ComparaTree: A Multi-Level Comparative Treebank Analysis Tool

Luka Terčon
Faculty of Arts, University of Ljubljana / Aškerčeva cesta 2, 1000 Ljubljana
Faculty of Computer and Information Science, University of Ljubljana / Večna pot 113, 1000 Ljubljana
luka.tercon@ff.uni-lj.si

Kaja Dobrovoljc
Faculty of Arts, University of Ljubljana / Aškerčeva cesta 2, 1000 Ljubljana
Jožef Stefan Institute / Jamova cesta 39, 1000 Ljubljana
kaja.dobrovoljc@ff.uni-lj.si

Abstract

ComparaTree is an open-source tool for comparative treebank analysis that combines various methods of quantitative linguistic analysis to provide a general overview of the differences and similarities between two treebanks. The comparison tool covers a range of subfields of linguistic analysis, providing a summary of the differences and similarities in terms of the lexical diversity, n-gram diversity, part-of-speech and dependency relation proportions, syntactic complexity, and syntactic diversity. We explain the various quantitative analyses performed on every level along with the generation of graphical visualizations, which add value by enabling user-friendly comparisons at a glance. We exemplify the comparison process by presenting the results produced by the tool when comparing two treebanks from the Universal Dependencies collection.

1 Introduction

The Universal Dependencies initiative (de Marneffe et al., 2021) has produced a large repertoire of treebanks featuring a consistent, cross-linguistically applicable grammatical annotation format. As of the latest 2.15 release of UD, the collection includes almost 300 treebanks in over 150 languages (Zeman et al., 2024), while at the same time boasting considerable diversity in terms of the various text genres included in the treebanks (Müller-Eberstein et al., 2021), containing also different language modalities such as spoken language (Dobrovoljc, 2022). Given this high degree of diversity, the UD collection is ideal for conducting both intra-linguistic as well as cross-linguistic comparisons, with cross-linguistic studies using UD treebanks becoming especially common (e.g., Nikolaev et al. (2020), Berdicevskis et al. (2018), Levshina et al. (2023)).

Although comparative studies based on Universal Dependencies are becoming increasingly common, there is still a lack of general-purpose tools that facilitate such analyses in a systematic way, as only a handful of specialized tools currently support comparative work. The QuanSyn Python package (Yang and Liu, 2025) supports the analysis of syntactic properties, such as the distribution of parts of speech and dependency relations, within and across treebanks. The STARK tool (Krsnik et al., 2024) enables the extraction of dependency (sub)trees from parsed corpora and supports frequency-based comparisons between datasets. The conllu-diff utility[1] generates statistical summaries of differences between CoNLL-U files, but is limited to individual token-level labels. Beyond these, various processing tools and programming packages are listed on the official UD website,[2] but they are typically not optimized for direct comparative analysis.

While each of the tools mentioned above offers valuable functionality, they are typically limited in scope—focusing on a single linguistic level, requiring programming expertise, or lacking support for user-friendly side-by-side comparison. To address this gap, the present paper introduces ComparaTree, a user-friendly tool for comparative treebank analysis that combines multiple methods of quantitative linguistic analysis. It supports comparisons across lexical diversity, n-gram diversity, part-of-speech and dependency relation distributions, syntactic complexity, and syntactic diversity. ComparaTree also generates visualizations in the form of graphs and diagrams, providing a clear visual overview of the similarities and differences between two treebanks.

In the present paper we first describe the different levels of linguistic analysis for which ComparaTree generates a comparison in Section 2. In Section 3 we exemplify the usage of the tool and

[1] https://pypi.org/project/conlludiff/

[2] https://universaldependencies.org/tools.html

Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025), pages 129–139
August 28-29, 2025 ©2025 Association for Computational Linguistics

present the format of the results by performing an analysis using two UD treebanks. Finally, in Section 4 we conclude with a discussion of the possible future improvements and extensions.

2 Treebank Comparison

ComparaTree is a tool written in the Python programming language that takes two treebanks in the CoNLL-U format[3] as input and calculates values for various linguistic measures. Although the tool was designed to be used with UD treebanks, in principle any dependency grammar formalism is supported, as long as the treebank used conforms to the standard CoNLL-U format. The source code of the tool is publicly available and can be accessed via a dedicated GitHub repository along with the documentation for its use.[4]

The calculated linguistic measures pertain to five different levels of linguistic analysis: lexical diversity, n-gram diversity, part-of-speech and dependency relation proportions, syntactic complexity, and syntactic diversity. For every level of comparison the tool also outputs a visualization of the results. In the following, we first describe a special process of segment-based averaging in Section 2.1 that is performed for several of the analysis levels. Next we describe the various measures calculated on each level in Section 2.2, while Section 2.3 introduces the various types of resulting visualizations.

2.1 Segment-Based Averaging

For three out of the five analysis levels—lexical diversity, n-gram diversity, and syntactic diversity—a similar methodology is employed to calculate the corresponding measures. The analysis procedure on these three levels involves first splitting the treebank into segments that contain approximately the same number of tokens[5] and subsequently calculating the ratio between the number of unique items and the total number of items in each segment. The final score is obtained by taking the mean of this ratio over all the segments. Each analysis level differs in terms of what is taken as the item for which the ratio is calculated.

This unified analysis method stems from the Type-Token Ratio, a well-established measure of lexical diversity which is obtained by taking the ratio between the number of distinct types (wordforms) and the total number of tokens in a corpus. Although this measure is very commonly used as the default measurement of lexical diversity in many comparative linguistic analyses (e.g., Muñoz-Ortiz et al. (2024) and André et al. (2023)), it is also very sensitive to text length (McCarthy and Jarvis, 2010) and thus might lead to unfair comparisons between treebanks of different sizes. Thus ComparaTree aims to counteract the effect of treebank size using the segment-based averaging technique.

2.2 Levels of Comparison

2.2.1 Basic Comparison

At the most basic level, the tool outputs an overview of the size of both treebanks in terms of the number of tokens contained, the mean sentence length in the number of tokens, and the sentence length standard deviation.

2.2.2 Lexical Diversity

Lexical diversity refers to the amount of variation in the vocabulary used in some corpus and is calculated within ComparaTree using the above-described Type-Token Ratio (henceforth *TTR*). The measure is first calculated for each segment individually and then averaged across all segments. The tool takes the total number of unique lemmas as the number of types in a segment, as this proves more robust when dealing with morphologically richer languages.

2.2.3 N-Gram Diversity

N-gram diversity refers to how prevalent established sequences of words are in a treebank. If a treebank contains fewer unique n-grams (i.e. sequences of *n* consecutive words), this indicates that the corpus is more formulaic and thus has a lower n-gram diversity.

To compute the level of n-gram diversity, the ComparaTree tool first extracts every n-gram[6] in every segment of each treebank along with its corresponding frequency. The tool then calculates the fraction of unique n-grams in the segment, a measure that is also known as the N-gram Diversity score (henceforth *NGD*) (Padmakumar and He, 2024) and averages the score across all segments.

[3] https://universaldependencies.org/format.html

[4] https://github.com/clarinsi/ComparaTree

[5] This segment length value can be adjusted by the user and is set to 1000 tokens by default.

[6] Several values of n can be defined by the user on input to be extracted in a single run of the comparison process.

2.2.4 UD Label Proportions

The tool also calculates the proportional representation for UD part-of-speech and dependency relation labels in each treebank. This involves calculating the ratio between the number of tokens that are assigned a certain label and the total number of tokens in the treebank. We consider the labels which occur more often in one treebank and for which the difference in proportions in both treebanks is the highest the most typical labels for one treebank with respect to the other. In the case of dependency relation labels, dependency subtypes are not counted together with their basic relation types, but are considered as separate categories. ComparaTree also calculates a chi-square test to determine whether the difference between label frequencies in the two treebanks is statistically significant.

2.2.5 Syntactic Complexity

A variety of different measures have been developed which aim to capture the level of syntactic complexity of a text. ComparaTree focuses on the notion of dependency distance as an indicator of syntactic complexity, supporting the calculation of both the Mean Dependency Distance measure (henceforth *MDD*) as well as the Normalized Dependency Distance (henceforth *NDD*) measure. The MDD measures the average distance between syntactically linked words and is a widely-used method that has been the subject of a number of syntactic complexity studies (Ferrer i Cancho, 2004; Futrell et al., 2015). The NDD is based on a similar principle to the MDD, but also takes into account sentence length during calculation and is consequently found to correlate much less with it (Lei and Jockers, 2020; Terčon, 2024). Both measures are calculated on the level of individual sentences and then averaged over the entire treebank.

2.2.6 Syntactic Diversity

The last dimension of analysis provided by ComparaTree is syntactic diversity. It refers to the number of different syntactic patterns that appear in a corpus (De Clercq and Housen, 2017). In the context of treebank comparison, diversity can be represented by the number of different syntactic trees and subtrees that are present. To this end, ComparaTree uses the aforementioned STARK tool for dependency tree extraction (Krsnik et al., 2024) in order to first extract all relevant syntactic trees from each treebank segment. STARK produces a list of all trees and subtrees in a segment along with their associated absolute and relative frequencies based on a number of configuration settings.[7] Once the extraction is complete, ComparaTree uses these lists to calculate a tree diversity score by dividing the number of unique trees in the segment by the total number of trees in the segment. As in the case of lexical diversity and n-gram diversity, the final syntactic diversity score is obtained by taking the mean of the tree diversity scores for all segments.

2.3 Result Visualization

ComparaTree outputs the results both in the form of various lists and tables pertaining to each analysis level, as well as a concise HTML-format summary which consists of two parts: the first is a result summary table containing all the most important measure calculations. For the second part, ComparaTree produces various diagrams in order to visualize the tendencies present in the analyzed data. Examples are given in Appendix A.

In the cases of lexical diversity, n-gram diversity, and syntactic diversity, histograms are generated for both treebanks which show the number of occurrences of each value of the calculated measure—the above-described TTR, NGD, and tree diversity scores—when measured on the level of segments. Similarly, for the basic average sentence length and syntactic complexity—the MDD and NDD scores—histograms are also generated with the values measured on the level of sentences.

In addition, for UD label proportion analysis, the tool generates a barchart showing the proportions for each analyzed UD label with the labels ordered according to the difference in proportion between the two treebanks, placing the labels that are most typical of each treebank at opposite ends of the barchart.

3 Example Comparison: SSJ-UD vs SST-UD

In this section we present an example comparison performed using the ComparaTree tool. The pair of compared treebanks consists of the Slovenian SSJ UD treebank (Dobrovoljc et al., 2017), which represents a balanced sample of written Slovenian, and the Slovenian SST UD treebank (Dobrovoljc and

[7] The STARK package supports various configuration options for tree extraction with the ability to adjust the desired tree size and the type of label that is taken as the tree node. By default, ComparaTree extracts trees of all sizes and considers UPOS tags as tree nodes. These settings can be adjusted by the user via a special configuration input file.

Nivre, 2016), which represents a balanced sample of spoken Slovenian. Both treebanks were provided to ComparaTree as input in the CoNLL-U file format and all the default levels of lingusitic analysis were included. The default segment length of 1000 tokens was used, while for the n-gram analysis only 3-grams were analyzed during this comparison session. In Appendix A, Table 1 presents the result overview table generated by the SSJ vs SST comparison, while Figures 1–8 show the visualizations generated at each individual level of analysis. While a detailed analysis of the results is beyond the scope of this paper, the results plainly illustrate the value of the tool for conducting such multi-level comparisons, as several clear tendencies can immediately be discerned from a single glance at the result summary.

On the **basic level**, Table 1 shows that, while the SSJ treebank contains on average longer sentences than the SST treebank, the sentence length tends to vary much more in SST than SSJ. In terms of **lexical diversity**, the mean TTR score suggests that the spoken language treebank is much less lexically diverse than the written treebank. A similar tendency can be seen in the results of the **n-gram diversity analysis**, where the NGD score for 3-grams is higher in the SSJ treebank compared to SST, indicating that the written treebank has a higher diversity of 3-grams. The differences in **UD label proportions** suggest that nominal phrases are more typical of the SSJ treebank, as nouns, adjectives and adpositions—which commonly occur within nominal phrases—tend to appear more prominently in SSJ. Conversely, particles, interjections, and adverbs—which are commonly connected with non-propositional lexica and other elements that reflect the flow of discourse—appear more typically in the spoken treebank. As for **syntactic complexity**, the values of the MDD and NDD measures exhibit opposite patterns, as the MDD appears to be higher in the SST treebank, while the NDD is higher in the SSJ treebank. Lastly, on the level of **syntacic diversity**, the tree diversity score values show a higher proportion of unique syntactic trees in the written treebank compared to the spoken treebank, suggesting a higher syntactic diversity.

4 Conclusion

In this article we introduced ComparaTree, a tool for comparative linguistic analysis which produces a multi-level comparison of two treebanks. We presented the various levels of linguistic analysis that ComparaTree offers and exemplified its use and output using two treebanks included in the Universal Dependencies treebank collection.

Many functionalities still remain to be added to ComparaTree, which will improve its analysis capabilities. Presently only the UD label proportion analysis is equipped with a statistical significance test that establishes the statistical significance of the observed patterns. In the future, various methods of statistical significance testing along with effect size calculations should be added to other analysis dimensions as well. Although the current presentation of results offers a good glimpse into the tendencies that can be observed in the data, rigorous statistical methods are required to give additional weight to the findings made using ComparaTree.

Additionally, the tool presently only supports pairwise comparisons of two treebanks. Important insights could be gained from comparing more than two treebanks simultaneously, so support for multiple comparisons should be implemented in the future. Such an expansion should also be accompanied by more advanced visualization techniques, which would shed light on different tendencies present in the data and complement the current assortment of histograms and barcharts.

There is also much room to expand the current inventory of measures calculated and range of analyses performed at each level (also in line with new methods that have recently been proposed, such as in Čibej (upcoming)) as well as the potential to expand into other dimensions of linguistic analysis, such as semantics, discourse analysis, etc. Future improvements to the tool in this regard should be dictated by the demand presented by the target users and the broader computational linguistics community.

Acknowledgements

The work described in the present paper was made possible by the Young researchers program, the SPOT project (A Treebank-Driven Approach to the Study of Spoken Slovenian, Z6-4617), the Language Resources and Technologies for Slovene research program (project ID: P6-0411), and the PoVeJMo research project (Adaptive Natural Language Processing with the Help of Large Language Models), all financed by the Slovenian Re-

search and Innovation Agency (ARIS), and the CLARIN.SI research infrastructure.

References

Christopher MJ André, Helene FL Eriksen, Emil J Jakobsen, Luca CB Mingolla, and Nicolai B Thomsen. 2023. Detecting AI authorship: Analyzing descriptive features for AI detection. In *Rome, 7th workshop on natural language for artificial intelligence. NL4AI*.

Aleksandrs Berdicevskis, Çağrı Çöltekin, Katharina Ehret, Kilu von Prince, Daniel Ross, Bill Thompson, Chunxiao Yan, Vera Demberg, Gary Lupyan, Taraka Rama, and Christian Bentz. 2018. Using Universal Dependencies in cross-linguistic complexity research. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 8–17, Brussels, Belgium. Association for Computational Linguistics.

Jaka Čibej. upcoming. A computational method for analyzing syntactic profiles: The case of the ELEXIS-WSD parallel sense-annotated corpus. In *SyntaxFest 2025*, Ljubljana, Slovenia.

Bastien De Clercq and Alex Housen. 2017. A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *The Modern Language Journal*, 101(2):315–334.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Kaja Dobrovoljc. 2022. Spoken language treebanks in Universal Dependencies: an overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

Kaja Dobrovoljc, Tomaž Erjavec, and Simon Krek. 2017. The Universal Dependencies treebank for Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 33–38, Valencia, Spain. Association for Computational Linguistics.

Kaja Dobrovoljc and Joakim Nivre. 2016. The Universal Dependencies treebank of spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1566–1573, Portorož, Slovenia. European Language Resources Association (ELRA).

Ramon Ferrer i Cancho. 2004. Euclidean distance between syntactically linked words. *Phys. Rev. E*, 70:056135.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Luka Krsnik, Kaja Dobrovoljc, and Marko Robnik-Šikonja. 2024. Dependency tree extraction tool STARK 3.0. Slovenian language resource repository CLARIN.SI.

Lei Lei and Matthew L. Jockers. 2020. Normalized dependency distance: Proposing a new measure. *Journal of Quantitative Linguistics*, 27(1):62–79.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Philip M McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021. How universal is genre in Universal Dependencies? In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 69–85, Sofia, Bulgaria. Association for Computational Linguistics.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, 57(10):265.

Dmitry Nikolaev, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Maria Saeboe, and Omri Abend. 2020. Fine-grained analysis of cross-linguistic syntactic divergences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1159–1176, Online. Association for Computational Linguistics.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? *Preprint*, arXiv:2309.05196.

Luka Terčon. 2024. Uporaba šestih mer skladenjske kompleksnosti za primerjavo jezika v govornem in pisnem korpusu. *Jezikovne tehnologije in digitalna humanistika*, pages 668–686.

Mu Yang and Haitao Liu. 2025. QuanSyn: A package for quantitative syntax analysis. *Journal of Quantitative Linguistics*, 0(0):1–18.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Matthew Andrews, and 633 others. 2024. Universal Dependencies 2.15. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Example ComparaTree Output

Metric	**SSJ**	**SST**
Basic		
Total # of tokens	267,097	98,393
Total # of sentences	13,435	6,108
Average tokens per sentence	19.881	16.109
Standard deviation of tokens per sentence	12.766	17.881
Lexical Diversity		
Average Segmental Type-Token Ratio	0.482	0.297
Segmental Type-Token Ratio standard deviation	0.039	0.050
N-Gram Diversity		
Average Segmental 3-gram Diversity Score	0.995	0.984
Segmental 3-gram Diversity Score standard deviation	0.006	0.010
UD Label Proportions		
Largest part-of-speech tag proportion differences	NOUN – 0.10 ADJ – 0.05 ADP – 0.03 PROPN – 0.02	PUNCT – 0.08 PART – 0.04 INTJ – 0.03 ADV – 0.03
Largest dependency relation proportion differences	nmod – 0.05 amod – 0.05 case – 0.04 obl – 0.02	punct – 0.08 advmod – 0.03 discourse – 0.03 root – 0.01
Syntactic Complexity		
Average Mean Dependency Distance	2.572	2.738
Mean Dependency Distance standard deviation	0.925	1.099
Average Normalized Dependency Distance	1.146	0.850
Normalized Dependency Distance standard deviation	0.509	0.452
Syntactic Diversity		
Average Segmental Tree Diversity Score	0.730	0.689
Segmental Tree Diversity Score standard deviation	0.033	0.052

Table 1: Table showing a summary of every measure calculated at each level of linguistic analysis as provided by ComparaTree for the comparison between the SSJ and SST UD treebanks. The *UD Label Proportions* subdivision presents the four labels for which the proportion difference between the two treebanks is the greatest, thus presenting the labels that are most typical of one treebank with respect to the other. The absolute values of the differences are provided next to the label names.

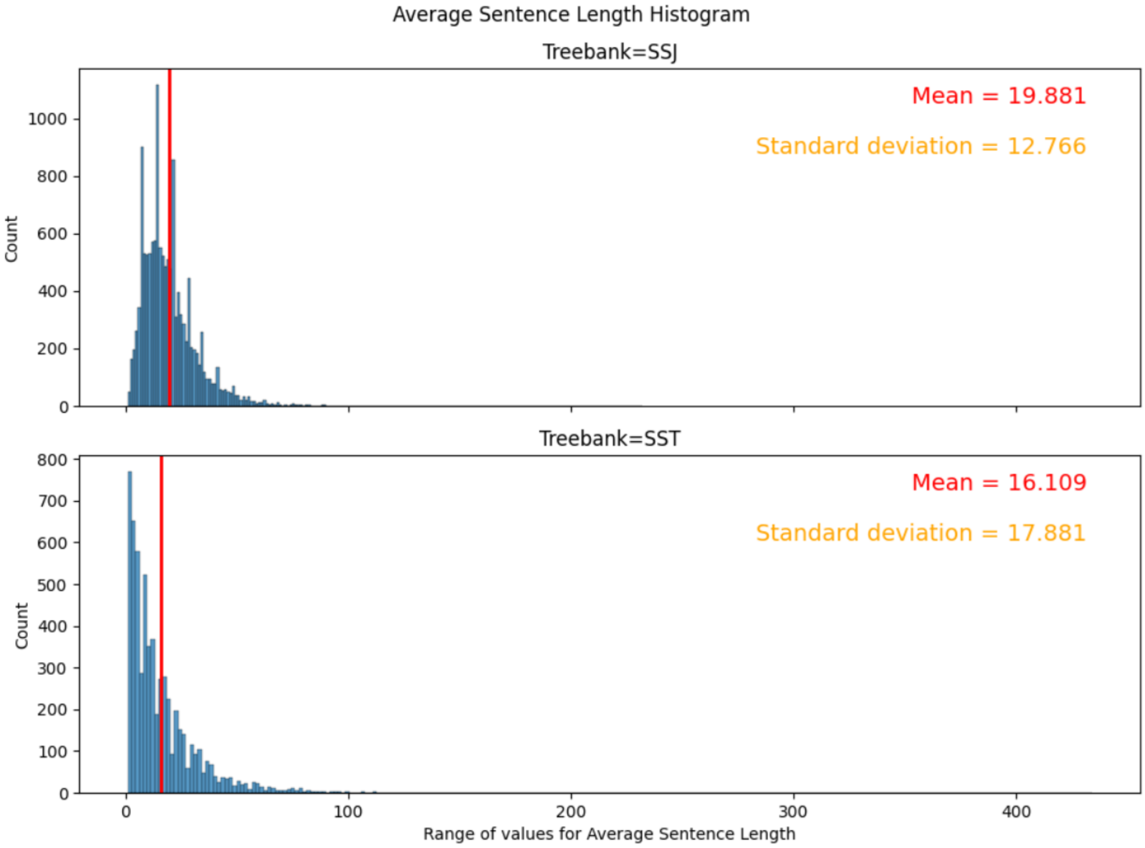

Figure 1: Histogram showing the frequency distribution for the sentence length in the number of tokens for both treebanks. The x axis represents the range of values for the sentence lengths. The blue bars represent the number of observations for each value of the measure. The red vertical line represents the mean sentence length.

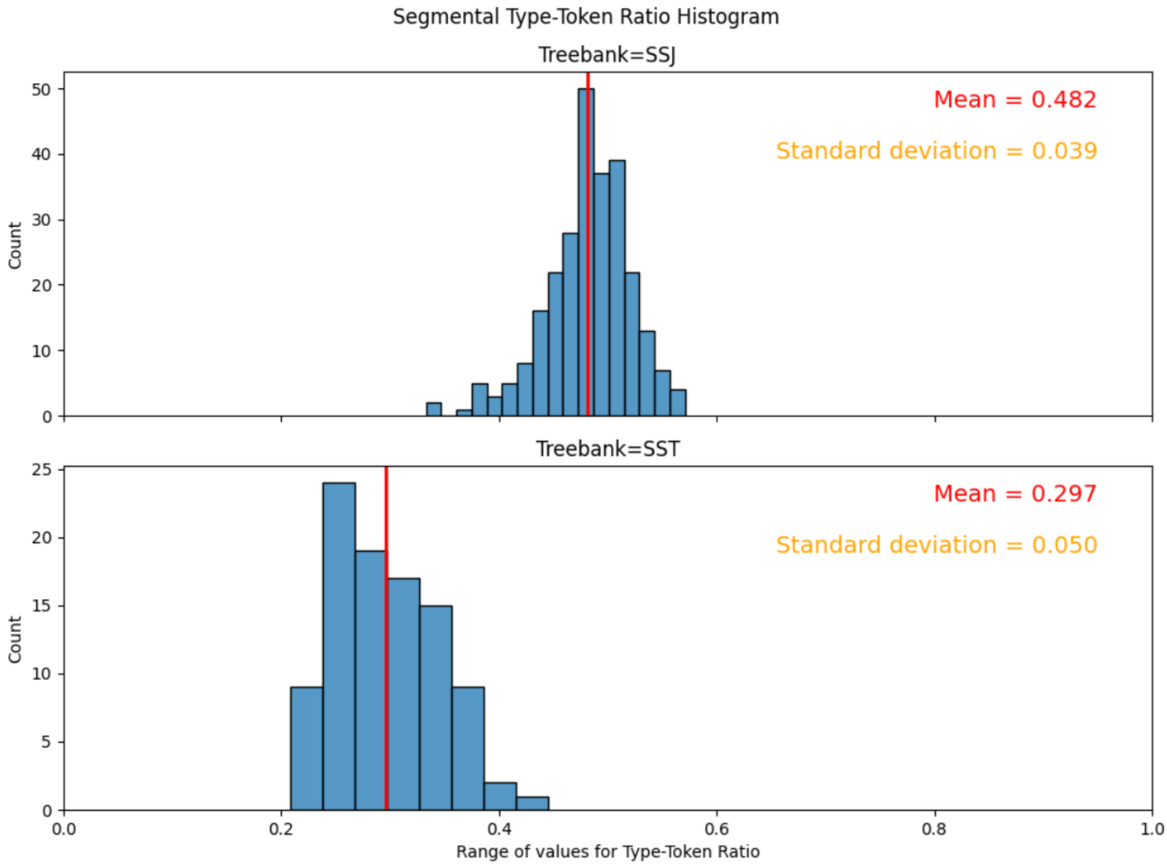

Figure 2: Histogram showing the frequency distribution for the per-segment Type-Token Ratio in both treebanks. The x axis represents the range of values for the Type-token Ratio. The blue bars represent the number of observations for each value of the measure. The red vertical line represents the mean of the Type-Token Ratio over all segments.

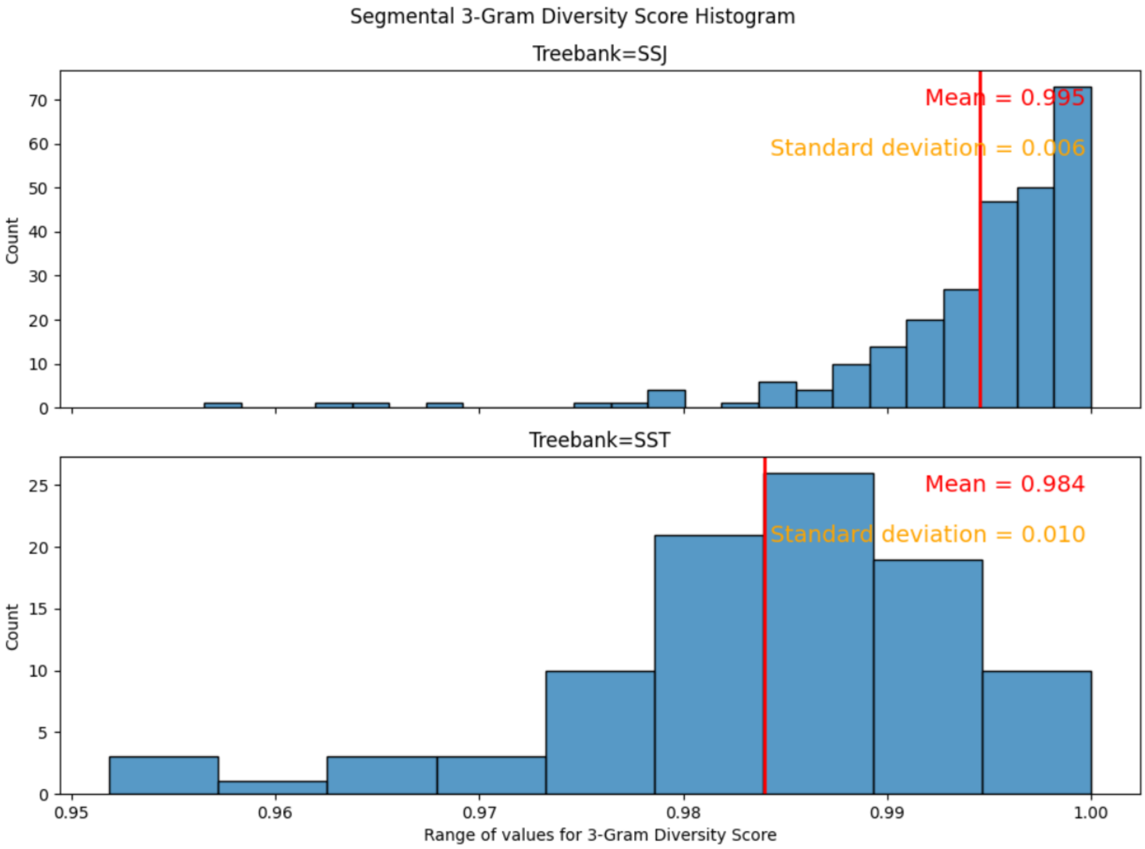

Figure 3: Histogram showing the frequency distribution for the per-segment 3-Gram Diversity Score in both treebanks. The x axis represents the range of values for the 3-Gram Diversity Score. The blue bars represent the number of observations for each value of the measure. The red vertical line represents the mean of the 3-Gram Diversity Score over all segments.

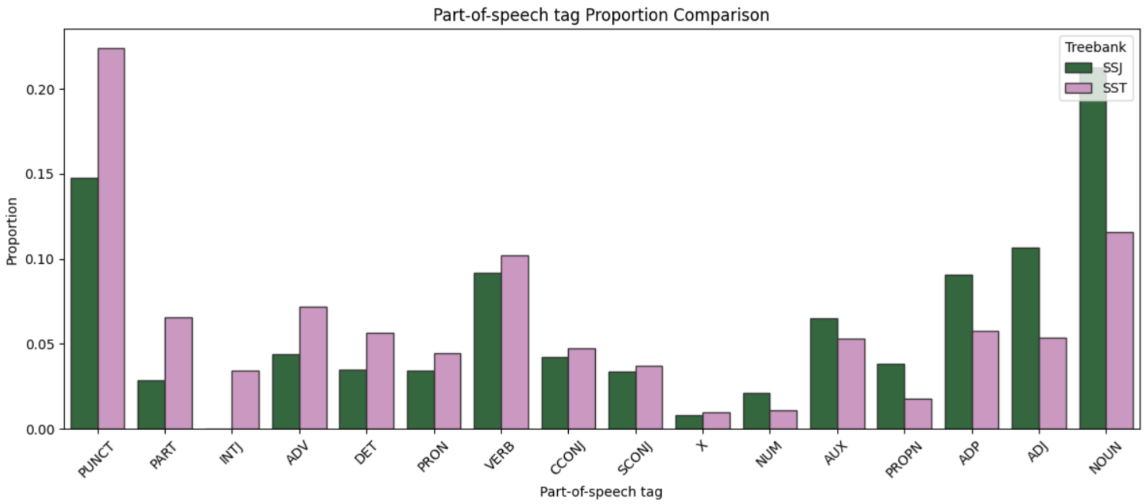

Figure 4: Barchart showing the proportion of every UPOS tag for each treebank. The ordering of the tags is determined by the difference between the tag proportions between the two treebanks, with the tags on the left end being more typical (i.e. occurring with a higher proportion difference) of SST, while the tags on the right end being more typical of SSJ.

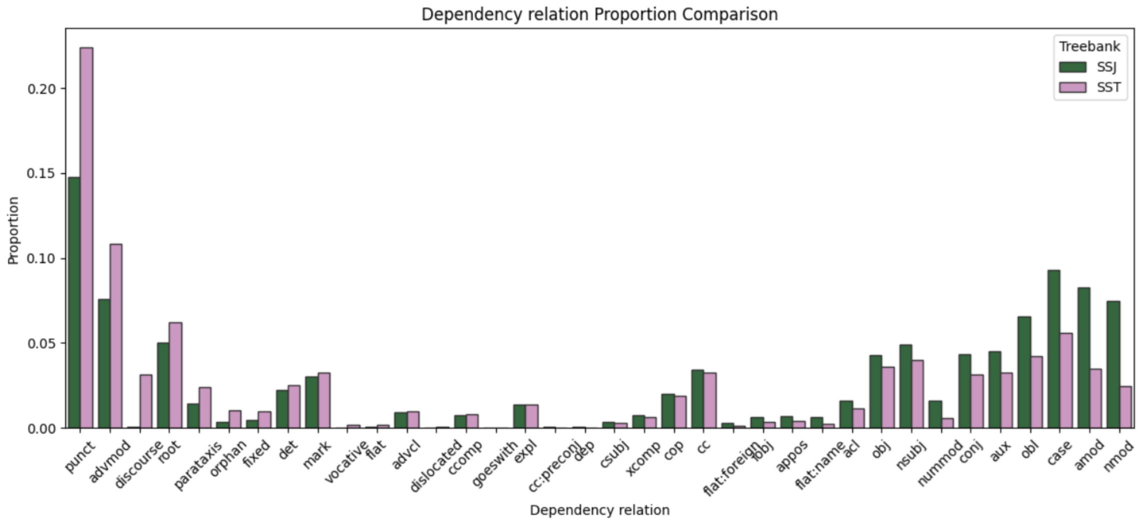

Figure 5: Barchart showing the proportion of every dependency relation tag for each treebank. The ordering of the tags is determined by the difference between the tag proportions between the two treebanks, with the tags on the left end being more typical (i.e. occurring with a higher proportion difference) of SST, while the tags on the right end being more typical of SSJ.

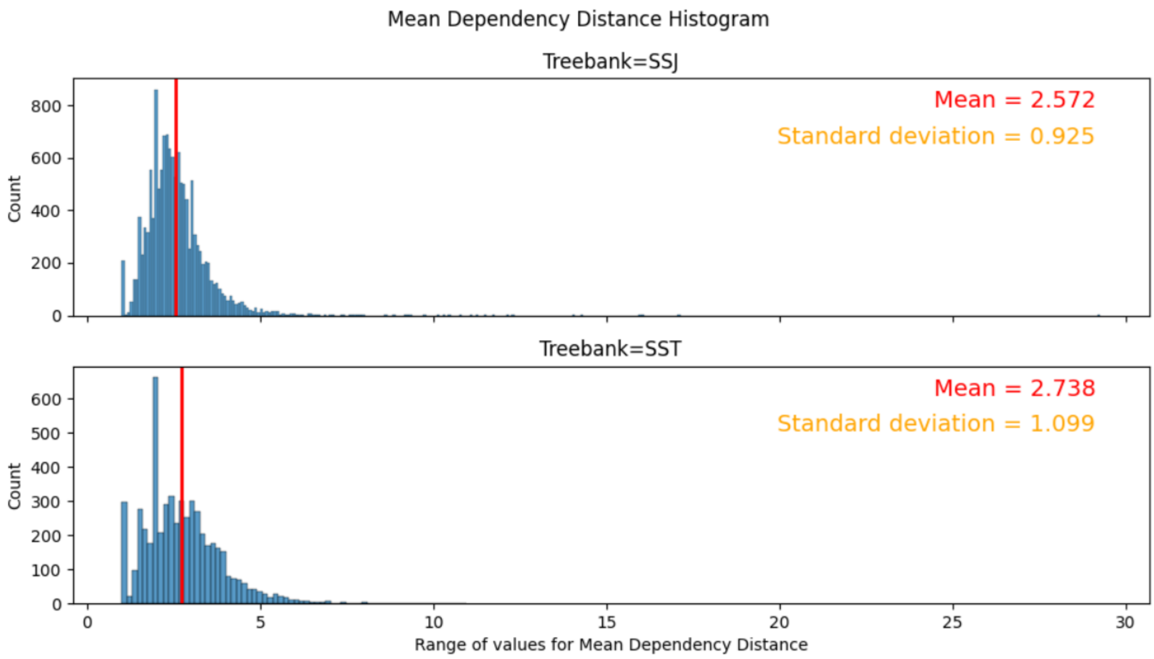

Figure 6: Histogram showing the frequency distribution for the per-sentence Mean Dependency Distance in both treebanks. The x axis represents the range of values for the Mean Dependency Distance. The blue bars represent the number of observations for each value of the measure. The red vertical line represents the mean of the Mean Dependency Distance over all sentences.

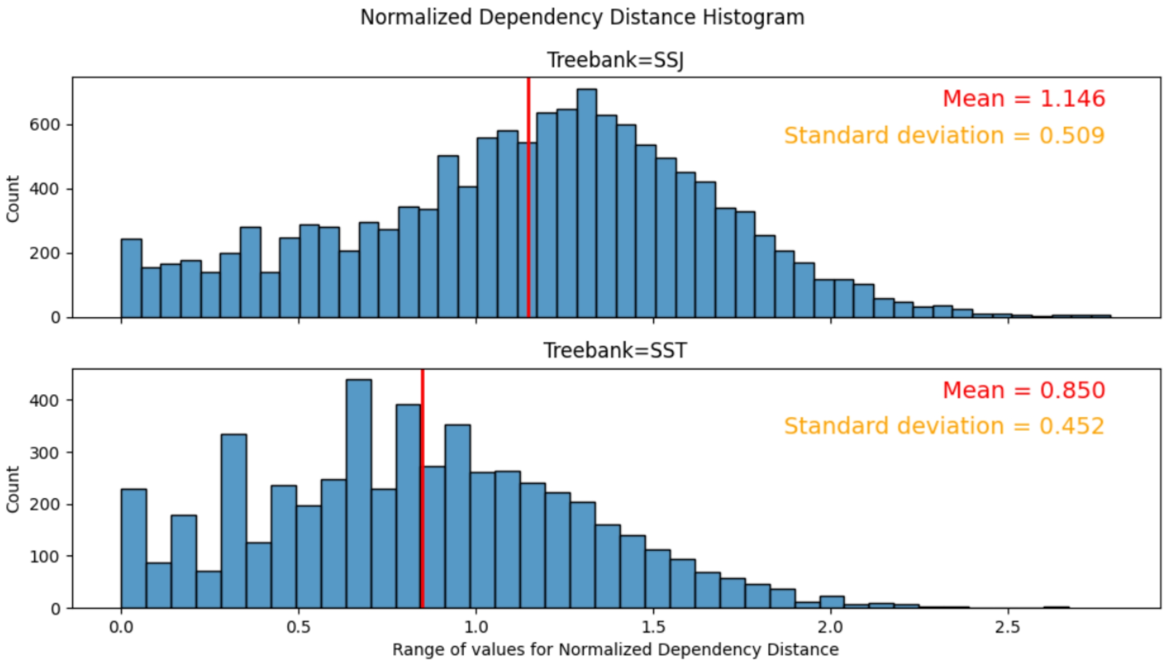

Figure 7: Histogram showing the frequency distribution for the per-sentence Normalized Dependency Distance in both treebanks. The x axis represents the range of values for the Normalized Dependency Distance. The blue bars represent the number of observations for each value of the measure. The red vertical line represents the mean of the Normalized Dependency Distance over all sentences.

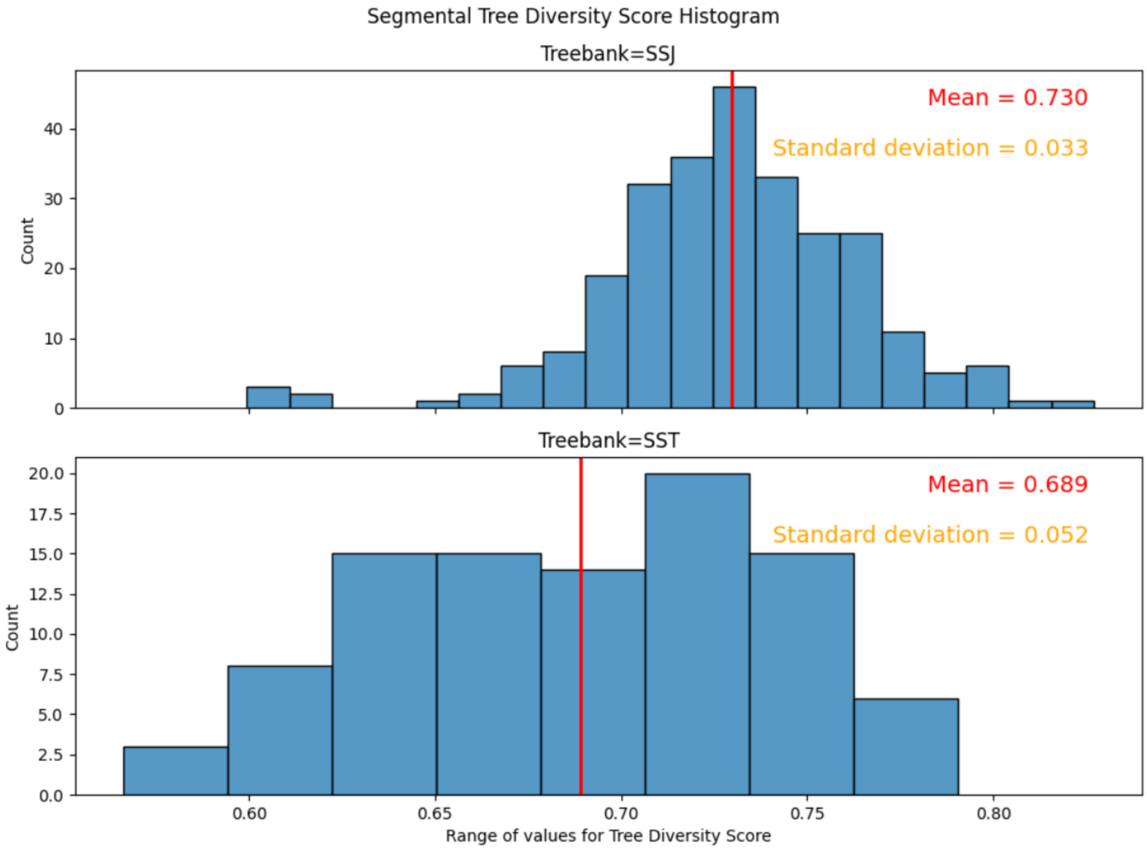

Figure 8: Histogram showing the frequency distribution for the per-segment Tree Diversity Score in both treebanks. The x axis represents the range of values for the Tree Diversity Score. The blue bars represent the number of observations for each value of the measure. The red vertical line represents the mean of the Tree Diversity Score over all segments.