual format. However, identifying these strategies requires addressing grey areas left unresolved by our guidelines analysis (such as the preference for high-frequency but ambiguous and polysemic words over context-specific technical terms) and determining how to deal with conflicting guideline recommendations (such as the ban on abstract concepts). To achieve this goal, we conducted a corpus analysis to identify expected and unexpected language-dependent phenomena that characterize professionally adapted E2U texts (forthcoming). The analysis and subsequent comparison with the guidelines results will help us determine which adaptation strategies we should pursue in order to reduce the verbal complexity of the 'Accessible Cues' that we intend to develop to address cognitive barriers posed by film narratives.

---

[26] For example: "After attending the function, everyone will reconvene at the hotel" becomes "You will meet the group. You will have dinner. You will go back to the hotel" (Scope, 2015). The example also highlights the use of syntactic structure reiteration strategies (simplification strategy).

[27] For example: "Beside the pilot, **who** was to guide the ship into the harbour**,** stood a young sailor, **leaning** against the railing" and "The young man stood and watched a small rowing boat **which** was hurrying towards the Pharaon".

[28] Constructed with *might happen* or *should do* (ILSMH, 1998; PLAIN, 2011a).

[29] See Table C in appendix.

6.   Limitations

We acknowledge that our framework, developed through a qualitative guidelines analysis is, to some extent, subjective and tailored to a project in Media Accessibility. The analysis was conducted using a limited sample of guidelines, as our focus was on guidelines issued by official bodies. Moreover, the selected guidelines originate from English-speaking countries, although their distribution is not uniform, as 5 guidelines were developed by British bodies, 2 by American officials, 1 by an Australian charity, and 2 by the European Union. This variation could affect the lexical and syntactic recommendations provided, considering the differences in English language usage. In our corpus analysis and 'Accessible Cues' all recommendations will be normalised to British English spelling and grammar.

7.   Bibliography

APA, American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders*. 5th ed. Arlington: APA.

Arfé, B., Mason, L. & Fajardo, I., 2018. Simplifying informational text structure for struggling readers. *Reading and Writing,* Issue 31 (1), pp. 2191-2210.

Štajner, S., 2015. *New data-driven approaches to text simplification - unpublished PhD thesis.* Available at: https://wlv.openrepository.com/bitstream/handle/2436/601113/Stajner_PhD+thesis.pdf?sequence=1

Bernabé Caro, R., 2020. New Taxonomy of Easy-to-Understand Access Services. : *Traducción y Accesibilidad en los medios de comunicación: de la teoría a la práctica.* s.l.:MonTI. Monografías de Traducción e Interpretación., pp. 345-380.

Bernabé Caro, R. & Cavallo, P., 2021. Easy-to-Uderstand Access Services: Easy Subtitles. : M. Antona & C. Stephanidis, éds. *Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments. HCII 2021. Lecture Notes in Computer Science.* Cham: Springer, pp. 241-254.

Bernabé Caro, R. & Orero, P., 2019. Easy to Read as Multimode Accessibility Service. *Hermēneus Revista de traducción e interpretación*, 20 12, Issue 21, pp. 53-74.

Bernabé Caro, R. & Orero, P., 2021. Easier audio description: exploring the potential of Easy-to-Read principles in simplifying AD. *Innovation in audio description research.* New York: Routledge, pp. 55-75.

Bhatia, V. K., 1983. Simplification vs. Easification – The case of Legal Texts. *Applied Linguistics,* Issue 4 (1), pp. 42-54.

British Council, 2022. *Teaching English.* Available at: https://www.teachingenglish.org.uk/article/esol

Change. 2016. *How To Make Information Accessible. A guide to producing easy read documents*. Available at :. https://www.changepeople.org/getmedia/923a6399-c13f-418c-bb29-   051413f7e3a3/How-to-make-info-accessible-guide-2016-Final

COGA, 2017. *Cognitive and Learning Disabilities Accessibility Task Force on Cognitive Accessibility User Research.* Available at: https://w3c.github.io/coga/user-research/

Crossley, S. A., Allen, D. & McNamara, D. S., 2012. Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research,* 16(1), pp. 89-108.

Disability Unit & Cabinet Office, 2021. *Accessible communication formats.* Available at: https://www.gov.uk/government/publications/inclusive-communication/accessible-communication-formats

Dumas, A., 1997. *The Count of Monte Cristo*. Ware: Wordsworth Editions.

European Commission, 2015. *The European Accessibility Act.* Available at: https://ec.europa.eu/social/main.jsp?catId=1202

European Commission, 2021. *Accessibility of ICT products and services.* Available at: https://joinup.ec.europa.eu/collection/rolling-plan-ict-standardisation/accessibility-ict-products-and-services

European Parliament, 2016. *EU Directive 2016/2102.* Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016L2102&from=EN

Fajardo, I. et al., 2014. Easy-to-read Texts for Students with Intellectual Disability: Linguistic Factors Affecting Comprehension. *Journal of Applied Research in Intellectual Disabilities,* Volume 27, pp. 212-225.

Government Digital Service (GDS). 2022. *Style guide and guidance on content design*. Available at : https://www.gov.uk/guidance/content-design/writing-for-gov-uk

García Muñoz, Oscar,. 2012. *Lectura fácil: Métodos de redacción y evaluación*. Madrid: Plena Inclusión Madrid.

Gibbs, A., 2020. *Supporting Your Children's and Teens' Home Learning: Using Context Clues to Understand New Words, Iowa Reading Research Center.* Available at: https://iowareadingresearch.org/blog/supporting-home-learning-context-clues

Greco, G. M., 2016. On Accessibility as a Human Right, with an Application to Media Accessibility. : P. O. Anna Matamala, éd. *Researching Audio Description. New Approaches..* s.l.:Palgrave, pp. 11-33.

Halliday, M. A. K., 2008. *Complementarities in language.* Beijing: The Commercial Press.

Hansen-Schirra, S. & Maaß, C., 2020a. *Easy Language Research: Text and User Perspectives.* Berlin: Frank&Timme.

Hansen-Schirra, S. & Maaß, C., 2020b. Easy Language, Plain Language, Easy Language Plus: Perspectives on Comprehensibility and Stigmatisation.. : H. Silvia & C. Maaß, éds. *Easy Language Research: Text and User Perspectives.* Berlin: Frank&Timme.

Hawthorne, K. & Loveall, S. J., 2020. Interpretation of ambiguous pronouns in adults with inellectual disabilities. *Journal of Intellectual Disability Research,* 65(2), pp. 125-132.

IFLA, 2010. *Guidelines for Easy-to-Read Materials.* Available at: https://www.ifla.org/wp-content/uploads/2019/05/assets/hq/publications/professional-report/120.pdf

Inclusion Europe. 2010. *Information for all European Standards for making information easy to read and undersand.* Available at : https://www.inclusion-europe.eu/wp-content/uploads/2017/06/EN_Information_for_all.pdf.

ILSMH European Association. (1998). *Make it Simple: European Guidelines for the Production of Easy-to- Read Information for People with Learning Disability.* Available at: https://docplayer.net/142050357-Ilsmh-european-association-make-it-simple-european-guidelines-for- the-production-of-easy-to-read-information-for-people-with-learning-disability.html

I.S. EN17161. 2019. *CEN/CLC/JTC 12– 'Design for All'.* Available at: https://universaldesign.ie/products-services/i-s-en-17161-2019-design-for-all-accessibility-following-a- design-for-all-approach-in-products-goods-and-services-extending-the-range-of-users/

Jordanova, V., Evans, R. & Cerga Pashoja, A., 2014. *D7.2: Benchmark report (results of piloting task),* Bruxelles: European Commission.

Kintsch, W. & Van Dijk, T. A., 1978. Toward a model of text comprehension and production.. *Psychological Review,* Volume 85, pp. 363-394.

Kusumarasdyati, A. A., 2001. Semantic and Syntactic Clues as Vocabulary Strategies in Reading Comprehension. *FSU in the Limelight,* 8(1).

Lindholm, C., & Vanhatalo, U., 2021. *Handbook of Easy Languages in Europe*. Berlkn: Frank&Timme

Maaß, C., 2020. *Easy Language – Plain Language – Easy Language Plus. Balancing Comprehensibility and Acceptability.* Berlin: Frank&Timme.

Maaß, C. & Hernández Garrido, S., 2020. Easy and Plain Language in Audiovisual Translation. : C. Maaß & S. Hansen-Schirra, éds. *Easy Language Research: Text and User Perspectives.* Berlin: Frank&Timme, pp. 131-161.

Mazur, B., 2000. Revisiting Plain Language. *Technical Communication,* 47(2), pp. 205-211.

McGee, J., 2010. *Toolkit For Making Written Material Clear And Effective.* U.S. Department of Health and Human Services Centers for Medicare and Medicaid Services. Available at : https://emilylinginfelter.com/writing-tips/toolkit-for-making-written-material-clear-and-effective/

McNamara, D. S., Louwerse, M. M., McCarthy, P. M. & Graesser, A. C., 2010. Coh-Metrix: capturing linguistic features of cohesion. *Discourse Processes,* Volume 47, p. 292– 330.

Meisel, J., 1980. Linguistic simplification. : S. Felix, éd. *Second Language: Trends and Issues.* Tubingen: Gunter Narr, p. 13– 140.

Mencap. 2000. *Am I making myself clear? Mencap's guidelines for accessible writing.* Available at : http://funding4sport.co.uk/downloads/guidelines_for_accessible_writing.pdf

Morton Gernsbacher, A. & Pripas-Kapit, R. S., 2012. Who's Missing the Point? A Commentary on Claims that Autistic Persons Have a Specific Deficit in Figurative Language Comprehension. *Metaphor and Symbol,* 27(1), pp. 93-105.

Nash, H. & Snowling, M., 2006. Teaching new words to children with poor existing vocabulary knowledge: A controlled evaluation of the definition and context methods. *International Journal of Language & Communication Disorders,* Volume 41, p. 335–354.

Oakhill, J. & Yuill, N., 1986. Pronoun resoluion in skilled and less skilled comprehenders: Effects of memory load and inferential complexity.. *Langauge and Speech,* 29(1), pp. 25-37.

ODI, 2010. *The Plain Langauge Act.* Available at: https://www.dni.gov/index.php/plain-language-act#:~:text=The%20Plain%20Writing%20Act%20of,and%20collaboration"%20in%20his%20Jan.

Perego, E., 2020. *Accessible Communication: A Cross-country Journey.* Berlin: Frank&Timme.

Perego, E. & Blaž, Z., 2018-2021. *EASIT project, Unit 3B: E2U and Audio Description.* Available at: https://transmediacatalonia.uab.cat/easit/unit-3b/

PLAIN. 2011a. Federal Plain Language Guidelines. Available at: https://www.plainlanguage.gov/guidelines/

PLAIN. 2011b. *Plain Language: Use simple words and phrases*. Available at : https://www.plainlanguage.gov/guidelines/words/use-simple-words-phrases/.

Reed, D. K., Petscher, Y. & Truckenmiller, A. J., 2017. The contribution of general reading ability to science achievement. *Reading Research Quarterly,* Volume 52, pp. 253-266.

Robinson, H. A., 1975. *Teaching Reading and Study Strategies: The Content Areas.* Boston: Allyn and Bacon.

Saggion, H., 2018. Text Simplification. : R. Mitkov, éd. *The Oxford Handbook of Computational Linguistics 2nd edition.* Oxford: Oxford University Press.

Saggion, H. et al., 2011. Text Simplification in Simplext: Making Texts more Accessible. *Procesamiento del Lenguaje Natural*, 09, Issue 47, pp. 341-342.

Scope Australia. 2015. *Clear Written Communications. The Easy English Style Guide.* Available at : . https://www.rch.org.au/uploadedFiles/Main/Content/ethics/Clear-Written-Communications.-The-Easy-English-Style-Guide.pdf

Tavares, G. et al., 2015. Who do you refer to? How young students with mild intellectual disabilities confront anaphoric ambiguities in texts and sentences. *Reserach in Developmental Disabilities,* Volume 38, pp. 108-124.

UNCRPD, 2006. *United Nations, Convention on the Rights of Persons with Disabilities.* Available at: https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html

WCAG 2.1, 2019. *The Web Content Accessibility Guidelines.* Available at: https://www.w3.org/WAI/WCAG21/quickref/

Yaneva, V., 2016. *Assessing text and web accessibility for people with autism spectrum disorder – unpublished Thesis.* Available at: https://wlv.openrepository.com/handle/2436/620390

8. Language Resource References

Onestopenglish, 2007. *News lessons. Macmillan English Campus.* Available at: http://www.onestopenglish.com

Orasan, C., Evans, R.J., & Mitkov, R. (2018). Intelligent Text Processing to Help Readers with Autism. *Intelligent Natural Language Processing : Trends and Applications*, Springer.

9. Appendix

TOOLKIT for Making Written Material Clear and Effective
SECTION 2 Detailed guidelines for writing and design
PART 4: Understanding and using the "Toolkit Guidelines for Writing"

CHAPTER 3: Guidelines for writing style   55

Instead of saying, "Get adequate rest," explain what you mean:

This first part gives the basic instruction to the patient

This phrase signals that an explanation will follow

For the next week, you need a lot of rest, | and that means

at least eight hours of sleep each night and a two-hour rest period lying down each afternoon.

The rest of the sentence explains what is meant by "a lot of rest."

Here are additional tips:

- Even after you have explained a new idea, continue to include some context to help readers remember what it means. Remember that readers need time and repetition to absorb new material.

- In addition, if the material is long, repeat the explanations to reinforce readers' understanding. When they read something they feel they have already learned, their confidence grows.

- Finally, make it easy on those who skim by repeating the explanations in each new section.

Figure A: Example of simplification strategies: using definition and reiteration in healthcare materials (McGee, 2010)
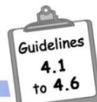
Use clear and simple text (plain English) with short sentences, simple punctuation and no jargon.

Use larger print (at least 12 point), a clear typeface and plenty of spacing.

Use bullet points or story boxes and fact boxes to make the main points clear.

2   Am I making myself clear?

Mencap (2000)

Guidelines 4.1 to 4.6

Engaging, supporting, and motivating your readers

Toolkit Part 4, Chapter 4 shows how to apply these guidelines

4.1   **Be friendly and positive.**
When your messages have a supportive tone, readers will be more receptive, especially if you are urging them to do something difficult or unfamiliar.

4.2   **Use devices that engage and involve your readers, such as stories and quotations, questions and answers, quiz formats, and blank spaces for them to fill in.**
When you get people actively involved with the material, they become more interested and learn more easily.

McGee (2010)

Figure B: Examples of easification strategies extracted from Mencap (2000) and McGee (2010)

Be cautious about using symbols in your explanations

When you introduce a complex concept, take the time to give a careful explanation. If you use symbols or pictures to represent a concept, be sure to explain what they mean in a caption or the text. Also, check on how members of your intended audience are interpreting them. It is surprisingly hard to create clear and effective picture symbols (see Toolkit Part 5, Chapter 6, *Guidelines for photographs, illustrations, and clip art*).

As shown below in Figure 4-3-f, a short cut summary puts too much burden on readers.

Figure   4-3-f.   Be cautious about using symbols to explain concepts.

Alcohol + Pregnancy = Fetal Alcohol Syndrome
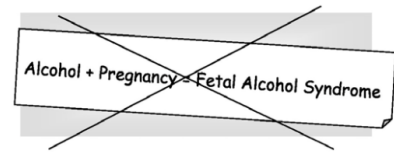
Putting the message in the form of a word equation makes it abstract, impersonal, and hard to understand. It's up to the reader to extract the meaning and figure out the personal implications. In addition, "fetal alcohol syndrome" is very difficult vocabulary that requires explanation.

If you are pregnant, do not drink alcohol. If you drink alcohol when you're pregnant, your baby could be born with a serious health problem called "fetal alcohol syndrome."
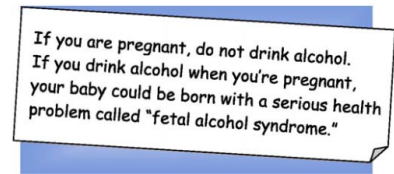
This version builds in the beginning of an explanation of fetal alcohol syndrome ("a serious health problem"). It explains the risk of drinking alcohol while pregnant and tells the reader directly not to do it.

Source: Adapted from *Simply Put* (CDC, 1999:7).

Figure C: Example of how inferences should be spelled out in healthcare materials (McGee, 2010)

Source	Domain & txt type	End-users	Validation status	Guidance on validation
Mencap (2000)	for service providers (local councils, Government departments, charities, hotels and restaurants, shops, leisure facilities, accountants, solicitors, churches, courts, hospitals and clinics)	people with learning disability	Guidelines have been validated	Seek advice from supporters and professionals who are familiar with client's needs. Focus groups with people with learning disability to provide feedback
McGee (2010)	to design healthcare material, provides appendix examples of questionnaires and forms	culturally diverse audiences, less skilled readers, elderly with age-related declines in vision, ability to read and process written info, regardless of literacy level	Guidelines have been validated	Validate with end users: usability testing by piloting material beforehand through interviews, questionnaires or forms. Look for feedbacks and work in teams.
PLAIN (2011)	regulations, law, administration	any audience	Unclear	involvement a priori and through iteration (while work is in progress) and retest after making changes of specific end-users
GDS (2022)	writing on the web (legal, administrative, GOV)	general audience (more than one user group – including specialists) living in the UK. Also mentions people with moderate learning disabilities	Guidance validated through style guides user testing	Check feedback left on GOV.UK or helplines and the proportion of users who found the page useful.
ILSMH (1998)	for beginner content producers (authors, editors, information providers, translators and other interested persons). For government, commerce, voluntary, service and media sectors. Formats: printed, audio tapes, video or interactive media.	those with limited skills in reading, writing and understanding: learning disabilities, disabilities, limited formal education, social problems, immigrants. These guidelines focus on learning disabilities.	Unclear	Consult people with learning disability during production process (from selection of relevant topics to writing the text and final layout of publication). When providing draft, allow enough time for reading, and clarify if they don't understand the contents, highlight confusing words or phrases and possible extra questions and information needed
Inclusion Europe (2009)	written information, websites, video with subs, AD or audio information (news, announcements). Not applicable to poetry or stories only general advice. Especially for lifelong learning programmes.	adults with ID, reading difficulties, L2, blind people with ID	never been tested	Involve people with ID in decision-making processes (about the subject, what to say on the subject, about where to make info available). Only validate target text not source text. Validate end-result with users.
IFLA (2010)	printed/electronic/audio/video editorial content: literature (fiction & non-fiction, original and adaptations); news; magazines; informational content (governmental or commercial, including on the web) For publishers.	people with special needs across different age groups (adults, YA, school-children). 2 groups: 1) people with ADHD, autism, Asperger & Tourette syndrome; ID; learning/reading difficulties (dyslexia & others); prelingually deaf, deafblind, aphasia, dementia 2) recent migrants, non-natives, children (<grade 4, approx. 9 y/o), functional illiterates (education, social issues, mental illness).	never been tested	Test the material before it goes to press with target groups
ODI (2010)	For public sector organisations (NHS & health related) to commission or create easy read materials. Text based but also other formats: video, talks, presentations, drama, murals, role-play or posters, even E2R booklets with work book sections where people answer questions and can send back to get checked.	Aimed at learning disabilities but also useful for BSL, English as L2, black and ethnic minorities	Unclear	Validate with end users to find how to make info accessible and useful. Do not use jargon when "consulta ting". During consultations adapt questions for audience. Read draft aloud. Use focus groups, scenarios and role-plays or questionnaires (if to be filled with handwriting, allow for big space). Involve end users from the start, provide information through different channels and formats, ensure info meets users' needs, signpost to other services, define responsibility for information provision and identify barriers.

Scope (2015)	Card, poster, information sheet or flayer, brochure, booklet, book or series of book, forms, survey, Websites, documents for websites, power point presentations.	Low literacy (difficulty with spoken and written language): learning disability, intellectual or cognitive disability, acquired disability (stroke, brain injury, degenerative condition), low literacy, ageing, culturally or linguistic diverse backgrounds (L2)	Unclear	validate with end users in groups or individually (consumer testing). Direct feedback to determine readability and usability of written material. Assist those that cannot read txt by themselves. Elicit feedback on: general layout and presentation of the information, is the language clear and easy to understand, images used make sense and support language, overall ease of use and readability
Change (2016)	for professionals and organisations that want to make their information accessible to provide clear instructions, facts and statements	learning disabilities, people that struggle with reading and writing (non-readers, low literacy skills, sensory disabilities), people with English L2	Unclear	involve people with learning disabilities *ad priori*, to understand what information they want. Use local advocacy groups, organisations run by disabled people. Face-to-face in steering groups, workshops, small focus groups. Provide background information so they can make informed comments. Get feedback on the final draft of your document. It is important to consider the feedback and make any necessary amendments before distributing.

Table A: Overview of main characteristics of the analysed guidelines

Source	Conversational style	Attention to register and grammar	Declare purpose	Declare target audience	Age appropriate	Culturally appropriate	Accurate information	Credible information	Relevance
Mencap (2000)	√				√				√
McGee (2010)	√		√	√	√	√	√	√	√
PLAIN (2011a)	√		√	√			√		
GDS (2022)	√	√					√		√
ILSMH (1998)	√				√				√
Inclusion Europe (2010)	√		√	√	√				
IFLA (2010)					√				√
ODI (2010)		X			√				√
Scope (2015)	√	√	√		√	√			√
Change (2016)	√	X			√		√	√	√

Table B: Overview of macro strategies suggested, rejected or not mentioned in the analysed guidelines

Linking words		
Pointing words	Echo links	Connectives
That, the, these, this those	Words or phrases that repeat previously mentioned ideas	**Transitions** (also, further, therefore)
		Adding a point (also, and, besides, further, in addition, similarly, what is more)
		Examples (for another thing, for example, for instance, for one thing)
		Restating (again, in other words, in short, put differently, that is)
		Results (accordingly, as a result, so, then, therefore, thus, when)
		Contrasting (but, conversely, however, nevertheless, on the other hand, still)
		Summing up (to conclude, in conclusion, in short, to summarise, to sum up)
		Sequencing ideas (finally, first, secondly, thirdly)

Table C: Linking words to be used according to PLAIN (2011a)

Source	Generic	Narrative-related	Space	Time	Terms	Inferences
Mencap (2000)						
McGee (2010)					After explaining a new idea, continue to include some context to help readers remember meaning. Reiterate terms by providing additional context as you move on. Use context to help understand abstract terms like "excessive bleeding", "regular exercise", "a variety of", by introducing "that means" or "if" and "when" clauses.	Spell out implications and be direct in saying what they should do. If you make readers do the work of identifying and interpreting the personal implications of the material, they may miss or misinterpret an important message.
PLAIN (2011a)	Present information in context without expecting background knowledge.			Present information in a chronological order.		
GDS (2022)			Write the full name of the area the first time you use it. Use a capital for a shortened version of a specific area or region if it's commonly known by that name.	Use "to" in time ranges, not hyphen. Use 12 hours with am and pm: 5:30 PM; 10am to 11am; midnight, midday (not 12, noon, or 12pm); 6 hours 30 minutes.	Use terms in context. The title should provide full context so that users can easily see if they've found what they're looking for. By being general about a topic, you leave the user asking, 'what is this in relation to?'. Give the user context around the topic and what this content will tell them. If the context is right, you read a short word faster than a single letter. By giving full information and using common words, you help people speed up their reading and understand information in the	

					fastest possible way. Content also needs to be in context. Contextualizing terms improves literacy.	
ILSMH (1998)	Don't assume previous knowledge.		Pictures of places to help locate rather than address or name of place	For dates use "a long time ago" and similar.		
Inclusion Europe (2010)	Provide context related to people or places.	Present the background voice before they start to speak.	Explain where new place is if place of filming changes. Explain each place in new scene. It can also be easier to see people going from one place to another rather than seeing someone here and then suddenly elsewhere without knowing why.	Present information in a chronological order.		
IFLA (2010)	Remove any additional details that audiences can't relate to. Provide background explanations of context.	Keep frames of reference into account. Action should be direct and simple without a long introduction and involvement of too many characters. Remove irrelevant characters. Remove plot irrelevant or obvious information Avoid lengthy aesthetic descriptions. Remove digressions. There is no need to use markers to introduce dialogues.	Write the name of the area and give context (Marseille, in the south of France). Remove any additional detail that audiences can't relate to. Keep it to a need-to-know basis.	Present events in a chronological order. Action should follow a single thread with logical continuity. Events take place in logical chronological order. Be specific with time and keep dates mentioned in the original.		Explain complicated relationships in a concrete and logical manner. Place facts in a specific context and provide background explanations to account for readers' frames of reference in terms of different cultural, religious or educational background.
ODI (2010)				Avoid the 24-hour clock. Use am & pm. Pictures using analogue or digital clocks can help explain time.	Provide explanation of technical terms in context.	
Scope (2015)				Be specific with dates, show a 12-hour clock image and a digital clock. Present events in a chronological order.		
Change (2016)	Avoid detailed background information and detailed explanations.				No subtle variations on the same theme.	Avoid multiple points of view, debates, discussions or variation on the same theme.

Table D: Overview of Contextualization strategies suggested by the analysed guidelines

Original version	Easy-to-read version
Marseille – Arrival	**In Marseilles**
On February 24, 1815, the lookout at Notre-Dame de la Garde signalled the arrival of the three-master *Pharaon*, coming from Smyrna, Trieste and Naples. As usual, a coastal pilot immediately left the port, sailed hard by the Château d´If, and boarded the ship between the Cap de Morgiou and the island of Riou.	On 24 February 1815 a French ship came sailing into the port of Marseilles in south of France. The name of the ship was Pharaon.
At once (as was also customary) the terrace of Fort Saint-Jean was thronged with onlookers, because the arrival of a ship is always a great event i Marseille, particularly when the vessel, like the *Pharaon,* has been built, fitted out and laded in the shipyards of the old port and belongs to an owner from the town.	Beside the pilot, who was to guide the ship into the harbour, stood a young sailor, leaning against the railing. He was at most twenty years old. He was tall and slim, he had beautiful dark eyes and his hair was black.
Meanwhile the ship was drawing near, and had successfully negotiated the narrows created by some volcanic upheaval between the islands of Calasareigne and Jarre; it had rounded Pomègue and was proceeding under its three topsails, its outer jib and its spanker, but so slowly and with such melancholy progress that the bystanders, instinctively sensing some misfortune, wondered what accident could have occured on board. Nevertheless, those who were experts in nautical matters acknowledged that, if there had been such an accident, it could not have affected the vessel itself, for its progress gave every indication of a ship under perfect control: the anchor was ready to drop and the bowsprit shrouds loosed. Next to the pilot, who was preparing to guide the *Pharaon* through the narrow entrance to the port of Marseille, stood a young man, alert and sharp-eyed, supervising every movement of ship and repeating each of the pilot´s commands.	He looked strong and steady. His name was Edmond Dantés. The young man stood and watched a small rowing boat which was hurrying towards the Pharaon. A man in the rowing boat waved eagerly to him. "Oh, it´s you, Edmond Dantés" he called. "Why do you look so sad, my young friend?" "We have suffered a great misfortune, Mr. Morrel", answered the young man. "We have lost our captain!"
One of the spectators on the terrace of Fort Saint-Jean had been particularly affected by the vague sense of unease that hovered among them, so much so that he could not wait for the vessel to come to land; he leapt into a small boat and ordered it to be rowed out to the *Pharaon*, coming alongside opposite the cove of La Réserve. When he saw the man approaching, the young sailor left his place beside the pilot and, hat in hand, came and leant on the bulwarks of the ship.	
He was a young man of between eighteen and twenty, tall, slim, with fine dark eyes and ebony-black hair. His whole demeanour possessed the calm and resolve peculiar to men who have been accustomed from childhood to wrestle with danger.	
"Ah, it´s you, Dantès!" the man in the boat cried. "What has happend, and why is there this air of dejection about all on board?"	
"A great misfortune, Monsieur Morrel!" the young man replied. "A great misfortune, especially for me: while off Civita Vecchia, we lost our good Captain Leclère."	

Table E: Standard and adapted version (IFLA, 2010) of an excerpt from *The Count of Monte Cristo* (Dumas, 1997)

Source	Familiar words	Consistency	Slang	Regional words	Short words	Foreign words	Explicitate numbers with words	Abbreviations	Acronyms	Abstract words	Jargon	Technical terms
Mencap (2000)	√	√					√	X		X	X	X
McGee (2010)	√		X	X	√		√	X	X	√	X	X
PLAIN (2011a)	√	√			√	X		X	X	X	X	X
GDS (2022)	√				√		X	√	√		X	√
ILSMH (1998)	√	√				X	√	X	X	X	X	X
Inclusion Europe (2010)	√	√				X	X	X	X			
IFLA (2010)		X			√			X		X		X
ODI (2010)	√	X			√		X	X	X		X	X
Scope (2015)	√	√	X		√		X	X	X		X	√
Change (2016)	√	√			√		X	X	X	√	X	X

Table F: Lexical recommendations – generic and noun-related

Source	Questions	Figures of speech	Personal pronouns as referents	Conversational pronouns	Noun strings	Adverbs	Compound adjectives
Mencap (2000)				√			
McGee (2010)		X	X	√			
PLAIN (2011a)	√	X	√	√	X	√	
GDS (2022)	X	X		√			
ILSMH (1998)		√		√			
Inclusion Europe (2010)		X	√	√			
IFLA (2010)		√	√			X	X
ODI (2010)	X	√					
Scope (2015)		X	X	√			
Change (2016)				√			

Table G: Lexical recommendations – noun-related, referents, adverbs and adjectives

Source	Present	Past	Future	Conditional	Progressive and compound tenses	Passive voice	Contractions	Modal verbs	Negation	Hidden verbs
Mencap (2000)						X				
McGee (2010)						X	√	X		
PLAIN (2011a)	√		X	X		X	√	√	X	X
GDS (2022)						X	√	√	√	
ILSMH (1998)			X			X		X	X	
Inclusion Europe (2010)	√	X				X	X		X	
IFLA (2010)	√	√		X	√		√			
ODI (2010)						X	√			
Scope (2015)						X	X			
Change (2016)							X		√	

Table H: Lexical recommendations – verbs

Inclusion Europe guidelines	Inclusion Europe examples
Use positive sentences rather than negative ones where possible. For example, say "You should stay until the end of the meeting" rather than "You should not leave before the end of the meeting".	Always use the right language for the people your information is for. For example, **do not use** language for children when your information is for adults. **Do not use** difficult ideas such as metaphors. A metaphor is a sentence that **does not** actually **mean** what it says. Make sure it is always clear who or what the pronoun is talking about. If it is **not clear** then use the proper name instead.
Avoid all abbreviations like "e.g." or "etc."	Instead, write My **son's** name is Michael. Yesterday, I bought a new bike for him. The new bike is green and yellow.
Where possible, use the present tense rather than the past tense.	We **did not have** the time to check if the standards to make stories or poetry easy to read and understand would be the same or slightly different. We **have made** these standards as part of a project that **took** place in Europe. People from 8 European countries **met** several times to write these standards. The project which **brought** these people together **was called** "Pathways to adult education for people with intellectual disabilities".
Use active language rather than passive language where possible. For example, say "The doctor will send you a letter" not "you will be sent a letter".	We have made these standards as part of a project that took place in Europe. People from 8 European countries met several times to write these standards. The project which brought these people together **was called** "Pathways to adult education for people with intellectual disabilities".

Table I: Example of incoherence in Inclusion Europe (2010) regarding negations, contractions, past tense and passives

Source	Hyphenation to split words	Sentence length	Order	Topicalization	Periods
Mencap (2000)	No	Short sentences.	92	One idea per sentence	Simple punctuation (no semicolon and colon). Break sentences with commas or *and*. No too many commas to break up a sentence.
McGee (2010)	No	Short sentences. Vary sentence length: 8-15 words per sentence.			Simple sentences or use simple conjunctions (*or, but, and*). Limit number of explanatory and qualifying clauses.
PLAIN (2011a)		Short sentences	Prefer SVO order. SVO followed by modifiers, phrases or clauses.	One idea per sentence.	Avoid wordy and dense constructions. Use lots of full stops. Avoid dependent clauses and exceptions. *If* for conditions, *when* to introduce other clauses after *if*. Complex phrases can be put into tables.
GDS (2022)		Short sentences: max. 25 words. Otherwise, split. For moderate learning disabilities best 5-8 words.	Marked order (front-load sentences) to emphasise words.		Don't use semicolon. Long sentences with semicolon should be broen.
ILSMH (1998)	No	Short sentences. One line per sentence. Otherwise split into separate lines at natural speech break.		One idea per sentence. New ideas should go on new page.	Simple punctuation (no commas, semicolon, hyphens). Break sentences at natural speech break. Avoid complex structures.
Inclusion Europe (2010)	No	Short sentences.		One idea per sentence. Use full stop before starting a new idea. One idea per line. New sentence on a new line.	Simple punctuation (no comma or *and*).
IFLA (2010)		Prefer one line per sentence.		Avoid several actions in a single sentence.	Break sentences at a natural speech break. Avoid subordinate clauses and express them with single sentences, *and,* clauses with commas and relative clauses.
ODI (2010)	No	Sentences as short as possible. Max 15 words per sentence. 10 to 15 preferable.	Can be marked.	Use full stop. One idea per verb.	No difficult punctuation (no colon). Use full stops. Use commas in lists of items. Sentences can end with prepositions or start with *and* or *but*.
Scope (2015)		Short sentences. Use 25-30 characters per line if paired with images. If not, no more than 50-60 characters per line.		One idea per sentence. No split words, complete sentence on the page where it starts.	Simple punctuation. No brackets, hyphens, &, slashes. Prefer simple sentences.
Change (2016)	No	Short sentences.		Key statements or key information per sentence. Identify keywords. One idea per sentence.	Single sentences.

Table J: Syntactic recommendations

LanguageTool as a CAT tool for Easy-to-Read in Spanish

Margot Madina[1], Itziar Gonzalez-Dios[2], Melanie Siegel[1]

[1]Darmstadt University of Applied Sciences (Hochschule Darmstadt)
Hochschule Darmstadt, Max-Planck Str.2, 64807 Dieburg
[2]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
Manuel Lardizabal, 1, 20018 Donostia-San Sebastian, Gipuzkoa
margot.madina-gonzalez@h-da.de, itziar.gonzalezd@ehu.eus, melanie.siegel@h-da.de

Abstract

Easy-to-Read (E2R) is an approach to content creation that emphasizes simplicity and clarity in language to make texts more accessible to readers with cognitive challenges or learning disabilities. The Spanish version of E2R is called *Lectura Fácil* (LF). E2R and its variants, such as LF, focus on straightforward language and structure to enhance readability. The manual production of such texts is both time and resource expensive. In this work, we have developed *LFWriteAssist*, an authoring support tool that aligns with the guidelines of LF. It is underpinned by the functionalities of *LanguageTool*, a free and open source grammar, style and spelling checker. Our tool assists in ensuring compliance with LF standard, provides definitions for complex, polysemic, or infrequently used terms, and acronym extensions. The tool is primarily targeted at LF creators, as it serves as an authoring aid, identifying any rule infringements and assisting with language simplifications. However, it can be used by anyone who seek to enhance text readability and inclusivity. The tool's code is made available as open source, thereby contributing to the wider effort of creating inclusive and comprehensible content.

Keywords: Lectura Fácil, readability, cognitive accesibility

1. Introduction

As our world becomes increasingly interconnected, the need for accessible and inclusive communication is more important than ever. This has led to the development of specialized linguistic strategies aimed at bridging communication gaps among diverse audiences. One such strategy is the use of simplified language variants, which enhance understanding and engagement for all. Among these innovations, the Easy-to-Read (E2R) initiative stands out as a key transformation in the way information is conveyed and comprehended in society. E2R refers to a simplified version of a standard language, designed to be less complex and thereby enhance the clarity and understanding of texts. Its purpose is to enhance the clarity and understanding of texts, particularly for individuals with communication challenges. It achieves this by using only the fundamental vocabulary and grammatical structures of the respective natural language. It should not be confused with *Plain Language*, as the Plain Language movement has the entire society as target audience, while E2R focuses on people with communication challenges. Each E2R variant receives a name depending on the standard language it is based on; in the case of Spanish, it is *Lectura Fácil* (LF).

Currently, the availability of E2R or LF texts is limited, as they are created from scratch or adapted from standard texts. The adaptation process is expensive in both time and financial resources, as it involves several steps, including incorporating auditory and/or visual aids, providing additional ex-planations, simplifying syntax and vocabulary, and summarizing content. It can be especially difficult to keep up with new content in time-sensitive materials such as news articles. On the other hand, translations are often carried out by individuals without formal translation training, resulting in inconsistent and poor quality texts (Hansen-Schirra et al., 2020).

One solution to help overcome these challenges is the use of Computer-Assisted Translation (CAT) tools to convert standard texts to E2R. Such tools would not only accelerate the translation process for professional translators but also provide a means for non-professional translators to verify the accuracy of their work. *LanguageTool* (Naber et al., 2003) is a writing assistant that checks grammar and spelling mistakes, and offers a nuanced analysis of texts, focusing on style, tone, and typography. This approach allows it to provide context-sensitive suggestions, helping users refine their writing[1]. *LanguageTool* finds errors based on rules. Their core technology is available as open-source software, and therefore, users can create their own custom rules and include them in their grammar. This paper explores the extent to which *LanguageTool* can be used as a CAT tool for E2R in Spanish and presents *LFWriteAssist*, an authoring support tool for LF based on *LanguageTool*.[2]

The rest of the paper is organised as follows: in

---

[1]*LanguageTool*. About us. https://languagetool.org/about?force_language=1

[2]The code and rules are available at https://github.com/margotmg/LFWriteAssist.git

section 2 we present the related work, section 3 introduces *LanguageTool* and its main characteristics, in section 4 we overview the guidelines for LF, in section 5 we present our tool, in section 6 we discuss the main limitations, and in section 7 the conclusions.

2. Related Work

Automatic Text Simplification (ATS) consists of lexical, syntactic, or discourse simplification levels (Chen et al., 2017). Lexical simplification involves Complex Word Identification (CWI) (rare, technical, or abstract words) and substituting them with simpler, more commonly used synonyms or providing their definitions, images, videos, or similar enhancements. Syntactic simplification focuses on simplifying complex sentence structures, such as passive constructions or lengthy sentences; this involves re-organizing, splitting, and adjusting sentence structures, reducing grammatical complexity, and omitting unnecessary information. Discourse simplification addresses coherence and coreference issues to ensure that no important information has been lost during lexical and syntactic simplifications. In E2R, visual or graphic adaptation is also taken into account, that is, the visual design and layout of the text. ATS usually follows three main approaches: rule-based approach, data-driven approach, and hybrid approach. However, there is a lack of tools based on neural approaches (Espinosa-Zaragoza et al., 2023).

Different tools and approaches have been proposed for Spanish text simplification and adaptation [3]:

- *LexSiS* (Bott et al., 2012a) is the first approach to lexical simplification in Spanish. It was created based on the empirical analysis of a sample of data from the *Simplext* corpus, and it relies on freely available resources, such as dictionaries and the web.

- *DysWebxia* (Rello et al., 2013) is the first model for people with dyslexia that presents synonyms for complex words in the text and includes changes in the design of text presentation based on quantitative studies with people with dyslexia.

- *OpenBook* (Barbu et al., 2015) is a rule-based tool that helps the Autistic Spectrum Disorder (ASD) carers and people with ASD to simplify the written documents on a discourse, syntactical and lexical level. It is multilingual, available for Spanish, English and Bulgarian, and it was developed in the scope of the *FIRST* project (Valdivia et al., 2014).

- *Simplext* (Bott et al., 2012b; Saggion et al., 2015a) is a rule-based prototype for syntactic simplification in Spanish, tackling sentence splitting, lexical substitution, and syntactic reordering. This was part of the *Simplext* project (Saggion et al., 2015b).

- *CASSA plug in* (Rello et al., 2015) was created based on the *CASSA* algorithm (Context-Aware Synonym Simplification Algorithm) (Baeza-Yates et al., 2015), a context-aware algorithm for generating simpler synonyms, using resources like Google Books Ngram Corpus and Spanish OpenThesaurus and real web frequencies of the complex word for disambiguation.

- The *Able2Include* project[4] (Saggion et al., 2017) aims at improving the living conditions of people with Intellectual or Developmental Disabilities (IDD) in key areas of society by introducing accessible web-based tools. Some of their tools are also available in languages other than Spanish.

- *MUSST* (Scarton et al., 2017) is a rule-based multilingual syntactic simplification tool, supporting sentence simplifications for Spanish, English, and Italian. It was implemented in the context of the European project SIMPATICO on text simplification for public administration texts.

- *NavegaFácil* (Bautista et al., 2018) is a web application aimed at facilitating the comprehension of text. It allows users to visualize and navigate through the original content of any web page, and provides definitions, synonyms and antonyms, lemmatisations, images, Google search, Wikipedia, translation and text to voice.

- *EASIER* (Alarcon et al., 2021) performs CWI following machine learning techniques and contextual embeddings using Easy Reading and Plain Language resources, and also provides definitions.

- The *ClearText* project[5] (Moreda et al., 2023)

---

[3]There is also *arText-claro* (http://sistema-artext.com/lenguaje-claro) (da Cunha, 2022), the first Spanish-assisted copywriter that helps to write texts in specialised fields and texts in *Lenguaje Claro* (the Plain Language equivalent of Spanish). It has not been included in this list because it focuses on *Lenguaje Claro* and not LF, and it is therefore beyond the scope of this paper.

[4]*Able2Include* project https://able-to-include.ccl.kuleuven.be/index.html

[5]*ClearText* project https://cleartext.gplsi.es

aims to create a tool that simplifies Spanish texts from the public administration, making them more accessible to people with mild to moderate cognitive impairment.

- *FACILE* ([Suárez-Figueroa et al., 2024](#)) is an AI-driven tool to aid, in a semi-automated way, in the E2R adaptation process of documents. It is still under development, but its primary objective is to assist E2R specialists in their routine activities, which include evaluating documents against E2R standards and modifying them in line with E2R principles.

In spite of the existence of these tools and resources, it is worth highlighting that only *arText*, *Simplext* and *EASIER* are operative (the latest being also open source). *MUSST* and *NavegaFácil* offer open source code, but are not operative, to our knowledge, at the writing of this paper. We did not find information on the operativeness of *DysWebxia*, *CASSA plug-in* and *OpenBook* are inoperative, and there is no information on *LexSiS*. In fact, the availability and accessibility of ATS tools is a recurrent problem in various languages ([Espinosa-Zaragoza et al., 2023](#)).

To the best of our knowledge, there is no previous study or tool that employs *LanguageTool* to aid in LF text adaptation. However, *LanguageTool* has been used as an authoring support tool for *Leichte Sprache* (the E2R equivalent for German) ([Siegel and Lieske, 2015](#)). In this study, they created *Leichte Sprache* rules and implemented them in *LanguageTool*. Our paper follows these steps, but also introduces some important changes (see section [5](#)).

3. LanguageTool as a CAT tool

LanguageTool is an open-sourced, multilingual proofreading tool. As of February 2024, it supports 30 languages and 20 language variants[6] (version 6.3, released October 4th, 2023). It detects spelling, grammatical, and stylistic errors, as well as ambiguities and opportunities for improvement in wording. It can also paraphrase text to improve clarity and fluency. *LanguageTool* integrates seamlessly with a variety of platforms and applications, including web browsers and Office programs, such as Google Docs and OpenOffice. The premium version of LanguageTool offers enhanced capabilities for more thorough and detailed proofreading of texts. *LanguageTool* is known for its focus on open source development, which allows anyone to access and contribute to its code[7], and anyone can

set up their own LanguageTool server locally or in the cloud. This approach encourages continuous improvement and adaptability of the software to different linguistic needs and contexts. Additionally, users can create custom rules to adjust the tool to their own writing styles or specific needs, which makes it a versatile and flexible tool suitable for a wide range of users and applications[8].

The integration of *LanguageTool* into applications offers several advantages. Customizability is a standout feature, allowing developers to tailor *LanguageTool*'s rules to specific needs or guidelines, such as adapting it for E2R content. As an open-source tool, *LanguageTool* invites a collaborative approach to development and improvement, offering transparency in its functionality and the flexibility to modify its code to fit different requirements. This open-source aspect ensures that it can evolve continually with contributions from a global community. The multilingual support of *LanguageTool* allows for integration in applications that cater to diverse user groups, ensuring accurate grammar and style checks across many languages. Finally, the consistency that LanguageTool brings to text is crucial for maintaining a coherent narrative in written text. This consistency is particularly important in E2R, as clear and uniform communication is paramount.

3.1. Custom Rules

The rules in *LanguageTool* follow a specific pattern, which has also been followed in *LFWriteAssist* (refer Table [1](#) for examples). These are the main elements of the rules and their attributes:

- `id`: an internal, unique identifier of the rule.

- `name`: short text displayed in the configuration, describing the rule.

- `antipattern`: complex exception to a rule (optional).

- `pattern`: part of the original text that should be marked as an error.

- `message`: text displayed to the user if the rule matches. Here, we include the sub-element `suggestion` to suggest a replacement to correct the error. If the `suggestion` sub-element is not included in the rule, the text will not be corrected.

- `url`: url to a page explaining the rule in more detail (optional).

- `short`: short description of the rule (optional)

- `example`: example with an incorrect sentence. The position of the error must be marked up with the sub-element `marker`.

At times, it may be necessary to employ multiple rules to identify all instances of an error. All these rules can be combined into a single `rulegroup` element. The `rulegroup id` and `name` attribute are used for all the rules belonging to that group. The rules can also be grouped into categories, depending on their purpose; this allows enabling and disabling those rules at the same time[9].

Custom rules can first be tested in the *LanguageTool* online rule editor[10]; this way, users can check if the rule has any errors and whether it covers all the desired linguistic features. After this, the rules must be included in the `grammar` XML file of the preferred language. The rules are different depending on the language; that means that even if the same language phenomena happen in two different languages (e.g. passive voice), the rules will be different. Therefore, each language has its own `grammar` XML file, and custom rules will only work for the language they were created for.

4. *Lectura Fácil Guidelines*

The Spanish language counts with the standard *Norma UNE 153101:2018 EX de Lectura Fácil. Pautas y recomendaciones para la elaboración de documentos* (UNE 153101 EX of Lectura Fácil. Guidelines and recommendations for the elaboration of documents) (UNE, 2018). This standard aims to guarantee the comprehension of written documents and the entitlement of all individuals to access information. This standard explains the process of adapting texts in LF, as well as the process of adapting standard texts into LF. It also contains guidelines and recommendations for writing text in LF, and guidelines and recommendations for the design of a document in LF. In this paper, we have focused on the former, which contain the following subsets of guidelines and recommendations:

1. Guidelines and recommendations related to orthotypography

2. Guidelines and recommendations related to vocabulary and expressions

3. Guidelines and recommendations related to phrases and sentences

4. Guidelines and recommendations related to text organisation and style

Within these rules, some of them are specific, clearly defined rules that are straightforward to implement. For instance, the rule that claims that "you should not use acronyms". Conversely, we encountered challenges with rules that are inherently vague or overly generic. An example of such a rule is "avoid the use of words that do not add information to the text and make it longer to read". The subjective nature and the broad scope of this rule make its implementation problematic. Consequently, these types of rules have been set aside in our current framework due to the difficulty in quantifying and codifying them into a set of parameters. Furthermore, our system does not incorporate rules that require a deeper understanding of the context and meaning such as "do not use nominal phrases and avoid nominal use of adjectives". These are beyond the scope of our current rule-based approach. The complexity of semantic interpretation presents significant challenges for rule-based systems. In an attempt to get the most out of our tool, we have integrated some other resources that align with our focus on lexicon and syntactic rules. We have utilized *Diccionario Fácil*[11], a dictionary that offers simplified definitions of complex, polysemic, or infrequently used terms. It is designed for individuals with reading comprehension difficulties and is an initiative by *Plena Inclusión Madrid*[12], the Madrid Community Federation supporting people with intellectual or developmental disabilities. We have extracted all dictionary entries and definitions, so that they are provided to our tool users. We have also employed the *EASIER* corpus (Alarcon et al., 2023)[13] to provide easier synonyms for the complex words encountered in the text. Additionally, we also created a list of acronyms and a list of abbreviations, which were extracted from *Wikilengua*[14], an open and participatory site for sharing practical information about the norm, usage and style of Spanish. On the other hand, we have also integrated rules that were created using Python in addition to the standard *LanguageTool* rules. These include the detection of long phrases and long words.

---

[9]For further details, please refer to *LanguageTool* development overview on custom rules `https://dev.languagetool.org/development-overview`

[10]*LanguageTool* online rule editor `https://community.languagetool.org/ruleEditor2/`

[11]*Diccionario Fácil* `https://www.diccionariofacil.org/diccionario`

[12]*Plena Inclusión Madrid* `https://plenainclusionmadrid.org`

[13]260 documents were annotated, from which they gathered 8,100 complex words. A total of 7,892 synonyms were proposed.

[14]*Wikilengua* acronyms `https://www.wikilengua.org/index.php/Lista_de_siglas_A` and *Wikilengua* abbreviations `https://www.wikilengua.org/index.php/Lista_de_abreviaturas_A`

Rule with suggestion (figurative meaning)

```
<rule>
 <pattern>
  <token min="0">hasta</token>
  <token>por</token>
  <token>los</token>
  <token>codos</token>
 </pattern>
 <message>Se debe evitar el uso de enunciados con sentido figurado.</message>
 <suggestion>mucho</suggestion>
 <example correction='mucho'>Ella habla <marker>hasta por los codos</marker>.</example>
 <example>Ella habla mucho.</example>
</rule>
```

Rule without suggestion (passive voice)

```
<rule>
 <pattern>
  <token regexp='yes'>asunto|cosa|algo</token>
 </pattern>
 <message>Se debería evitar el uso de palabras de contenido indeterminado como "cosa", "algo" o "asunto".</message>
 <example correction='problema'>Era un <marker>asunto</marker> complicado</example>
 <example>Era un problema complicado.</example>
</rule>
```

Table 1: Examples of a rule with `suggestion`, and a rule without `suggestion`.

5. *LanguageTool* for *Lectura Fácil*

As mentioned in Section 1, to the best of our knowledge, *LanguageTool* has not been previously employed to aid in the LF text adaptation process. There is one study that implements *Leichte Sprache* rules on *LanguageTool*. We follow this work by Siegel and Lieske (2015) and create LF rules to be used with *LanguageTool*.

As mentioned in subsection 3.1, the inclusion or exclusion of a `suggestion` determines whether the text will be automatically changed or not. Some of the UNE rules we have adapted include a `suggestion`, while others do not. Examples of a rule with `suggestion`, and a rule without `suggestion` are provided in Table 1. The rule with a `suggestion` applies for a phrase with figurative meaning. In this case, the rule matches the figurative phrases *hasta por los codos* (even through the elbows) and *por los codos* (through the elbows), which refer to a person that talks a lot. This part of the text is then changed to *mucho* (a lot). Therefore, when having the sentence *él habla hasta por los codos* (he talks even through the elbows), the text will be automatically changed to *él habla mucho* (he talks a lot). The rule without a `suggestion` applies for the rule that claims that "the use of words with indeterminate content such as *thing*, *something* or *issue* should be avoided". The rule matches any of these words, but does not make any changes nor offer any alternative terms, as more information on the context is necessary, and the phrase might need to be rephrased.

The decision to include a `suggestion` that would automatically correct the text depended on various factors:

- **Need for context:** some text may require a deeper understanding of context to make an appropriate correction. The tool flags these areas to prompt the user to review them, as automatic correction could alter the intended meaning or not be feasible without additional context.

- **Limitations of *LanguageTool* rules:** the correction may not fit within the intrinsic nature and structure of *LanguageTool* rules. Sometimes it might not be possible to structure and capture all linguistic nuances or complex structures.

- **Extensive rewriting required:** in cases where a text segment requires significant rephrasing, it may be more efficient to rewrite the entire section rather than attempt to correct it piece by piece. This approach can help maintain the coherence and flow of the text.

In these situations, the tool provides guidance by highlighting the areas of concern in orange, but leaves the decision and manner of revision to the human user, who can take into account the nuance, context, and complexity of the content. Whenever automatic changes have been applied, the areas are highlighted in green. We have named our tool *LFWriteAssist*, and will refer to it as such for now on.

5.1. *LFWriteAssist* Structure

Currently, *LFWriteAssist* functions primarily as an interface, serving as a user-friendly front-end for the more technical aspects of the *LanguageTool* framework and allowing users to interact with the tool's functionalities. The structure of the interface consists of the following parts (see Figure 1):

- Input panel, named *Campo de entrada*, which is a text entry field where users write their source text.

- The second panel, named *Resumido y revisado*, shows the summarized text, the tool's
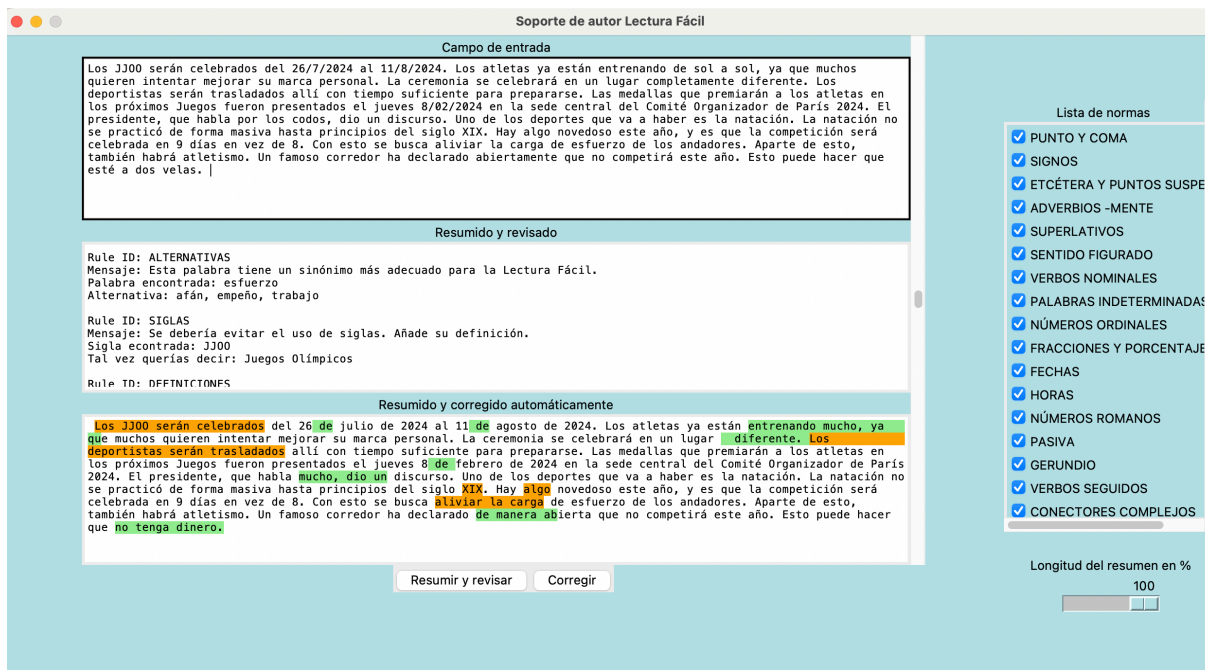
Figure 1: Interface of the *LFWriteAssist* tool displaying three text panels: original text input, automated suggestions for acronyms and complex word definitions, as well as explanations of the broken rules encountered, and the final output. Underlined in green, the changes that have been performed automatically, and underlined in orange, something that should be changed or revised. On the right, a side panel listing various language rules for text simplification. Below the side panel, a slider control to adjust the summary length of the text. This particular example shows a text which has not been summarized, in an aim to show as many error corrections as possible in the output.

suggestions on the original text (both for automatically corrected parts, and areas that need revision), as well as definitions of complex, polysemic, or infrequently used terms and acronym expansions.

- The third panel, named *Resumido y corregido automáticamente*, shows the summarized and automatically corrected version of the source text. Those parts of the text that have been automatically corrected are underlined in green. Those parts underlined in orange are parts of the text that violate some LF rule, but have not been automatically corrected.

- The side panel, named *Lista de normas*, lists the rules for LF that we have created.

- The slider control, named *Longitud del resumen en %*, allows the user to adjust the length of the summary. When choosing 100%, the output text will keep all the information in the source text.

The interface is in Spanish, but if needed, it can be localised to other languages.

6. Limitations

In spite of the strengths our tool offers in the realm of ATS, particularly LF texts, it is important to acknowledge certain limitations and areas for future development. Primarily, our current focus is on the Spanish language, with future research planned for other languages. Notably, a German version exists, but lacks the suggestion feature; therefore, automatic changes are not applied in the final output. Our work is grounded in the guidelines for writing text in LF, but it is important to recognize that LF and E2R texts encompass more than just language simplification. Factors like layout are also vital, which our current tool does not address. All rules have been created manually, which may result in inadvertently missing certain linguistic elements such as figurative phrases, abbreviations, and other language nuances. However, the collaborative nature of *LanguageTool*, on which *LFWriteAssist* is based, allows for the potential addition of more rules by the community, progressively enriching its capabilities. This aspect underscores the tool's evolving nature and the scope for continuous improvement through community involvement. As this is the initial prototype of *LFWriteAssist*, it is important to acknowledge that it may exhibit some errors or limitations. However, it's crucial to note that our pri-

mary audience is LF developers, not the end-users themselves. This distinction is significant because any inaccuracies or shortcomings in the *LFWrite-Assist*'s current iteration are less likely to directly impact the target audience. The developers, being more familiar with LF principles and guidelines, can identify and mitigate these issues during the content creation process. Therefore, while the tool aims to aid in producing more accessible texts, its current prototype status implies a phase of testing and refinement primarily within a professional context.

7. Conclusions and Future Work

The traditional process of producing E2R and LF texts is notably resource-intensive, both in terms of time and financial investment. Despite the existence of some ATS tools, including some targeting LF texts, many lack full operational capability. We have proposed *LFWriteAssist*, an authoring support tool based on *LanguageTool*. We perform extractive summarization, cover different language phenomena and provide definitions when needed based on already existing LF resources, such as dictionaries and guidelines. A distinctive feature of *LFWriteAssist* is its ability to perform automatic alterations in the text, which are highlighted in green for ease of recognition. This visual cue assists users in quickly identifying the modifications made for simplicity and clarity. Moreover, the tool also highlights sections that require manual review. The combination of these features makes our tool a comprehensive assistant in the creation of E2R and LF texts. We advocate for the involvement of target users in the creation and evaluation of ATS tools, therefore, future developments include conducting surveys with LF translators to refine the tool according to their needs. Additionally, we aim to enhance accessibility for LF professionals by implementing this tool on a web page, eliminating the current installation requirements. The open-source nature of this tool invites collaboration and continuous improvement, potentially leading to further advancements in this field. It opens up opportunities for other developers and users to contribute to its development, ensuring that the tool remains adaptable and up-to-date with the evolving needs of its user base.

Our tool aims to enhance the overall simplicity of documents, reduce human effort, and ensure adherence to E2R guidelines. Although a specific evaluation method for *LFWriteAssist* has not yet been finalised, a strategic approach is in place. The plan is to involve professional E2R translators in a comprehensive review process. This approach will involve selecting a diverse group of translators, providing them with various texts, and asking them to use the tool in their translation and proofreading tasks. After using the tool, translators will be asked to provide feedback through surveys and interviews. The feedback will focus on the tool's usability, effectiveness in simplifying texts, and integration into their workflow. The feedback will be critically analysed to assess the tool's performance in terms of accuracy, time efficiency, and overall user satisfaction. The evaluation insights will refine the tool, meeting practical needs of professional translators and aiding in creating high-quality E2R content. The expert-driven process enhances functionality and provides valuable research data, demonstrating real-world applicability and impact.

In addition to considering professional feedback, we are exploring the possibility of conducting a false positive/false negative analysis as part of the evaluation for *LFWriteAssist*. This method involves assessing how accurately the tool identifies E2R issues. A false positive occurs when the tool incorrectly flags a piece of text as non-compliant with E2R guidelines when it is compliant, while a false negative is when the tool fails to identify an E2R issue in the text. By analysing these occurrences, we can measure the precision and accuracy of our tool, providing critical insights into its effectiveness and reliability.

8. Acknowledgements

9. Ethical Considerations

The primary objective of this study is to foster understanding and inclusivity through our focus on E2R and LF. Our use of these terms is strictly for descriptive purposes. The used terminology carries no judgement on the value of languages, dialects, or linguistic styles. We hold all forms of linguistic expression in high regard and are mindful of the sensitivities surrounding discussions about language. Should any part of our discourse or the terminology we have employed unintentionally imply otherwise, we offer our sincere apologies.

10. Bibliographical References

Rodrigo Alarcon, Lourdes Moreno, and Paloma Martínez. 2021. Lexical Simplification System to Improve Web Accessibility. *IEEE Access*, 9:58755–58767.

Rodrigo Alarcon, Lourdes Moreno, and Paloma Martínez. 2023. EASIER corpus: A Lexical Simplification Resource for People with Cognitive Impairments. *Plos one*, 18(4):e0283622.

Ricardo Baeza-Yates, Luz Rello, and Julia Dembowski. 2015. CASSA: A Context-Aware Synonym Simplification Algorithm. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1385.

Eduard Barbu, M Teresa Martín-Valdivia, Eugenio Martinez-Camara, and L Alfonso Urena-López. 2015. Language Technologies Applied to Document Simplification for Helping Autistic People. *Expert Systems with Applications*, 42(12):5076–5086.

Susana Bautista, Raquel Hervás, Pablo Gervás, Axel Bagó, and Javier García-Ortiz. 2018. Taking Text Simplification to the User: Integrating Automated Modules into a Web Browser. In *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*, pages 88–96.

Stefan Bott, Luz Rello, Biljana Drndarevic, and Horacio Saggion. 2012a. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *COLING*.

Stefan Bott, Horacio Saggion, and Simon Mille. 2012b. Text Simplification Tools for Spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1665–1671.

Ping Chen, John Rochford, David N Kennedy, Soussan Djamasbi, Peter Fay, and Will Scott. 2017. Automatic Text Simplification for People with Intellectual Disabilities. In *Artificial Intelligence Science and Technology: Proceedings of the 2016 International Conference (AIST2016)*, pages 725–731. World Scientific.

Iria da Cunha. 2022. Un redactor Asistido para Adaptar Textos Administrativos a Lenguaje Claro. *Procesamiento del Lenguaje Natural*, 69:39–49.

Isabel Espinosa-Zaragoza, José Abreu-Salas, Elena Lloret, Paloma Moreda Pozo, and Manuel Palomar. 2023. A Review of Research-Based Automatic Text Simplification Tools. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 321–330.

Silvia Hansen-Schirra, Jean Nitzke, Silke Gutermuth, Christiane Maaß, and Isabel Rink. 2020. Technologies for Translation of Specialised Texts into Easy Language. *Easy language research: Text and user perspectives*, 2:99.

Marcin Miłkowski. 2012. Translation Quality Checking in LanguageTool. *Corpus Data across Languages and Disciplines. Peter Lang, Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien*, pages 213–223.

Paloma Moreda, Beatriz Botella, Isabel Espinosa-Zaragoza, Elena Lloret, Tania Josephine Martin, Patricio Martínez-Barco, Armando Suárez Cueto, Manuel Palomar, et al. 2023. CLEAR. TEXT Enhancing the Modernization Public Sector Organizations by Deploying Natural Language Processing to Make Their Digital Content CLEARER to Those with Cognitive Disabilities.

Daniel Naber et al. 2003. A Rule-Based Style and Grammar Checker.

Luz Rello, Ricardo Baeza-Yates, and Horacio Saggion. 2013. DysWexia: Textos Más Accesibles Para Personas con Dislexia. *Procesamiento del lenguaje natural*, 51:205–208.

Luz Rello, Roberto Carlini, Ricardo Baeza-Yates, and Jeffrey P Bigham. 2015. A Plug-In to Aid Online Reading in Spanish. In *Proceedings of the 12th International Web for All Conference*, pages 1–4.

Horacio Saggion, Daniel Ferrés, Leen Sevens, Ineke Schuurman, Marta Ripollés, and Olga Rodríguez. 2017. Able to Read My Mail: An Accessible e-mail Client with Assistive Technology. In *Proceedings of the 14th International Web for All Conference*, pages 1–4.

Horacio Saggion, Montserrat Marimon, and Daniel Ferrés. 2015a. Simplificación automática de textos para la accesibilidad de colectivos con discapacidad: experiencias para el español y el inglés. *IX Jornadas Científicas Internacionales de Investigación sobre Personas con Discapacidad*.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015b.

Making it Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Carolina Scarton, Alessio Palmero Aprosio, Sara Tonelli, Tamara Martín Wanton, and Lucia Specia. 2017. MUSST: A Multilingual Syntactic Simplification Tool. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 25–28.

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 *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE.

Melanie Siegel and Christian Lieske. 2015. Beitrag der Sprachtechnologie zur Barrierefreiheit: Unterstützung für Leichte Sprache. *Zeitschrift für Translationswissenschaft und Fachkommunikation*, 8(1):40–78.

Mari Carmen Suárez-Figueroa, Isam Diab, Edna Ruckhaus, and Isabel Cano. 2024. First steps in the development of a support application for easy-to-read adaptation. *Universal Access in the Information Society*, 23(1):365–377.

UNE. 2018. UNE 153101:2018 EX Lectura fácil. Pautas y Recomendaciones para la Elaboración de Documentos. Madrid: Asociación Española de Normalización.

María-Teresa Martín Valdivia, Eugenio Martínez Cámara, Eduard Barbu, L Alfonso Ureña López, Paloma Moreda, and Elena Lloret. 2014. Proyecto FIRST (Flexible Interactive Reading Support Tool): Desarrollo de una Herramienta para Ayudar a Personas con Autismo Mediante la Simplificación de Textos. *Procesamiento del Lenguaje Natural*, 53:143–146.

Paying attention to the words: explaining readability prediction for French as a foreign language

Rodrigo Wilkens, Patrick Watrin, Thomas François

CENTAL, IL&C, University of Louvain, Belgium

{rodrigo.wilkens, patrick.watrin, thomas.francois}@uclouvain.be

Abstract

Automatic text Readability Assessment (ARA) has been seen as a way of helping people with reading difficulties. Recent advancements in Natural Language Processing have shifted ARA from linguistic-based models to more precise black-box models. However, this shift has weakened the alignment between ARA models and the reading literature, potentially leading to inaccurate predictions based on unintended factors. In this paper, we investigate the explainability of ARA models, inspecting the relationship between attention mechanism scores, ARA features, and CEFR level predictions made by the model. We propose a method for identifying features associated with the predictions made by a model through the use of the attention mechanism. Exploring three feature families (i.e., psycho-linguistic, word frequency and graded lexicon), we associated features with the model's attention heads. Finally, while not fully explanatory of the model's performance, the correlations of these associations surpass those between features and text readability levels.

Keywords: readability, model explainability, linguistic features, attention maps

1. Introduction

A significant proportion of the population suffers from poor reading skills in their everyday life (Schleicher, 2019, 2022). According to the results of international surveys on reading abilities like PISA (Schleicher, 2019), approximately 20% of 15-year-old students are ranked as poor readers. This highlights the widespread nature of reading difficulties among young individuals globally and reminds us of the importance of improving literacy skills and assisting those struggling with reading difficulties. Poor reading skills may make day-to-day life difficult, e.g., restricting access to medical information (Friedman and Hoffman-Goetz, 2006) or complicating administrative tasks (Kimble, 1992). Automatic Readability Assessment (ARA) has long been seen as a means of combating these difficulties, for example, by automating recommendations of texts suited to a specific audience to support reading practice and the development of reading skills (Pera and Ng, 2014; Sare et al., 2020).

Research on readability assessment traces back to the 1920s' when Lively and Pressey (1923) used statistical models for predicting the reading difficulty of texts.[1] These models are commonly named readability formulas. At the time, readability formulas were computed by hand and designed as a trade-off between reliability and minimization of effort (e.g., (Flesch, 1948; Dale and Chall, 1948)). Later, the first automatized formulas appeared, such as the Automated Readability Index (Smith

and Senter, 1967). In addition, readability formulas incorporate features (Bormuth, 1966; Coleman and Liau, 1975; Kintsch and Vipond, 1979).

With the advent of the 21st century, the use of Natural Language Processing (NLP) techniques enabled researchers to capture complex textual features automatically, and sophisticated Machine Learning (ML) algorithms allowed them to combine them better through feature engineering (see François and Miltsakaki, 2012; Crossley and Mc-Namara, 2012; Collins-Thompson, 2014; Vajjala, 2021). These models rely on linguistic features exploiting knowledge about the reading process from cognitive psychology (Chall and Dale, 1995), offering insights on how textual characteristics affect readers (Javourey-Drevet et al., 2022). For instance, Collins-Thompson and Callan (2005) showed that taking into account word distributions across grade levels within a multinomial Naïve Bayes classifier outperforms classic readability formulas such as (Flesch, 1948). Schwarm and Ostendorf (2005) captured several syntactic features based on parsing trees, whereas Pitler and Nenkova (2008) designed various semantic and discourse features for capturing properties of lexical chains and discourse relations. In addition, the relatively good interpretability of features allows them to be included in tools that help writers simplify a text by analyzing the reading difficulties of the text (François et al., 2020).

Current ARA work relies on distributed representations of texts (i.e. embeddings) (Cha et al., 2017; Filighera et al., 2019) and Deep Learning (DL) (Nadeem and Ostendorf, 2018; Azpiazu and Pera, 2019; Martinc et al., 2021), yielding improve-

---

[1]Readability should not be confused with Text Simplification that aims to modify a text, making it simpler (Saggion, 2017).

ment over linguistic feature-based systems (e.g., Deutsch et al. (2020); Martinc et al. (2021) for English and Yancey et al. (2021a) for French). Consequently, DL has become the standard in ARA. Contrary to feature-based approaches, the interpretability needs to be improved.

That being said, researchers have been making progress in developing methods to provide explanations for DL models, thus making them more transparent (see Danilevsky et al., 2020; Liang et al., 2021; Sun et al., 2021; Saleem et al., 2022). These methods can provide *global* explanations – i.e., an "overall understanding of deep neural networks model features and each of the learned components such as weights and structures providing" (Liang et al., 2021, 1) – or *local* explanations that try to understand how the model makes a decision based on individual observations. In this paper, we will be concerned with the second class of methods, including saliency maps, explanation generation, probing, and attention scores. Attention scores have been a popular interpretation technique. However, it is subject to some criticisms[2]. Nevertheless, the association between attention head, model's predictions and the linguistic features remains an open question.

In this work, we aim to narrow this gap by identifying if the scores from a attention head in a fine-tuned transformer model for readability are related to ARA features. Our work concentrated on French as a Foreign Language (FFL) readability, using the Common European Framework of Reference for Languages (CEFR) scale (Council of Europe, 2001). Specifically, our objective in this paper is to inspected whether the scores assigned to the tokens by the attention mechanism may relate the ARA features and the CEFR level predictions made by the model. In this work, we focus on the attention mechanism of the transformer model (i.e., self-attention) since it is one of the main keys to the high performance of these models. The main contributions of this work are two. A method for identifying features associated with the prediction made by a model through the attention mechanism. This allows the generation of an explanation of the model's decision from the point of view of linguistic features, which enables a justification of the predicted level to the model's user. The second is contribution consists of the identification that filtering by attention seems to magnify the correlation between feature and text level.

The structure of this paper is as follows. In Section 2, we introduce the standard modeling approach for ARA and discuss related interpretability approaches. Section 3 outlines the features, corpus, and model utilized in this study, accompanied by a detailed description of the proposed method.

Our findings, including an analysis of the features related to model's prediction and a feature-based description of model's decision process, are presented in Section 4. Finally, we offer concluding remarks and suggest avenues for future research in Section 5.

2. Related Work

As this paper combines different research lines, this section first explores the work investigating readability features, identifying informative features for ARA and focusing on those that are explored in this paper (Section 2.1). In Section 2.2, we examine the current literature to predict text readability, focusing on their model's architectures. Finally, in Section 2.3, we discuss frameworks for explaining models.

2.1. Linguistic Features for ARA

There exists a plethora of linguistic features for readability (e.g., 484 are described by Kyle and Crossley (2015), 154 by Chen and Meurers (2016), 380 by Kyle (2016), 16 by Crossley et al. (2016), 400 by Okinina et al. (2020) and 427 by Wilkens et al. (2022)). These may be grouped in different ways. For example, François and Fairon (2012) grouped them by level of information (i.e., lexical, syntactic, semantic and specific) and Wilkens et al. (2022) grouped them by families (e.g., word length, lexical frequency, graded lexicons and lexical norms). From those, our work focuses on lexical norms, lexical frequency and graded lexicons.

Psycho-linguistics explores the relationship between the human mind and language (Field, 2003), where psycho-linguistics norms (or lexical norms) describe how human beings process and understand language and words. These norms are also associated with the reading comprehension of young readers (Crossley et al., 2017; Beinborn et al., 2014), and their scores have been associated with writing quality and development (Sadoski et al., 1995; Crossley et al., 2019; Crossley, 2020). The most commonly explored psycho-linguistic norms in readability research are age of acquisition (AoA), subjective frequency (or familiarity), and concreteness (sometimes conflated with imageability).

Age of acquisition refers to the average age at which individuals acquire a particular word in their vocabulary. This norm is related to readability because earlier acquired words tend to be easier to recognize and understand (Juhasz, 2005). As regards subjective frequency, it measures the perceived frequency of words as a result of individual's experience (i.e. reading experience, oral input, etc.). Initially identified by Solomon and

---

[2]See Bibal et al. (2022).

Postman (1952), the familiarity effect explains that more familiar words to a given reader tend to be processed more quickly and accurately (Balota et al., 2004). Gernsbacher (1984) showed that (1) frequency effects coexists with familiarity effects and (2) word familiarity is fairly stable from one individual to another, at least for high and median frequency items, which justified building lists of familiar words. In ARA, texts containing predominantly familiar words are generally easier to read and comprehend. The last lexical norms we focus on is word concreteness. Neuroscientists have found that concrete and abstract words are processed differently in the brain, and that concreteness gives an advantage in recognition and recall tasks due to their higher degree of imageability (Jessen et al., 2000; Steacy and Compton, 2019).

Lexical frequency strongly predicts lexical complexity and readability (Rayner and Duffy, 1986). Howes and Solomon (1951) first identified the frequency effect, which was subsequently confirmed by numerous studies in psychology (Monsell, 1991; O'Regan and Jacobs, 1992). This effect corresponds to a more frequent word being recognized more quickly. At the text level, a higher reading speed puts less demand on memory resources, which can be allocated to higher-level processes related to comprehension. This explains why word frequency also indirectly affects the comprehension rate of a text (Crossley et al., 2008).

Finally, commonly used for foreign language teaching, graded lexical resources relate a vocabulary to a proficiency scale, assigning each word of the vocabulary to a given proficiency level, at which the word is considered known by most learners of this level. It can be built based on expert perceptions, such as the reference level descriptors for the CEFR (Beacco et al., 2008; Capel, 2010), or derived from an annotated corpus, as in the CEFRLex project (François et al., 2014). Graded lexicons have been already used in ARA as a way to help readability models to encode readers' expected knowledge (Xia et al., 2016; Yancey et al., 2021a).

2.2. ARA models

Recent literature on ARA has consistently demonstrated the superiority of DL methods over conventional feature engineering approaches. Martinc et al. (2021) compared these methods across multiple manually labeled English and Slovenian corpora, concluding that deep neural networks are effective for both supervised and unsupervised readability prediction tasks. However, they noted that the choice of architecture depends on the dataset. Similarly, Deutsch et al. (2020) evaluated various models including conventional machine learning (ML) methods (e.g., SVMs, Linear Models, Logistic Regression), Convolutional Neural Networks, Transformers, and Hierarchical Attention Networks, and also found that the optimal architecture varies depending on the corpus being tested. However, achieving superior performance with DL models in readability assessment requires fine-tuning the model; otherwise, its performance would be comparable to that of a feature-based model (Imperial, 2021).

Targeting French as foreign language readability, Yancey et al. (2021b) compared linguistic, cognitive and pedagogical features and deep learning models. Despite their efforts, non fine-tuned transformers model (i.e., CamemBERT (Martin et al., 2020)) failed to surpass the baseline model by François and Fairon (2012). However, fine-tuning CamemBERT led to a significant improvement, outperforming the previous state-of-the-art model for French.

2.3. Model Explainability

We begin this section by distinguishing interpretability (or comprehensibility) from explainability, to avoid the confusion existing in the literature (Rudin et al., 2022; Broniatowski et al., 2021). In this work, we follow the definitions outlined by Broniatowski et al. (2021): an *interpretable* model offers only the essential information required to make significant decisions, ensuring that the information provided is justified based on the system's functional objectives, while an *explainable* model elucidates the intricate mechanisms by which a particular implementation produced a specific output, without considering the significance of that output to the decision-maker. Our work thus falls under explainability.

In the context of explainability, Rogers et al. (2020) review several papers investigating how BERT encode linguistic information (e.g, represent phrase-structures (Reif et al., 2019), dependency relations (Jawahar et al., 2019), semantic roles (Kovaleva et al., 2019), and lexical semantics (Garí Soler and Apidianaki, 2020). Most studies on linguistic information in transformers uses the probing (or probing-like) method, thus training a classifier ("probe") to map LLM-states to linguistic target labels (Tenney et al., 2019; Niu et al., 2022). Although this allows inferring the linguistic knowledge of a model, this method does not tell us whether the model actually uses information associated with these features in a given prediction.

Alternatively, Clark et al. (2019) proposed methods to analyze the attention mechanisms of pretrained models. They found that certain attention heads process information in such a way that corresponds well to linguistic notions of syntax and coreference. They also demonstrated that a substantial amount of BERT's attention focuses on a

limited number of tokens (e.g., the special token *[SEP]*). Indeed, the inspection of attention heads and attention weights assigned to words is a common method applied in explanatory visualization systems such as Vig (2019); Braşoveanu and Andonie (2020).

Diving deeper into the specifics of the Transformer architecture, it is important to note that not all attention heads are equally important, and some of them can be pruned with marginal performance degradation (Hao et al., 2021). Moreover, it is unclear what relationship exists between attention weights and model outputs (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Bibal et al., 2022). Therefore, the association between attention, prediction and the linguistic properties of the model remains an open question.

The only other existing work that focuses specifically on explainability of readability models, to the best of our knowledge, is Imperial and Ong (2021). Using ELI5[3], they analyzed the weights that classic ML models assign to the features that are part of the model's input vector. The explanation is an interpretation of the features based on their meaning and models' weights.[4]

3. Methodology

Given our goal of identifying how ARA features could explain the predictions of a transformer model fine-tuned for ARA, our starting point is to fine-tune such a model. In this work, we follow the methodology described by Yancey et al. (2021a) for fine-tuning CamemBERT (Martin et al., 2020).[5] Then, we use this model to study the association between ARA features and the tokens on which the model's attention mechanism focuses on. CamemBERT is a model based on the RoBERTa architecture, so it is made up of 12 layers, each with 12 heads of attention. As in all transformers, each attention head uses an attention mechanism to assign weights to the tokens and multiplies these weights by the embeddings of the tokens, thus weighting them. This process is carried out when multiplying *value* by the *softmax* (i.e., a matrix of words by words where values indicate the attention score) in Equation 1. The results of these weightings are concatenated and fed the

next layer. The result of this process passes from one layer to the next until, in the last layer, it is sent to an Multi-layer Perceptron (MLP) which performs the classification.

$$attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

The method explored in this paper relies exclusively on linguistic features (see Section 3.1) and on the attention scores that the model assigns to each token. To identify the attention score of each token, we use the attention heads from the last encoding layer since these are the closest to the classification layer. Thus, we obtained 12 attention scores for each token, each one corresponding to a different head from CamemBERT.

It should be noted that the information produced by an attention head is a matrix of tokens by tokens produced by a self-attention mechanism. The values of this attention matrix indicate the weight of attention to be given to all tokens when another is processed. This mechanism is the core element for creating contextual embeddings in the transformer's architecture. Since an attention matrix indicates weights for all tokens, identifying which tokens receive the most attention is an important question. A simple answer would be to use the n biggest values. However, this method always indicates the same number of tokens. As the model may concentrate the attention scores on a few tokens, which often are punctuation marks, we follow Clark et al. (2019) by considering that a token receives significant score attention only if it is greater than the scores assigned to the punctuation marks and special tokens. In this way, we can distinguish the tokens that receive attention from the others for each attention head. For example, given the output of *softmax* illustrated in Figure 1, our method analyzes row by row, selecting the tokens that have an attention score higher than the highest attention score between *‹s›*, *‹/s›* and punctuation. Therefore, for the token *vous*, in the second row, the selected tokens are *vous*, *étudier*, *un*, *pays*, *european* and *pas*. Next, in our method, we annotate the select tokens with linguistic features (see Section 3.1). In this way, given a feature f, we weight the token by the feature value.[6] For example, lets consider f as word length, the tokens selected in the previous example would therefore be $f(vous) = 4$, $f(étudier) = 7$, $f(un) = 2$, and so on.

---

[4]The main difference between Imperial and Ong (2021)'s work and ours is the type of model used. While we focus on one type of transformer, Imperial and Ong (2021) focuses on classic ML models.

[5]Note that we explore CamemBERT in this work, but the proposed methodology could be applied in any transformer encoder such as BERT (Devlin et al., 2018) or RoBERTa (Liu et al., 2019).

[6]The annotation process consists of a tokenization normalization step, due to the fact that the tokenizer used by the transformer model is different from the one used by the lexical resources in which the linguistic features are stored.
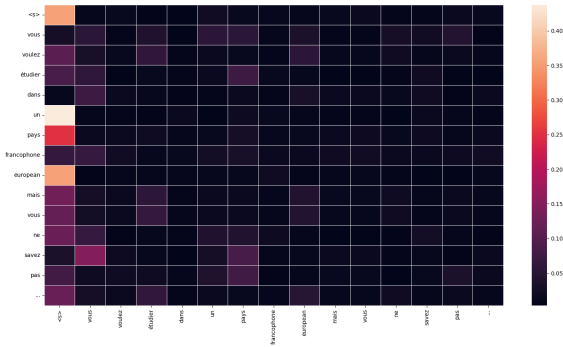
Figure 1: Example of matrix from *softmax*

Then, we use the Equation 2 to calculate the Spearman correlation (ρ) between all tokens that receive attention and the level predicted by the model.[7] More precisely, this correlation is computed based on the predicted CEFR level (l) of a text and the average token score ($score$; see Equation 3), which is the average, for each selected token, of the value of linguistic feature (f) corresponding to the token ($f(token)$) weighted by the attention score assigned to it (α). Similarly, we calculate the correlation for tokens whose attention score were lower than the threshold. In other words, we measure the correlation between the features and the difficulty levels based on the words either considered important to the model or not.

$$\rho = corr(average(score(token)), l) \quad (2)$$

$$score(token) = \alpha(token) \times f(token) \quad (3)$$

As the final step of our analysis, we investigate whether some attention heads tend to specialize towards specific features. We attribute a feature to a specific attention head when the correlation between the feature and the predicted level is higher in the group of tokens selected by the attention threshold than in the group of non-selected tokens.

3.1. Linguistic Features

We explored three families of linguistic features: psycho-linguistic norms, frequency score and graded lexicon. These are widely used in readability studies, as outlined in Section 2. For the annotation of features associated with these families, we used the FABRA toolkit (Wilkens et al., 2022), thus obtaining 19 features:

psycho-linguistic norms: age of acquisition (*AoA)*, word *concreteness*, and word *subjective* frequency (also know as subjective

word familiarity). These scores are based on (Ferrand et al., 2008; Alario and Ferrand, 1999) for AoA, (Desrochers and Thompson, 2009; Ferrand et al., 2008; Bonin et al., 2003; Desrochers and Bergeron, 2000) for subjective frequency, and (Bonin et al., 2018, 2011; Desrochers and Thompson, 2009; Bonin et al., 2003; Desrochers and Bergeron, 2000) for concreteness.

frequency score: word frequency and word frequency band. The latter identifies to which frequency band each word belongs, based on its rank in a reference frequency list. So, as opposed to the word's frequency, we consider the value of the associated band in this feature (e.g., 1000 for the 1000 most frequent words and 2000 for words with a frequency between 1000 and 2000). Since this feature could also be considered as a proportion of words belonging to a frequency band, we chose to use this feature in two ways: the value of the frequency band and the proportion of a band in the text. For the latter, the proportion of each band is named *freq. band"band value"* (e.g., *freq. band_{1000}*).

graded lexicon: proportion of words at one of the 6 CEFR levels (between A1 to C2). These features are named *word level"CEFR level"*. In this work, we use FLELex (François et al., 2014) are reference for the expected CEFR level of a word.

3.2. Corpus

A common way to build readability corpora is to collect textbooks and label each extracted text with the level of the textbook it comes from (e.g., Sato et al. (2008); Volodina et al. (2014)). In this work, we focus on French as a Foreign Language readability, using the CEFR scale (Council of Europe, 2001), which includes six levels: A1 (Breakthrough); A2 (Waystage); B1 (Threshold); B2 (Vantage); C1 (Effective Operational Proficiency) and C2 (Mastery). We used the same corpus as Yancey et al. (2021a), which is composed of 2.734 texts with a balanced distribution of texts in each of the target levels, as described in Table 1.

This corpus is build upon pedagogical materials published after 2001 indicate which CEFR level they are intended for. It was originally proposed by François and Fairon (2012) who creates an initial version of 1.793 texts. Later, Yancey et al. (2021a) expanded their collection into a larger and more diverse corpus extracted from 47 FFL textbooks published between 2001 and 2018. In this corpus, the level of a text is the level indicated in the textbook it was extracted from; with the exception of the C1

---

[7]We used the level predicted by the model because, in this study, we aim to explain the readability model and not the readability phenomenon.

and C2 levels that the authors have grouped into a single level.

Target	Texts	Words
A1	572	60,022
A2	574	83,294
B1	580	119,048
B2	442	130,877
C1 and C2	566	198,517
Total	**2734**	**591,758**

Table 1: Description of the corpus compiled by Yancey et al. (2021a)

4. Results

The first result to report in this paper is the performance of the readability prediction model. After training, the fine-tuned model achieved an accuracy of 0.57 and an F-score of 0.54 (0.74 for level Al, 0.53 fro A2, 0.48 for B1, 0.26 for B2, and 0.72 for C), estimated with a five-fold cross-validation. These results are similar to those reported by Yancey et al. (2021a). As the model is not the focus of this work, we are looking for a model close to the state of the art in terms of architecture and performance. This being achieved, this model can serve as the cornerstone for the results reported in the rest of this section.

4.1. Discrimination power

Before we start to study the applicability of the feature to explain the model, we assess their discrimination power. So, we computed the correlation between each of the 20 features studied and the target levels, as is usually done in ARA studies. Although these values are not connected with our model, they will serve as a reference. As can be seen at column "true label (0)" in Table 2, we found correlations ranging from -0.65 (*word level$_{A1}$*) to 0.55 (*word level$_{C1}$*) when relating the true readability level with the average feature value of all tokens in a text. These correlations confirms that some of our features are good predictor of the CEFR level of a text. In addition, in column "pred (1)", we also calculate the correlation between the predicted readability level with the average feature value, since our ultimate goal is to identify whether the model might be explained by the features. We observe tiny increases when comparing these correlations, which suggest that the approximation made by the model is closer from these features than these features are from the real readability level.

The model explainability analysis starts by considering the relationship between the features and

the model's predictions. This is done without distinguishing the attention heads, meaning that we calculate the attention for each head, but we do not differentiate which head generates the association. We calculated the correlation between the level predicted by the model and each feature, but, this time, we removed the words that had a small attention score (see Section 3). These values can be seen in the selected words column (2) of Table 2, and the absolute difference between these correlations and the original correlations is in column "(1) - (2)". The latter shows an increase in correlation for all the features, except for *word level$_{A1}$*, which had a decrease of 0.23 in its correlation with the predicted level. This already allow us to identify that attention scores acts as a sort of filter that magnifies the correlation between ARA features and predictions, possibly by removing noise (i.e., word embeddings unnecessary for the classification).

Although this analysis already reveals an association between the features and the predictions, it does not indicate how the model measures the features (as they are not provided to the model). We, therefore, explored an alternative version of the correlation between the predicted levels and the values of the features in the list of selected words. In this version, we weighted the features' values by the attention score assigned by the model. These results are shown in column "selected words weighted by attention (3)" of Table 2. As can be seen, the weight of attention does not affect the intensity of the correlation for most of the features[8], except *AoA* (increase of 0.16 points), *concreteness* (0.24), *subjective frequency* (0.31) and *frequency band* (0.08). We therefore observed that the attention-based word filter has a greater impact than the combination of attention weights.

In order to complement the analysis of the correlation between the features and the readability levels, we also analyzed the impact of the predictive capacity of a simple machine learning model to identify the readability level of the text using only the words selected by the attention filter. The goal of this analysis is to identify how the reduction in the text length caused by the proposed filter would affect the performance of a classification model based purely on linguistic features. For that end, we compared the performance of Random Forest classifiers trained using all tokens in the document with RF classifiers using only the tokens selected by the proposed filter. Moreover, we also assess the impact of training the RF classifiers on the true labels and the transformer predictions. This allowed us to further confirm the relation existing between the linguistic variables and the predictions of transformer that are not leveraging any of these

---

[8]Absolute value of column "(2) - (3)" ≤ 0.05.

Features		Correlation				Difference		
		entire corpus		selected words				
		true label (0)	pred (1)	(2)	wgt att (3)	(0) - (1)	(1) - (2)	(2) - (3)
AoA		0.31	0.33	0.36	-0.52	0.02	0.03	0.16
Concreteness		-0.31	-0.34	-0.39	-0.63	0.03	0.05	0.24
Subjective F.		-0.15	-0.17	-0.27	-0.58	0.02	0.10	0.31
Word Freq.		0.23	0.26	0.39	-0.34	0.03	0.13	-0.05
Freq. Band		0.34	0.37	0.39	-0.47	0.03	0.02	0.08
Freq. Band	1000	-0.40	-0.45	0.47	0.45	0.05	0.02	-0.02
	2000	0.26	0.31	0.54	0.58	0.05	0.23	0.04
	3000	0.18	0.20	0.54	0.53	0.02	0.34	-0.01
	4000	0.24	0.28	0.55	0.53	0.04	0.27	-0.02
	5000	0.15	0.16	0.53	-0.05	0.01	0.37	-0.05
	6000	0.20	0.21	0.51	0.46	0.01	0.30	-0.05
	7000	0.24	0.24	0.51	0.46	0.00	0.27	-0.05
	8000	0.27	0.27	0.50	0.45	0.00	0.23	-0.05
Word Level	A1	-0.65	-0.73	0.41	0.46	0.08	-0.32	0.05
	A2	0.25	0.28	0.54	0.51	0.03	0.26	-0.03
	B1	0.27	0.32	0.58	0.58	0.05	0.26	0.00
	B2	0.16	0.17	0.51	0.45	0.01	0.34	-0.06
	C1	0.55	0.60	0.66	0.70	0.05	0.06	0.04
	C2	0.38	0.44	0.63	0.63	0.06	0.19	0.00

Table 2: Correlation between features and CEFR target levels of documents. The last two columns indicate the absolute difference between the correlations of the other three columns.

Target	Attention Filter	F1	Acc
true label	no	0.43	0.45
true label	yes	0.41	0.43
prediction	no	0.48	0.51
prediction	yes	0.47	0.51

Table 3: The ability of a feature to predict the target features.

As can be seen in Table 3, the result of the predictive capacity shows a reduction of 0.02 of F1 and 0.01 of accuracy when using the word filter for predicting the document readability level and 0.01 of F1 and accuracy when predicting the transformer predictions. These results point out that the reduction of a considerable part of the words in the documents does not strongly impact the model's performance, suggesting that the filter is removing possible duplicated or unnecessary words. In other words, the filter allows us to train models with similar performance with less input. However, it is essential to note that this experiment aims to assess whether the selected words can still be used for the task, not to propose an explanation of the transformer model.

4.2. Features and Attention heads

Moving on in our study, we compared the attention head level. This analysis found that psycho-linguistic features tend to be associated with the same attention heads. Similarly, the features related to frequency tend to be grouped in the same way. Following the same behavior but with fewer associated heads, the graded lexicon features tend to be found in the same attention heads.

4.2.1. Base Method

The association between attention heads and features is shown in Table 4. In this table, we can see that several heads are related to at least one feature of the three families of features. However, some heads are associated with several features from the same family. Furthermore, some of them are associated with more than one family. For example, *Head 5* is associated with *psycho-linguistic* and *frequency* features, *Head 9* with *graded lexicon* and *frequency* features, and *Head 7* is associated with all three groups of features. Considering the perspective of features, the psycho-linguistics features are related with, on average, 6.5 attention heads, 2.8 for frequency features, and 2.5 for graded lexicon. In addition, *psycho-linguistics* features are also associated with *Head 4*, *7* and *10*, and the *frequency* features are also associated with *Head 2* and *3*. In general, these results are in line with those of Clark et al. (2019), where it was identified that only a few heads are related to the model's decision.

	Psycholinguistic	Frequency	Graded lexicon	Count
Head 1	-	-	-	0
Head 2	subj.Freq. (-0.49)	freq. band$_{6000}$ (0.44) freq. band$_{8000}$ (0.45)	**word level$_{B2}$** (0.45)	6
Head 3	subj.Freq. (-0.51)	freq. band$_{6000}$ (0.43) freq. band$_{8000}$ (0.44)	-	5
Head 4	**aoa** (-0.49) **concreteness** (-0.58) **subj.Freq.** (-0.54)	freq. band$_{2000}$ (0.58)	word level$_{C1}$ (0.7)	6
Head 5	**aoa** (-0.52) **concreteness** (-0.58) **subj.Freq.** (-0.54)	**freq. band** (-0.42) **wordFreq** (-0.34) freq. band$_{2000}$ (0.56)	word level$_{C1}$ (0.69)	9
Head 6	subj.Freq. (-0.52)	freq. band$_{3000}$ (0.53)	word level$_{C1}$ (0.7)	4
Head 7	aoa (-0.51) **concreteness** (-0.57) **subj.Freq.** (-0.56)	**freq. band** (-0.47) freq. band$_{1000}$ (0.45) freq. band$_{3000}$ (0.51) freq. band$_{5000}$ (0.48) **freq. band$_{6000}$** (0.46) freq. band$_{8000}$ (0.45)	word level$_{B2}$ (0.45) word level$_{C2}$ (0.63)	14
Head 8	-	freq. band (-0.43) **freq. band$_{6000}$** (0.46)	**word level$_{A1}$** (0.46) word level$_{B2}$ (0.44)	6
Head 9	concreteness (-0.54) **subj.Freq.** (-0.53)	**freq. band$_{3000}$** (0.53)	word level$_{B1}$ (0.58) word level$_{C1}$ (0.69)	6
Head 10	**aoa** (-0.51) **concreteness** (-0.63) **subj.Freq.** (-0.58)	freq. band$_{2000}$ (0.58)	word level$_{C1}$ (0.69)	6
Head 11	**aoa** (-0.51) **concreteness** (-0.59) **subj.Freq.** (-0.53)	freq. band$_{2000}$ (0.57)	-	5
Head 12	-	-	-	0

Table 4: Association between attention heads and features. The number in brackets indicates the correlation between the predicted CEFR level and feature weighted by attention score for each attention head. Items in bold are those selected with a threshold of 0.02.

4.2.2. Acceptance threshold

The results we have presented so far rely on the assumption that a feature is related to an attention head if the correlation between the feature and the level predicted is higher in the group of words selected based on attention scores. In order to better understand the method explored in this paper, we relaxed this assumption. To do this, we defined a simple acceptance threshold based on the difference in correlation between the groups of words (selected v. non-selected). When this threshold is set to zero, the results described above in this section (with 67 associations between features and heads) are obtained, while no association is observed when it is set to 0.14. The other values explored in this threshold show 53 heads selected for 0.01, 35 for 0.02, 25 for 0.03, 19 for 0.04, 19 for 0.05, 16 for 0.06, 14 for 0,07, 8 for 0,08, 6 for 0,09, 4 for 0,10, 2 for 0,11, 2 for 0,12, and 1 for 0,13. This trend towards a reduction in the method's selectivity should be considered in light of the range of the correlation values. These have an average value of 0.43. Thus, the 0.1 limit range explored accounts for 23% of the correlation range available for exploration. Taking a closer look at the distribution of the distance between the absolute correlation values of selected and non-selected words, we see a median of 0.05 (variance of 0.004, Q1 of 0.02, Q3 of 0.09 and max of 0.32).

4.2.3. Base Method with Acceptance threshold

We therefore revisited the association between the attention heads and the features, setting a threshold of 0.02. These values are in bold in Table 4.

The application of the threshold allows us to see a clearer picture of the data. It can be seen that *psycho-linguistic* family is the one most associated with the attention heads, contrary to the previous perspective marked by a similar presence of all types of features. In fact, *psycholinguistic* features are most related with 6 heads (*Heads 4, 5, 7, 9, 10* and *11*). Surprisingly, the features of family *graded lexicon*, which represent features most associated with the task the model was fine-tuned for, were not associated with most of the heads. They were

only associated with *Heads 2* and *8*. For *Head 2*, the feature identified was *word level$_{B2}$*, which had the lowest correlation with the corpus of features in its family. Finally, the *frequency* family, previously the most relevant feature, now is associated with 4 heads. However, it only has few relevant features per head, in contrast to family *psycho-linguistic* where there are several features associated with the same head. In this family, the most relevant features were *Frequency Band$_{6000}$*, which indicates words of medium complexity, and *Frequency Band$_{3000}$* which indicates easy words.

5.   Conclusion

The field of ARA has evolved a lot recently due to recent advances in NLP: it has shifted from models based on theoretically-grounded linguistic features to more accurate black-box DL models. As a consequence, the relationship between readability models and the literature about the cognitive processes involved in reading has been weakened. Thus, it could be possible for a model to identify the expected level of a text, but for the wrong reasons.

Aiming to narrow the gap opened by the widespread use of black-box models, we proposed a method to investigate whether the transformer architecture, when fine-tuned on the readability task, is sensitive to word characteristics that traditional readability features were capturing. We also explore whether attention heads might get specialized to some ARA features. For that, we correlated the level of the predictions made by the model with the ARA features on tokens to which the model is paying attention.

In our finding, we identified that the filtering of word information by the attention layer seems to magnify the correlation between features and the predicted text level. In addition, we were able to identify that attention heads are more sensitive to some linguistic features than others, and describe those which are associated to most of the ARA features explored in this work. Despite being able to identify a relationship between attention heads and linguistic features, these do not explain 100% of the model's behavior as well as the ARA features cannot fully explain the readability level in the corpus. This might indicate that the method is not capable of indicating the feature precisely, but rather something more interesting: the readability effect that the feature seeks to approximate.

As future work, we foreseen the extension of the proposed method to include other than lexical features, such as grammatical or discursive properties. We could also reproduce the analysis to the other layers of the transformer, as it is expected than some layers might be more sensitive to some kind of information than others. Finally, it would be necessary to assess our method on other corpora and using more diverse transformer architectures in order to assess its robustness.

6.   Acknowledgements

7.   Bibliographical References

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

David A Balota, Michael J Cortese, Susan D Sergent-Marshall, Daniel H Spieler, and Melvin J Yap. 2004. Visual word recognition of single-syllable words. *Journal of experimental psychology: General*, 133(2):283.

J.-C. Beacco, S. Lepage, R. Porquier, and P. Riba. 2008. *Niveau A2 pour le français: Un référentiel*. Didier.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Readability for foreign language learning: The importance of cognates. *ITL-International Journal of Applied Linguistics*, 165(2):136–162.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900.

J.R. Bormuth. 1966. Readability: A new approach. *Reading research quarterly*, 1(3):79–132.

Adrian MP Braşoveanu and Răzvan Andonie. 2020. Visualizing transformers for nlp: a brief survey. In *2020 24th International Conference*

Information Visualisation (IV), pages 270–279. IEEE.

David A Broniatowski et al. 2021. Psychological foundations of explainability and interpretability in artificial intelligence. *NIST, Tech. Rep.*

A. Capel. 2010. A1-b2 vocabulary: Insights and issues arising from the english profile wordlists project. *English Profile Journal*, 1(1):1–11.

M. Cha, Y. Gwon, and H.T. Kung. 2017. Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2003–2006. ACM.

J.S. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge.

Xiaobin Chen and Detmar Meurers. 2016. Ctap: A web-based tool supporting automatic complexity analysis. In *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)*, pages 113–119.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

M. Coleman and T.L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

K. Collins-Thompson and J. Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

S. Crossley, J. Greenfield, and D. McNamara. 2008. Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3):475–493.

Scott Crossley. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3).

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48(4):1227–1237.

Scott A Crossley and Danielle S McNamara. 2012. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2):115–135.

Scott A Crossley, Stephen Skalicky, and Mihai Dascalu. 2019. Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, 42(3-4):541–561.

Scott A Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359.

E. Dale and J.S. Chall. 1948. A formula for predicting readability. *Educational research bulletin*, 27(1):11–28.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. 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*, pages 447–459.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

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*.

John Field. 2003. *Psycholinguistics: A resource book for students*. Psychology Press.

Anna Filighera, Tim Steuer, and Christoph Rensing. 2019. Automatic text difficulty estimation using embeddings and neural networks. In *European Conference on Technology Enhanced Learning*, pages 335–348. Springer.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

T. François. 2015. When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, 20(2):79–97.

T. François, N. Gala, P. Watrin, and C. Fairon. 2014. FLELex: a graded lexical resource for French foreign learners. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3766–3773.

T. François and E. Miltsakaki. 2012. Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the 2012 Workshop on Predicting and improving text readability for target reader populations (PITR2012)*.

Thomas François. 2011. *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*. Ph.D. thesis, Ph. D. thesis, Université Catholique de Louvain. Thesis Supervisors: Cédrick ….

Thomas François and Cédrick Fairon. 2012. An "ai readability" formula for french as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477.

Thomas François, Adeline Müller, Eva Rolin, and Magali Norré. 2020. Amesure: a web platform to assist the clear writing of administrative texts. 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: System Demonstrations*, pages 1–7.

Daniela B Friedman and Laurie Hoffman-Goetz. 2006. A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Education & Behavior*, 33(3):352–373.

Aina Garí Soler and Marianna Apidianaki. 2020. BERT knows Punta Cana is not just beautiful, it's gorgeous: Ranking scalar adjectives with contextualised representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7385, Online. Association for Computational Linguistics.

M.A. Gernsbacher. 1984. Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, 113(2):256–281.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971.

D. Howes and R. Solomon. 1951. Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, 41(40):1–4.

Joseph Marvin Imperial. 2021. Knowledge-rich bert embeddings for readability assessment. *arXiv preprint arXiv:2106.07935*.

Joseph Marvin Imperial and Ethel Ong. 2021. Under the microscope: Interpreting readability assessment models for filipino. *arXiv preprint arXiv:2110.00157*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Ludivine Javourey-Drevet, Stéphane Dufau, Thomas François, Núria Gala, Jacques Ginestié, and Johannes C Ziegler. 2022. Simplification of literary and scientific texts to improve reading fluency and comprehension in beginning readers of french. *Applied Psycholinguistics*, 43(2):485–512.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

F. Jessen, R. Heun, M. Erb, D.-O. Granath, U. Klose, A. Papassotiropoulos, and W. Grodd. 2000. The concreteness effect: Evidence for dual coding and context availability. *Brain and Language*, 74:103–112.

Barbara J Juhasz. 2005. Age-of-acquisition effects in word and picture identification. *Psychological bulletin*, 131(5):684.

J. Kimble. 1992. Plain english: A charter for clear writing. *TM Cooley L. Rev.*, 9:1.

W. Kintsch and D. Vipond. 1979. Reading comprehension and readability in educational practice and psychological theory. In L.G. Nilsson, editor, *Perspectives on Memory Research*, pages 329–365. Lawrence Erlbaum, Hillsdale, NJ.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Kristopher Kyle. 2016. *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. Ph.D. thesis, Georgia State University, Atlanta, Georgia.

Kristopher Kyle and Scott A Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, 49(4):757–786.

Yu Liang, Siguang Li, Chungang Yan, Maozhen Li, and Changjun Jiang. 2021. Explaining the blackbox model: A survey of local interpretation methods for deep neural networks. *Neurocomputing*, 419:168–182.

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*.

B.A. Lively and S.L. Pressey. 1923. A method for measuring the "vocabulary burden" of textbooks. *Educational Administration and Supervision*, 9:389–398.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, Benoît Sagot, et al. 2020. Camembert: a tasty french language model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

S. Monsell. 1991. The nature and locus of word frequency effects in reading. In D. Besner and G. Humphreys, editors, *Basic processes in reading : Visual word recognition*, pages 148–197. Lawrence Erlbaum Associates Inc., Hillsdale, NJ.

Farah Nadeem and Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 45–55.

Jingcheng Niu, Wenjie Lu, and Gerald Penn. 2022. Does bert rediscover a classical nlp pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3143–3153.

Nadezda Okinina, Jennifer-Carmen Frey, and Zarah Weiss. 2020. Ctap for italian: Integrating components for the analysis of italian into a multilingual linguistic complexity analysis tool. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7123–7131.

J. O'Regan and A. Jacobs. 1992. Optimal viewing position effect in word recognition : A challenge to current theory. *Journal of Experimental Psychology : Human Perception and Performance*, 18(1):185–197.

Maria Soledad Pera and Yiu-Kai Ng. 2014. Automating readers' advisory to make book recommendations for k-12 readers. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 9–16.

E. Pitler and A. Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195.

Keith Rayner and Susan A Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85.

Mark Sadoski, Ernest T Goetz, and Enrique Avila. 1995. Concreteness effects in text recall: Dual coding or context availability? *Reading Research Quarterly*, pages 278–288.

Horacio Saggion. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

Rabia Saleem, Bo Yuan, Fatih Kurugollu, Ashiq Anjum, and Lu Liu. 2022. Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing*.

Antony Sare, Aesha Patel, Pankti Kothari, Abhishek Kumar, Nitin Patel, and Pratik A Shukla. 2020. Readability assessment of internet-based patient education materials related to treatment options for benign prostatic hyperplasia. *Academic Radiology*, 27(11):1549–1554.

Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. 2008. Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Andreas Schleicher. 2019. Pisa 2018: Insights and interpretations. *OECD Publishing*.

Andreas Schleicher. 2022. How the european schools compare internationally pisa for schools 2022. *OECD Publishing*.

Sarah E Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530.

E.A. Smith and R.J. Senter. 1967. Automated Readability Index. Technical report, AMRL-TR-66-220, Aerospace Medical Research Laboratories, Wright-Patterson Airforce Base, OH.

R.L. Solomon and L. Postman. 1952. Frequency of usage as a determinant of recognition thresholds for words. *Journal of Experimental Psychology*, 43(3):195–201.

L.M. Steacy and D.L. Compton. 2019. Examining the role of imageability and regularity in word reading accuracy and learning efficiency among first and second graders at risk for reading disabilities. *Journal of Experimental Child Psychology*, 178:226–250.

Xiaofei Sun, Diyi Yang, Xiaoya Li, Tianwei Zhang, Yuxian Meng, Han Qiu, Guoyin Wang, Eduard Hovy, and Jiwei Li. 2021. Interpreting deep learning models in natural language processing: A review. *arXiv preprint arXiv:2110.10470*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Sowmya Vajjala. 2021. Trends, limitations and open challenges in automatic readability assessment research. *arXiv preprint arXiv:2105.00973*.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 128–144.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P Yancey, and Thomas François. 2022. Fabra: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233.

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.

Kevin Yancey, Alice Pintard, and Thomas Francois. 2021a. Investigating readability of french as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e Linguaggio*, 2021(2):229–258.

Kevin Yancey, Alice Pintard, and Thomas Francois. 2021b. Investigating readability of french as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e linguaggio*, 20(2):229–258.

8. Language Resource References

Alario, F-Xavier and Ferrand, Ludovic. 1999. *A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition*. Springer.

Bonin, Patrick and Méot, Alain and Aubert, Louis-F and Malardier, Nathalie and Niedenthal, Paula and Capelle-Toczek, M-C. 2003. *Normes de concrétude, de valeur d'imagerie, de fréquence subjective et de valence émotionnelle pour 866 mots*. Persée-Portail des revues scientifiques en SHS.

Bonin, Patrick and Méot, Alain and Bugaiska, Aurélia. 2018. *Concreteness norms for 1,659 French words: Relationships with other psycholinguistic variables and word recognition times*. Springer.

Bonin, Patrick and Méot, Alain and Ferrand, Ludovic and Roux, Sébastien. 2011. *L'imageabilité: normes et relations avec d'autres variables psycholinguistiques*. NecPlus.

Desrochers, Alain and Bergeron, Mylène. 2000. *Valeurs de fréquence subjective et d'imagerie pour un échantillon de 1,916 substantifs de la langue franĉaise.* Canadian Psychological Association.

Desrochers, Alain and Thompson, Glenn L. 2009. *Subjective frequency and imageability ratings for 3,600 French nouns*. Springer.

Ferrand, Ludovic and Bonin, Patrick and Méot, Alain and Augustinova, Maria and New, Boris and Pallier, Christophe and Brysbaert, Marc. 2008. *Age-of-acquisition and subjective frequency estimates for all generally known monosyllabic French words and their relation with other psycholinguistic variables*. Springer.

François, T. and Fairon, C. 2012. *An "AI readability" formula for French as a foreign language*.

François, Thomas and Gala, Núria and Watrin, Patrick and Fairon, Cédrick. 2014. *FLELex: a graded lexical resource for French foreign learners.*

Author Index