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

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

**Abstract**

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

**Keywords:** simplification, evaluation, out-of-domain

## 1. Introduction

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

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

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

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

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



*3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)*, pages 1–14
20-25 May, 2024May 20th, 2024. 

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

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

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

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

## 2. Related Work

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

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

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

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

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

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

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

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

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

## 3. Experimental Setup

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

### 3.1. Data

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

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

### 3.2. Simplification Systems

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

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

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

Table 1 provides a summary of the model attributes.

[1] We use the same document-level test set as Cripwell et al. (2023b).

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

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

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

### 3.3. Evaluating Faithfulness

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

**SummaC (Summary Consistency)** (Laban et al., 2022) first works by using an out-of-the-box NLI model[2] to compute an NLI entailment matrix over a document. This is an $M \times N$ matrix of entailment scores between each of the $M$ input sentences and $N$ output sentences. This is transformed into a histogram form of each column and a convolutional layer is used to convert the histograms into a single score for each output sentence, which are then averaged. As such, this metric is naturally more precision-oriented and therefore could favour shorter, lexically conservative simplifications. In response, we also compute a recall-oriented version, whereby scores are calculated for each input sentence (i.e. high scores will require generating a simple document that retains as much source information as possible).

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

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

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

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

[2] In our case, we use an implementation that uses the version of ALBERT-xlarge from Schuster et al. (2021) finetuned on the Vitamin C and MNLI datasets, available at https://huggingface.co/tals/albert-xlarge-vitaminc-mnli.

[3] https://spacy.io

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

### 3.4. Evaluating Simplicity

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

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

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

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

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

where $y_i$ is the $i$th sentence of document $Y$. We further adapt this to our task by deriving the simplicity level error ($\epsilon$**SLE**$_{doc}$) of a system as the mean absolute error (MAE) between the predicted and target document reading levels.

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

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

## 4. Results and Discussion

### 4.1. Newsela Performance

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

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

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

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

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

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

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

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

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

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

### 4.2. Out-of-Domain Performance

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

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

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

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

## 5. Human Evaluation

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

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

### 5.1. Human Evaluation Results

Table 6 shows the results of the human evaluation.

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

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

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

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

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

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

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

| System | FKGL ↓ | $\epsilon SLE_{doc}$ ↓ |
|---|---|---|
| Input | 10.07 | - (0.89) |
| $PG_{Dyn}$ | 4.72 | **0.21** (2.92) |
| $LED_{para}$ | 4.92 | 0.29 (2.78) |
| $\hat{O} \rightarrow LED_{para}$ | 5.02 | 0.31 (2.76) |
| $\hat{O} \rightarrow$ ConBART | **4.58** | **0.21** (3.00) |

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

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

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

| System | Flu | Faith | Simp | Mean |
|---|---|---|---|---|
| $PG_{Dyn}$ | 0.898 | 0.732 | **0.820** | **0.817** |
| $LED_{para}$ | **0.932** | 0.632* | 0.664* | 0.743* |
| $\hat{O} \rightarrow LED_{para}$ | 0.890 | 0.684 | 0.760 | 0.778* |
| $\hat{O} \rightarrow$ ConBART | 0.890 | **0.760** | 0.764 | 0.805 |

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

## 6. Conclusion

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

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

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

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

## 7. Limitations

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

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

## 8. Acknowledgements

We thank the anonymous reviewers for their feedback and gratefully acknowledge the support of the Agence Nationale de la recherche, of the Region Grand Est, and Facebook AI Research Paris (Gardent; xNLG, Multi-source, Multi-lingual Natural Language Generation. Award number 199495).

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

## 9. Bibliographical References

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

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

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

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

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

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

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

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

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

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

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

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

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

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

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

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

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

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

7943–7960, Online. Association for Computational Linguistics.

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

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

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

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

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6365–6378, Online. Association for Computational Linguistics.

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

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

Zhe Lin, Yitao Cai, and Xiaojun Wan. 2021. Towards document-level paraphrase generation with sentence rewriting and reordering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1033–1044, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuan Ma, Sandaru Seneviratne, and Elena Daskalaki. 2022. Improving text simplification with factuality error detection. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 173–178, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

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

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

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

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

Akifumi Nakamachi, Tomoyuki Kajiwara, and Yuki Arase. 2020. Text simplification with reinforcement learning using supervised rewards on grammaticality, meaning preservation, and simplicity. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 153–159, Suzhou, China. Association for Computational Linguistics.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. Controllable text simplification with deep reinforcement learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 398–404, Online only. Association for Computational Linguistics.

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

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

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

## A. WikiData Article Annotation

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

## B. SLE In-Group Distributions

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

## C. Human Evaluation Details

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

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

Table 7: Distribution of Wikipedia article categories.

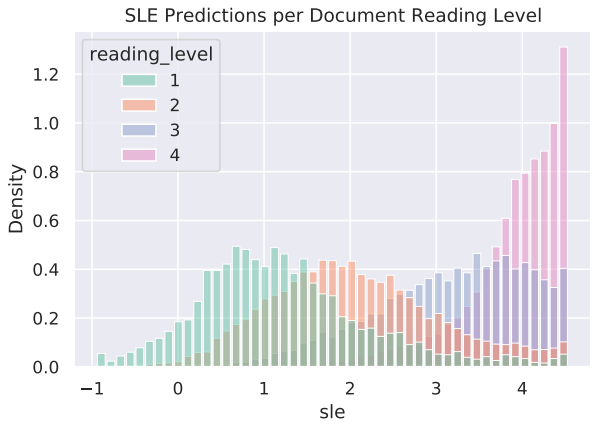


Figure 1: Distribution of SLE scores for reference sentences within each Newsela reading level group.

## D. Extra Evaluation Results

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

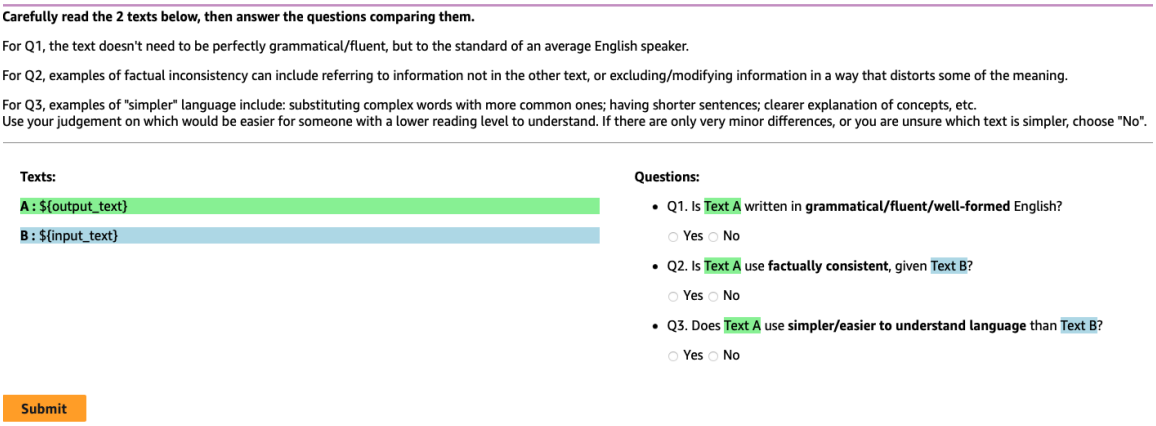


Figure 2: Submission form presented to annotators during the human evaluation.

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

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