

# Dancing Between Success and Failure: Edit-level Simplification Evaluation using SALSA

David Heineman, Yao Dou, Mounica Maddela, Wei Xu

School of Interactive Computing

Georgia Institute of Technology


{david.heineman, douy, mmaddela}@gatech.edu; wei.xu@cc.gatech.edu

## Abstract

We introduce SALSA, an edit-based human annotation framework that enables holistic and fine-grained text simplification evaluation. We develop twenty one linguistically grounded edit types, covering the full spectrum of success and failure across dimensions of conceptual, syntactic and lexical simplicity. Using SALSA, we collect 19K edit annotations on 840 simplifications, revealing discrepancies in the *distribution* of simplification strategies performed by fine-tuned models, prompted LLMs and humans, and find GPT-3.5 performs more quality edits than humans, but exhibits frequent errors. Our data, and annotation toolkit are available at <https://salsa-eval.com>. **This work has appeared previously at EMNLP 2023.**

## 1 Introduction

Text simplification aims to improve a text’s readability or content accessibility while preserving its fundamental meaning (Stajner, 2021; Chandrasekar et al., 1996). Traditional human evaluation for text simplification often relies on individual, shallow sentence-level ratings (Sulem et al., 2018; Alva-Manchego et al., 2021), easily affected by the annotator’s preference or bias. Maddela et al. (2023) recently proposes a more reliable and consistent human evaluation method by ranking and rating multiple simplifications altogether. However, as text simplification involves performing a series of transformations, or *edits*, such as paraphrasing, removing irrelevant details, or splitting a long sentence into multiple shorter ones (Xu et al., 2012), sentence-level scoring remains difficult to interpret since it is not reflective of detailed information about the types of edits being performed.

We introduce SALSA  – Success and FAilure-driven Linguistic Simplification Annotation – an *edit-level* human evaluation framework capturing a broad range of simplification transformations. SALSA is built on a comprehensive typology (§2)

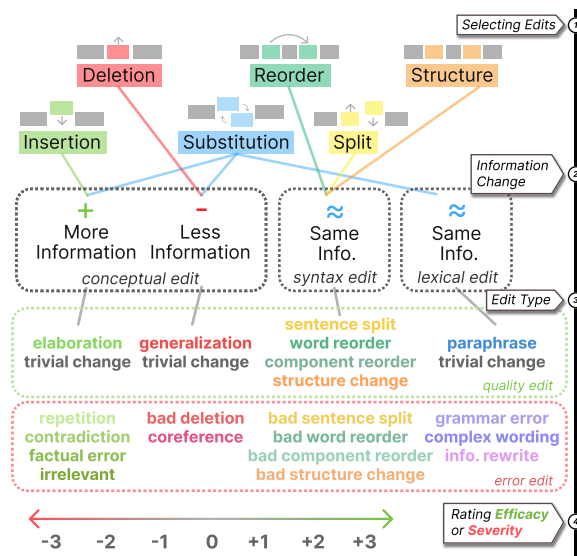


Figure 1: The multi-stage SALSA edit evaluation framework. Spans are classified into twenty one **success** and **failure** types.

containing 21 quality and error edit types. Using SALSA, we develop an interactive interface and collect 19K edit annotations of 840 simplifications written by eleven state-of-the-art language models and two humans. With these annotations, we conduct a large-scale analysis of model and automatic metric performance, and further introduce the automatic word-level quality estimation task for text simplification. Our results demonstrate that SALSA provides an interpretable and exhaustive evaluation of text simplification.

## 2 SALSA Framework

We introduce SALSA, an edit-based human evaluation framework for text simplification. SALSA is defined by a typology of 21 linguistically-grounded edit types with the aim of capturing both successes and failures (i.e., quality changes and errors). The annotation methodology of SALSA is structured as a decision tree and implemented via an easy-to-use interface, with the decision tree shown in Figure 1.

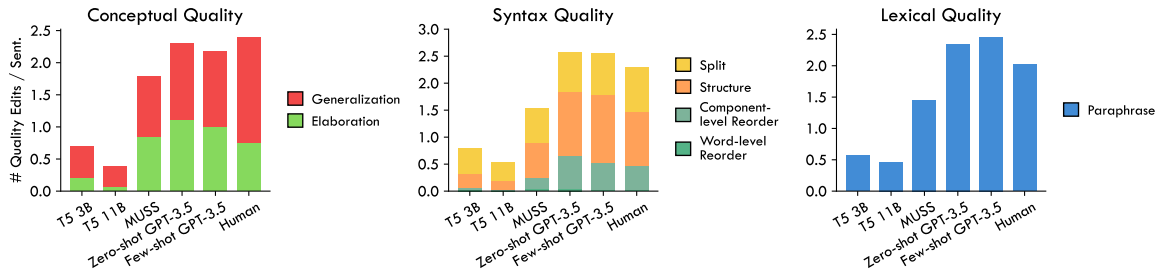


Figure 2: Successful edits per-model, organized by edit type. MUSS outperforms fine-tuned T5 but fails to capture more complex simplification techniques. Compared to GPT-3.5, human written simplifications have more generalization ■, a similar distribution of syntax edits, and slightly less paraphrasing ■.

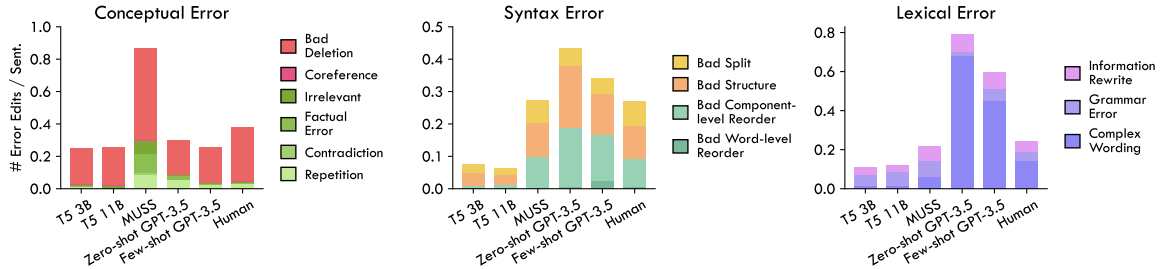


Figure 3: Failure edits per-model, organized by edit type. Compared to humans, both GPT-3.5 setups make more syntax and lexical errors. Although humans perform bad deletion ■ errors at a higher frequency than GPT-3.5, this is reflective of the inherent ambiguity in judging the relevancy of the deleted content.

## 2.1 Edit Selection

Annotation begins with *edit selection*, where annotators identify the edits performed by the simplification and select the corresponding spans for each edit. We define six types of edit operations: single-operation **insertion**, **deletion**, **substitution**, word-/clause-**reorder**, and multi-operation sentence **split** and **structure** changes. An insertion or deletion edit exclusively modifies content, while a substitution either modifies or paraphrases content. Reorder, split, or structure edits perform a context-free syntax transformation. As split and structure edits are multi-operation (i.e., require a combination of single operations), they are defined by a set of underlying single-operation *constituent* edits.

## 2.2 Categorizing by Information Change

Each selected edit is then labeled with its impact on the underlying sentence information: *less*, *same*, *more* or *different* information. Given the type of operation and change to information, we subsequently organize each edit into three linguistic families as defined by Siddharthan (2014):

**Lexical edits** perform simple changes in “wording”. This includes paraphrasing (i.e., substitution that keeps the same information) and inconsequential trivial changes (e.g., inserting ‘the’).

**Syntax edits** capture transformations to the *distribution* of information, rather than substance. A

split converts a candidate sentence to two sentences, a re-order edit re-arranges clauses or wording within a clause, and a structural edit modifies the voice, tense or clausal structure.

**Conceptual edits** modify underlying ideas conveyed by the text. A conceptual edit requires elaboration to add clarifying information or generalization to delete unnecessary/complicated ideas.

## 3 Results

We use SALSA to evaluate state-of-the-art simplification by collecting annotations on our extended version of the SIMPEVAL corpus (Maddela et al., 2023), which includes fine-tuned, LLM- and human-written simplifications. Our resulting data collection includes 19K edit annotations across 840 simplifications. We present our primary results in Figures 2, 3, which illustrate the frequency of quality and error edit types. Few-shot GPT-3.5 far surpasses existing models, particularly in making syntax and content edits. However, its simplifications are not *aligned* to the types of operations performed by human. Some fine-tuned models such as the MUSS (Martin et al., 2022) produce more diverse edits than GPT-3.5, yet suffer from incredibly high errors, while others (T5, Raffel et al., 2020) learn to minimize loss by making very few changes. Despite low conceptual and syntactic simplification, MUSS paraphrases at a human rate.

## Limitations

Our annotation only represents a single use case of text simplification and we encourage an extension of SALSA to domain-specific simplification, such as medical (Joseph et al., 2023), legal (Garimella et al., 2022), or multi-lingual text (Ryan et al., 2023), and annotations by groups of specific downstream users (Stajner, 2021). Incorporating higher-level operations such as sentence fusion, paragraph compression, and reordering would require an extension to SALSA and presents unique analytical challenges. Finally, detailed human evaluation inherently requires greater resources to produce a high granularity of annotations. While we show this process can be streamlined with a robust annotator training, SALSA requires a similar amount of resources as widely used fine-grained evaluation in other tasks such as MQM (Lommel et al., 2014) or FRANK (Pagnoni et al., 2021).

## Ethics Statement

Our annotations were performed using the SIMPEVAL<sub>2022</sub> corpus, originally collected from publicly available Wikipedia articles (Maddela et al., 2023) and we further extend the dataset with complex sentences collecting using the same methodology from publicly available Wikipedia articles. As discussed in §B.2, we perform data collection with in-house annotators from a US university. Annotators were all native English speakers and paid \$15-\$18/hour. We took care to manually review all data prior to annotation as to exclude any triggering or sensitive material from our annotation data. Annotators were informed that any data they felt uncomfortable with was not required to annotate. Our interface was built using the open-source Vue.js<sup>1</sup> library, and training of our added T5-11B system was implemented using the open-source Hugging Face Transformers<sup>2</sup> library.

## Acknowledgements

We thank Tarek Naous, Nghia T. Le, Fan Bai, and Yang Chen for their helpful feedback on this work. We also thank Marcus Ma, Rachel Choi, Vishnesh J. Ramanathan, Elizabeth Liu, Govind Ramesh, Ayush Panda, Anton Lavrouk, Vinayak Athavale, and Kelly Smith for their help with human annotation. This research is supported in part by the NSF

awards IIS-2144493 and IIS-2112633, ODNI and IARPA via the HIATUS program (contract 2022-22072200004). 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 NSF, 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

- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. [Data-driven sentence simplification: Survey and benchmark](#). *Computational Linguistics*, 46(1):135–187.
- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. [The \(un\)suitability of automatic evaluation metrics for text simplification](#). *Computational Linguistics*, 47(4):861–889.
- 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*.
- Jiangjie Chen, Rui Xu, Wenxuan Zeng, Changzhi Sun, Lei Li, and Yanghua Xiao. 2023. [Converge to the truth: Factual error correction via iterative constrained editing](#). In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI’23*. AAAI Press.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%\\* chatgpt quality](#).
- Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. [Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.
- Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nandkishore Kambhatla. 2022. [Text simplification for legal domain: Insights and challenges](#). In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 296–304, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Sian Gooding and Manuel Tragut. 2022. [One size does not fit all: The case for personalised word complexity](#)

<sup>1</sup><https://vuejs.org/>

<sup>2</sup><https://huggingface.co/>

- models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 353–365, Seattle, United States. Association for Computational Linguistics.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Comput. Surv.*, 55(12).
- Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh J Ramanathan, Wei Xu, Byron C Wallace, and Junyi Jessy Li. 2023. [Multilingual simplification of medical texts](#). *arXiv preprint arXiv:2305.12532*.
- Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. [Multidimensional quality metrics \(MQM\): A framework for declaring and describing translation quality metrics](#). *Revista Tradumàtica: tecnologies de la traducció*, 12:455–463.
- Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. [LENS: A learnable evaluation metric for text simplification](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.
- Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. [MUSS: Multilingual unsupervised sentence simplification by mining paraphrases](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. [Training language models to follow instructions with human feedback](#). *arXiv preprint arXiv:2203.02155*.
- 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.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *J. Mach. Learn. Res.*, 21(140):1–67.
- Michael Ryan, Tarek Naous, and Wei Xu. 2023. [Revisiting non-English text simplification: A unified multilingual benchmark](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.
- Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. [Beyond fair pay: Ethical implications of NLP crowdsourcing](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.
- Advait Siddharthan. 2014. [A survey of research on text simplification](#). *ITL-International Journal of Applied Linguistics*, 165(2):259–298.
- Sanja Stajner. 2021. [Automatic text simplification for social good: Progress and challenges](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.
- Elior Sulem, Omri Abend, and Ari Rappoport. 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.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. [LLaMA: Open and efficient foundation language models](#). *arXiv preprint arXiv:2302.13971*.
- Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. [Paraphrasing for style](#). In *Proceedings of COLING 2012*, pages 2899–2914, Mumbai, India. The COLING 2012 Organizing Committee.

## A Defining the SALSA 🧩 Framework

We provide detail into the SALSA framework, including qualitative examples which helped guide design decisions when building the typology. Table 1 illustrates each final edit type, as organized by Figure 1. During development, we adjusted our scheme based on preliminary annotations with the final goal of SALSA’s ability to evenly represent all modes of simplification and the full space of errors.

### A.1 Quality Evaluation

We organize quality edits by their approach to simplification, as real-world application and models’ capability to simplify falls into tiers of *conceptual*, *syntactic* and *lexical* simplification (Stajner, 2021). An ideal simplification system demonstrates a balance of these ‘tiers’ and incorporates different techniques depending on the original text, context and users (Gooding and Tragut, 2022). Automatic simplification research initially focused on lexical paraphrasing (Siddharthan, 2014), but has since evolved to emphasize the importance of syntactic and conceptual editing (Alva-Manchego et al., 2020).

### A.2 Error Evaluation

We describe the SALSA error typology, with examples of each type in Table 1. Although despite their sparsity, errors have a far greater impact on fluency and adequacy than individual quality edits (Chen et al., 2023). We refined our definition of errors by focusing on minimizing the amount of error types while retaining the ability to capture the full possibility of simplification ablations. Notably, we specifically exclude a *hallucination* due to its ambiguous definition in related work (Ji et al., 2023), and instead define our error categories to capture any possible hallucination.

## B Data Collection

We describe our use of SALSA to collect 19K edit annotations covering 11.6K spans on 840 model-generated and human-written simplifications.

### B.1 Simplification Data

Data collection is performed on an extended version of SIMPEVAL<sub>2022</sub> (Maddela et al., 2023), including a train set covering state-of-the-art simplification systems and held-out test set of recent LLMs.

The screenshot displays the SALSA annotation interface, which is divided into four main sections corresponding to the steps in the caption:

- 1. Selecting Edits:** This section shows the original sentence: "Civil unrest in northern Italy spawns the medieval musical form of Geisslerlieder, penitential songs sung by wandering bands of Flagellants." Below it is the simplified sentence: "Geisslerlieder was created from civil unrest in Northern Italy through songs sung by travelling bands of Flagellants." The interface allows selecting spans from both sentences. The selected spans are "X spawns" from the original and "X was created from" from the simplified sentence.
- 2. Identifying Information Change:** This section compares the original phrase "spawns" with the new phrase "was created from". It asks: "Compared to the original phrase, the new phrase expresses:". The options are "the same meaning", "less information", "more information", and "totally different meaning". The selected option is "less information".
- 3. Classifying Edit Type:** This section asks: "Select the impact of this paraphrase edit on the sentence simplicity." The options are "Negative Impact", "No Impact", and "Positive Impact". The selected option is "Positive Impact".
- 4. Rating Efficacy/Severity:** This section asks: "Rate the efficacy (how much it helps you to read and understand the sentence?)". The options are "1 - minor", "2 - somewhat", and "3 - a lot". The selected option is "2 - somewhat".

Figure 4: The SALSA annotation process consists of (1) selecting edits, (2) identifying information change, (3) classifying edit type and (4) rating efficacy/severity.

**SALSA Train.** We first extend the 360 simplifications from SIMPEVAL<sub>2022</sub> to 700 simplifications based on 100 complex sentences from Wikipedia articles dated between Oct 2022 and Dec 2022. The complex sentences are unseen during the training of the LLMs and were selected to be intentionally difficult (avg. length of 37.3 words) to enable an evaluation of the models’ full capabilities in performing diverse simplification edits. Simplifications are generated by five models including fine-tuned T5-3B and T5-11B (Raffel et al., 2020), MUSS (Martin et al., 2022), a controllable BART-large model trained with unsupervised, mined paraphrases, zero- and few-shot GPT-3.5 (Ouyang et al., 2022), and two human-written references.

**SALSA Test.** We further gather 20 more complex sentences from Wikipedia articles published in Mar 2023 and generate 140 simplifications using recent LLMs including GPT-3.5, ChatGPT, GPT-4, Alpaca-7B (Touvron et al., 2023) and Vicuna-7B (Chiang et al., 2023), along with T5-3B and T5-11B fine-tuned with control tokens.

### B.2 Annotation

As crowd-sourced annotators have shown to have inconsistent quality (Shmueli et al., 2021), we hire 6 undergraduate students from a US university. An-

	Type	Description	Example
<b>Quality Evaluation</b>			
Conceptual	Elaboration	Meaningful and correct information which enumerates the main idea	Many volatile organic chemicals, <b>which harm our environment</b> , are increasing in abundance in the lower troposphere.
	Generalization	Removes unnecessary, irrelevant or complicated information	Many volatile organic chemicals are increasing in the lower troposphere. ( <i>in abundance was removed</i> )
Syntax	Word-level Reorder	Order of words within a phrase is swapped	Many <b>organic volatile</b> chemicals are increasing in abundance in the lower troposphere.
	Component-level Reorder	Order of phrases within a sentence is swapped	<b>In the lower troposphere</b> , many volatile organic chemicals are increasing in abundance.
	Sentence Split	Independent information converted to two separate sentences	Many volatile organic chemicals are increasing. <b>They are found</b> in abundance in the lower troposphere.
Lexical	Structure Change	Rewrites voice, tense or structure	<b>The abundance of</b> many volatile organic chemicals <b>is</b> increasing in the lower troposphere.
	Paraphrase	Lexical complexity of the phrase decreases, while the meaning is unchanged	Many volatile organic chemicals are <b>being seen more</b> in the lower troposphere.
	Trivial Change	Adds clarity or removes verbosity, while the lexical complexity and meaning is unchanged	Many volatile organic chemicals are <b>currently</b> increasing in abundance in the lower troposphere.
<b>Error Evaluation</b>			
Conceptual	Bad Deletion	Deleted necessary and relevant content	Many chemicals are increasing in abundance in the lower troposphere. ( <i>volatile organic was removed</i> )
	Coreference	A reference to a named entity critical to understanding the main idea is removed	<b>They</b> are increasing in abundance in the lower troposphere.
	Repetition	Phrase added or changed but fail to contain novel information or insight	Many volatile organic chemicals, <b>which are chemicals</b> , are increasing in abundance in the lower troposphere.
	Contradiction	Phrase added or changed but clearly contradicts information presented in the original sentence	Many volatile organic chemicals, <b>which are decreasing in our troposphere</b> , are increasing in abundance in the lower troposphere.
	Factual Error	Externally verifiable incorrect claim is made by the phrase	Many volatile organic chemicals are increasing in abundance in the lower troposphere <b>when they decide to</b> .
Syntax	Irrelevant	New information is introduced which is unrelated to the main idea	Many volatile organic chemicals, <b>unlike low vapor pressure chemicals</b> , are increasing in abundance in the lower troposphere.
	Bad Word-level Reorder	Presented a new word order with less clarity within a clause	Many volatiles <b>are having their abundance increasing</b> in the lower troposphere.
	Bad Component Reorder	Presented a new clausal order with less clarity	<b>In abundance in the lower troposphere</b> , many volatile organic chemicals are increasing.
	Bad Structure	A failed attempt to modify the voice, tense or structure	Many volatile organic chemicals <b>have been</b> increasing in abundance in the lower troposphere.
	Bad Split	Split at an inappropriate location or interrupted the flow of ideas	Many volatile organic chemicals are <b>increasing</b> . <b>They</b> are increasing in abundance in the lower troposphere.
Lexical	Complex Word-ing	Lexical complexity of the phrase increases, while the meaning is retained	Many volatile organic chemicals are <b>proliferating throughout</b> the lower troposphere.
	Information Rewrite	All information was removed from the phrase and replaced with new information	Many volatile organic chemicals are <b>decreasing</b> in abundance in the lower troposphere.
	Grammar	Violation of conventional grammar	Many volatile organic chemicals <b>which</b> are increasing in abundance in the lower troposphere.

Table 1: Overview of the SALSA edit-level evaluation typology. Original text for the examples: *Many volatile organic chemicals are increasing in abundance in the lower troposphere.*

notators were trained with an in-depth tutorial consisting of broad explanations of simplification concepts, over 100 examples covering each of the 21 SALSA edit types and interactive exercises, completed two rounds of onboarding annotations and were provided continuous feedback by the authors. To concretely measure agreement for each stage of the SALSA framework, we collect annotations in three stages: (1) we have three annotators select edits, (2) a fourth annotator adjudicates the edits into a single selection and (3) the initial three annotators classify and rate the adjudicated edits. Figure 4 illustrates our annotation interface, with further

screenshots of our tutorial included in Appendix C.

### B.3 Inter-Annotator Agreement

We calculate edit selection agreement (i.e. agreement prior to adjudication) by each token, with Table 2 reporting agreement per edit, further broken down by their type of information change. We observe edit agreement is highly dependent on the edit type and type of information change being performed. High agreements are seen for deletion ( $\alpha=0.75$ ), paraphrase (**substitution** with the same information,  $\alpha=0.53$ ), and sentence splits ( $\alpha=0.66$ ). **Substitution** that introduces more information, how-

Edit	Sub-type	Kripp. $\alpha$	3 Agree%	2 Agree%
Insertion	More Information	0.45	14%	40%
Deletion	Less Information	0.75	42%	65%
Substitution	More Information	0.15	1%	11%
	Less Information	0.31	7%	26%
Reorder	Word-level	0.12	0%	13%
	Component-level	0.41	11%	38%
Split	Sentence Split	0.66	32%	55%
Structure	Structure	0.25	5%	25%
Substitution	Same Information	0.53	21%	51%

Table 2: Edit selection inter-annotator agreement measured per token. As Krippendorff’s  $\alpha$  (2018) includes unlabeled tokens, we also report the percentage of annotated tokens where at least 2 and 3 annotators agree.

ever, exhibits lower agreement ( $\alpha=0.15$ ), due to the subjectivity among annotators on determining whether new tokens contain ‘novel’ information, as was often mixed up with insertion. Reordering ( $\alpha=0.12$ ) and structure edits ( $\alpha=0.25$ ) also report lower agreements. Additionally, we find our % rates for annotator agreement are similar to fine-grained evaluation frameworks in other text generation tasks (Dou et al., 2022).

## C Annotation Tutorial

We include screenshots to highlight the diversity of exercises and interactive elements in our detailed interface tutorial.

# Text Simplification Annotation Tutorial



## Introduction

Welcome! In this project you will read simplified sentences generated by Artificial Intelligence and rate their quality.

This qualification HIT will train you to perform this task. You must be able to:

- Find the changes our AI made (i.e. "**selecting spans**")
- Evaluate the quality of each change
- Identify errors in each change

Here's a big picture of what we're doing:

The ID of the current sentence & total for this task: < Hit X / Y >

Download and upload sentences to annotate

A sentence before it's simplified: Original Sentence (Human Written): Civil unrest in northern Italy spawns the medieval musical form of Geisslerlieder, penitential songs sung by wandering bands of Flagellants.

A sentence before after simplified by a human or an AI model: Simplified Sentence (Human or AI Model Written): Geisslerlieder was created from civil unrest in Northern Italy through songs sung by travelling bands of Flagellants.

Comment

Submit extra notes about your annotations

EDITS ANNOTATIONS (5/5)

+ Add Edit Add another edit to annotate

We've found 5 edits for this sentence. 4 are "good" and 1 is "bad"

substitute spawns to was created from : good paraphrase

delete the medieval musical form of : good deletion

insert through : bad trivial insertion

delete penitential : good deletion

substitute wandering bands to travelling bands : good paraphrase

Modify or delete the annotations for each edit

Next

Figure 5: Landing page introducing annotators to each part of the task. The 10 stages organize different concepts in the SALSA typology.



## Examples

Observe this original sentence:

Original Sentence (Human Written):  
**Born into slavery in Virginia in 1856,** Booker T. Washington became an influential African American leader at the outset of the Progressive Era.

Simplified Sentence (Human or AI Model Written):  
Booker T. Washington became an influential African American leader at the outset of the Progressive Era.

This sentence communicates many different facts. Here are just a few:

- Booker T. Washington was born in Virginia
- Booker T. Washington was born in 1856
- **Booker T. Washington was born into slavery**
- Booker T. Washington was an influential leader
- Booker T. Washington was a African American leader
- Booker T. Washington was a influential African American leader as a result of being born into slavery
- Booker T. Washington was a leader at the beginning of the Progressive Era

Hover over each piece of information to see which part of the sentence could be deleted to remove that information from the sentence. As you can see, sentences which communicate many ideas may be hard to narrow down whether a span is *significant*.

In this case, the *main idea* of the sentence is *Booker T. Washington is an influential African American leader*. Deletions are necessary for text simplification, we just want to ensure this *main idea* is still being communicated.

Let's put it together with a few other sentences:

Original Sentence (Human Written):  
**Like so many hyped books before it,** *The Midnight Library* excited me and gave me pause.

Simplified Sentence (Human or AI Model Written):  
*The Midnight Library* excited me and gave me pause.

Like so many hyped books before it,

Is the deleted span significant to the main idea of the original sentence?

1 - not at all    2 - minor    3 - somewhat    4 - very much

Figure 6: Example interactive allowing annotators to see different spans to understand different amounts of relevancy to the *main idea* of the sentence.

## Phrase vs. Syntax Edits

As you can see *phrase edits* can only capture how individual pieces of information or a small set of words change. When the AI creates an edit which modifies the sentence as a whole, this *must* be captured by a structural edit. Here's some examples of overlapping phrase and syntax edits:

Original Sentence (Human Written):

The amount that Phoenix spends on criminal enforcement is difficult to quantify<sup>1</sup>, in part because<sup>2</sup> the City does not classify arrests by housing status.

Simplified Sentence (Human or AI Model Written):

Because<sup>2</sup> the City of Pheonix, in part, refuses to use an individual's housing to organize arrests, it has become hard to find out how much it spends on policing<sup>1</sup>.

structure 🌲 - 1

How does the edit impact the simplicity of the phrase?

positive no impact **negative**

Rate the severity:

1 - Minor **2 - Somewhat** 3 - A lot

*Explanation: This edit changes the order of the clauses in the sentence. Because the sentence is trying to communicate cause and effect (quantifying spending is difficult because a lack of data), the edit moves the effect before the cause.*

structure 🌲 - 2

How does the edit impact the simplicity of the phrase?

**positive** no impact negative

Rate the efficacy:

**1 - Minor** 2 - Somewhat 3 - A lot

*Explanation: This edit simply moves 'because' before 'in part'. This adds clarity to the order information is being presented.*

Figure 7: One of the 100 sentence examples provided to annotators, highlighting different types of structure edits existing within the same sentence.