# Report on Natural Language Processing: Semantic Similarity, Word Embeddings, and Linguistic Reconstruction for Text Transformation and Analysis

Georgios Mamantzis (P22090)
Kostantinos Petmezas (P22140)
Petros Mantaios (P22092)

## Abstract

This report investigates the application of Natural Language Processing (NLP) techniques for transforming disorganized, raw texts into structured and semantically consistent versions. The study integrates both rule-based correction mechanisms and transformer-driven paraphrasing pipelines to enhance clarity, grammar, and cohesion. To evaluate meaning preservation, we employed word embeddings, cosine similarity, and Principal Component Analysis (PCA). The analysis highlights how different reconstruction strategies affect semantic fidelity, readability, and tone. Challenges such as nuance retention, ambiguity handling, and scalability are discussed. We conclude that a hybrid framework, combining rule-based precision with the adaptability of transformer models, provides the most promising path forward for large-scale text refinement.

## Introduction

Natural Language Processing (NLP) has become indispensable in modern computing, bridging the gap between unstructured human communication and structured machine interpretation. From conversational agents and search engines to automatic summarization and translation tools, NLP underpins much of the technology that shapes our digital world. Despite tremendous advances, particularly with neural networks and transformer architectures, language remains inherently complex. Humans rely heavily on implicit meaning, context, and cultural nuance. For computational systems, these layers of interpretation remain difficult to capture reliably.

One of the areas where this complexity is most evident is text reconstruction. Emails, informal notes, and transcribed speech often appear fragmented, ambiguous, or ungrammatical. While humans can usually infer the intended message, machines tend to misinterpret or overlook such signals. The aim of this project is to examine methods of converting such texts into coherent and semantically faithful forms that can be analyzed computationally. The central problem lies in balancing two requirements: improving readability while preserving meaning.

This report begins by outlining the methodologies employed, which include both rule-based sentence refinement and transformer-based paraphrasing pipelines. It then turns to the experiments conducted, followed by a discussion of the results, the main challenges encountered, and the opportunities for future automation. The study concludes by synthesizing the findings and emphasizing the potential of hybrid approaches that combine rules with deep learning models.

## Methodology

The methodology followed in this project consisted of two main phases: text reconstruction and computational analysis. Each phase employed distinct techniques, yet both were connected by the overarching goal of improving text clarity while safeguarding semantic integrity.

The first phase, text reconstruction, used two strategies. In the first strategy, a rule-based mechanism was applied to carefully selected sentences. This approach emphasized grammatical correctness and syntactic clarity. By implementing direct replacements for common errors, such as transforming “bit delay” into “a bit of delay,” we ensured that glaring issues were addressed. To handle more nuanced improvements, the Stanza library was incorporated, offering tokenization, part-of-speech tagging, and lemmatization. This allowed for deeper corrections, such as inserting missing subjects, removing redundant particles, and adjusting punctuation and capitalization. Although highly precise, this method was limited in scope, as it relied on hand-crafted rules that could not scale to large or varied datasets.

The second strategy employed transformer-based paraphrasing pipelines. These models were applied to entire texts, allowing us to examine their ability to restructure larger portions of language while preserving meaning. Three models from Hugging Face were selected for their different characteristics. The first, Vamsi/T5_Paraphrase_Paws, tended to introduce structural variation without sacrificing coherence. The second, ramsrigouthamg/t5_paraphraser, produced conservative outputs that stayed very close to the original, offering strong semantic fidelity but less stylistic change. The third, prithivida/parrot_paraphraser_on_T5, generated more creative alternatives, often producing smoother text but sometimes omitting subtle details. By comparing these models, we were able to observe how different paraphrasing strategies influence the balance between fluency and semantic preservation.

Once reconstruction was complete, the project moved to computational analysis. Here, we turned to Sentence-BERT, specifically the all-MiniLM-L6-v2 model, to generate dense embeddings for sentences. These embeddings capture contextual meaning rather than simple word-level associations, making them well-suited for measuring semantic similarity. We used cosine similarity to quantify how closely the reconstructed texts aligned with the originals. In addition, PCA was applied to reduce the high-dimensional embeddings into two dimensions, allowing us to visualize how the different reconstructions clustered around their source material.

## Experiments and Results

The experiments began with the custom rule-based reconstruction of two problematic sentences. For example, the original phrase “Hope you too, to enjoy it as my deepest wishes” was transformed into “I hope you enjoy it too. My best wishes.”

This reconstruction achieved a cosine similarity score of 0.8669, indicating that the core meaning had been retained despite significant grammatical restructuring. Similarly, the sentence “Although bit delay and less communication at recent days, they really tried best for paper and cooperation” was reconstructed as “Although a bit of delay and less communication in recent days, they really tried their best on the paper and in our cooperation.” This version achieved an even higher similarity score of 0.9133, reflecting its success in correcting grammar while preserving semantics.

The transformer-based models offered a broader perspective on full-text reconstruction. Vamsi/T5_Paraphrase_Paws often reorganized sentences, sometimes making them flow more naturally while still conveying the original meaning. Ramsrigouthamg/t5_paraphraser produced outputs that were strikingly similar to the originals, with minimal rewording. In contrast, prithivida/parrot_paraphraser_on_T5 generated smoother paraphrases but occasionally left out details, creating subtle semantic drift.

The computational analysis confirmed these observations. Cosine similarity scores ranged from 0.8767 to 0.9624 across the models, indicating strong semantic preservation overall. The PCA plots showed reconstructed texts clustering close to their originals, though the degree of proximity varied by model. For example, Vamsi/T5_Paraphrase_Paws achieved the highest similarity for one text, while ramsrigouthamg/t5_paraphraser excelled in another. Prithivida/parrot_paraphraser, while generating fluent alternatives, consistently drifted further from the originals in both quantitative and visual analyses.

## Discussion

The results highlight both the promise and the limitations of current NLP methods for text reconstruction. Sentence-BERT embeddings proved to be highly effective at capturing semantic relationships, as demonstrated by the high similarity scores and clear clustering patterns in the PCA visualizations. These tools provide a robust means of quantifying the subtle shifts that occur when text is paraphrased or restructured.

However, several challenges emerged. Preserving tone and nuance proved particularly difficult, as transformer-based models often altered the emotional undertones of a message even when the literal meaning was preserved. Ambiguities in the source texts also posed problems, since informal phrases could be interpreted in multiple ways, leaving reconstructions somewhat inconsistent. Furthermore, variability among transformer models demonstrated that model selection has a significant impact on output quality. Rule-based approaches, while useful, were too rigid to handle the diversity of language found in real-world data, highlighting the need for more scalable solutions.

Automation in this domain is promising but requires balance. Purely rule-based systems cannot scale effectively, while purely neural approaches risk semantic drift. A hybrid solution seems best, where rules act as a safeguard for predictable issues, transformers handle general reconstruction, and post-processing steps ensure consistency. This layered pipeline could deliver both precision and scalability, which are crucial for practical deployment.

When comparing methods, it became clear that each had distinct strengths. Rule-based corrections excelled at precision but lacked generality. Ramsrigouthamg/t5_paraphraser was most reliable in preserving meaning but contributed little stylistic improvement. Vamsi/T5_Paraphrase_Paws provided a balance between structure and accuracy, while prithivida/parrot_paraphraser produced the most fluent outputs at the cost of fidelity. The choice of method, therefore, depends on whether the priority is strict semantic preservation, stylistic refinement, or scalability.

---

## Conclusion

This study demonstrated that NLP techniques can be effectively applied to reconstruct unstructured text into clearer, more coherent forms. Rule-based methods delivered targeted corrections for known issues, while transformer pipelines provided scalable yet varied paraphrasing. By combining embeddings, similarity measures, and visualizations, we showed that meaning can be preserved even in the face of significant linguistic restructuring. At the same time, challenges remain in maintaining tone, resolving ambiguity, and ensuring consistency across models. The evidence suggests that hybrid methods, which bring together rule-based precision and neural flexibility, offer the most promising approach for large-scale applications. As NLP continues to evolve, the ability to refine text while preserving its intended meaning will remain a key capability for both academic and industrial applications.

**Bibliography**

- Apostol, Elena, et al. *SIMPLEX: A Lexical Text Simplification Architecture*. arXiv preprint (2023).
- Camacho-Collados, José. *Embeddings in Natural Language Processing*. Draft book overview (2019).
- *Impact of Word Embedding Models on Text Analytics in Deep Learning*. PMC Journal review (2023).
- Vahtola, Teemu, et al. *Analysing Semantic Similarity with Antonyms and Negation Using Sentence Embeddings*. ACL Anthology (2022).
- Simusid, et al. “Why Cosine Similarity for Transformer Text Embeddings?” Reddit discussion (2023).
- *A Deep Learning Approach for Paragraph-Level Paraphrase Generation and Detection*. Springer journal (2025).