# Do Text Simplification Systems Preserve Meaning? A Human Evaluation via Reading Comprehension

**Sweta Agrawal***
Instituto de Telecomunicações
swetaagrawal20@gmail.com

**Marine Carpuat**
University of Maryland, USA
marine@umd.edu

## Abstract

Automatic text simplification (TS) aims to automate the process of rewriting text to make it easier for people to read. A pre-requisite for TS to be useful is that it should convey information that is consistent with the meaning of the original text. However, current TS evaluation protocols assess system outputs for simplicity and meaning preservation without regard for the document context in which output sentences occur and for how people understand them. In this work, we introduce a human evaluation framework to assess whether simplified texts preserve meaning using reading comprehension questions. With this framework, we conduct a thorough human evaluation of texts by humans and by nine automatic systems. Supervised systems that leverage pre-training knowledge achieve the highest scores on the reading comprehension tasks among the automatic controllable TS systems. However, even the best-performing supervised system struggles with at least 14% of the questions, marking them as ''unanswerable'' based on simplified content. We further investigate how existing TS evaluation metrics and automatic question-answering systems approximate the human judgments we obtained.

## 1 Introduction

Rewriting text so that it is easier to understand has the potential to help a wide range of audiences including non-native speakers (Petersen and Ostendorf, 2007; Allen, 2009; Crossley et al., 2014), children (Watanabe et al., 2009), or people with reading or cognitive disabilities (Alonzo et al., 2020) access information more easily (Chandrasekar et al., 1996; Stajner, 2021). Online resources such as Newsela Inc (2023) and OneStopEnglish (Macmillian Education, 2023) or the Cochrane systematic reviews (Cochrane Collaboration, 2023), provide text articles simplified by human editors so that they are easier to understand by K-12 students, English speakers with limited proficiency, and lay people seeking to understand medical literature, respectively. This has motivated a wealth of Natural Language Processing research on text simplification, framed as the task of rewriting an input text into a simpler version while preserving the core meaning of the original (Chandrasekar and Srinivas, 1997), which has been addressed with approaches ranging from dedicated supervised systems (Specia, 2010; Zhang and Lapata, 2017; Scarton and Specia, 2018; Martin et al., 2020; Jiang et al., 2020; Devaraj et al., 2021; Sheang and Saggion, 2021; Agrawal and Carpuat, 2022; Martin et al., 2022) to prompting black-box pre-trained models (Feng et al., 2023; Kew et al., 2023).

However, texts that are easier to read are not helpful if they mislead readers by providing information that is not consistent with the original document. This can happen with automatic text simplification (TS) outputs where deletions or inaccurate rewrites can change how a text is understood (Devaraj et al., 2022). Assessing to what extent the meaning of the original text is preserved should therefore be a critical dimension of TS evaluation (Stajner, 2021), and a pre-requisite to determining whether and how TS can be used in practice. Additionally, evaluating individual sentences out of context may not be sufficient to establish whether model-generated texts preserve meaning, as human simplifications often occur at the document or the passage level (Devaraj et al., 2022). Yet, TS outputs are primarily evaluated intrinsically, with automatic metrics that compare system outputs with human-written reference simplifications and/or the original source (Papineni et al., 2002; Xu et al., 2016; Maddela et al., 2023), or with generic human assessments of simplicity and meaning preservation of individual sentences outside of a context of use (Schwarzer and Kauchak, 2018). While these evaluations can

*Work done while at the University of Maryland.



Transactions of the Association for Computational Linguistics, vol. 12, pp. 432–448, 2024. https://doi.org/10.1162/tacl_a_00653
Action Editor: Ehud Reiter. Submission batch: 7/2023; Revision batch: 11/2023; Published 4/2024.

guide model development, they do not address whether readers get information from the simplified text that is consistent with the original content.

In this work, we conduct a human evaluation of the ability of state-of-the-art TS systems to preserve the meaning of the original text by measuring how well people can answer questions about key facts from the original text after reading a simplified version. We design reading comprehension (RC) tasks to directly assess meaning preservation in TS, different from prior uses of reading comprehension to assess people's reading efficiency (Angrosh et al., 2014; Laban et al., 2021). This framework lets us conduct a controlled comparison of simplified texts, whether written by humans or by TS systems: We compare people's ability to answer questions about the original text, a simplified version written by humans, and nine TS-generated versions that represent a diverse set of supervised and unsupervised approaches from the recent TS literature.[1]

We first discuss relevant literature for TS evaluation and the use of RC exercises to assess simplified or other model-generated texts in Section 2. Next, Section 3 elaborates on our RC-centered human evaluation framework, and Section 4 delves into the various design choices we made. Section 5 demonstrates the robustness of our evaluation and presents the main results. As we will see, supervised systems that utilize pre-training knowledge achieve the highest level of accuracy in RC tasks compared to other automatic controllable TS systems (§5.2). However, at least 14% of the questions remain unanswerable even for the best-performing system due to the errors introduced by these systems (§5.3). In Section 6, we shift our focus towards a meta-evaluation of existing automatic TS evaluation metrics which indicates that the 3-way comparison used in SARI makes it a reliable metric for system-level evaluation at the paragraph level. Finally, we include a preliminary discussion and analysis of the potential for automating the RC-based evaluation through the application of model-based question-answering techniques in Section 7.

[1]Collected annotations and code are released at https://github.com/sweta20/ATS-EVAL.git.

## 2 Background

How to design human and automatic evaluation protocols for TS is a research question unto itself. While automatic metrics are key to system development, commonly used metrics like BLEU (Papineni et al., 2002), SARI (Xu et al., 2016), or the Flesch-Kincaid Grade Level (Flesch, 2007) have low correlation with human judgments of simplicity (Sulem et al., 2018; Alva-Manchego et al., 2021; Maddela et al., 2023; Tanprasert and Kauchak, 2021). This suggests that these metrics can fail to capture meaningful differences between simplified texts. Furthermore, there is no standardized framework for measuring the **adequacy** of simplified outputs (Stajner, 2021; Grabar and Saggion, 2022), where adequacy refers to the degree to which the generated text accurately conveys the meaning from the original text (Blatz et al., 2004).[2]

Prior work highlights the importance of manually evaluating TS systems. For instance, Maddela et al. (2023) introduce RANK & RATE, a human evaluation framework that rates simplifications from several models at the sentence level by leveraging automatically annotated edit operations that are verified by annotators. These edit operations are then used in rating the output texts on a scale of 0 to 100. However, this rating is meant to jointly account for meaning preservation, simplicity, and fluency. Devaraj et al. (2022) show that factual errors often appear in both human and automatically generated simplified texts (at the sentence level), and define an error taxonomy to account for both the nature and severity of errors. Yet, these intrinsic evaluations do not directly tell us whether people correctly understand key facts conveyed in the original after reading a simplified text. Moreover, the evaluation is only performed at the sentence level without accounting for the context in which they appear which can impact the overall assessments as noted by Devaraj et al. (2022).

Reading comprehension tests are standard tools used by educators to assess readers' understanding of text materials, and thus provide an assessment of TS that is more in line with its intended use. They have been used to show that human-simplified

[2]We use the terms ''adequacy'' and ''meaning preservation'' interchangeably to convey whether the information from the original is preserved in the simplified throughout this paper.

texts are easier to comprehend by L2 learners (Long and Ross, 1993; Tweissi, 1998; OH, 2001; Crossley et al., 2014; Rets and Rogaten, 2021), as well as secondary and post-secondary students (Heydari et al., 2013). For instance, Long and Ross (1993) conduct a reading comprehension study with 483 Japanese students with varying English language proficiency levels and found that participants who had access to linguistically simplified or elaborated texts scored higher on the RC tasks than those who read the original text. Similarly, Crossley et al. (2014) showed a linear effect of text complexity on comprehension even when accounting for individual language and reading proficiency differences as well as their background knowledge. Rets and Rogaten's (2021) study with 37 adult English L2 users showed better comprehension and faster recall for participants with low English proficiency levels with simplified texts. Finally, in a within-subject study with four original and four simplified texts involving 103 participants with varying levels of English proficiency (beginner to native), Temnikova and Maneva (2013) show that utilizing Controlled Language (Temnikova et al., 2012) for TS improves reading comprehension. All these studies are conducted at the paragraph or the document level based on human-written simplifications which are implicitly assumed to be correct.

The use of reading comprehension for the evaluation of automatically simplified texts has been more limited. Angrosh et al. (2014) first used reading comprehension to evaluate *automatically* simplified texts from multiple TS models, with non-native readers of English. They conduct a multiple-choice test using five news summaries chosen from the Breaking News English website, originally at reading level 6 (hard) and simplified manually or via automatic TS systems. Their study found no significant differences between the comprehension accuracy of different user groups when reading automatically generated simplifications. However, they note that the drop in comprehension scores for some of these systems could be accounted to the content removal which can make some questions unanswerable. Hence, it is not clear whether the differences are non-significant due to user understanding, errors introduced by TS systems, or the effectiveness of simplifications. Laban et al. (2021) also conduct a reading comprehension study to evaluate the usefulness of automatic TS outputs with automatically generated questions that can be answered by original text and human-written references. They found that shorter passages generated by automatic TS systems lead to a speed-up in the RC task completion time regardless of simplicity. However, the automatic generation of questions mostly limited them to factoid, thus limiting the scope of understanding tests, and it is unclear whether the TS errors could render the RC questions unanswerable.

Evaluating text generation via automatic question answering (QA) has also received much attention, including for machine translation (Han et al., 2022), and for text summarization, where it has been used to assess the factuality (Wang et al., 2020) or faithfulness (Durmus et al., 2020) of model-generated summaries. For summarization evaluation, questions have been automatically created based on key information from the model-generated summary, such as important nouns or entities. An automatic QA system is then employed to generate answers to these questions using the original document as a reference. The quality of the generated summary is determined by comparing these answers using metrics that measure semantic similarity or exact matches. However, unlike summarization, where the primary goal is to condense a text (either in an extractive or abstractive fashion), text simplification also involves making structural and linguistic changes so that the text is easier to comprehend which the existing automatic QA-based evaluations are not equipped to assess.

In this work, we design a reading comprehension task to assess the ability of TS systems to preserve meaning via carefully constructed multiple-choice questions targeting language comprehension and use it to conduct a thorough controlled evaluation of a diverse set of state-of-the-art TS systems. We conduct our human evaluation at the paragraph level as humans naturally tend to simplify complex text at this granularity, and utilizing complete texts for measuring RC would yield more accurate results compared to relying on individual sentences (Leroy et al., 2022).

## 3 A Reading Comprehension-based Human Evaluation Framework

**Overview** Our human evaluation is based on the following task: Participants are presented with

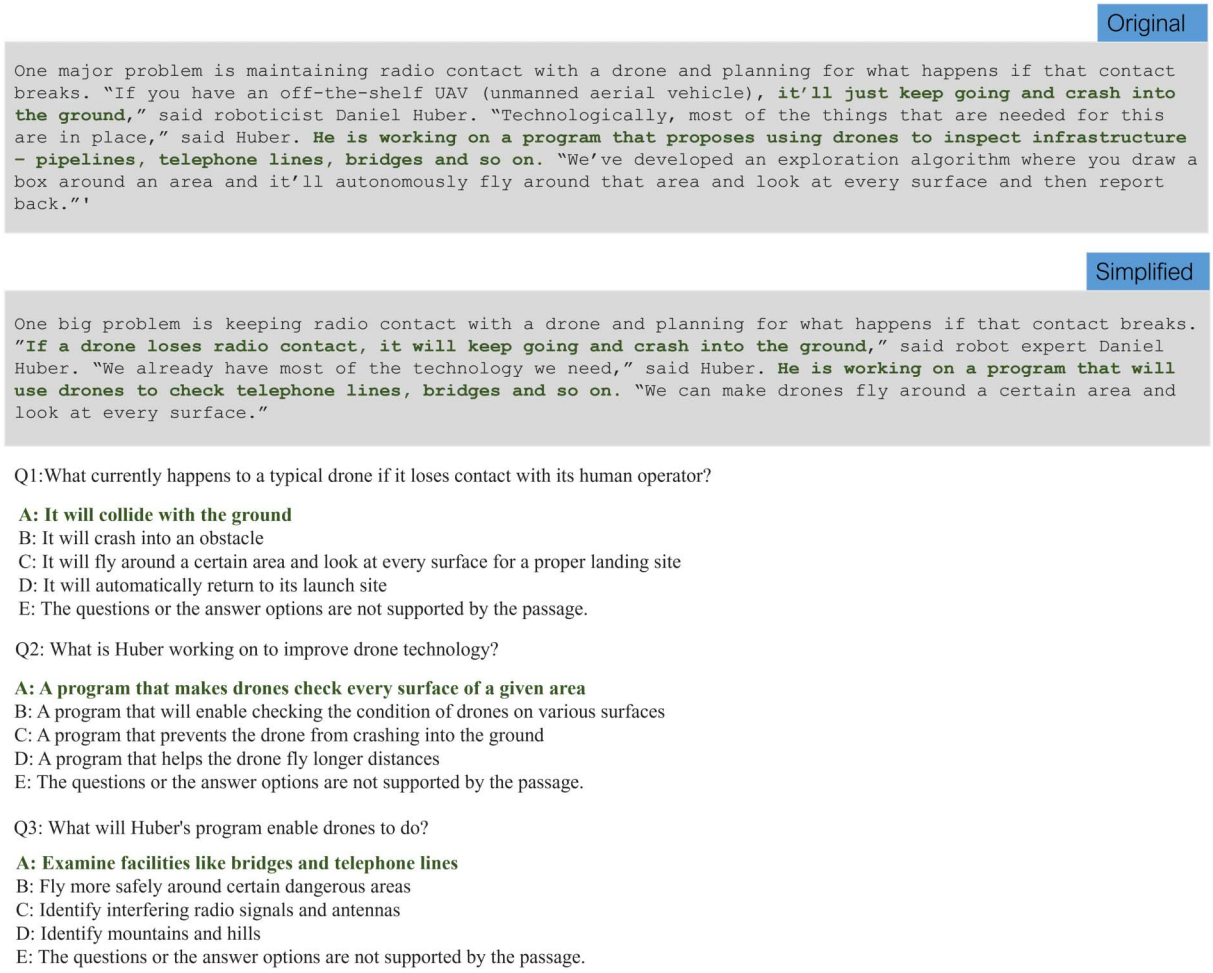


Figure 1: RC questions to be answered after reading either the original or the simplified text. The answer options include the correct answer (A), three incorrect options of varying difficulty (B, C, D), and an option (E) that captures questions rendered unanswerable after automatic TS.

text and then are asked questions to test their understanding of some of the information conveyed in the text, as illustrated in Figure 1. We seek to measure whether participants who read simplified versions of the original paragraph can answer questions as well as those who read the original. However, our goal is not to assess the participants but the TS systems: When working with participants who are proficient in the language tested, we assume that differences in reading comprehension accuracy indicate differences between the quality of TS systems that produce the different simplifications.

**OneStopQA** Within this simple framework, the design of the RC questions and answers is critical to directly evaluate the correctness of automatic TS systems. We build on the OneStopQA reading comprehension exercises created using the STARC (Structured Annotations for Reading Comprehension) annotation framework (Berzak et al., 2020), which is well suited to our task since it targets the real-world need of supporting readers with low English proficiency, and there is already evidence that it is a sound instrument to capture differences in reading comprehension from human-written text.

Specifically, OneStopQA is based on texts from the onestopenglish.com English language learning portal (Vajjala and Lučić, 2018), which are drawn from The Guardian newspaper. Questions are designed to assess language comprehension rather than numerical reasoning or extensive external knowledge. More importantly, these questions cannot be answered with simple string-matching and guessing strategies. Furthermore, the answer options under the STARC annotation framework follow a structured format that reflects four fundamental types of responses, ordered by miscomprehension severity: A indicates

correct comprehension, B shows the ability to identify essential information but not fully comprehend it, C reflects some attention to the passage's content, and D shows no evidence of text comprehension (Berzak et al., 2020). Participants are presented with the answer options in a randomized order to minimize any potential bias or pattern recognition. The correct answer typically is not present verbatim in the critical span, a text span from the passage upon which the question is formulated.[3] We note that the questions only target a subset of the information conveyed in a passage, and hence, our evaluation framework does not provide a measure of completeness. In other words, correctly answering the RC questions does not require understanding every piece of salient information from the original.[4]

Further, prior work suggests that OneStopEnglish text and OneStopQA questions provide a sound basis for evaluating automatic TS, as they can capture differences in reading comprehension from manually simplified text: Gooding et al. (2021) found a statistically significant difference between users scrolling interactions and the text difficulty level in a 518-participant study and Vajjala and Lucic (2019) showed that the nature of the reading comprehension questions can impact text understanding.

**Targeting Answerability** We augment the OneStopQA answer candidates with a fifth option motivated by the failure modes of automatic TS. For each question, participants have the option to pick ''unanswerable'' (UA), which they are instructed to select when ''The questions or the answer options are not supported by the passage.''. This lets us directly measure how often readers judge that there is no support for answering the question based on the input text, which is a more salient problem when presenting participants with automatic than human-written simplifications. The resulting reading comprehension problems are illustrated in Figure 1.

**Text Granularity** Participants are presented with a paragraph of text before answering each question, thus moving away from the prior focus on evaluating TS at the sentence level. In real world settings, people are unlikely to use text simplification on isolated sentences and might be able to understand important information by making inferences from the context. Thus evaluating text simplification outputs at the paragraph level strikes a good balance between providing a realistic amount of context to readers without making the task too long.

**Measures** Given $M$ paragraphs from one of the following: original text, human-written simplification, or the nine automatic TS systems, each paragraph $P \in M$, is accompanied by a set of $Q$ questions with the 5 multiple-choice answers $\{q, a_1^5\}_1^q$, as described above. We measure the adequacy of the simplified texts ($Acc$) using the number of questions answered correctly for that system by human participants. Formally,

$$Acc = \frac{1}{M \times Q} \sum_{m=1}^{M} \sum_{q=1}^{Q} 1[Selected == Correct] \tag{1}$$

where Selected is the answer marked by human participants for a given passage $P$. We rank the automatic TS systems based on the ranking induced by the above scores. Systems with higher scores produce simplifications that help people answer reading comprehension questions correctly.

We compute **answerability** ($Ans$) using questions that were marked ''UA'' by the participants:

$$Ans = 1 - \frac{1}{M \times Q} \sum_{m=1}^{M} \sum_{q=1}^{Q} 1[Selected == UA] \tag{2}$$

Systems often produce outputs that do not support answering the question, perhaps due to over-deletion or other serious output pathology (Devaraj et al., 2022). Systems with high $Ans$ scores produce outputs that are not necessarily correct but still support answering the questions.

## 4 Experimental Setup

First, we describe experiment details including data, participant selection, and study design. Then, we outline the selected TS systems for evaluation.

### 4.1 Study Design

**Data** The OneStopQA dataset includes 30 articles containing 162 paragraphs in total at three difficulty levels: Elementary, Intermediate, and

[3]We do not use the gold or distractor spans in the evaluation study or when generating the TS outputs.

[4]70% of the passages have critical spans (over the three questions) of at least 60%, showing that the questions generally cover most information conveyed in the original text.

Advanced. Each passage is accompanied by three multiple-choice questions that can be answered at all levels of difficulty. The simplified versions include common text simplification operations such as text removal, sentence splitting, and text rewriting. We select the first two paragraphs from each of the 30 articles and associated questions resulting in 60 unique passages and 180 questions in total. Unlike prior studies that evaluate the impact of human-generated simplifications on various target audiences using only a limited number of articles (typically 1-5) and questions (around 3-5) (Long and Ross, 1993; Tweissi, 1998; OH, 2001; Crossley et al., 2014; Rets and Rogaten, 2021), our evaluation is on a larger scale (180 diverse passage-question pairs), which provides more statistical support to rank different systems.

**Participants** The participants are paid directly through the crowd-sourcing platform at an average rate of USD 15/hour. The task is conducted on the Prolific crowd-sourcing platform.[5] We recruit 112 native speakers of English between ages 18 and 60 years identified by their first language and with an approval rating of at least 80% for evaluating the correctness of TS systems.

**Task Design** Each participant is provided with the following instruction: *In this study, you will be presented with 6 short excerpts of English text, accompanied by three multiple-choice questions. You are asked to answer the questions based on the information presented in the text.* A participant is presented with a random subset of 6 texts from one of the 11 conditions: original, simplified by humans, or simplified by one of the nine TS systems. Each passage-question pair is annotated by one native English speaker resulting in 1980 annotations. Annotations collected were manually spot-checked for straightlining (pattern where participants consistently select the same response option) and time differences to ensure that the participants were paying attention to the RC task.

### 4.2 Models for Evaluation

We generate simplified outputs for the selected passages at ''Advanced'' difficulty, i.e., the Original text, using the systems described below as they are representative of the variety of architectures and learning algorithms (supervised, unsupervised, black-box) proposed in the TS literature:

1. Keep-it-simple (KIS) (Laban et al., 2021) is an unsupervised TS system trained using a reinforcement learning framework to enforce the generation of simple, adequate, and fluent outputs at the paragraph level.[6]

2. MUSS (Martin et al., 2022) finetunes a BART-large (Lewis et al., 2020) model with control tokens (Martin et al., 2020) extracted on paired text simplification datasets and/or mined paraphrases to train both supervised and unsupervised TS systems.[7] We use the suggested hyperparameters from the original paper to set the control tokens during simplification generation.[8]

3. ControlT5-Wiki (Sheang and Saggion, 2021) is a supervised controllable sentence simplification model that finetunes a T5-base model with control tokens. Again, we use the suggested hyperparameters from the original paper.[9]

4. ControlSup (Scarton and Specia, 2018) is a controllable supervised TS model that trains a transformer-based sequence-to-sequence model with U.S. target grade as a side-constraint to generate audience-specific simplified outputs. We generate simplified outputs corresponding to Grades 7 and 5 to match the target complexity of the human-written Elementary simplified texts and to assess the impact of the degree of simplification on correctness.

5. EditingCurriculum (EditCL), proposed by Agrawal and Carpuat (2022) trains a supervised edit-based non-autoregressive model that generates a simplified output for a desired target U.S. grade level through a sequence of edit operations like deletions

[5] prolific.co.

[6] github.com/tingofurro/keep_it_simple.

[7] The control tokens are added to the beginning of the input acting as side constraints (Sennrich et al., 2016) and specify the text transformation, like the compression (via the length ratio between the source and the target) or degree of paraphrasing (via the character-level Levenshtein similarity). Please refer to Martin et al. (2020) for more details.

[8] github.com/facebookresearch/muss.

[9] github.com/KimChengSHEANG/TS_T5.

| MODEL | TEXT | | | CONFIG | | METRICS (P) | | |
|---|---|---|---|---|---|---|---|---|
| | # (WORDS | # (SENTS) | FKGL | ARCH | DATA | SARI | BERTSCORE | |
| ORIGINAL | 137.4 | 5.1 | 10.5 | – | – | – | – | P |
| ELEMENTARY | 118.1 | 5.7 | 7.4 | – | – | – | – | P |
| MUSS-SUP | 126.8 | 7.3 | 7.0 | BART | WIKILARGE | 45.07 | 0.940 | S |
| CONTROLT5-WIKI | 135.0 | 7.7 | 6.6 | T5 | | 44.76 | 0.938 | S |
| CONTROLSUP-Grade7 | 132.8 | 5.9 | 9.0 | TRANSFORMER | NEWSELA | 29.27 | 0.946 | S |
| CONTROLSUP-Grade5 | 124.1 | 7.3 | 6.8 | | | 38.35 | 0.939 | S |
| EDITCL-Grade7 | 134.7 | 6.1 | 9.0 | | | 30.49 | 0.939 | S |
| EDITCL-Grade5 | 131.0 | 8.3 | 6.1 | | | 39.69 | 0.929 | S |
| CHATGPT | 123.0 | 5.1 | 10.5 | GPT-3.5 | – | 41.41 | 0.927 | P |
| MUSS-UNSUP | 124.7 | 5.7 | 9.1 | BART | – | 40.67 | 0.937 | S |
| KIS | 73.6 | 3.3 | 9.1 | GPT-2 | – | 33.06 | 0.893 | P |

Table 1: Simplified texts are shorter and include more sentences than the Original. Automatic TS models use various architectures and datasets, generate simplified texts at either the sentence (S) or the paragraph (P) level, and show different tradeoffs in adequacy-simplicity (measured using BERTScore and SARI computed at the paragraph level (P)). A 0.005 difference in BERTScore is significant (p-value = 0.00).

and insertions applied to the complex input text. We generate simplified outputs corresponding to Grades 7 and 5.[10]

6. We generate paragraph-level simplified outputs using ChatGPT with the following prompt:[11]

```
{Text}

Rewrite the above text so that
it can be easily understood by a
non-native speaker of English:
```

We also include the Elementary version of the text from the OneStopEnglish corpus to compare the reading comprehension of the original and model-generated simplified texts against a ground truth reference as a control condition.

Statistics for the human-written and automatically generated passages as well as model summary are presented in Table 1. Automatically generated or manually written simplified texts are shorter and include more sentences (due to sentence splitting) than the original unmodified text. Systems that use pre-trained knowledge (`MUSS, T5, ChatGPT`) receive a higher simplicity (SARI) score than models trained from scratch (`ControlSup, EditCL`) except `KIS`, which achieves low simplicity and adequacy scores according to automatic metrics.[12] Both `ControlSup` and `EditCL` models generate simplified outputs at a higher complexity level than intended (Average FKGL for Grades 7 and 5 are Grades 9 and 7, respectively). Furthermore, the outputs span a wide range of adequacy and simplicity scores where some systems trade-off adequacy for simplicity with low BERTScore and high SARI values (e.g., `ChatGPT, EditCL-Grade5`) and vice-versa (e.g., `ControlSup-Grade7`). While the range of BERTScore values appears small, differences of $> 0.005$ are statistically significant suggesting that the 0.4+ wide range includes meaningful differences within this set of systems.

## 5 Results

We first analyze the results to show the validity of the evaluation set-up, before comparing TS systems on the accuracy and answerability metrics.

### 5.1 Validity of the Human Evaluation

**Results on Human-written Texts Align with the Literature.** As can be seen in Table 2, human-written texts (`Original, Elementary`) achieve the highest accuracy scores of approximately 78%. This is consistent with a study by Berzak et al. (2020) who report that Prolific crowd workers achieve a score of 80.7% when tested on all 162 passages from OnestopQA. As expected, even with human written texts, participants do not answer all questions perfectly,

[10] `github.com/sweta20/EditingCL`.

[11] `openai.com/blog/chatgpt`.

[12] SARI measures lexical simplification based on the words that are added, deleted, and kept by the systems by comparing system output against references and the input text.

| Type | Model | Pre-trained | % Correct | B | C | D | Rank | |
|---|---|---|---|---|---|---|---|---|
| Human | Original | – | 78.33 | 6.11 | 2.22 | 1.11 | **1** | |
| | Elementary | – | 77.22 | 5.56 | 2.78 | 0.00 | **2** | |
| Supervised | MUSS-Sup | ✓ | 76.11 | 6.67 | 1.67 | 1.67 | **3** | |
| | ControlT5-Wiki | ✓ | 74.44 | 6.11 | 2.78 | 1.67 | **4** | ∗ |
| | ControlSup-Grade7 | × | 70.56 | 3.89 | 2.78 | 2.78 | **7** | |
| | EditCL-Grade7 | × | 69.44 | 10.56 | 2.22 | 0.56 | **8** | ∗∗ |
| | EditCL-Grade5 | × | 69.44 | 10.00 | 2.22 | 0.00 | **8** | ∗∗ |
| | ControlSup-Grade5 | × | 67.78 | 11.11 | 3.89 | 0.00 | **10** | |
| Black Box | ChatGPT | ✓ | 74.44 | 9.44 | 1.11 | 0.00 | **4** | ∗ |
| Unsupervised | MUSS-Unsup | ✓ | 73.33 | 6.67 | 2.78 | 1.11 | **6** | |
| | KIS | ✓ | 20.50 | 7.22 | 3.89 | 3.89 | **11** | |

Table 2: Supervised systems that leverage pre-trained knowledge achieve the highest accuracy on the RC tasks. ∗ and ∗∗: Systems that attain the same rank due to the same overall accuracy scores.

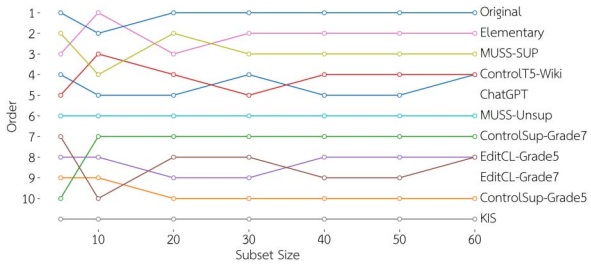


Figure 2: The accuracy scores per condition are averaged over 50 runs for each subset-size $k$: Rankings stabilize with a sample size of 40 passages.

reflecting individual differences in reading proficiency, background knowledge, and familiarity with the topic (Young, 1999; Rets and Rogaten, 2021) as well as the difficulty of the questions. These results thus provide an upper bound to contextualize the scores obtained by TS systems. Furthermore, the scores obtained on the `Original` and `Elementary` versions are very close, as expected when working with native speakers.

**Inter-annotator Agreement (IAA).** We collect a second set of annotations for a subset of 6 passages, covering all 11 conditions, and compute the IAA using Cohen's kappa (McHugh, 2012). The IAA score for selecting the correct answer indicates moderate agreement (0.437) despite the high subjectivity (individual comprehension differences) and complexity (5 answer options) of the task.

**System Rankings are Stable.** Sampling 50 random subsets of $k$ passages for $k \in \{5, 10, 20, 30, 40, 50, 60\}$, we aggregate the mean accuracy score for each subset size and show the rankings for the systems in Figure 2. Using $> 40$ unique passages for each condition, i.e., approximately 120 questions, stabilizes the rankings among the 11 systems, with `ChatGPT, T5,` and `EditCL-Grade7, Grade5` system pairs achieving the same rank.

Taken together, these findings suggest that the evaluation framework is sound and provides a valid instrument to evaluate and compare systems.

### 5.2 TS Adequacy Findings

Table 2 shows the *Acc* scores for human-written texts (`Original, Elementary`) and automatic simplifications generated from supervised (edit and non-edit based), unsupervised, and black-box LLMs. Systems achieve a wide range of scores, starting as low as 20% to approaching within 1% of the accuracy achieved on human-written text. We discuss the main findings below.

Results first show that systems based on unsupervised pre-training yield more correct answers. This is the case for `MUSS-Sup` which achieves the highest accuracy among all systems. `ChatGPT` attains a similar score to that of `ControlT5-Wiki`, a supervised sentence-level TS model, showing the benefits of large scale pre-training, and of reinforcement learning with human feedback—even though it is unfortunately unknown whether `ChatGPT` was trained on TS or related tasks. Overall, the scores show that the best performing TS

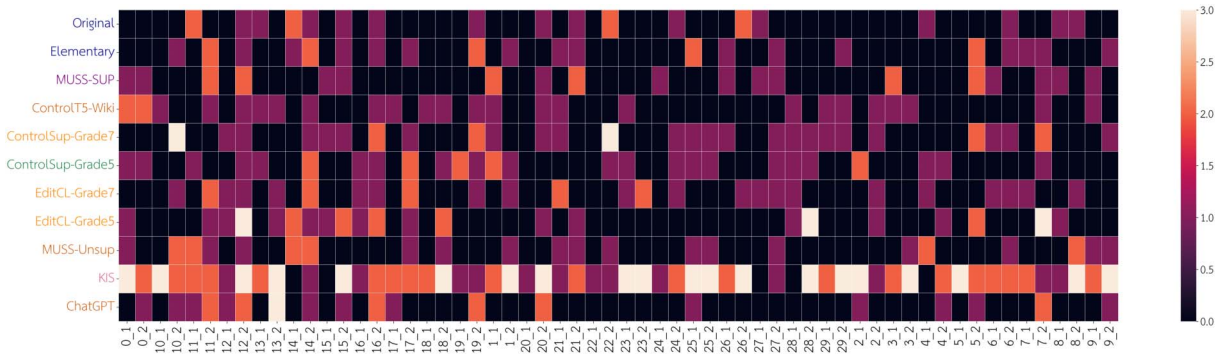


Figure 3: Number of questions marked with UA per paragraph: Different systems have different passage-questions pairs marked as unanswerable, suggesting different deletion errors.

systems rewrite content so that people understand the information tested as well as in human-written text. This suggests that those systems are worth including in usability testing in future work—thus asking not only whether rewrites are *adequate* as we do here, but also whether they are useful to readers that need simplified text.

At the other end of the spectrum, the texts simplified by KIS lead to answering only 20% of questions correctly. This is consistent with the low BERTScore for this system in Table 1, and manual inspection which suggests that KIS is prone to deletions and hallucinations which do not preserve the meaning of the original. We will study the impact of deletions in more depth in the next section.

In the middle of the pack, among systems for grade-specific TS, edit-based models outperform autoregressive models. The autoregressive model `ControlSup` exhibits a 3% decrease in accuracy, due to a more aggressive deletion (Table 1) when simplifying to Grade 5, whereas edit-based models like EDITCL maintain their accuracy score even when generating simpler outputs at both grade levels 7 and 5. However, these edit-based models also result in miscomprehension as suggested by the relatively high percentage of questions marked with option B by the human participants. Note that option B represents a plausible misunderstanding of the critical span upon which the question is based (Section 3). We hypothesize that this could be due to the reduced fluency of the model-generated simplifications via edit-based models.

### 5.3 TS Answerability Findings

We show the answerability score, $Ans$, for all evaluation conditions in Figure 5.

Human-written text does not achieve perfect scores. Using the STARC annotation framework should ideally yield answerable questions, yet in practice, participants still mark 12%–14% of questions with UA. Manual inspection shows that these questions require making complex inferences and hypotheses about the plausibility of the various options. As a result, when given the UA option, participants are more conservative in selecting the four other alternatives.

Most systems achieve 83%–86% answerability for questions, except for KIS, which scores the lowest at 35.56%. On the subset of questions answerable by both `Original` and `Elementary` texts, scores range from 53%–92%. This indicates that errors in model-generated texts hinder question answerability beyond individual comprehension differences. Models, except KIS, achieve similar scores but make different errors, as shown in Figure 3, where no passage-question pairs are correctly answered by all models.

Building upon the finding of Devaraj et al. (2022), who show the prevalence of deletion errors in TS system outputs and our own manual inspection, we hypothesize that over-deletion is the key culprit that makes questions unanswerable. To test this hypothesis automatically, we examine how the unigram overlap (after stop word removal) between the question and the passage (`Support(Q)`) and the answer options and the passage (`Support(A)`) influence question answerability when model-generated outputs are used (Sugawara et al., 2018). While, in most cases, the correct answer does not appear as is in the critical span, we expect the unigram overlap to still provide a useful signal as the rephrased version often shares at least some unigrams with the critical span.

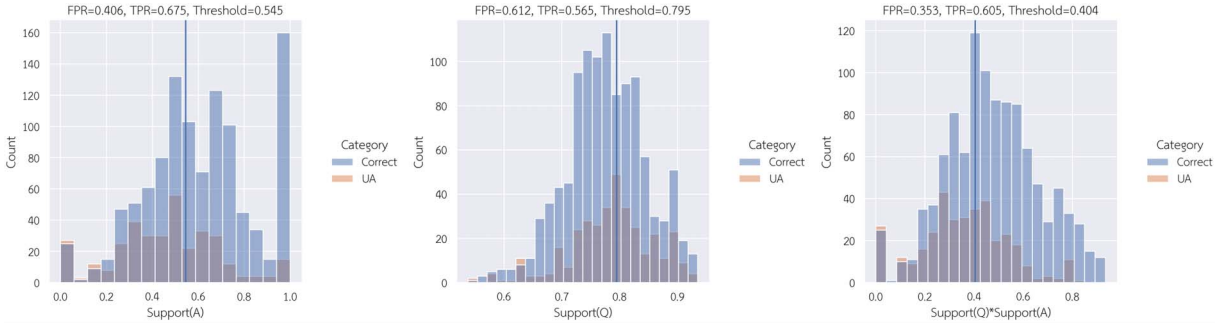


Figure 4: Unigram overlap between the answer options and the passage (Support (A)), question and the passage (Support(Q)) and product of the two (Support (A)*Support (Q)) indicates content deletion as a major factor in making a question unanswerable. TPR: True Positive Rate; FPR: False Positive Rate.

| METRICS | MEANING (REF) BLEU | MEANING (REF) BERTSCORE (P) | MEANING (REF) BERTSCORE (R) | MEANING (REF) BERTSCORE (F1) | MEANING (REF) LEVDIST | MEANING (SRC) BERTSCORE (P) | MEANING (SRC) BERTSCORE (R) | MEANING (SRC) BERTSCORE (F1) | MEANING (SRC) LEVDIST | SIMPLICITY SARI (A) | SIMPLICITY SARI (K) | SIMPLICITY SARI (D) | SIMPLICITY SARI (Avg.) | READABILITY FKGL | QAFACTEVAL (F1) | QAFACTEVAL (EM) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALL | −0.193 | 0.418 | 0.292 | 0.310 | −0.142 | 0.084 | 0.033 | 0.033 | −0.159 | 0.686 | −0.134 | 0.301 | **0.728** | 0.126 | 0.126 | 0.025 |
| ALL − {KIS} | 0.157 | 0.167 | −0.012 | 0.012 | −0.634 | −0.311 | −0.383 | −0.383 | −0.659 | 0.719 | −0.622 | 0.707 | **0.778** | 0.335 | −0.252 | 0.395 |

Table 3: Evaluation of automatic metrics computed at the paragraph-level using all (9) and all but KIS (8) automatic TS systems: SARI achieves the highest correlation with human judgments, followed by LevDist, surpassing meaning preservation and readability metrics: BLEU, BERTScore, and FKGL.

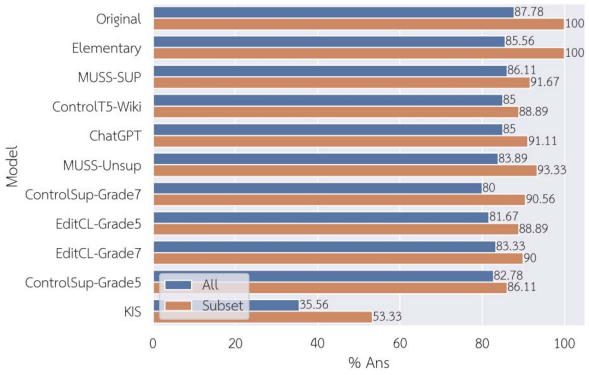


Figure 5: Participants mark 14%–65% questions with UA, suggesting that the meaning of the original text is not entirely preserved in the simplified texts.

Figure 4 shows that `Support(A)` is a more reliable predictor of UA with a true positive rate (TPR) of 0.675 at a false positive rate (FPR) of 0.406 than `Support(Q)` (TPR: 0.565, FPR: 0.612). Answers that appear verbatim in the passage (Support(A)=1.0) are correctly answered 93% of the time. However, when the question lacks support in the passage, the unigram overlap with just the answer becomes an insufficient signal. Therefore, we also report the distribution of the product of `Support(A)` and `Support(Q)` in the same figure to directly capture the support for both the question and the answer options in the passage, i.e., the UA option. The resulting TPR rate for predicting UA is 0.605 at an FPR of 0.353, indicating that the incorrect deletion of partial or complete phrases by the systems affects the support for both the question and the answer options making RC question unanswerable.

These results temper the adequacy results, suggesting that even the best-performing systems delete content. Taken together these findings call for more research on calibrating the deletion tendencies of TS systems, and for human subject studies to develop machine-in-the-loop workflows to validate automatically simplified content before it is presented to readers.

## 6 Evaluating Automatic TS Evaluation Metrics

We now turn towards investigating to what extent automatic TS evaluation metrics frequently used in the literature capture the system rankings obtained via the RC task (Al-Thanyyan and Azmi, 2021; Maddela et al., 2023; Devaraj et al., 2022). We compute the Spearman-Rank correlation of the system-level scores using selected automatic metrics and the RC accuracy scores in Table 3. For meaning preservation, we evaluate BLEU, BERTScore (Zhang et al., 2020) and the Levenshtein distance computed between the system output and the Elementary text (REF) or the system output and the Original text (SRC). For simplicity and readability dimensions, we report correlation

scores with SARI, and FKGL, respectively. SARI measures lexical simplicity based on the n-grams that are kept (K), added (A), and deleted (D) by the system relative to the original text and to the reference simplified (elementary) texts. Note that all metrics are computed at the paragraph level, just like in the RC task, unlike prior evaluation which uses and evaluates these metrics for sentence-level simplification. We also report the correlation scores with QAFactEval, a QA-based metric designed to evaluate factual consistency in summaries (Fabbri et al., 2022).[13]

Overall, SARI achieves the best correlation across the board with or without including the outlier system, i.e., `KIS`. The addition component (A) of SARI that rewards the insertion of n-grams present in the simplified reference but absent from the original text achieves a moderate-high correlation score (0.686–0.719) in both settings. The Levenshtein edit distance of the system output with the `Original` (−0.659) and the `Elementary` (−0.634) text receives a negative moderate-high correlation with human judgments, outperforming both surface-form (BLEU) and embedding-based metric (BERTScore) after removing the outlier system (KIS). We hypothesize that metrics that focus on similarity to only the original or the simplified text do not fully capture the balance between simplicity and adequacy. SARI's 3-way comparison between the input, the output, and the reference is key in yielding system rankings that are consistent with those based on our accuracy results, which could be further repurposed to more directly align evaluation metrics with the accuracy scores.

Furthermore, QAFactEval exhibits only a weak correlation (0.395) at best with human judgments. This is consistent with the current findings by Kamoi et al. (2023), who discuss and show how automatically extracting facts from summaries could lead to a fundamental problem in the evaluation where current QA-based frameworks not only struggle to accurately identify errors in the generated summaries but also perform worse than straightforward exact match comparisons.

## 7 Model-based Question Answering

Our evaluation so far has relied on human-written questions answered by crowd workers, using either model-generated or human-written texts. Automating one or both components would help scale the evaluation and port it to new settings more flexibly. Recent work suggests that this might be plausible: Krubiński et al. (2021) show that automatically generated questions and answers can be used to evaluate Machine Translation systems at the sentence level, and automatic QA techniques (Fabbri et al., 2022; Wang et al., 2020) have been used to assess the factuality and faithfulness of summarization systems.

Here, we assess the performance of a state-of-the-art QA system in recovering the gold-standard ranking induced by human judgments, leaving the more complex study of multiple-choice RC question generation to future work. We use `UnifiedQA v2`[14] a QA model, trained to answer questions in 4 different formats using 20 different datasets. This model has been shown to support better generalization to unseen datasets compared to models specialized for individual datasets. We use the format recommended in the original paper: `{question} \n (A) {choice 1} (B) {choice 2} ... \ n {paragraph}` to generate answers for all the conditions. The Spearman-rank correlation between Exact Match (EM) and ground truth accuracies (C) for all systems is 0.838 and 0.744, excluding KIS. However, we note that the system's ability to distinguish closely competing systems (highlighted by the same color) is limited, as shown in Figure 6.

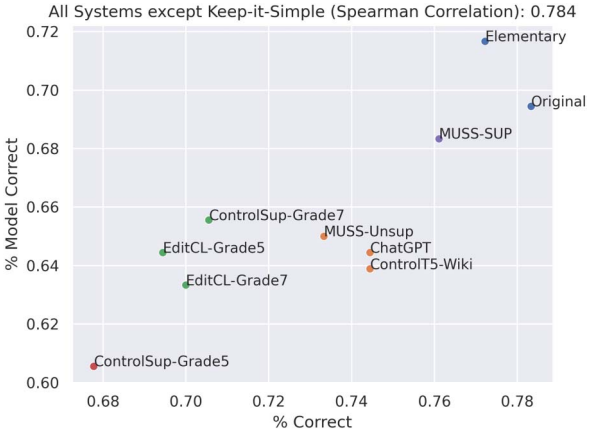


Figure 6: Model-based QA achieves moderate-high correlation with human judgments but fails to distinguish closely competing systems.

[13] `https://github.com/salesforce/QAFactEval`.

[14] `allenai/unifiedqa-v2-t5-3b-1363200`.

Interestingly, QA using human-simplified text achieves higher accuracy than using original unmodified text. This finding is in line with prior work where TS has been shown to improve the performance of multiple downstream NLP tasks such as information extraction (Miwa et al., 2010; Schmidek and Barbosa, 2014), parsing (Chandrasekar et al., 1996), semantic role labeling (Vickrey and Koller, 2008), machine translation (Gerber and Hovy, 1998; Štajner and Popovic, 2016; Hasler et al., 2017; Štajner and Popović, 2018; Miyata and Tatsumi, 2019; Mehta et al., 2020), among others (Van et al., 2021). This suggests that automating part of the evaluation framework is a direction worth investigating in more depth in future work.

## 8 Conclusion

We introduced an evaluation framework based on reading comprehension to directly assess whether TS systems correctly convey salient information from the original texts to readers. This framework lets us conduct a thorough human evaluation of the adequacy of 10 simplified texts: a human-written version and outputs from nine TS systems.

Supervised systems that leverage pre-trained knowledge (MUSS, T5) produce texts that lead to the highest reading comprehension accuracy, approaching the scores obtained on human-written texts. Prompted LLMs (ChatGPT) perform well but are not as accurate as supervised systems. However, we find that even those systems do not preserve the meaning of the original text, with at least 14% of questions marked as ''unanswerable'' on the basis of the text they generate.

When human evaluation is not practical, our analysis suggests that SARI is a better metric than meaning-preservation metrics such as BERTScore and BLEU to rank systems by adequacy, and that model-based QA can approximate system rankings but at the cost of reduced discriminative power across systems and can introduce other confounding factors.

Overall, these results confirm the importance of directly evaluating the accuracy of the information conveyed by TS systems, and suggest that while some systems are overall correct enough to warrant usability studies, all systems still make critical errors. This motivates future work on machine-in-the-loop workflows to let editors and readers rely on TS appropriately (Leroy et al., 2022), and on improving the over-deletion of content by current TS systems. Our human evaluation framework provides a blueprint for evaluating whether correct TS outputs improve reading comprehension for people who have difficulty understanding complex texts, which we intend to investigate in future work.

## Acknowledgments

We thank our TACL action editor, the anonymous reviewers, and the members of the UMD CLIP lab for their helpful and constructive comments on the paper. We also want to thank J. Jessy Li, Ani Nenkova, Philip Resnik, Jordan Boyd-Graber and Abhinav Shrivastava for their feedback on the earlier versions of the work. This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200006, the NSF grant 2147292, funding from Adobe Research, the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI) and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

## References

Sweta Agrawal and Marine Carpuat. 2022. An imitation learning curriculum for text editing with non-autoregressive models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7550–7563, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.520

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Computing Surveys*, 54(2). https://doi.org/10.1145/3442695

David Allen. 2009. A study of the role of relative clauses in the simplification of news texts for learners of English. *System*, 37(4):585–599. https://doi.org/10.1016/j.system.2009.09.004

Oliver Alonzo, Matthew Seita, Abraham Glasser, and Matt Huenerfauth. 2020. Automatic text simplification tools for deaf and hard of hearing adults: Benefits of lexical simplification and providing users with autonomy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3313831.3376563

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. https://doi.org/10.1162/coli_a_00418

Mandya Angrosh, Tadashi Nomoto, and Advaith Siddharthan. 2014. Lexico-syntactic text simplification and compression with typed dependencies. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1996–2006, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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*, pages 5726–5735, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.507

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING. https://doi.org/10.3115/1220355.1220401

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*. https://doi.org/10.3115/993268.993361

Raman Chandrasekar and Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10(3):183–190. https://doi.org/10.1016/S0950-7051(97)00029-4

The Cochrane Collaboration. 2023. Cochrane – trusted evidence. informed decisions. better health. https://www.cochrane.org/ [Accessed: 2023-07-21].

Scott A. Crossley, Hae Sung Yang, and Danielle S. McNamara. 2014. What's so simple about simplified texts? A computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, 26(1):92–113.

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. https://doi.org/10.18653/v1/2021.naacl-main.395, PubMed: 35663507

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. https://doi.org/10.18653/v1/2022.acl-long.506, PubMed: 36404800

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. https://doi.org/10.18653/v1/2020.acl-main.454

Macmillian Education. 2023. Onestopenglish – A teacher resource site. https://www.onestopenglish.com/ [Accessed: 2023-07-21].

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. https://doi.org/10.18653/v1/2022.naacl-main.187

Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models. *arXiv preprint arXiv:2302.11957*.

Rudolf Flesch. 2007. Flesch-Kincaid readability test. *Retrieved October*, 26(3):2007.

Laurie Gerber and Eduard Hovy. 1998. Improving translation quality by manipulating sentence length. In *Conference of the Association for Machine Translation in the Americas (AMTA)*. https://doi.org/10.1007/3-540-49478-2_40

Sian Gooding, Yevgeni Berzak, Tony Mak, and Matt Sharifi. 2021. Predicting text readability from scrolling interactions. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 380–390, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.conll-1.30

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.

HyoJung Han, Marine Carpuat, and Jordan Boyd-Graber. 2022. SimQA: Detecting simultaneous MT errors through word-by-word question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5598–5616, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.378

Eva Hasler, Adrià de Gispert, Felix Stahlberg, Aurelien Waite, and Bill Byrne. 2017. Source sentence simplification for statistical machine translation. *Computer Speech & Language*, 45(Supplement C):221–235. https://doi.org/10.1016/j.csl.2016.12.001

Maryam Heydari, Morteza Khodabandehlou, and Shahrokh Jahandar. 2013. On the effectiveness of strategy-based instruction of textual simplification on efl learners' reading comprehension ability. *Indian Journal of Fundamental and Applied Life Sciences*, 3(2):176–183.

Newsela Inc. 2023. Newsela - online education platform for content. https://newsela.com/ [Accessed: 2023-07-21].

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*. https://doi.org/10.18653/v1/2020.acl-main.709

Ryo Kamoi, Tanya Goyal, and Greg Durrett. 2023. Shortcomings of question answering based factuality frameworks for error localization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 132–146, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.11

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.821

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021. Just ask! Evaluating machine translation by asking and answering questions. In *Proceedings of the Sixth Conference on Machine Translation*, pages 495–506, 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. https://doi.org/10.18653/v1/2021.acl-long.498

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):ooac044. https://doi.org/10.1093/jamiaopen/ooac044, PubMed: 35663117

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. https://doi.org/10.18653/v1/2020.acl-main.703

Michael H. Long and Steven Ross. 1993. Modifications that preserve language and content.

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. https://doi.org/10.18653/v1/2023.acl-long.905

Louis Martin, Éric Villemonte de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4689–4698.

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.

Mary L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282. https://doi.org/10.11613/BM.2012.031, PubMed: 23092060

Sneha Mehta, Bahareh Azarnoush, Boris Chen, Avneesh Saluja, Vinith Misra, Ballav Bihani, and Ritwik Kumar. 2020. Simplify-then-translate: Automatic preprocessing for black-box translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8488–8495. https://doi.org/10.1609/aaai.v34i05.6369

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.

Rei Miyata and Midori Tatsumi. 2019. Evaluating the suitability of human-oriented text simplification for machine translation. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*.

SUN-YOUNG OH. 2001. Two types of input modification and efl reading comprehension: Simplification versus elaboration. *TESOL Quarterly*, 35(1):69–96. https://doi.org/10.2307/3587860

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. https://doi.org/10.3115/1073083.1073135

Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: A corpus analysis. In *Workshop on Speech and Language Technology in Education*. https://doi.org/10.21437/SLaTE.2007-20

Irina Rets and Jekaterina Rogaten. 2021. To simplify or not? Facilitating English l2 users' comprehension and processing of open educational resources in english using text simplification. *Journal of Computer Assisted Learning*, 37(3):705–717. https://doi.org/10.1111/jcal.12517

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2113

Jordan Schmidek and Denilson Barbosa. 2014. Improving open relation extraction via sentence re-structuring. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3720–3723, Reykjavik, Iceland. European Language Resources Association (ELRA).

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints, pages 35–40. Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1005

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.inlg-1.38

Lucia Specia. 2010. Translating from complex to simplified sentences. In *International Conference on Computational Processing of the Portuguese Language*, pages 30–39. Springer. https://doi.org/10.1007/978-3-642-12320-7_5

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. https://doi.org/10.18653/v1/2021.findings-acl.233

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.

Sanja Štajner and Maja Popović. 2018. Improving machine translation of English relative clauses with automatic text simplification. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 39–48, Tilburg, the Netherlands. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-7006

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1453

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1081

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. https://doi.org/10.18653/v1/2021.gem-1.1

Irina Temnikova and Galina Maneva. 2013. The C-score – proposing a reading comprehension metrics as a common evaluation measure for text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 20–29, Sofia, Bulgaria. Association for Computational Linguistics.

Irina Temnikova, Constantin Orasan, and Ruslan Mitkov. 2012. CLCM - a linguistic resource for effective simplification of instructions in the crisis management domain and its evaluations. In *Proceedings of the Eighth International Conference on Language Resources and*

*Evaluation (LREC'12)*, pages 3007–3014, Istanbul, Turkey. European Language Resources Association (ELRA).

Adel I. Tweissi. 1998. The effects of the amount and type of simplification on foreign language reading comprehension.

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. https://doi.org/10.18653/v1/W18-0535

Sowmya Vajjala and Ivana Lucic. 2019. On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 349–359, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4437

Hoang Van, Zheng Tang, and Mihai Surdeanu. 2021. How may I help you? Using neural text simplification to improve downstream NLP tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4074–4080, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.343

David Vickrey and Daphne Koller. 2008. Sentence simplification for semantic role labeling. In *Proceedings of ACL-08: HLT*, pages 344–352, Columbus, Ohio. 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. https://doi.org/10.18653/v1/2020.acl-main.450

Willian Massami Watanabe, Arnaldo Candido Junior, Vinícius Rodriguez Uzêda, Renata Pontin de Mattos Fortes, Thiago Alexandre Salgueiro Pardo, and Sandra Maria Aluísio. 2009. Facilita: Reading assistance for low-literacy readers. In *Proceedings of the 27th ACM international conference on Design of communication*, pages 29–36. https://doi.org/10.1145/1621995.1622002

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. https://doi.org/10.1162/tacl_a_00107

Dolly N. Young. 1999. Linguistic simplification of sl reading material: Effective instructional practice? *The Modern Language Journal*, 83(3):350–366. https://doi.org/10.1111/0026-7902.00027

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

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. https://doi.org/10.18653/v1/D17-1062