# ARTS: Assessing Readability & Text Simplicity 🎨

**Björn Engelmann[1], Christin Katharina Kreutz[2,3], Fabian Haak[1], Philipp Schaer[1]**
[1]TH Köln - University of Applied Sciences, Germany
[2]TH Mittelhessen - University of Applied Sciences, Germany
[3]Herder Institute for Historical Research on East Central Europe, Germany
bjoern.engelmann@th-koeln.de, ckreutz@acm.org, fabian.haak@th-koeln.de,
philipp.schaer@th-koeln.de

## Abstract

Automatic text simplification aims to reduce a text's complexity. Its evaluation should quantify how easy it is to understand a text. Datasets with simplicity labels on text level are a prerequisite for developing such evaluation approaches. However, current publicly available datasets do not align with this, as they mainly treat text simplification as a relational concept (*"How much simpler has this text gotten compared to the original version?"*) or assign discrete readability levels.

This work alleviates the problem of **A**ssessing **R**eadability & **T**ext **S**implicity. We present ARTS, a method for language-independent construction of datasets for simplicity assessment. We propose using pairwise comparisons of texts in conjunction with an Elo algorithm to produce a simplicity ranking and simplicity scores. Additionally, we provide a high-quality human-labeled and three GPT-labeled simplicity datasets. Our results show a high correlation between human and LLM-based labels, allowing for an effective and cost-efficient way to construct large synthetic datasets.

## 1 Introduction

Text simplification enables participation, promotes equal opportunities, and removes barriers in the digital society. Large language models offer effective means of simplifying texts (Ermakova et al., 2023; Engelmann et al., 2023). However, simplification evaluation lacks in comparison (Grabar and Saggion, 2022; Stajner, 2021; Kreutz et al., 2024).

Ideally, text simplification evaluation should follow the principles of readability assessment (Vajjala, 2022) and quantify how easy it is for a certain group of readers to understand a given text. This quantification should be independent of an original version of a text. We define this task of quantifying a text's simplicity in this manner as *simplicity assessment*.

Current text simplification evaluation measures mostly do not align with this, as they mainly treat text simplification as a relational concept. Human-labeled datasets for constructing text simplification evaluation measures typically contain labels based on the extent to which a simplified text is easier to understand than an original text (Alva-Manchego et al., 2020; Alva-Manchego et al., 2021; Maddela et al., 2023; Scialom et al., 2021; Sulem et al., 2018b). Therefore, they do not quantify *simplicity* of a text but rather the degree of *simplification* between an original and modified version of a text. Despite their applicability to single texts, automated readability assessment (ARA) approaches are not well suited for text simplification evaluation. ARA mostly assigns distinct readability levels (Vajjala, 2022) and poorly correlates with simplicity scores assigned by human annotators (e.g., FKGL (Maddela et al., 2023)).

A dataset with manually determined simplicity scores for single texts would be most desirable to enable the development of a text simplicity assessment approach. However, this is a laborious and demanding task. Quantification of the simplicity of a text is complex and subjective for annotators; for instance, a person with medical expertise may perceive the simplicity of a medical text differently than a layperson (Kauchak et al., 2022). Manually assigning a score for a text's simplicity is cognitively demanding because it requires determining what constitutes high and low simplicity. Even for advanced large language models, the task of simplicity assessment is non-trivial. This work strives to overcome this issue by introducing a method for facilitating simplicity assessment dataset creation.

Our contribution, ARTS, lies in *i)* the application of pairwise comparison and an Elo algorithm for dataset creation, including a rating interface, *ii)* a dataset containing manually rated simplicity scores for texts with the use case of text simplicity assessment, and *iii)* an evaluation of the use of large



*Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14925–14942
November 12-16, 2024 

language models for the rating task by comparing the rankings with human annotations. Finally, we show that ARTS datasets can be used to train a model for simplicity assessment.

ARTS is language-independent and can accommodate user characteristics such as domain knowledge, English skill level, and subjectivity. We show that using ARTS, automatically generated simplicity scores correlate highly with human-annotated data, making ARTS ideal for generating synthetic training data.

The code for our project can be found on GitHub[1], the datasets can be found on Zenodo (Engelmann et al., 2024). For modified texts, the licenses of the original texts apply; everything else is published under an MIT license.

## 2 Related Work

### 2.1 Assessing Simplicity

Text simplification is currently evaluated by reporting on BERTScore (Zhang et al., 2020), LENS (Maddela et al., 2023), SAMSA (Sulem et al., 2018a), BLEU (Papineni et al., 2002) or SARI (Xu et al., 2016) computed on the original and simplified texts and in some cases additional reference texts. All these measures have been developed, calibrated, or evaluated using human labels indicating the difference between complex and simplified texts. Therefore, these approaches do not necessarily indicate text simplicity but quantify the relative degree of simplification.

Text readability is generally defined as the ease at which a text can be read and comprehended by a certain group of readers (Collins-Thompson, 2014; Vajjala, 2022). However, most readability datasets and assessment methods use distinct categories, such as reading levels or general binary indications of complexity (Arshad et al., 2023; Vajjala, 2022). Traditional readability formulae, also referred to as readability indices (Arshad et al., 2023), like Flesch Reading Ease (Flesch, 1948), Flesch-Kincaid Grade Level (Kincaid et al., 1975), or SMOG grade (Harry and Laughlin, 1969) generally do not perform well at reflecting human simplicity evaluation (Maddela et al., 2023), and have severe limitations such as ignoring semantics and user aspects (Collins-Thompson, 2014).

ARA is treated mostly as a classification task in NLP (Lee and Vajjala, 2022). Current ARA approaches primarily utilize neural networks and deep learning (Liu et al., 2023; Vajjala, 2022). Most frequently encoder-based transformer models such as BERT (Ibañez et al., 2022; Li et al., 2022; Mohtaj et al., 2022; Zeng et al., 2024) and RoBERTa (Lee et al., 2021) are used. Since most current ARA approaches have been developed to differentiate between distinct categories of readability, such as text difficulty or learner grade level (Li et al., 2023), they are often not fine-grained enough to measure simplification effectiveness and better suited for identifying texts that need to be simplified (Vajjala and Meurers, 2014).

In some cases, pairwise comparisons of texts are used to overcome problems posed by a lack of high-quality training data and the general complexity of assessing simplicity. Pairwise comparisons can be used in pre-processing or as part of training (Liu et al., 2023; Xia et al., 2016; Zeng et al., 2022), most notably by Lee and Vajjala (2022).

Some works explicitly tackle the quantification of text's simplicity. Kreutz et al. (2024) indicate the difficulty of a text through rules a text complies with or breaks. These rules are constructed from literature and either an indicator for simplicity or complexity. Brunato et al. (2018) consider simplicity or complexity of text on a 7-point scale, producing a discrete assessment of sentences similar to the Flesch-Kincaid grade level. This assessment is different from the continuous scores we are producing with ARTS. In contrast, the work by Schumacher et al. (2016) also uses an Elo-based rating system, Trueskill, resulting in continuous scores. While their main objective is the assessment of sentences' contexts, we employ our method for automatic dataset creation, capturing the subjective features of annotators.

The ARTS methodology shares commonalities with pairwise approaches yet extends their principles by incorporating additional layers of complexity. We utilize multiple rounds of pairwise comparisons as well as an Elo algorithm to assign simplicity scores to a set of texts. Further, we do not classify texts categorically but assign a continuous score that reflects the simplicity of a text compared to a large set of other texts.

### 2.2 Datasets with Human Labels

Currently, few publicly available English datasets contain human labels on simplicity, which are used for assessing text simplification and evaluating measures. They all contain pairs of complex

[1] https://github.com/irgroup/ARTS

texts and their corresponding simplifications[2]. The datasets often only stem from a small number of sources and annotators rate their agreement with a variation of the statement: *"The Simplified sentence is easier to understand than the Original sentence"*. Thus, such datasets capture the amount of simplification that has occurred between an original and a simplified version of a text.

In contrast, ASSET$_{preference}$ (Alva-Manchego et al., 2020) holds information on 718 triples of texts consisting of one source text and two simplified versions. Annotators indicated which of the two simplifications is easier to read and understand, is more fluent, and best expresses the original text's meaning. As texts in triples were on the same topic, a comparison of the simplicity of texts across topics is not possible.

ARA models are usually trained and evaluated with datasets that have distinct levels of readability, such as (learner) grade level[3]. These datasets are typically built from sources with inherent readability, and texts are assigned discrete levels. For example, datasets constructed from textbooks or other graded texts that have been produced for a certain level of reading expertise assign grade levels to texts (Martinc et al., 2021; Vajjala, 2022).

With ARTS we provide datasets spanning different domains and target audiences found in publicly available English datasets (see Table 1) for text simplification. The datasets contain human- and LLM-generated continuous simplicity scores for single texts.

## 3 Methodology

Figure 1 gives an overview of our approach. We first select and prepare texts and match them into pairs. These pairs are rated and an Elo algorithm is applied to the rated pairs. Based on that, the ARTS simplicity ranking is composed.

### 3.1 Rationale behind Elo System

Boubdir et al. (2023) describe that the Elo rating system was increasingly used to compare Large Language Models (LLMs) through "A vs. B" paired comparisons. We build on this idea and evaluate

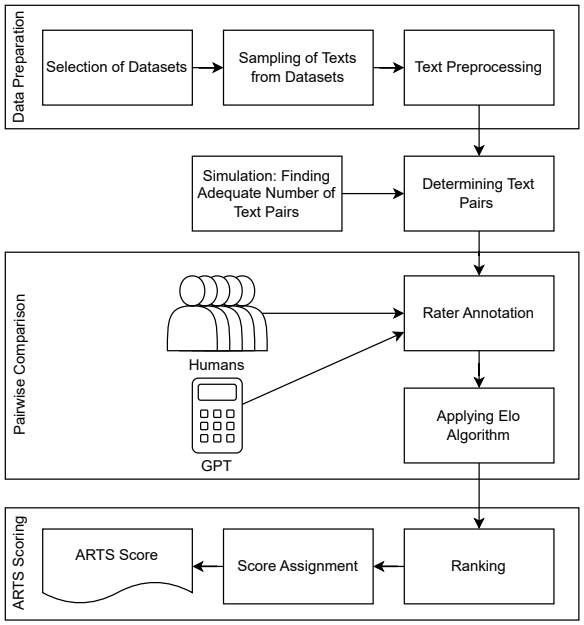


Figure 1: Simplified overview of the ARTS method.

texts instead of LLMs to assign them a score indicating their simplicity. To assess the simplicity of texts, a score must be in an appropriate relationship to other text scores. For this reason, we prefer an annotation environment in which only relative judgments are made. We use an algorithm from the chess domain to derive a score from these pairwise relative judgments. The assumption when determining a (chess) player's score is that the probability that a stronger player will win against a weaker player is higher than that they will lose. However, a player can still lose to a weaker player. After an increasing number of matches, the Elo rating of a chess player converges with their actual playing strength (Boubdir et al., 2023). The more matches a player has played, the more reliable the score assigned to him is. In our setting, the chess players and the texts are to be understood analogously. Two texts always compete against each other in terms of their simplicity. This leads to the rating of more difficult texts increasing and that of simpler texts decreasing in each round. The rules for the round-by-round updates of the scores are presented in Section 3.2. We also present an analysis in Section 3.3 in which we use a simulation to assess the robustness of our method depending on the number of annotations.

[2]Table 4 in the Appendix gives information on all datasets' origins, the number of contained pairs, a brief description of pairs, the number of simplicity ratings per pair and the total number of simplicity ratings.

[3]Table 5 in the Appendix gives an overview of the most frequently used datasets containing human labels for assessing readability, along with the number of classes and number of texts in the datasets.

### 3.2 Rating through Pairwise Comparison

Instead of calculating the Elo rating for players, we apply the Elo system to texts. A text $T_1$ wins against $T_2$ if it is more difficult to understand. To calculate the Elo rating, it is assumed that the probability of text $T_1$ winning against text $T_2$ results from the difference in Elo ratings. If $T_1$ and $T_2$ have identical ratings, the estimated probability of $T_1$ winning is 50%. In the beginning, all Elo scores of our texts are initialized with the same predefined value. Without considering the case of draws and multiple matches, the expected probability $E_{T_1}$ that $T_1$ with Elo rating $R_1$ wins against $T_2$ with Elo rating $R_2$ is defined as follows (Good, 1955; Boubdir et al., 2023):

$$E_{T_1} = \frac{1}{1 + 10^{(R_2 - R_1)/400}}. \quad (1)$$

After a victory of $T_1$ over $T_2$ ($T_1 > T_2$), the new rating $R'_1$ of $T_1$ is calculated as follows:

$$R'_1 = R_1 + k(S_{T_1} - E_{T_1}). \quad (2)$$

The constant $k$ controls how strong the change should be after a game. In our case, $k$ is set to 16. $S_T$ is 1 if $T$ has won and 0 otherwise. The expected probability of winning for $T_2$ and the calculation of the score update are calculated analogously. In the beginning, the ratings of all texts are initialized as 1200. A decision between texts then results in two updates for the scores of the texts per match. The number of matches controls the quality of the resulting Elo ratings. After completing all matches, we transform the Elo scores into the range 0 to 1. We achieve this by ranking the texts in ascending order based on their Elo ratings. Highly-ranked texts are therefore rated as simpler than low-ranked texts.[4] The rank of a text $r(T)$ is then mapped to the simplicity score in the following way:

$$score(T) = \frac{r(T) - 1}{N}, \quad (3)$$

where $N$ is the total number of texts. We use this type of rank-based scaling to produce equidistant distances.

### 3.3 Simulation as Proof of Concept

To ensure a certain level of quality in our pairwise Elo rating updates of texts, we need to determine how many annotations (decisions for a pair of texts) are required for a given number of total texts. As this is currently unclear, we conducted a simulation to find out the parameters. Apriori, the trade-off between annotation effort and rank quality needs to be clarified. To assess this, we make several assumptions for our simulation. We assume we know the ground truth simplicity scores for a hypothetical set of texts. Furthermore, we assume that an oracle knows the correct decision in each match and annotates the text accordingly. We now produce a ranking with the algorithm from Section 3.2 for the hypothetical texts and compare it with the ground truth ranking. With this, we can estimate the quality of our method for a given number of annotations. As a measure for determining the ranking quality, we use the rank correlation of the predefined ranking and the estimated ranking. We also introduce two users $u_1$, $u_2$ who agree with the correct order in 90% and 75% of the annotations to model subjectivity $P(T_1 <_{u_1} T_2 | T_1 > T_2) = .1$ and $P(T_1 <_{u_2} T_2 | T_1 > T_2) = .25$. We also assume that an annotation takes 15 seconds to complete. In the simulation, we only vary the number of annotations to obtain an estimate of the effort and quality for this parameter. To assess the robustness of the simulation, 20 runs are performed for each setting, and the resulting mean values are shown. In addition to the users and the oracle, we simulate a group of 16 users who all make an individual decision and from which a joint judgment based on a majority vote (with ties broken randomly) emerges. Suppose we also have 16 raters for our annotation with real humans. In that case, the majority vote with 16 simulated users can be considered as an upper bound since real raters are probably subject to bias. As a lower bound for the ranking quality of the group of 16 real raters, we consider the individual simulated user with 75% correct decisions. It is important to emphasize that the subjective decision is only modeled with incorrect choices in the simulation. We do not interpret the contradictory annotations of the human raters as incorrect but as subjective.

Based on Figure 2, it can be estimated that the ranking quality increases for an increasing number of annotations. While a linear curve can be seen for the annotation effort, a flattening effect can be seen for the ranking quality. The ranking quality of the simulated users $u_1$ and $u_2$ also approaches the quality of the oracle for an increasing number of annotations. While the difference of user $u_1$ is about

[4] The methodology of ARTS can also be employed in different contexts, for example to evaluate the bias contained in texts (Haak et al., 2024).

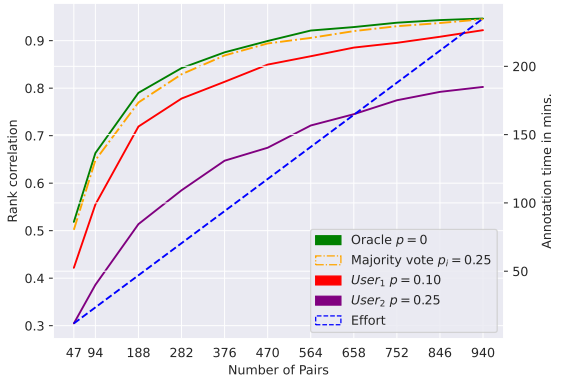


Figure 2: Trend of diminishing returns for linearly increasing effort (dashed blue line, annotation time in minutes on right y-axis) and flattening quality of ranking (green/orange/red/purple line, rank correlation on left y-axis) depending on the number of annotated pairs.

0.09 at the beginning (0.21 respectively for $u_2$), it drops to 0.02 at 940 pairs (0.12 respectively).The ranking quality of the majority vote is almost equal to that of the oracle. Based on this simulation analysis, we decided to have 376 annotations made by 16 users. Based on our assumptions, the resulting rank correlation would be between 0.65 and 0.87. The estimated effort per person would be 94 minutes. The comparison of $u1$ and the majority vote shows how well a majority decision guarantees robustness, even if the individual decision deviates significantly more from the oracle.

## 4 Constructing ARTS

### 4.1 Underlying Data Sources

For creating ARTS, we use the 26 publicly available English parallel corpora for text simplification contained in Table 1[5]. These datasets contain texts simplified for different target audiences and domains, indicated in the table. **ARTS**$_{94}$ contains 48 texts extracted from the source part of datasets and 46 texts extracted from simplified parts of datasets. We randomly extracted four texts from all 26 datasets: we selected two texts each from the simplified part of the datasets and any two texts each from the source part of the datasets and made sure not to select text pairs. This produced 104 unique candidate texts. We excluded the upper 10% of texts in terms of text length, resulting in 94 texts (and slightly imbalanced classes), the longest

[5]Information on licensing, numbers of pairs, and text lengths can be found in Table 6 in the Appendix.

| Dataset information | | | # texts in ARTS | | | |
|---|---|---|---|---|---|---|
| **Name** | **TA** | **D** | **94** | **300** | **3000** | **160** |
| ASSET | - | W | 4 | 20 | 116 | |
| AutoMeTS | - | 🏥 | 4 | 19 | 116 | 40 |
| BenchLS | - | - | 4 | 16 | 116 | |
| Britannica | 🧒 | W | 4 | | 116 | 40 |
| D-Wikipedia | 🧒🌐 | W | 1 | 7 | 117 | |
| EW-SEW-gmpm | 🧒🌐 | W | 4 | 20 | 116 | |
| EW-SEW-Turk | - | W | 4 | 20 | 116 | |
| HTSS | 🧑‍💼 | 🔬 | | | 4 | |
| HutSSF | - | 📰 | 4 | 14 | 116 | |
| MASSAlign | - | 🔬 | 3 | 15 | 116 | 40 |
| METAeval | - | W | 4 | | 116 | |
| MTurkSF | 🧑‍💼 | 🏥 | 4 | | 116 | |
| NNSeval | 🌐 | W | 4 | | 116 | |
| OneStopEnglish | 🌐 | 📰 | 3 | 9 | 116 | |
| PWKP | 🧒🌐 | W | 4 | 20 | 116 | |
| QuestEval | - | W | 4 | | 116 | |
| SemEval-2007 | - | - | 4 | | 116 | |
| SimPA | 🌐 | 🔨 | 4 | 20 | 116 | 40 |
| SimpEval | - | W | 4 | 4 | 116 | |
| SSCORPUS | 🧒🌐 | W | 4 | 19 | 116 | |
| TurkCorpus | 🧒🌐 | W | 4 | 20 | 116 | |
| Wiki-Auto | - | W | 4 | 19 | 116 | |
| Wiki-Manual | - | W | 4 | 1 | 116 | |
| Wikipedia_v1 | 🧒🌐 | W | 4 | 20 | 116 | |
| Wikipedia_v2 | 🧒🌐 | W | 3 | 17 | 118 | |
| WikiSplit | - | W | 4 | 20 | 209 | |

Table 1: Names, target audience (TA), and domains (D) of datasets. Target audiences indicate if the texts are intended for a more specific person group instead of "general": 🧒- children, 🌐- language learners, 🧑‍💼- non-experts. The domain indicates if the texts in the dataset are intended for a more specific domain instead of "general": W- encyclopedia, 🏥- medical, 🔬- science, 📰- news, 🔨- legislative administration.

of which was 306 characters. Formatting of all texts in ARTS$_{94}$ (correction of capitalization as some datasets contain texts in lowercase, unwanted text conversion artifacts, homogenization of spacing and quotes) was manually corrected. **ARTS**$_{300}$ contains 150 source and 150 simplified texts. We randomly extracted five source texts and five simplified texts from all 26 datasets (not contained in ARTS$_{94}$), resulting in 520 unique candidate texts. Out of these, we randomly include 300 texts shorter than 400 characters while making sure to keep the same number of source texts and simplified texts. Again, the formatting of all texts in ARTS$_{300}$ was manually corrected. **ARTS**$_{3000}$ contains 1500 source texts and 1500 simplified texts. We randomly included (not contained in both ARTS$_{94}$ or ARTS$_{300}$) 58 source texts and 58 simplified unique texts from all of the 26 datasets, which were at most 400 characters long. For HTSS (which contains complete documents instead of only short texts or sentences), we were only able to include four texts. The remaining 96 texts were randomly collected from all datasets, where most of them (93) were selected from WikiSplit, the biggest dataset

by far. Formatting was done in batches using ChatGPT 3.5 (prompt in Appendix A.2.1). **ARTS**$_{160}$ contains 80 source and 80 simplified texts not yet used in ARTS$_{94}$, ARTS$_{300}$ and ARTS$_{3000}$, 20 each with a medical (AutoMeTS), geography (Britannica, contains texts from Wikipedia on cities), philosophy (MASSAlign) and legislative administration (SimPA) focus. Texts were 400 characters long at most, formatting was again corrected using ChatGPT 3.5.

Table 7 in the Appendix contains detailed information on the numbers of texts from source and simplified sentences as well as their average lengths in characters from all datasets.

### 4.2 Rating Guidelines and Interface

We use a rating interface as the annotation environment to collect the relative simplicity judgments of the pairs of text from the ARTS$_{94}$ dataset. This allows for an effective and reliable way of collecting the labels needed to rank the text using the approach described in Section 3. In a web interface (shown in Figure 3 in the Appendix), the raters are provided with the instruction to *"click on the text which is easier to understand"*, followed by the text pair. We further provide detailed descriptions of what we define as simplicity that acts as a guideline for the raters (full descriptions in Appendix A.3). All raters were given the same deterministic set of pairs, asserting that we can perform majority votes. Progress in the rating process is saved for each user so that they can perform the task in multiple sessions.

### 4.3 Reuse

We publish the re-formatted texts under the same licenses of the datasets the texts stem from. We publish our scores and source code under MIT License. Used packages are given in A.4.

## 5 Evaluation

### 5.1 User Annotations

Based on our simulation from Section 3.3, we have decided that each annotation run contains 376 annotations, and we perform a total of 16 annotation runs with 16 different human raters. Based on these annotation runs, we compose majority labels where ties are broken randomly. Here, we assume that decisions with a higher agreement of several raters are more robust than those of individual raters. These majority labels are used as the basis for the evaluation. For our majority labels, we achieved a moderate inter-rater agreement (Krippendorff's $\alpha$: .4231). All raters are non-native English speakers but members of the scientific/academic community at different levels from students to professors, between 21 and 43 years old. Our annotators conducted an university-offered English language test which scored ten annotators as level B and six annotators on level C. Using the rating interface, we also logged the time of the rater decisions. Based on these time stamps, we report a median effort duration of 20 seconds per annotation. We did not receive feedback from our annotators of anyone not being able to *break ties*, so, to decide on which text is simpler. Even if annotators considered two texts to be similar in simplicity, they always were able to make a decision. In addition to this, the Elo algorithm is well-suited for ambiguous cases. The assumption is that similarly complex texts differ slightly in their Elo score. Therefore, the difference after the decision will be relatively small.

As shown in Table 9, the agreement rate for the majority labels ranges between .69 and .90. To the best of our knowledge, this range results from the subjectivity of the raters' judgments. Based on the assessment (Landis and Koch, 1977), we can speak of substantial agreement with the $\kappa$-values in the range of .37 to .79. The measures Spearman's $\rho$ and Kendall's $\tau$ provide information on how strong the correlation is between the rankings of the individual raters and the ranking resulting from the majority labels. We can conclude a clear correlation between the ranked scores with a minimum rank correlation $\rho$ of .52 (or Kendall's $\tau$ of .37). Table 2 holds two examples for text pairs with a high human rater disagreement.

### 5.2 Automatic Scoring Approaches

In addition to our human raters, we performed simplicity annotation with different GPT strategies and two well-established readability measures as baselines: **Flesch Reading Ease (FRE)** is a measure that evaluates the readability of texts (Flesch, 1948). Texts that contain relatively few words and few syllables per word receive lower scores.

**Dale Chall Formula (Da_Ch)** produces a numerical value that indicates how difficult it is to understand a text (Dale and Chall, 1948). Based on a list of words that are easy to understand, the score increases for each word from the text that does not appear in the list.

With **GPT3.5** we instruct GPT3.5 (more specifi-

| Case | Text A | Text B |
|---|---|---|
| 1 | Fives is a British sport. It comes from the same origin as many Racquet sports. | "Ohio state's library system includes twenty-one libraries located in the city of Columbus." |
| 2 | An operation to the nasal septum is known as a septoplasty. | Convinced that the grounds were haunted, they decided to publish their findings in a book An Adventure (1911), under the pseudonyms of Elizabeth Morison and Frances Lamont. |

Table 2: Examples for text pairs with a high human rater disagreement.

cally *gpt-3.5-turbo-1106*) to choose the one that is easier for two given texts. This decision is based on the same guidelines that the human raters were given[6]. The prompt can be found in Appendix A.2.2.

With **GPT4**$^R$, the idea is that a request to GPT contains all texts, and we instruct GPT to sort the given texts in ascending order of simplicity. Here, we used the *gpt-4-1106-preview* model. The prompt can be found in Appendix A.2.3.

The **GPT4**$^S$ strategy individually observes a single text with the instruction to rate this text on a scale from 0 to 1 in terms of simplicity. A ranking is then created based on the resulting scores (model used: *gpt-4-1106-preview*). The prompt can be found in Appendix A.2.4.

The **GPT4** approach is analogous to the GPT3.5 strategy with the use of the *gpt-4-1106-preview* model and the same prompt.

Table 3 indicates inter-rater reliabilities. Since GPT4$^R$ and GPT4$^S$ do not make pairwise decisions[7], we do not report agreement rate and Cohen's $\kappa$ for these variants. It can be seen that GPT4 has the highest agreement with the majority labels across all measures.

Compared to human raters, the GPT4 strategy performs on par with an agreement rate of 81%. This strategy is, therefore, a good candidate for making further high-quality annotations with lower effort than human annotations and low processing costs of ∼$1[8]. With a rank correlation of 0.79, this strategy even achieves a ranking with higher quality than the simulated user from Section 3.3. It should also be emphasized here how the results of GPT4 differ from GPT4$^R$ and GPT4$^S$. Since a ranking process also created the majority labels, comparing the strategies is not entirely fair. However, the comparison shows that a pairwise comparison is a good annotation strategy in this setting.

| | Da_Ch | FRE | GPT3.5 | GPT4$^R$ | GPT4$^S$ | GPT4 |
|---|---|---|---|---|---|---|
| A | - | - | .6702 | - | - | **.8085** |
| $\kappa$ | - | - | .437 | - | - | **.6293** |
| $\rho$ | .4234 | .519 | .6081 | .3759 | .7128 | **.7949** |
| $\tau$ | .2992 | .3514 | .4445 | .2562 | .5727 | **.6189** |

Table 3: Inter-Rater Reliability between automatic scoring approaches and human majority vote for Dale Chall Formula (Da_Ch), Flesch Reading Ease (FRE) and GPT variants: Agreement (A), Cohen's $\kappa$, Spearmans' $\rho$, Kendall's $\tau$.

Traditional readability measures show a relatively low rank correlation compared to the GPT4 strategy. Da_ch, in particular, shows a weak correlation with the majority vote of human annotators. However, it is striking that FRE performs better than GPT3.5. Thus, we can demonstrate that the GPT4 strategy provides more reliable simplicity scores in our setting than two established classical readability measures.

Based on these results for GPT4, we create two further data sets for simplicity scores for 300 and 3000 texts. Section A.7 in the Appendix describes the average ratings for originally simple and complex texts produced by ChatGPT for ARTS$_{94}$, ARTS$_{300}$ and ARTS$_{3000}$.

### 5.3 Qualitative Analysis

To evaluate the ranking produced by ARTS$_{94}$, we looked at the overall ranking at different levels to see if the ranking produced by the Elo algorithm produced an agreeable ranking.

We further looked into text pairings, where the human raters' disagreement was high. Table 2 shows examples of two common reasons for disagreement. We assume that in the first sentence pair, the level of simplicity is very similar, which is

[6]For LLM-based ratings we re-run the prompt if an LLM returned anythong other than *A* or *B*, indicating which text was chosen as the simpler one.

[7]In general the Elo algorithm is always used when a system or a person creates orders between pairs. This ordering induces a winner and a loser, which is necessary for applying Equation 1 and Equation 2. Accordingly, we used the algorithm for all human raters, GPT3.5 and GPT4. But not for the runs Da_Ch, FRE, GPT4$^R$, and GPT4$^S$.

[8]Approximate costs for using ChatGOT to annotate all datasets are given in A.5.

also reflected by the overall score of the two texts in the final $ARTS_{94}$ ranking. The second example represents a sentence pair, in which one sentence shows a high degree of lexical complexity while the other text is syntactically complex. We also analyzed cases in which GPT and human ratings disagreed but could not identify any relevant patterns.

### 5.4 Subjectivity of Annotators

**English Level of Annotators.** The overall inter-annotator agreement on the annotation of $ARTS_{94}$ by our human annotators was moderate (Krippendorff's $\alpha$=.4231, $n$=16). Separating annotators by their English level (ones with some sort of B level and ones with some sort of C level) and observing the inter-annotator agreement of these two groups leads to the highest agreement of annotaotors within C level ($\alpha$=.4596, $n$=6) and a considerably lower agreement of those in group B ($\alpha$=.3996, $n$=10).

**Domain Expertise.** We use the GPT4 pairwise approach and extend it to incorporate domain expertise (prompt can be found in Appendix A.2.5) on the $ARTS_{160}$ dataset. We conduct 3 per domain expertise, which align with the domains contained in the dataset: medical, legal administration, philosophy and geography. The overall inter-annotator agreement of the twelve GPT annotations is almost perfect (Krippendorff's $\alpha$=.8831, $n$=12) but among the different GPT-expert groups, the agreement is considerably higher (medical: $\alpha$=.9333, legislative administration: $\alpha$=.9459, philosophy: $\alpha$=.9146, geography: $\alpha$=.9563; $n$=3 for all groups). Section A.8 in the Appendix describes two example text pairs with synthetic domain expert annotations.

From these two experiments we conclude that our approach is suitable of capturing subjective differences between human or simluated annotaors. Therefore the approach could be used for domain- or target audience-specific dataset generation for both automatic readability and text simplification evaluation.

### 5.5 Simplicity Regression Model

We investigate applicability of ARTS by training a regression model predicting simplicity scores for texts represented with state-of-the-art embeddings. We use the OpenAI embeddings client[9] with the *text-embedding-3-small* model for embedding the texts. The maximum number of input tokens is 8191, and the length of the embedding vectors is 1536. We train a Gradient Boosting regressor (without changes in parameters) using embeddings and scores from the $ARTS_{300}$ and $ARTS_{3000}$ dataset and predict scores using the embeddings of $ARTS_{94}$. We compare predicted scores against actual majority scores from human annotators. We also compare against a random baseline in which we draw random values from [0, 1].

The regressor trained on $ARTS_{3000}$ performs better (MSE=.0608, $R^2$=.3781[10]) than the regressor trained on $ARTS_{300}$ (MSE=.0731, $R^2$=.2522) and our random baseline (MSE=.1683, $R^2$=-.7216).

### 5.6 Discussion

Due to the low difference in the mean absolute simplicity scores of source and simplified texts in current text simplification datasets (as presented in Section A.7 in the Appendix), these datasets are not adequately suitable for developing an automated approach quantifying text simplicity. Our evaluation shows promising results, indicating that models trained on ARTS datasets are very capable of quantifying simplicity. This is especially true since the size of the training datasets positively correlates with the quality of the result, and due to the scalability of our ranking approach, larger datasets can easily be developed. The approach presented in this work could also be applied to other tasks where data availability is a problem and the measurement is subjective, for example, persuasiveness, appeal, bias, funniness, or offensiveness quantification.

However, we also identified some minor disadvantages of our methodology. The annotators are all non-native speakers and despite very advanced English skills, might decide differently from native speakers in some labeling decisions. Further, all have an academic background and could be more familiar with some topics, formulations, or concepts than an average person. The selection of sentences from the text simplification datasets could also have induced bias based on the current state of text simplification datasets. Domains, tar-

[9] https://platform.openai.com/docs/guides/embeddings/embedding-models

[10] Mean squared error (MSE) penalizes larger errors between predicted and actual simplicity scores. In our case the actually predicted scores do not matter as much (is a text's simplicity 0.00342 or 0.0037) but the scores should rather indicate if a text is simple or complex and should not be vastly different. The closer the MSE is to 0, the better. $R^2$ quantifies a regressor's ability to predict the actual data as the proportion of the variance for the complexity score that can be explained by the input embeddings. The closer $R^2$ is to 1, the better.

get audiences and other aspects have not been balanced in the selection of sentences for $ARTS_{94}$ and $ARTS_{300}$. For $ARTS_{3000}$ the balancing was better and $ARTS_{160}$ contains four balanced domains. We do not expect this to impact our results significantly since the vast majority of texts are taken from general target audience and domain datasets. There is also a range of adjustments and exchangeable components in our approach that might improve the overall effectiveness. We have not evaluated the use of text triplet comparison instead of pairs, different parametric approaches to pairwise comparison, or other ways to sample the texts we rank.

Despite the headroom for optimization, we deem our results to suffice as evidence for the applicability of ARTS as a foundation for a shift from a relative simplification-centric view toward simplicity quantification.

## 6 Conclusion

In this publication, we present ARTS, a method for constructing datasets with continuous simplicity scores that can be used to develop text simplicity evaluation approaches. Additionally, we provide four different-sized datasets with texts spanning multiple domains and target audiences as well as scores derived from human- and LLM-based annotations. Using a pairwise comparison approach and an Elo ranking system, we facilitate simplicity assessment. Our approach is language-independent and is able to accommodate readers' subjective characteristics. We found that using ARTS, automated LLM-based labeling produces similar results to human labels, thus allowing for an effective and cost-efficient way of constructing large synthetic datasets.

Future work will focus on the construction of larger simplicity datasets. We further plan to use the datasets in developing simplicity quantifying models, that are based on ARTS. As another line of work we want to extend the application of the methodology to the quantification of other subjective phenomena, e.g., offensiveness, bias, or appeal.

## 7 Limitations

Even though we have carried out an extensive process of textual preprocessing, we include texts whose data quality is heterogeneous. A comparatively low diversity for our human annotators must be noted: they are all based in the same regional area. They differ in age, English language level and highest obtained degree.

Due to the high annotation effort, the number of labels is also limited. However, we have tried to estimate the quality of our labels using a simulation.

Another limitation is that the distribution of our simplicity scores is unclear, as the underlying distribution is lost through the ranking. Furthermore, the score produced by our method is always relative, so it depends on which texts are used for comparison. We have tried to minimize distortions due to relativity by using a wide variety of comparison texts (cf. Table 1). Our method's shortcoming is its lack of interpretability. However, since our score's central use is to evaluate simplification models, interpretability is of secondary significance in contrast to systems that produce outputs with which humans interact directly. Nevertheless, integrating this point is an important aspect of future work, for example, to enable comprehensible model debugging.

As base for our ARTS approach we used texts obtained from parallel corpora for evaluation of text simplification which has two classes: source texts and simplified texts. This was done as one of the goals of our work was the construction of a dataset which can be used in the evaluation of text simplicity. Asides from OneStopEnglish we did not use typical datasets from ARA in which text is grouped into several categories. We also did not experiment solely on such ARA-specific datasets.

This work does not explore the distribution of scores produced by ChatGPT with the individual scoring prompt. While we could have created a typical ARA dataset with discrete readability levels we instead focused on composing datasets where the scores of texts are dependent on the ARTS corpus they are part of.

We did not conduct experiments with $ARTS_{160}$ and human experts from the four contained domains as annotators. Instead we solely used multiple runs where we prompted ChatGPT to annotate while assuming to have domain expertise in one of the domains.

## References

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. *Preprint*, arXiv:2005.00481.

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

Muddassira Arshad, Muhammad Murtaza Yousaf, and Syed Mansoor Sarwar. 2023. Comprehensive readability assessment of scientific learning resources. *IEEE Access*, 11:53978–53994.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *EMNLP 2003*.

Jan A. Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. 2018. Learning to split and rephrase from Wikipedia edit history. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 732–737, Brussels, Belgium. Association for Computational Linguistics.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo uncovered: Robustness and best practices in language model evaluation. *Preprint*, arXiv:2311.17295.

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699, Brussels, Belgium. Association for Computational Linguistics.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. In *ITL - International Journal of Applied Linguistics*, volume 165, pages 97–135. ISSN: 0019-0829, 1783-1490 Issue: 2 Journal Abbreviation: ITL.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

Björn Engelmann, Fabian Haak, Christin Katharina Kreutz, Narjes Nikzad-Khasmakhi, and Philipp Schaer. 2023. Text simplification of scientific texts for non-expert readers. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2987–2998. CEUR-WS.org.

Björn Engelmann, Christin Katharina Kreutz, Fabian Haak, and Philipp Schaer. 2024. Arts datasets - arts94, arts300, arts3000, arts160.

Liana Ermakova, Eric SanJuan, Stéphane Huet, Olivier Augereau, Hosein Azarbonyad, and Jaap Kamps. 2023. CLEF 2023 simpletext track - what happens if general users search scientific texts? In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III*, volume 13982 of *Lecture Notes in Computer Science*, pages 536–545. Springer.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233. Place: US Publisher: American Psychological Association.

I. J. Good. 1955. On the marking of chess-players. *The Mathematical Gazette*, 39(330):292–296.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of Automatic Text Simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 453–463, Avignon, France. ATALA.

Fabian Haak, Björn Engelmann, Christin Katharina Kreutz, and Philipp Schaer. 2024. Investigating bias in political search query suggestions by relative comparison with llms. In *Companion Publication of the 16th ACM Web Science Conference, WebSci Companion 2024, Stuttgart, Germany, May 21-24, 2024*, pages 5–7. ACM.

G. Harry and Mc Laughlin. 1969. SMOG Grading - A New Readability Formula. *The Journal of Reading*.

Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2022. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a lexical simplifier using wikipedia. volume 2, pages 458–463.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Michael Ibañez, Lloyd Lois Antonie Reyes, Ranz Sapinit, Mohammed Ahmed Hussien, and Joseph Marvin Imperial. 2022. On Applicability of Neural Language Models for Readability Assessment in Filipino. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium*, pages 573–576, Cham. Springer International Publishing.

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 and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *COLING 2016*, pages 1147–1158. ACL.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Annual Meeting of the Association for Computational Linguistics*.

David Kauchak, Jorge Apricio, and Gondy Leroy. 2022. Improving the quality of suggestions for medical text simplification tools. *AMIA Jt Summits Transl Sci Proc*, 2022:284–292.

J. Kincaid, Robert Fishburne, Richard Rogers, and Brad Chissom. 1975. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. *Institute for Simulation and Training*.

Christin Kreutz, Fabian Haak, Björn Engelmann, and Philipp Schaer. 2024. BATS: BenchmArking text simplicity . In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11968–11989, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Wenbiao Li, Wang Ziyang, and Yunfang Wu. 2022. A Unified Neural Network Model for Readability Assessment with Feature Projection and Length-Balanced Loss. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7446–7457, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhenzhen Li, Han Ding, and Shaohong Zhang. 2023. Cross-Corpus Readability Compatibility Assessment for English Texts. *IEEE Access*, 11:101985–101997.

Yuliang Liu, Zhiwei Jiang, Yafeng Yin, Cong Wang, Sheng Chen, Zhaoling Chen, and Qing Gu. 2023. Unsupervised Readability Assessment via Learning from Weak Readability Signals. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 1324–1334, New York, NY, USA. Association for Computing Machinery.

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.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

Salar Mohtaj, Babak Naderi, Sebastian Möller, Faraz Maschhur, Chuyang Wu, and Max Reinhard. 2022. A Transfer Learning Based Model for Text Readability Assessment in German. *arXiv preprint*. ArXiv:2207.06265 [cs].

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.

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

Gustavo Paetzold and Lucia Specia. 2016a. Benchmarking lexical simplification systems. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3074–3080, Portorož, Slovenia. European Language Resources Association (ELRA).

Gustavo H. Paetzold and Lucia Specia. 2016b. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 3761–3767. AAAI Press.

The pandas development team. 2020. pandas-dev/pandas: Pandas.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

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).

Elliot Schumacher, Maxine Eskenazi, Gwen Frishkoff, and Kevyn Collins-Thompson. 2016. Predicting the relative difficulty of single sentences with and without surrounding context. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1871–1881, Austin, Texas. Association for Computational Linguistics.

Max Schwarzer, Teerapaun Tanprasert, and David Kauchak. 2021. Improving human text simplification with sentence fusion. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 106–114, Mexico City, Mexico. Association for Computational Linguistics.

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

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

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

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. 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 *EMNLP 2021*, pages 7997–8013. Association for Computational Linguistics.

Sowmya Vajjala. 2022. Trends, Limitations and Open Challenges in Automatic Readability Assessment Research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2014. Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL - International Journal of Applied Linguistics*, 165(2):194–222. Publisher: John Benjamins.

Hoang Van, David Kauchak, and Gondy Leroy. 2020. AutoMeTS: The autocomplete for medical text simplification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1424–1434, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

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, San Diego, CA. Association for Computational Linguistics.

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

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

Farooq Zaman, Matthew Shardlow, Saeed-Ul Hassan, Naif Aljohani, and Raheel Nawaz. 2020. Htss: A novel hybrid text summarisation and simplification architecture. *Information Processing & Management*, 57:102351.

Jinshan Zeng, Xianchao Tong, Xianglong Yu, Wenyan Xiao, and Qing Huang. 2024. InterpretARA: Enhancing Hybrid Automatic Readability Assessment with Linguistic Feature Interpreter and Contrastive Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19497–19505. Number: 17.

Jinshan Zeng, Yudong Xie, Xianglong Yu, John Lee, and Ding-Xuan Zhou. 2022. Enhancing Automatic Readability Assessment with Pre-training and Soft Labels for Ordinal Regression. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4557–4568, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR 2020*. OpenReview.net.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *COLING 2010*, pages 1353–1361. Tsinghua University Press.

# A Appendix

## A.1 Datasets

### A.1.1 Text Simplification Measures

Table 4 gives insights into the datasets containing human labels, which are currently used developing text simplification evaluation measures. Table 5 holds information on the most common (Arshad et al., 2023; Martinc et al., 2021) datasets used in assessing readability.

### A.1.2 Text Simplification

Table 6 contains information on the used 26 publicly available English parallel datasets for text simplicity evaluation which we used parts of to run our ARTS approach.

Table 7 contains information on the used numbers and lengths of texts from the different parallel datasets as part of the four ARTS datasets.

## A.2 Prompts

The following sections describe the prompts used throughout this work. All prompts are given in typewriter font while additional descriptions or explanations are given without special formatting.

### A.2.1 Formatting Prompt

`Check the following` $x$ `texts given as a Python list. Please correct the capitalization, spacing, and character errors:`

This prompt was then followed by a list of texts; $x$ was substituted by the number of texts.

### A.2.2 Pairwise Prompt

```
I'm going to present you with two texts and I want you to decide which one is simpler. The following guidelines should be taken into account for the decision: Imagine you are writing an exam where you are allowed to google and where the task is to understand the two given texts. Which of the two texts: generates less cognitive load?, can you understand more quickly?, are you more confident to answer questions about?, is easier for you to reformulate without changing the meaning? Both Texts are delimited by "'

Text A:
"'
TEXT_A
"'

Text B:
"'
TEXT_B
"'
The answer should be either A or B, depending on which of the texts is easier to understand. Please answer without any further text, just one letter.
```

### A.2.3 Ranking Prompt

```
I will present you with a numbered list of texts. It is important that each text has a corresponding id. I would like you to give me back a sorted list of these ids. The criterion for the sorting should be the simplicity of the texts. Please use the following guidelines to evaluate the simplicity of the texts: Imagine you
```

| Dataset | Origin of source | \|P\| | Description | $R_p$ | $R_t$ |
|---|---|---|---|---|---|
| ASSET (Alva-Manchego et al., 2020) | TurkCorpus (Xu et al., 2015, 2016) | 100 | 1 system's simplification, 100 source texts | 15 | 1500 |
| metaeval (Alva-Manchego et al., 2021) | TurkCorpus (Xu et al., 2015, 2016) | 600 | 6 systems' simplifications of (not the same) 100 source texts | 15 | 9000 |
| SimpEval$_{Past}$ (Maddela et al., 2023) | TurkCorpus (Xu et al., 2015, 2016) | 2400 | 24 systems' simplifications, 100 source texts | 5 | 12,000 |
| SimpEval$_{2022}$ (Maddela et al., 2023) | Wikipedia | 360 | 6 systems' simplifications, 60 source texts | 3 | 1080 |
| QuestEval (Scialom et al., 2021) | metaeval$_{Likert}$ (Alva-Manchego et al., 2021) | 100 | 100 unique simplifications | 30 | 3000 |
| simplification-acl (Sulem et al., 2018b) | TurkCorpus (Xu et al., 2015, 2016) | 1750 | 25 simplifications, 70 source texts | 3 | 5250 |
| ASSET$_{preference}$ (Alva-Manchego et al., 2020) | ASSET (Alva-Manchego et al., 2020), TurkCorpus (Xu et al., 2015, 2016), HSplit (Sulem et al., 2018a) | 718 | 2 simplifications each, 359 source texts | 1 | 718 |

Table 4: Datasets containing human labels for assessing text simplification evaluating measures; number of pairs (|P|), numbers of ratings on simplicity per paper ($P_p$) and in total ($P_t$).

| Dataset | \|C\| | \|T\| | Description |
|---|---|---|---|
| Cambridge Exams (Xia et al., 2016) | 5 | 3125 | Corpus for English learners' language assessment |
| CLEAR (Heintz et al., 2022) | 3 | 4785 | English Language Arts education corpus |
| Newsela (Xu et al., 2015)* | 5 | 10,786 | 1911 English news articles with up to 4 manual re-writes |
| OneStopEnglish (Vajjala and Lučić, 2018) | 3 | 567 | Parallel corpus based on English language resource site for teachers |
| WeeBit (Vajjala and Meurers, 2012)* | 5 | 6388 | Contains newspaper articles (WeeklyReader) and educational resources (BBC-Bitsize) for children and teenagers |

Table 5: Datasets containing human labels for assessing readability; number of classes (|C|), number of texts (|T|). Datasets marked with * have restricted public access.

| Name | License | # P | \|src\| | \|simp\| |
|---|---|---|---|---|
| ASSET (Alva-Manchego et al., 2020) | CC BY-NC | 23590 | 116.77 | 98.52 |
| AutoMeTS (Van et al., 2020) | MIT | 4280 | 205.14 | 154.07 |
| BenchLS (Paetzold and Specia, 2016a) | CC BY-SA 4 | 6846 | 152.76 | 150.08 |
| Britannica (Barzilay and Elhadad, 2003) | - | 600 | 88.15 | 147.35 |
| D-Wikipedia (Sun et al., 2021) | GPL 3.0 | 143,546 | 781.38 | 414.01 |
| EW-SEW-gmpm (Nisioi et al., 2017) | MIT | 301,036 | 137.2 | 102.37 |
| EW-SEW-Turk (Horn et al., 2014) | - | 7330 | 146.97 | 147.29 |
| HTSS (Zaman et al., 2020) | - | 5205 | 36,443.31 | 4184.09 |
| HutSSF (Schwarzer et al., 2021) | - | 5245 | 161.89 | 136.43 |
| MASSAlign (Paetzold et al., 2017) | LGPL 3.0 | 8252 | 417.38 | 398.93 |
| METAeval (Alva-Manchego et al., 2021) | CC BY-NC-SA 4.0 | 600 | 123.85 | 99.4 |
| MTurkSF (Kauchak et al., 2022) | - | 221 | 168.26 | 171.24 |
| NNSeval (Paetzold and Specia, 2016b) | CC BY-SA 4.0 | 1791 | 145.27 | 143.54 |
| OneStopEnglish (Vajjala and Lučić, 2018) | CC BY-SA 4.0 | 6321 | 341.11 | 285.96 |
| PWKP (Zhu et al., 2010) | CC BY 4.0 | 108,016 | 128.67 | 103.63 |
| QuestEval (Scialom et al., 2021) | - | 366 | 116.55 | 93.08 |
| SemEval-2007 (McCarthy and Navigli, 2007) | - | 1208 | 150.6 | 153.42 |
| SimPA (Scarton et al., 2018) | - | 6600 | 165.76 | 160.5 |
| SimpEval (Maddela et al., 2023) | - | 2570 | 155.85 | 135.17 |
| SSCORPUS (Kajiwara and Komachi, 2016) | - | 492,993 | 130.27 | 89.41 |
| TurkCorpus (Xu et al., 2015, 2016) | GPL 3.0 | 21,231 | 118 | 110.98 |
| Wiki-Auto (Jiang et al., 2020) | - | 488,332 | 136.86 | 93.47 |
| Wiki-Manual (Jiang et al., 2020) | - | 1054 | 106.83 | 97.23 |
| Wikipedia_v1 (Coster and Kauchak, 2011) | - | 137,362 | 127.55 | 110.83 |
| Wikipedia_v2 (Kauchak, 2013) | - | 227,432 | 2438.79 | 298.13 |
| WikiSplit (Botha et al., 2018) | CC BY-SA 4.0 | 1,004,944 | 175 | 193.19 |

Table 6: Names, references, licenses, number of pairs (# P) and average length in characters of source texts (|src|) and simplifications (|simp|) contained in datasets.

are writing an exam where you are allowed to google and where the task the given texts. Which of the texts: generates less cognitive load?, can you understand more

| | $\text{ARTS}_{94}$ | | | | $\text{ARTS}_{160}$ | | | | $\text{ARTS}_{300}$ | | | | $\text{ARTS}_{3000}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | simplified | | source | | simplified | | source | | simplified | | source | | simplified | | source | |
| **Name** | # | $\lvert len \rvert$ | # | $\lvert len \rvert$ | # | $\lvert len \rvert$ | # | $\lvert len \rvert$ | # | $\lvert len \rvert$ | # | $\lvert len \rvert$ | # | $\lvert len \rvert$ | # | $\lvert len \rvert$ |
| ASSET | 2 | 98.5 | 2 | 88 | 0 | 0 | 0 | 0 | 10 | 79.9 | 10 | 151.5 | 58 | 97.98 | 58 | 118.28 |
| AutoMeTS | 2 | 97 | 2 | 155 | 20 | 115.6 | 20 | 171.55 | 9 | 126.11 | 10 | 171.5 | 58 | 136.93 | 58 | 199.69 |
| BenchLS | 2 | 114 | 2 | 162 | 0 | 0 | 0 | 0 | 8 | 127.12 | 8 | 114.75 | 58 | 137.88 | 58 | 154.09 |
| Britannica | 2 | 169 | 2 | 63.5 | 20 | 106.95 | 20 | 77.75 | 0 | 0 | 0 | 0 | 58 | 147.78 | 58 | 89.76 |
| D-Wikipedia | 0 | 0 | 1 | 116 | 0 | 0 | 0 | 0 | 7 | 252.71 | 0 | 0 | 59 | 174.41 | 58 | 197.9 |
| EW-SEW-gmpm | 2 | 134.5 | 2 | 88 | 0 | 0 | 0 | 0 | 10 | 101.1 | 10 | 142.4 | 58 | 92.22 | 58 | 131.81 |
| EW-SEW-Turk | 2 | 154 | 2 | 139 | 0 | 0 | 0 | 0 | 10 | 141.8 | 10 | 143.1 | 58 | 153.52 | 58 | 139.19 |
| HTSS | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 251.75 | 0 | 0 |
| HutSSF | 2 | 89.5 | 2 | 141 | 0 | 0 | 0 | 0 | 9 | 152.67 | 5 | 133.8 | 58 | 144.09 | 58 | 160.03 |
| Massalign | 1 | 306 | 2 | 147.5 | 20 | 145.85 | 20 | 140.35 | 8 | 191.5 | 7 | 190.57 | 58 | 183.98 | 58 | 186.98 |
| Metaeval | 2 | 61 | 2 | 70 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 58 | 99.29 | 58 | 124.5 |
| MTurkSF | 2 | 153.5 | 2 | 132 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 58 | 142.72 | 58 | 137.31 |
| NNSeval | 2 | 154 | 2 | 167 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 58 | 142.66 | 58 | 144.79 |
| OneStopEnglish | 2 | 195 | 1 | 165 | 0 | 0 | 0 | 0 | 7 | 203 | 2 | 318.5 | 58 | 238.34 | 58 | 266.86 |
| PWKP | 2 | 105 | 2 | 177.5 | 0 | 0 | 0 | 0 | 10 | 124.6 | 10 | 118.7 | 58 | 101.22 | 58 | 139.78 |
| QuestEval | 2 | 84.5 | 2 | 179.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 58 | 96.45 | 58 | 114.55 |
| SemEval_2007 | 2 | 178.5 | 2 | 169 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 58 | 146.81 | 58 | 149.36 |
| SIMPA | 2 | 132 | 2 | 146 | 20 | 139.35 | 20 | 142.65 | 10 | 172.4 | 10 | 168 | 58 | 156.95 | 58 | 163.02 |
| SimpEval | 2 | 120.5 | 2 | 136.5 | 0 | 0 | 0 | 0 | 4 | 171.5 | 0 | 0 | 58 | 128.16 | 58 | 162.62 |
| SSCORPUS | 2 | 81.5 | 2 | 161.5 | 0 | 0 | 0 | 0 | 9 | 85.22 | 10 | 153.7 | 58 | 85.21 | 58 | 124.62 |
| TurkCorpus | 2 | 112.5 | 2 | 99.5 | 0 | 0 | 0 | 0 | 10 | 110.8 | 10 | 98 | 58 | 123.48 | 58 | 113.24 |
| Wiki-Auto | 2 | 78.5 | 2 | 152.5 | 0 | 0 | 0 | 0 | 10 | 115.6 | 9 | 157 | 58 | 96.5 | 58 | 125.24 |
| Wiki-Manual | 2 | 95 | 2 | 123.5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 112 | 58 | 107.28 | 58 | 111.59 |
| Wikipedia_v1 | 2 | 102.5 | 2 | 110.5 | 0 | 0 | 0 | 0 | 10 | 123.7 | 10 | 113.4 | 58 | 94.24 | 58 | 123.6 |
| Wikipedia_v2 | 1 | 209 | 2 | 100 | 0 | 0 | 0 | 0 | 8 | 136.12 | 9 | 102.11 | 60 | 123.55 | 58 | 138.4 |
| WikiSplit | 2 | 228 | 2 | 209 | 0 | 0 | 0 | 0 | 10 | 187.6 | 10 | 166.8 | 101 | 198.83 | 108 | 174.19 |

Table 7: Numbers (#) of texts and average lengths in characters of texts ($|len|$) used from the simplified and source parts of parallel corpora for the four ARTS datasets.

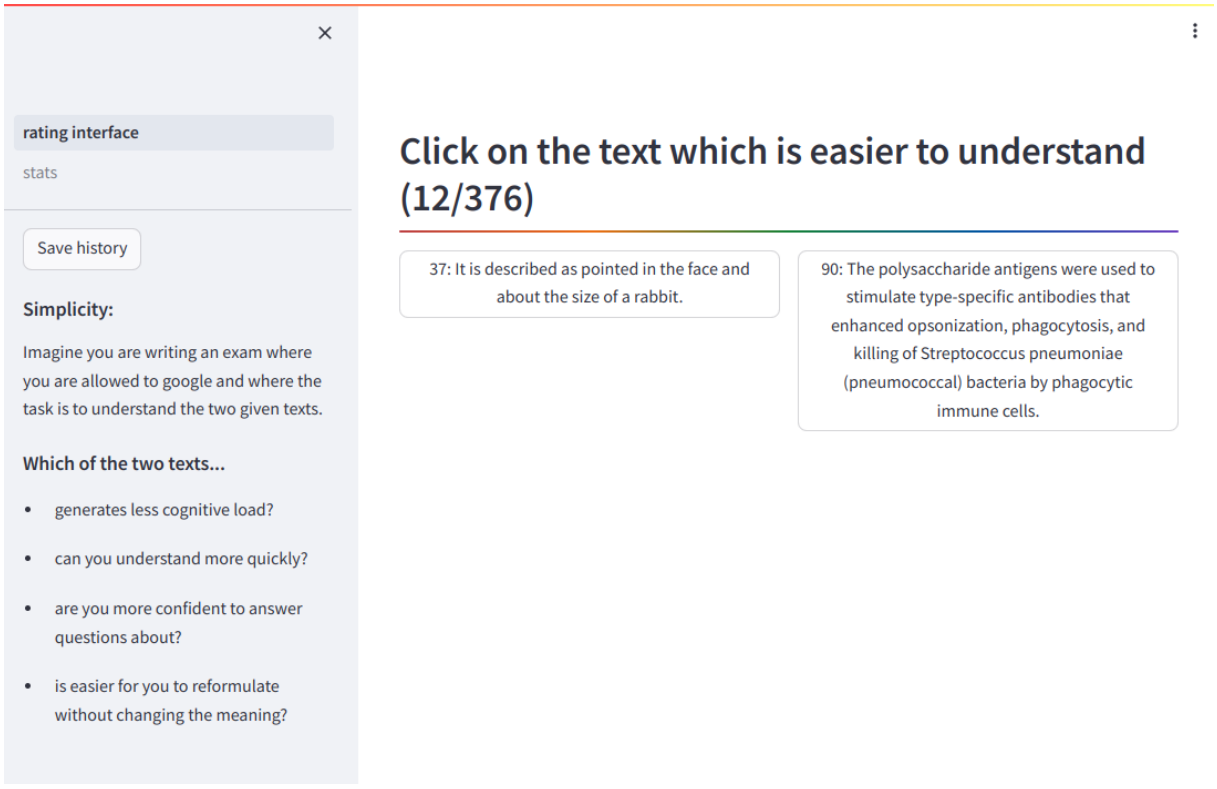


Figure 3: Rating interface for human annotators.

quickly?, are you more confident to answer questions about?, is easier for you to reformulate without changing the meaning? The texts will be delimited by “TEXT“ The answer should only be a list of ids sorted by simplicity. Please answer without any further text.

### A.2.4 Individual Score Prompt

I will present you a text, and I want you to score its simplicity between 0 and 1. A text that is very easy to

understand should receive a score of 0.0; a very challenging text should have a score of approximately 1.0. The following guidelines should be taken into account for the decision: Imagine you are writing an exam where you are allowed to google and where the task is to understand the given text. Keep in mind the following criteria for the simplicity of a text: Generation of cognitive load, time to understand, confidence to answer questions about, difficulty to reformulate without changing the meaning. The text to score is delimited by “‘
Text:
“‘
TEXT
“‘
The answer should only contain a number between 0 and 1. Please answer without any further text, just one number with high precision.

#### A.2.5 Pairwise Prompt with Domain

**You are an adult, a native speaker of English, and an average reader of US newspapers and an expert on the topic "TOPIC".** I'm going to present you with two texts and I want you to decide which one is simpler. The following guidelines **as well as your domain knowledge on the texts' topics** should be taken into account for the decision: Imagine you are writing an exam where you are allowed to google and where the task is to understand the two given texts. Which of the two texts: generates less cognitive load?, can you understand more quickly?, are you more confident to answer questions about?, is easier for you to reformulate without changing the meaning? Both Texts are delimited by “‘

Text A:
“‘
TEXT_A
“‘

Text B:
“‘
TEXT_B
“‘
The answer should be either A or B, depending on which of the texts is easier to understand. Please answer without any further text, just one letter.

The bold parts of this prompt highlight its difference from the pairwise prompt Section A.2.2. As TOPIC, one of the following four was inserted: medical, legislative administration, philosophy, geography.

### A.3 Rating Interface

Figure 3 depicts the rating interface. The Text shown in the sidebar of the rating interface giving the instructions for annotators is formulated as:
*Simplicity: Imagine you are writing an exam where you are allowed to google and where the task is to understand the two given texts.*
*Which of the two texts...*

- *generates less cognitive load?*
- *can you understand more quickly?*
- *are you more confident to answer questions about?*
- *is easier for you to reformulate without changing the meaning?*

### A.4 Used Python Packages

- matplotlob (Hunter, 2007)
- openai (https://platform.openai.com/docs/api-reference?lang=python)
- pandas (Wes McKinney, 2010; pandas development team, 2020)
- scipy (Virtanen et al., 2020)
- seaborn (Waskom, 2021)
- sklearn (Pedregosa et al., 2011)
- streamlit (https://streamlit.io)

### A.5 Costs of Annotating Datasets with LLMs

Running the experiments using ChatGPT which are part of our evaluation would lead to costs of $\sim$ \$61 (see Table 8). The costs for running the prompts to clean the $\text{ARTS}_{160}$, $\text{ARTS}_{300}$ and $\text{ARTS}_{3000}$ datasets are negligible.

### A.6 Human Interrater Agreements

Table 9 contains the inter-rater agreements of single raters with the majority vote decisions on $\text{ARTS}_{94}$.

| | $ per run | \|runs\| | total $ |
|---|---|---|---|
| $ARTS_{94}$ | 1 | 4 | 4 |
| $ARTS_{160}$ | 2 | 12 | 24 |
| $ARTS_{300}$ | 3 | 1 | 3 |
| $ARTS_{3000}$ | 30 | 1 | 30 |
| total cost | | | 61 |

Table 8: Compilation of approximate costs our running the prompts with ChatGPT to compose the annotations for the four ARTS datasets with the cost per run in dollar, the number of runs (|runs|) and the total costs.

| IRR | A | $\kappa$ | $\rho$ | $\tau$ |
|---|---|---|---|---|
| $R1_{94}$ | .8963 | .7928 | .8786 | .7232 |
| $R2_{94}$ | .8138 | .6279 | .7665 | .5804 |
| $R3_{94}$ | .8271 | .6552 | .8051 | .6257 |
| $R4_{94}$ | .8271 | .6552 | .8014 | .6211 |
| $R5_{94}$ | .7899 | .5803 | .7398 | .5443 |
| $R6_{94}$ | .8245 | .6492 | .7775 | .5841 |
| $R7_{94}$ | .7739 | .5491 | .7567 | .5548 |
| $R8_{94}$ | .7606 | .5236 | .7062 | .5145 |
| $R9_{94}$ | .8085 | .6173 | .7924 | .5923 |
| $R10_{94}$ | .8378 | .6759 | .8334 | .6486 |
| $R11_{94}$ | .7899 | .5805 | .7278 | .5415 |
| $R12_{94}$ | .7793 | .5599 | .6943 | .5397 |
| $R13_{94}$ | .8005 | .6013 | .7352 | .5479 |
| $R14_{94}$ | .7793 | .5588 | .7129 | .5255 |
| $R15_{94}$ | .6862 | .3736 | .5198 | .3731 |
| $R16_{94}$ | .8191 | .6384 | .8231 | .6477 |

Table 9: Inter-Rater Reliability (IRR) between human rater $R_i$ and majority vote: Agreement (A), Cohen's $\kappa$, Spearmans' $\rho$, Kendall's $\tau$.

### A.7 Statistics

In $ARTS_{94}$ average ratings composed by ChatGPT for texts from sources are 0.5448 while the average rating for simplified texts is 0.4532. For ratings composed by human raters, source texts' simplicity is on average 0.5289 while simplifications have an average of 0.4698. In $ARTS_{300}$ average ratings for texts from sources are 0.5447 while the average rating for simplified texts is 0.4603. In $ARTS_{3000}$ average ratings for texts from sources are 0.5241 while the average rating for simplified texts is 0.4759.

We found no significant differences between the two groups of ratings using Mann-Whitney U tests[11] for both $ARTS_{94}$ settings, for $ARTS_{300}$ (p=.0119) and $ARTS_{3000}$ (p=5.06e-06) we found significant differences.

[11] Shapiro-Wilk's tests did not find normally distributed data in all cases.

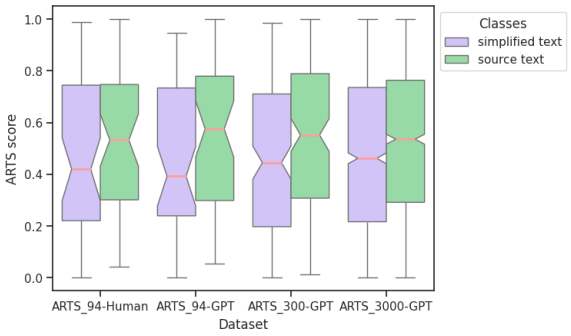


Figure 4: Distribution of ARTS scores for the different datasets in the two classes of source and simplified texts.

Figure 4 shows boxplots of scores, the origin of source and simplified texts stems from the respective datasets the texts are extracted from. The even lower difference of mean scores of source and simplified texts in the larger ARTS datasets hints at the sampling procedure not causing any of the observed low differences between the mean scores of the two groups.

### A.8 Domain Expertise

Table 10 contains annotations for pairs of texts A and B for the different groups of synthetic domain experts using $ARTS_{160}$ and the prompt given in Section A.2.5.

In example 1 texts from the legislative administration and geography are compared. The synthetic annotators disagree on the simpler text out of the two options. Five out of six synthetic annotators consider the text which stems from their own domain of expertise as the easier text.

In example 2 the texts stem from the legislative administration and philosophy. Except for synthetic annotators with expertise in philosophy, the text from philosophy is considered the easier one. Contrasting this, all annotators from philosophy agree on the legislative adminstration text being the one with higher simplicity. The synthetic philosophy experts might consider the text from their domain more complex as they know about the real complexity of the text's topic whereas the other annotators might not recognize the potential underlying assumtions behind the seemingly less difficult words.

### A.9 ARTS Examples

Table 11 gives examples of texts contained in $ARTS_{94}$.

| Example | Text A | Text B | G | L | M | P |
|---|---|---|---|---|---|---|
| 1 | *(L)* It has both residential and commercial/retail areas with a lot of pedestrians and cyclists, who should feel safer from the risk of vehicle collisions. | *(G)* After World War I (1914-18), the Hapsburg Empire came to an end. | 0 | 2 | 1 | 3 |
| 2 | *(P)* I answer that the only thing an idea can resemble is another idea; a color or shape can't be like anything but another color or shape. | *(L)* The main difference in the procedure is that Building Control does not formally assess for approval the information supplied under the Building Notice route. | 3 | 3 | 3 | 0 |

Table 10: Pairs of texts from $\text{ARTS}_{160}$ with an indication of the domain of the dataset they stem from (in italics, not part of the original text) as well as the number of synthetic domain experts who believed text A was simpler than text B; geography (G), legislative administration (L), medical (M) and philosophy (P)

| Text | Part | $\text{ARTS}_{94}$ score | FRE |
|---|---|---|---|
| She was born in Detroit, Michigan. | simplified | .0215 | 90.77 |
| Ohio state's library system includes twenty-one libraries located in the city of Columbus. | simplified | .0538 | 49.82 |
| Dauenhauer died at her home in Juneau, Alaska on September 25, 2017 at the age of 90. | simplified | .2473 | 88.06 |
| "Typically, the biggest difference between film and stage musicals is the use of lavish background scenery which would be impractical in a theater." | source | .5269 | 39.67 |
| Executive power is to be exercised by the Governor-General, advised by the Federal Executive Council. | source | .5484 | 5.49 |
| Its status as an official geological period was ratified in 2004 by the international union of geological sciences, making it the first new geological period declared in 120 years. | source | .7312 | 8.2 |
| In some mollusks the mantle cavity is a brood chamber, and in cephalopods and some bivalves such as scallops, it is a locomotory organ. | simplified | .8925 | 55.58 |

Table 11: Example texts, part of the parallel corpus the text originates from, simplicity scores, and Flesch Reading Ease (FRE) from $\text{ARTS}_{94}$ dataset. The higher the score, the more complex the text. As shown by the comparison to the Flesch-Kincaid reading ease score (the lower, the harder to read), the $\text{ARTS}_{94}$ complexity score better reflects the overall text complexity.