Analysis of Subject-Object-Verb Word Order Patterns in Slovenian Spoken and Written Corpora

Final Project Report

Nives Hüll

Faculty of Humanities and Social Sciences, University of Zagreb

January 19, 2025

Contents and List of Figures

Contents

List of Figures

1 Summary

This project initially aimed to enhance large language models (LLMs) for advanced grammatical analysis of multilingual corpora, with a focus on Slovenian syntax. The qualitative analysis investigated spoken and written Slovenian word order patterns in relation to typological features outlined in the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013b). Two key questions were addressed: the alignment of Slovenian spoken word order with WALS typologies and the differences between spoken and written modalities.

Utilizing Universal Dependencies (UD) (Zeman et al., 2024) frameworks and tools like STARK (Krsnik et al., 2024), the analysis revealed that spoken Slovenian exhibits greater syntactic variability compared to written Slovenian, which aligns more closely with the SVO pattern. Significant statistical differences were identified, emphasizing the influence of modality and pragmatics on syntax.

Outcomes included a refined methodology for word order analysis, quantitative findings on Slovenian syntax, and a prototype interactive table for typological data. This research highlights the importance of corpus-based approaches for linguistic typology and provides a foundation for future studies on cross-linguistic syntactic variation and pragmatic influences in language.

2 Motivation and Background

This project, part of the Gravitacija[1] – LLM4DH initiative (Task T2.3), aims to enhance LLMs for grammatical analysis across languages. My expertise in Universal Dependencies aligned with the project, initially focusing on WALS for its reputation and documentation. However, discrepancies between WALS data and corpus-based findings for spoken Slovene highlighted the need to bridge typological databases with real-world language data.

3 Research Questions and Hypotheses

The study investigates three key research questions:

1. To what extent do spoken Slovenian word order patterns align with WALS typological features? **Hypothesis:** Spoken Slovenian will deviate from WALS word order typologies due to its greater variability in Subject-Verb-Object arrangements.

[1] More about the project can be found here.

2. Are there significant differences in word order patterns between spoken and written Slovenian in relation to WALS typological features? **Hypothesis:** Spoken Slovenian will display greater variability in Subject-Object-Verb word order patterns compared to written Slovenian, which adheres more closely to WALS's SVO classification. Do WALS's binary categories sufficiently capture the syntactic nuances

3. Do WALS's binary categories sufficiently capture the syntactic nuances found in spoken Slovene? **Hypothesis:** WALS's binary categorization lacks the nuance needed to represent complex syntactic patterns in spoken Slovene, which often display more variability than WALS's categories allow.

4 Scope and Limitations

The project consists of two main phases. The first involved creating a comprehensive table of WALS features and generating linguistic questions for each feature, followed by translating these questions into UD queries using the STARK tool. The resulting output includes a Google Sheets table[2] and a demo of an interactive feature table, which was a personal experimental addition beyond project expectations.

The second phase focused on analyzing the subject-object-verb word order in Slovenian using two corpora: SSJ[3] and SST[4]. While both corpora are publicly available and pre-annotated, the analysis required additional processing, such as normalizing punctuation in SSJ to align with SST's format. Limitations included the labor-intensive manual extraction of WALS features and the challenges of balancing broad linguistic inquiries with the need for focused, feasible research within the project's timeframe.

5 Team Roles

While this project was carried out individually, I received support from colleagues Kaja Dobrovoljc, who leads the work package, and Luka Terčon, who provided guidance on using STARK.

[2] The final table for all WALS features can be found at this link.

[3] The SSJ corpus and other related data are available at this link.

[4] The SST corpus and other related data are available at this link.

6 Methodology

Data Collection

Phase 1 involved manually extracting linguistic features from WALS due to challenges with automating data from its dynamic website and disorganized GitHub repository. Despite being time-consuming, the manageable data size made this feasible. Phase 2 utilized two publicly available Slovenian corpora, SSJ and SST, pre-annotated in the Universal Dependencies framework and ethically suitable for research.

Theoretical Background

UD is a framework for consistent syntactic annotation across languages (Nivre et al., 2020). Slovenian UD treebanks, SSJ and SST, include unique features like discourse:filler and clitic annotations, supporting comparative syntax studies (Dobrovoljc et al., 2023). While the WALS classifies languages by static word order patterns such as SVO (Dryer and Haspelmath, 2013a), corpus-based approaches provide more nuanced insights (Baylor et al., 2024). Studies using corpus data reveal gradient syntax patterns and inconsistencies in word order classifications (Choi et al., 2021, Yan and Liu, 2021). Slovenian, a flexible South Slavic language, shows greater syntactic variability in speech than in writing, with spoken Slovenian often lacking a dominant order compared to the SVO dominance in written forms (Zuljan and Kumar, 2019, Groselj, 2004, Stegovec and Marvin, 2012, Choi et al., 2021). This study investigates these modality-based differences.

Data Processing and Analysis

The analytical part of the project comprised two phases. Phase 1 required minimal processing, handled with Python scripts[5] sourced from StackOverflow or generated via ChatGPT. Phase 2 focused on analyzing subject-object-verb word order using pre-annotated SSJ and SST corpora, with additional preprocessing to normalize SSJ punctuation to match SST. STARK extracted dependency trees using the query `upos=VERB >nsubj _ >obj|iobj _`, identifying verbs with nominal subjects and direct or indirect objects. The data was cleaned by removing unnecessary columns, assigning correct word order patterns, and merging both corpora into a single dataset for further analysis.

[5] All scripts are available in the scripts subfolder on the GitHub repository. See: this link.

Categorization and Pipeline Preparation

Each example was manually categorized into one of six word order patterns (SVO, SOV, VSO, OSV, OVS, VOS) since STARK output lacked labels. The annotated dataset was then formatted for quantitative analysis using a Python pipeline[6].

7 Methodology

The methodology was carried out in two main phases and followed a structured process to ensure clarity and precision in the analysis.

The analysis followed six steps: word order patterns were normalized and visualized as proportions using bar plots, and dominant orders were identified based on strict frequency criteria, with patterns labeled as ”No Dominant Order” (NDO) if no dominant pattern emerged. Proportional distributions between SST and SSJ were compared using cosine similarity, and differences were visualized with heatmaps. A chi-square test validated the statistical significance of word order differences. Direct and indirect relations were handled using Python scripts to clean data, assign correct word order patterns, and merge outputs into a unified dataset for analysis.

The project involved 2-3 collaborative meetings to refine methodologies and address challenges. Phase 1 focused on extracting WALS features, performing quick Python-based statistics, and creating a comprehensive feature table. Phase 2 included defining a linguistic feature, conducting a literature review, developing a STARK query for word order patterns, preprocessing data, assigning patterns, and performing quantitative analysis with visualizations. The findings were summarized in a paper, and an interactive WALS feature table demo was created as a supplementary tool.

8 Decision-Making Process

Key decisions ensured the project remained focused and feasible. WALS was chosen over Grambank for its intuitive structure and subcategories, enabling detailed analysis. Manual extraction of WALS features was preferred due to technical challenges with automation. WALS Feature 82A (Order of Subject and Verb)[7] was selected for its relevance to Slovenian word order. For dependency extraction, STARK was favored over Python libraries for its user-friendly interface and efficiency, despite limited output flexibility. These choices balanced feasibility, scope, and alignment with project goals.

[6] All used Python packages are listed on the GitHub repository in the `README.md` file.

[7] More about the feature can be found at this link.

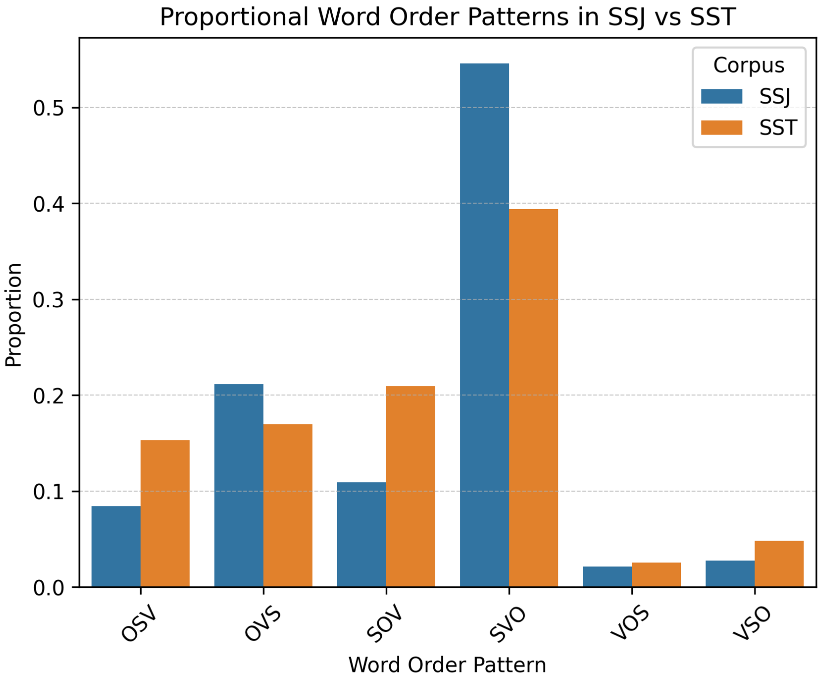

Figure 1: Distribution of Word Order Patterns in Written and Spoken Slovenian

9 Results

Frequency and Proportional Distribution

We compared word order distributions between spoken Slovenian (SST) and written Slovenian (SSJ) by calculating the normalized proportions for each word order pattern, as shown in Figure 1. The SVO (Subject-Verb-Object) pattern is dominant in SSJ, accounting for 54.6% of all occurrences, reflecting the formal and structured nature of written language. In contrast, SST exhibits greater variability, with SVO at 39.4% and marked word orders like SOV (20.9%) and OSV (15.3%) occurring more frequently. This variability indicates the flexibility of spoken language, where word orders are often context-driven. The Chi-Square Test of Independence revealed significant differences ($\chi^2 = 133.59, p < 0.001$) between the corpora, confirming that modality plays a key role in shaping word order preferences in Slovenian.

Evaluation of Hypothesis 1: Supported. The SST corpus displays more variability, with greater frequencies of marked word orders, which deviates from the WALS typological norm of SVO.

Dominant Word Orders

The analysis confirmed that SVO is the dominant word order in SSJ, meeting the criterion of being at least twice as frequent as the next most common pattern. However, the

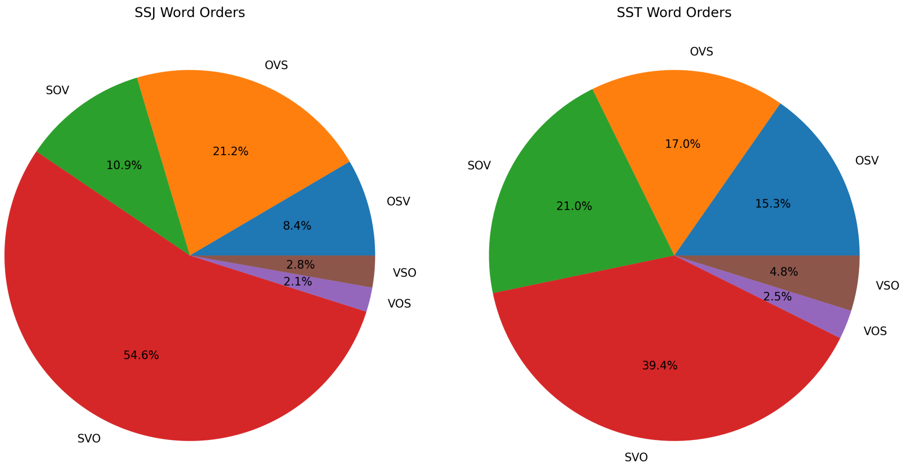

Figure 2: Frequency and Proportional Distribution in Written and Spoken Slovenian

SST corpus lacked a clear dominant order, leading to its classification as No Dominant Order (NDO).

Evaluation of Hypothesis 2: Supported. While SVO dominates in SSJ, SST shows more variability with no clear dominant order.

Distributional Comparisons

We employed distributional metrics to compare the word order patterns across SST and SSJ:

- **Jensen-Shannon Divergence:** 0.148, indicating moderate divergence in proportional patterns.
- **Entropy:** SST (1.52) has higher variability compared to SSJ (1.29), reflecting the greater flexibility in spoken syntax.
- **Euclidean Distance:** 0.20, indicating a noticeable but not overwhelming difference between the distributions.
- **Pearson Correlation:** 0.93, and **Spearman Rank Correlation:** 0.94, showing strong alignment in the overall ranking of patterns across the two modalities.

Evaluation of Hypothesis 3: Partially supported. While SSJ and SST show strong alignment in their overall ranking of word orders, SST displays greater variability and markedness, diverging more significantly from the typological norm.

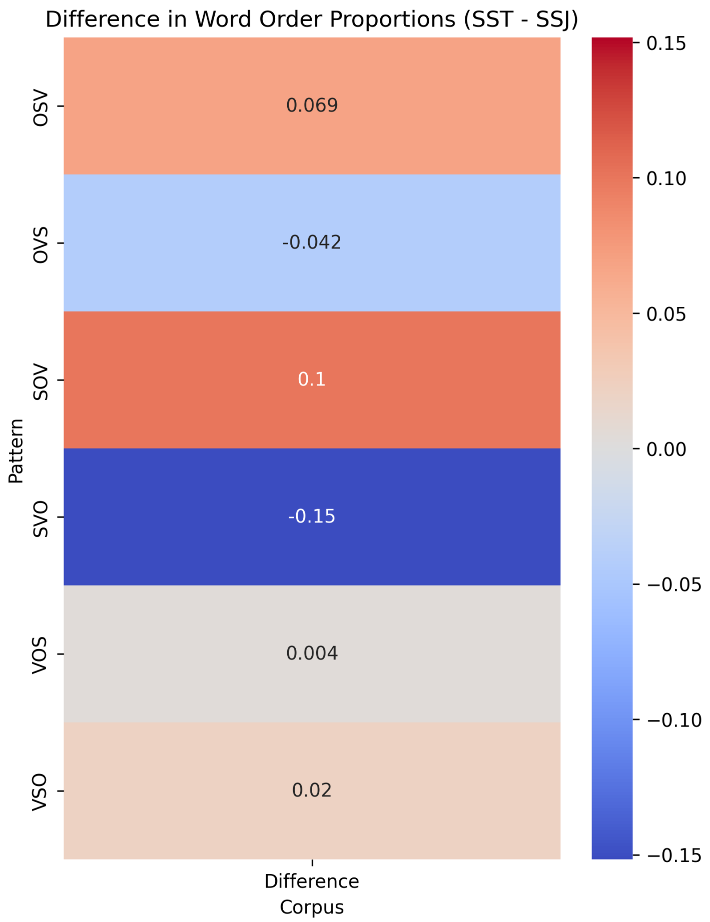

Figure 3: Comparison of Word Order Variability in Written and Spoken Slovenian

Differences in Proportions

Figure 3 illustrates the proportional differences between SST and SSJ. Positive values indicate patterns more frequent in SST, while negative values highlight those more common in SSJ. The results reveal that marked word orders like OSV and SOV are more frequent in SST, reflecting the flexible, context-driven nature of spoken language. In contrast, SVO is the dominant pattern in SSJ, emphasizing the structured nature of written language. Less common patterns, such as VOS and VSO, show only minor differences, indicating their limited role in both corpora.

10 Discussion and Analysis

This study shows how modality influences Slovenian syntax, following cross-linguistic patterns. Written language favors unmarked SVO for clarity, while spoken language allows more flexibility, similar to German (Kempen and Harbusch, 2005).

Marked word orders like SOV and OSV are more common in spoken Slovenian, driven by pragmatics for emphasis or topicalization. Similar trends in Czech and Russian highlight how flexible word orders manage information structure (Sgall et al., 1986, Slioussar, 2011), aligning with Greenberg's Universal 6 and underscoring the need for ongoing corpus-based typological research (Baylor et al., 2024). The frequent use of marked word orders in SST demonstrates how pragmatics shapes syntax, supporting Lambrecht's (Lambrecht, 1994) view on the interaction between syntax and discourse in spoken contexts.

The first phase of the project presented several challenges. One of the most significant obstacles was data extraction, as scraping WALS data proved difficult. The dynamic nature of the website forced me to resort to a manual process, which was time-consuming and less efficient than automating the extraction. Tool limitations also posed challenges, particularly with STARK. While it was effective for analyzing direct relations, it struggled with handling indirect relations, such as unconnected subject-object pairs. Additionally, implementing probabilistic modeling proved to be complex, suggesting that more advanced techniques may be needed in future work to improve the accuracy and efficiency of the models. I also encountered some difficulties while deploying the Streamlit app.

This study advocates for ongoing, corpus-based typological research, challenging static classifications. While confirming SVO dominance in written Slovenian, it highlights the flexibility of spoken language, shaped by modality and pragmatics. Word order is shown to be a gradient phenomenon, influenced by syntax, pragmatics, and discourse. Using Universal Dependencies and detailed statistical analyses, the study offers a replicable framework for analyzing word order variation across languages and modalities. The findings are relevant to other morphologically rich, free word order languages like Russian, Finnish, and Hungarian, where modality-specific differences also occur, emphasizing the need for continuous typological research.

11 Conclusion

Through this project, I gained a deeper understanding of linguistic typology, particularly the nuances of Slovenian word order patterns in spoken and written modalities, and the practical applications of the Universal Dependencies framework. I honed my skills in data processing, cleaning, and statistical analysis using tools like STARK and Python, while also overcoming challenges such as manual data extraction and tool limitations. The project enhanced my problem-solving abilities, adaptability, and proficiency in project management, particularly through systematic documentation and collaborative discussions with colleagues. Beyond technical skills, I learned to bridge theoretical and applied linguistics, underscoring the importance of corpus-based research to complement traditional typological resources like WALS. This experience highlighted areas for future growth, such as advanced coding skills and expanding the analysis to other linguistic features and languages, providing a strong foundation for future work in digital linguistics.

Building on the findings of this study, future research could explore sociological and discourse-oriented aspects, focusing on the pragmatic functions of marked word orders like OSV and SOV in spoken Slovenian. Expanding the analysis to informal writing (e.g., blogs, social media) or formal speech (e.g., political addresses) could offer insights into how word order interacts with context and discourse conventions. Cross-linguistic comparisons within the South Slavic family could reveal regional and modality-driven word order variation, contributing to a broader understanding of Slovenian word order in cultural and communicative contexts. Additionally, future contextual analysis could integrate qualitative insights, investigating how discourse markers, ellipsis, or interruptions influence word order in spoken language, further enriching the exploration of how language structure adapts in real-time communication. On the other hand, linguistic research could also examine other features of word order in WALS, such as the positioning of objects, obliques, adjectives, genitives, numerals, and relative clauses. The placement of question particles, interrogative phrases, and negative morphemes could also provide valuable insights into Slovenian syntax. Although these directions are not direct extensions of this study, they represent underexplored areas that could advance both linguistic typology and discourse analysis.

12 Acknowledgements

ChatGPT was used to generate and debug Python code for data extraction and preprocessing and to refine the grammar and language style of the documentation. Additionally, Perplexity was employed to assist in the literature review by identifying and summarizing relevant sources. All research design, data analysis, and conclusions were independently reviewed and finalized by me and my colleagues.

References

Baylor, R., et al. (2024). Multilingual gradient word-order typology from universal dependencies. *Journal of Linguistic Research*, *35*, 215–230.

Choi, Y., et al. (2021). Cross-linguistic insights from dependency treebanks. *Language Science*, *43*, 220–238.

Dobrovoljc, K., et al. (2023). Expanding the slovenian-ssj treebank with ssj500k and elexis-wsd. *Slovene Linguistics Review*, *29*, 102–118.

Dryer, M. S., & Haspelmath, M. (2013a). *World atlas of language structures*. Oxford University Press.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013b). *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology. Retrieved January 19, 2025, from https://wals.info

Groselj, A. (2004). *Stylistic word order in slovenian literature*. Slovene Academic Press.

Kempen, G., & Harbusch, K. (2005). Syntactic flexibility in german: A quantitative analysis. *Journal of Germanic Linguistics*, *17*, 121–140.

Krsnik, L., Dobrovoljc, K., & Robnik-Šikonja, M. (2024). Dependency tree extraction tool STARK 3.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1958

Lambrecht, K. (1994). *Information structure and sentence form*. Cambridge University Press.

Nivre, J., et al. (2020). Universal dependencies: A cross-linguistic syntactic annotation framework. *Computational Linguistics*, *46*, 115–135.

Sgall, P., et al. (1986). *The syntax of sentence structure in czech*. Charles University Press.

Slioussar, N. (2011). Flexible word order in russian: A typological perspective. *Slavic Linguistics*, *22*, 233–247.

Stegovec, M., & Marvin, T. (2012). Ditransitive constructions in slovenian: A corpus-based study. *Syntax and Semantics*, *24*, 89–103.

Yan, X., & Liu, Y. (2021). Probabilistic word order patterns in universal dependencies. *Linguistic Typology*, *25*, 100–115.

Zeman, D., Nivre, J., Abrams, M., & Ackermann, E. e. a. (2024). Universal dependencies 2.14 [LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics, Charles University]. http://hdl.handle.net/11234/1-5502

Zuljan, I., & Kumar, M. (2019). Word order and adjective modification in spoken slovenian. *Journal of Slavic Linguistics*, *18*, 145–162.