Why Reconstructing Scrambled Texts Fails: A Structural Analysis of Reconstruction Outputs

Kei Du, Christof Schöch

University of Trier
Universitätsring 15, 54296 Trier
{duk, schoech}@uni-trier.de

Abstract

This paper explores the limitations of reconstructing scrambled text within the context of Derived Text Formats (DTFs). While previous research has treated reconstruction as a technical challenge, this study shifts the focus to investigating the causes of reconstruction failure. Through a detailed analysis of outputs generated by language models on non-literary (IMDb reviews) and literary (Gutenberg texts) datasets, several systematic patterns were identified. First, reconstructed texts are generally shorter than the originals, indicating that the generated results are often incomplete. Second, models simplify expressions by omitting specific modifiers, thereby producing more general outputs. Third, high similarity at the string level does not guarantee semantic equivalence, revealing fidelity-related issues in text reconstruction. In literary texts, chunk-based segmentation poses additional challenges; this approach disrupts syntactic and contextual coherence, leading to sentences that are structurally correct but semantically distorted. These findings suggest that reconstruction difficulty is not merely a matter of model performance but also reflects the importance of higher-level textual organization. This study highlights the fundamental limitations of current language models and reframes reconstruction failure as an analytical perspective for understanding how meaning is constructed in text.

Keywords: derived text formats, scrambled text, reconstructibility

1. Introduction

In Digital Humanities, derived text formats (DTFs), also sometimes called extracted features, have been proposed for the storage, publication, and reuse of datasets built from in-copyright texts (Jett et al. 2020, Schöch et al. 2020). One approach is to scramble the order of the words in the text so that it becomes unreadable to humans but can still be used for text and data mining. An important precondition for applying this DTF, of course, is that the original text cannot be reconstructed. This question has attracted increasing attention. For example, Du et al. 2025 largely framed this issue as a problem of performance: to what extent can large language models (LLMs) successfully restore an original sequence of a text once its linear structure has been disrupted. While such an approach yields valuable empirical insights, it also risks narrowing the scope of inquiry by treating reconstruction primarily as a technical challenge.

This paper adopts a different perspective. Rather than asking whether disordered texts can be reconstructed, it focuses on why reconstruction often fails. It reveals the extent to which textual meaning depends on higher-level structural organization — such as logical sequencing and thematic coherence — which cannot be easily recovered once disrupted. By reinterpreting reconstruction failure as an analytical resource, this study seeks to reposition the problem within a broader theoretical framework. In doing so, it aims to demonstrate that reconstructability is not merely a measure of technical capability, but a reflection of the structural constraints that determine how texts generate and maintain meaning.

2. Previous work

Research on reconstructing scrambled texts draws on several interconnected strands of work in natural language processing and computational literary studies. In Du et al. 2025, reconstruction has been framed as a problem of recovering linguistic structure from transformed representations. Experiments have been conducted in which a language model (T5-base) was fine-tuned to take scrambled texts as input and generate outputs that resemble the original text. Two datasets including IMDb reviews (non-literary texts) and Gutenberg novels (literary texts) were used.

For the experiments on IMDb reviews, each review was treated as a single data point and was converted into DTF format by shuffling the word order within sentences while keeping sentence order unchanged. Three datasets (25k, 50k, and 75k reviews) were used for training the T5-model, with separate validation and test sets of 5,000 unseen reviews each. For the experiments using Gutenberg novels, texts from four genres were randomly selected: detective, historical, romance, and science fiction. Two datasets were built (12 and 60 novels) to study the effect of size of training data, with texts split into chunks where words in each chunk were shuffled but chunk order in each novel preserved. These chunks (50, 100, and 500 words) were used as data points and divided into training, validation, and test sets in an 80/10/10 ratio.

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 63–66

The reconstruction quality was evaluated using string similarity metrics (word error rate, rouge scores and sacreBLEU) between the reconstructed and original texts. The results showed that reconstruction performance is generally poor: similarity scores remain low across most cases, with only a few outliers achieving higher similarity. Longer and more complex literary texts are especially difficult to reconstruct. While models trained with more data offer slight improvements, they do not significantly change the overall outcome. The models tend to recover only fragments of vocabulary rather than accurate sentence structure or full semantic content. Overall, the study concludes that reconstructing original texts from scrambled texts is still a challenging task.

3. Analysis of reconstructed texts

While the previous study evaluated reconstructed texts by comparing the string similarity between the original and reconstructed texts, the present study reports on a detailed analysis of the reconstructed (non-literary and literary) texts and provides an in-depth examination of the reasons for reconstruction failure. The primary objective is to identify the differences between the reconstructed text and the original text across various levels, and to explore whether the quality of the reconstruction can be improved.

3.1 Non-literary texts (IMDb-reviews)

Based on the evaluation of the reconstruction results in Du et al. 2025, the model trained using 75,000 reviews produced better results than other models. Therefore, we have carefully compared the differences between the reconstructed text using this model and the original text and can make the following three observations.

First, the text length of the original and reconstructed IMDb-reviews was compared. The difference in length between the original and the reconstructed texts was calculated as follows:

$$difference = \frac{len(orig) - len(recon)}{len(orig)}$$

In Figure 1, the Y-axis represents the percentage difference in length, and the X-axis shows three models trained on 25,000, 50,000, and 75,000 texts, respectively. As can be seen, most of the values are larger than 0, indicating that most of the original texts are longer than their reconstruction. It is particularly noteworthy that, regardless of which model was used for text reconstruction, more than 75% of the reconstructed texts are at least 20% shorter than the original text. In contrast, only a very small number of reconstructed texts are as long as or longer than their original counterparts. This reflects a systemic issue with the model when reconstructing text: it tends to generate outputs that are shorter than the original text. To address this issue, a possible solution is to require that the length of the reconstructed text by the model must match that of the original text.

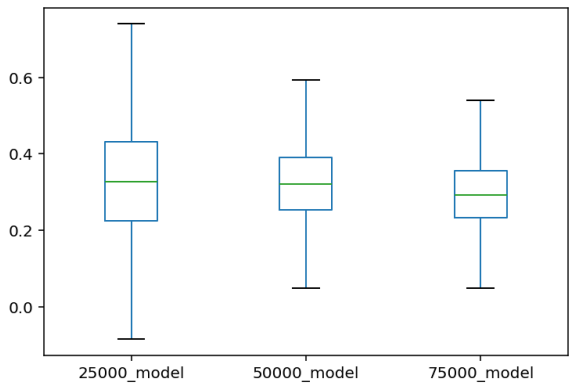

Figure 1: Distribution of the difference in length between the original and the reconstructed IMDb-texts (outliers are not visualized).

Second, the model tends to simplify sentences in reconstruction. Here are some examples:

- “A good solid piece” was reconstructed as “A solid piece”. (review no. 23)
- “Typical American movie” was reconstructed as “Typical movie”. (review no. 1)
- “This is one of the worst Stephen King movies I've ever seen” was reconstructed as “This is one of the worst movies I've ever seen”. (review no. 2073)

As we can see, the main subject (e.g. “piece”, “movie”) in these examples remains unchanged, and the overall sentiment or evaluation is preserved. However, specific modifiers such as adjectives, authorship, or origin are missing after the reconstruction. This is a common phenomenon in NLP text generation, and it is mainly caused by probability-based generation favoring high-frequency tokens, the loss of long-tail information, and decoding strategies that bias toward safer outputs. Prior works such as Guo et al. 2024 have shown that standard language model training tends to overemphasize high-frequency, low-information tokens, leading to fluent but generic outputs that lack linguistic diversity. This issue can be mitigated through controllable generation techniques such as incorporating weighting mechanism conditioned on token frequency (Jiang et al. 2019).

Third, string similarity does not fully indicate whether the reconstruction was successful, because similar texts may have completely different meanings. Here are some examples:

- “I hope everyone had a good time making this mess” was reconstructed as “I had a good time making this mess”. (review no. 4050)

- “After the first 5 minutes there is nothing worth watching in this film” was reconstructed as “After watching this film there is nothing worth watching”. (review no. 4690)
- “Predictable soaps, you've seen every soppy story full of sad” was reconstructed as “Predictable soaps, you've seen every soppy story full of sad”. (review no. 4884)

In fact, this issue is known as the “faithfulness” problem in text generation tasks and has been discussed in e.g. Maynez et al. 2020. In this work, textual entailment measures were suggested to address the need of developing evaluation and training methods that go beyond lexical overlap and can accurately model semantic faithfulness.

3.2 Literary texts (Gutenberg texts)

Compared to reconstructing non-literary texts, reconstructing literary texts presents both similar and additional challenges.

First of all, the comparison of the lengths of the reconstructed text and the original text reveals a different pattern to that observed before. Only when the Gutenberg texts are divided into longer 500-words-chunks, the original chunks are often 40% longer than the reconstructed text, or even more (see Figure 2). In contrast, the reconstructed text of 50-words-chunks and 100-words-chunks are similar in length to the originals or (in some cases) even longer than the originals.

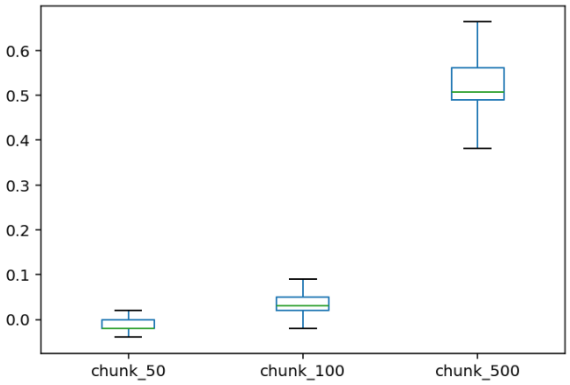

Figure 2: Distribution of the difference in length between the original and the reconstructed Gutenberg chunks (outliers are not visualized).

The evaluation results in Du et al. 2025 showed that reconstructing Gutenberg texts was more challenging than IMDb reviews. When we examined the reconstructed texts, we uncovered a key factor that led to this outcome: the different segmentation methods: The IMDb reviews were split into sentences and the word order within each sentence was shuffled. In comparison, the Gutenberg texts were split into chunks with exact number of words (50, 100 or 500). As a result, many chunks begin and end with incomplete sentences. Reconstructing such language chunks is more challenging than reconstructing a shuffled sentence, as it requires the model to maintain syntactic and semantic correctness while taking context into account and selectively discarding some information. We randomly selected a number of 50-words-chunks and carefully reviewed their reconstructed versions, and found that the beginnings and endings of nearly all the reconstructed chunks differed from those of the original text. There are three examples in Table 1. An interesting observation is that the first token in all three reconstructed examples is a punctuation mark, which is very common in reconstruction results. The reasons behind this phenomenon clearly require further analysis and research.

original text	reconstruction
room door . The diagram of this portion of the hotel will give you an idea of these connecting rooms . There are three of them , as you will see , all reception - rooms . Mr. Ransom had passed through them all in looking for his wife .	. There are three rooms in this hotel , all of them connecting . Mr. Ransom had an idea of his wife , as you will see . The reception - room will give you an idea of all these rooms , as you pass through the door .
to talk to Sir Andrew , if only for a moment . He felt lonely and desperately anxious . He had hoped to tire out his nerves as well as his body , but in this he had not succeeded . As soon as he had given up his tools	. As soon as he had succeeded in talking to Sir Andrew , he felt as if he had given up his nerves . He had not only given up his tools , but he had desperately hoped to tire out his body in this lonely moment , as well as
no other man could have come to him in that place ; and his whole body was wrung with torturing pains , and he was in the very article of death . And so it was , my prudence leading me to speak few and simple words , and my	; and he was so simple in his words that he could have no other article to speak of ; and in torturing him , and in leading me to my death , my whole body was wrung with pains and prudence . And so it was with my man ,

Table 1. Three examples of Gutenberg text chunks and their reconstruction

Another phenomenon worth noting is that, since each chunk may contain more than one sentence, and the words in each chunk were shuffled, the reconstructed text — while grammatically correct — often combines words of phrases from different sentences in the original chunk into a single sentence. This results in reconstructed text that may be very similar to the original at the word or n-gram level, but whose content differs significantly from the original. This phenomenon

is very common even in the reconstruction of short chunks containing just 50 words, let alone longer chunks containing 100 or 500 words. To avoid such errors, the model likely needs sufficient contextual information to accurately determine which subject corresponds to which verb and object within a sentence. However, since the order of the words within each chunk has been scrambled, the precise contextual information found in the original text is no longer present in the scrambled text.

4. Conclusion

In this study, we have examined the results of reconstructing scrambled text in detail. The results indicate that significant further improvements would be necessary for the successful reconstruction of DTFs, with areas of improvement ranging from the methods used to reconstruct the text to the evaluation of the quality of the reconstructed results. Given the current state of the art, reconstructing DTFs remains a highly challenging task. This also implies that reconstructability is not currently a factor that hinders the widespread use of DTFs for the publication of in-copyrighted text as research data.

5. Acknowledgments

This work was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460033370. The authors would like to thank for the funding and support. Furthermore, thanks also include all institutions and actors who are committed to the association and its goals.

Author contributions:

Keli Du: Conceptualization, Methodology, Investigation, Visualization, Writing - original draft, Writing - review & editing

Christof Schöch: Funding acquisition and Supervision, Writing - review & editing

6. Bibliographical References

Du, K., Ackerschewski, S., Navruz, U., Sınır, N., Valline, J. & Schöch, C., (2025) “Reconstructing Shuffled Text. Bad Results for NLP, but Good News for Using In-Copyright Texts”, *Journal of Computational Literary Studies* 4(1). DOI: https://doi.org/10.48694/jcls.4163.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2024.findings-naacl.228.

Jett, Jacob, Boris Capitanu, Deren Kudeki, Timothy Cole, Yuerong Hu, Peter Organisciak, Ted Underwood, Eleanor Dickson Koehl, Ryan Dubnicek, and J. Stephen Downie (2020). *The HathiTrust Research Center Extracted Features Dataset (2.0)*. DOI: http://doi.org/10.13012/R2TE-C227.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2019. Improving Neural Response Diversity with Frequency-Aware Cross-Entropy Loss. In The World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, 2879–2885. DOI: https://doi.org/10.1145/3308558.3313415.

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. DOI: https://doi.org/10.18653/v1/2020.acl-main.173.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke (2020). “Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen”. In: *Zeitschrift für digitale Geisteswissenschaften* 5. DOI: http://doi.org/10.17175/2020_006.