Faculty of Humanities and Social Sciences, University of Zagreb
Digital Linguistics Project 2025, Final Report

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

Nives Hüll

Abstract

This report details the work undertaken to analyze Slovenian word order using Universal Dependencies and the World Atlas of Linguistic Structures database. The project was structured into several phases: feature extraction and data table preparation, literature review and statistical analysis, and the drafting of the paper. The primary focus was on identifying linguistic patterns in Slovenian word order, particularly the relationship between subject, object, and verb.

Keywords
Word Order, Slovenian, Linguistic Typology, Universal Dependencies, Corpus Analysis

Repository: https://github.com/hulln/UD-WALS-Linguistic-Patterns
Advisors: Kaja Dobrovoljc, Luka Terčon, Petra Bago

1. Introduction

This report provides a comprehensive overview of the work conducted in the analysis of Slovenian word order using Universal Dependencies (UD) and the World Atlas of Linguistic Structures (WALS) database. The central aim of this project was to explore linguistic patterns in Slovenian word order, specifically focusing on subject-verb-object (SVO) order, and to assess the applicability of probabilistic modeling techniques to typological data.

The project provided an opportunity to apply both theoretical linguistic concepts and practical computational methods, resulting in valuable insights into Slovenian word order and methods for linguistic data analysis using UD.

2. Data Preparation

2.1 Table Creation

The first phase of the project focused on extracting linguistic features from the WALS database, initially concentrating on those relevant to Slovenian. These features were organized into categories based on their typological properties, with particular emphasis on the Word Order category, which had the highest representation in Slovenian. While the original plan was to create a feature table exclusively for Slovenian, this approach was later expanded to encompass features from all languages in the WALS database for a more comprehensive analysis.

Throughout this phase, I primarily relied on my previous knowledge of UD. However, as a non-comparative linguist, understanding the intricacies of each feature and its linguistic background took longer than anticipated. Although this phase was not central to my university project, it was crucial for my colleagues in Slovenia, who are developing Large Language Models (LLMs) designed to interpret UD queries more effectively. These models aim to simplify the use of UD queries for researchers and the public. Given that current tools like GrewMatch, Drevesnik, and STARK are specialized and not user-friendly, integrating UD knowledge into LLMs would greatly enhance accessibility and usability.

The final feature table, which includes the UD queries and linguistic questions, is available here. Additionally, a simple interactive app based on this table can be found here.

2.2 Data Extraction and Expansion

One of the major challenges encountered during this phase was extracting the feature data from the WALS database. Initially, I planned to automate the extraction process using web scraping. However, this approach proved difficult due to the data being embedded within a dynamic JavaScript element on the WALS website. As a result, I resorted to a manual process, which involved copying and pasting the relevant data into a '.txt' file and manually formatting it for analysis.

While this process was time-consuming, it allowed me to create a comprehensive list of features from WALS and focus specifically on those related to Slovenian. Initially, I aimed to generate a feature table for Slovenian alone. However, as I progressed, I decided to expand the table to include features for all languages in the WALS database. This expansion provided a broader context and allowed for more robust comparisons across languages. The final feature table contained 192 features, of which 43 (22.4%) were applicable to Slovenian.

The feature distribution suggested that Slovenian has strong representation in Word Order and Verbal Categories, with minimal presence in other categories, such as Phonology and Lexicon. These findings were important in shaping the direction for the subsequent phases of the project.

3. Theoretical Background

The second phase included reviewing existing research and doing my own analysis. I focused on the relationship between subject, object, and verb in two Slovenian corpora: SSJ (written) and SST (spoken). Both corpora are available on the UD website and already pre-annotated, which made the further work easier, but I had to clean up the punctuation in SSJ to make it comparable to SST that doesn't include punctuation.

3.1 Literature Review

The theoretical background of this project work is grounded in key linguistic frameworks and prior research. UD is a syntactic annotation framework that facilitates cross-linguistic analysis by encoding grammatical relations such as subject and object, and has been widely adopted for typological research. Two UD treebanks have been developed for Slovenian: Slovenian-SSJ and Slovenian-SST, with specific adaptations to reflect language characteristics like filled pauses and clitic clusters. WALS provides a typological classification of languages, with a focus on the Subject-Verb-Object (SVO) pattern, but it has limitations in capturing intra-language variations, as seen in Slovenian. Recent studies advocate for continuous typological representations, such as those from corpus-based analysis, to better account for linguistic diversity. Corpus-based studies have shown the value of dependency treebanks in revealing inconsistencies in word order patterns across languages, stressing that language should be viewed as a gradient phenomenon rather than fixed categories. Slovenian, with its flexible word order due to rich inflectional morphology, shows greater syntactic flexibility in spoken forms compared to written ones, as demonstrated by previous studies. For example, in spoken Slovenian, adjective modifiers can follow the noun, and word order in ditransitive constructions can vary. These variations are not fully captured by WALS's rigid typology. Previous research using UD corpora suggests that while written Slovenian typically follows an SVO order, spoken Slovenian shows more variability, often classified as No Dominant Order (NDO). This variability, particularly in spoken language, forms the basis for this study's exploration of how word order patterns differ between spoken and written Slovenian.

4. Analysis

Building on the insights from the literature review, I decided to focus on analyzing WALS Feature 82A: Order of Subject and Verb. This feature was selected because it offered a manageable challenge and directly related to the word order analysis in Slovenian. The feature involves determining the order of subject and verb in a language's basic declarative clauses.

To investigate this feature in Slovenian, I chose two pre-annotated corpora: SSJ and SST. Both corpora are available on the UD website and provided a solid foundation for further analysis. However, there was a need to clean up the punctuation in SSJ to make it comparable with SST, which does not include punctuation.

I used STARK, a tool designed to extract dependency trees from text, to analyze the corpora. One major difficulty was handling cases where multiple words, such as a subject and object, were not directly connected by a syntactic relation. For direct relations (e.g., verb-object), I was able to specify the direction of the dependency, but indirect relations required more complex handling.

4.1 Data Processing

Once the corpora were processed and the relevant data extracted, I proceeded to clean the data using Python. This involved removing unnecessary columns from the dependency tree output, assigning the correct word order patterns to each sentence based on the subject-verb-object structure, and combining data from both corpora into a single file for further analysis. These steps ensured that the data was well-organized and ready for the next stages of the project.

A significant challenge in this phase was assigning the correct parts of speech (POS) tags to words in the dependency trees. Initially, I struggled with tagging, but after some adjustments, I managed to correctly identify the parts of speech for each word.

I also tried using Python to automate the entire process, aiming to extract entire sentences instead of just isolated words from STARK. However, parsing complex clauses proved to be difficult, and I eventually returned to the method suggested by my colleagues, which yielded better results.

4.2 Statistical Analysis

The study revealed significant differences in word order patterns between spoken and written Slovenian, highlighting the flexibility of spoken language compared to the rigidity of written language. In the written Slovenian corpus (SSJ), the Subject-Verb-Object (SVO) word order was dominant, accounting for 54.6% of occurrences. In contrast, the spoken Slovenian corpus (SST) exhibited greater variability, with SVO occurring at 39.4% and marked word orders like Subject-Object-Verb (SOV, 20.9%) and Object-Subject-Verb (OSV, 15.3%) occurring more frequently. The written corpus had a clear dominant order (SVO), while spoken Slovenian showed no dominant word order, classified as No Dominant Order (NDO). Statistical analysis, including a Chi-square test ($\chi^2 = 133.59, p < 0.001$), confirmed significant differences between the two corpora. These differences reflected the influence of pragmatic factors such as emphasis and topicalization in spoken Slovenian, where marked word orders like SOV and OSV were more common. Proportional differences and visualizations revealed that while SVO was still dominant in both corpora, spoken Slovenian displayed significantly more syntactic variability. The findings supported the hypothesis that spoken Slovenian deviates from WALS typologies, with the spoken corpus showing more variability and marked orders. The study also partially supported the hypothesis that WALS's binary classifications fail to capture the full range of word order patterns in spoken Slovenian. This research underscores the importance of considering modality and pragmatic factors in typological studies, as well as the limitations of rigid typological categories in capturing the complexity of spoken language syntax.

4.3 Drafting the Paper

The final phase of the project involved drafting the research paper which can be found here. The writing process was an opportunity to refine my academic writing skills and learn to format and structure the document using LaTeX. This included inserting figures, managing references, and ensuring that the paper adhered to academic standards.

While the research did not lead to groundbreaking findings, it provided a valuable foundation for future work. The paper also serves as a starting point for further exploration of Slovenian word order and other WALS features.

5. Additional Contributions

5.1 Repository

Throughout the project, I documented the code, data, and feature tables in a GitHub repository, which serves as a central resource for the research. This repository ensures transparency, facilitates collaboration on future projects, and includes all references, sources, and tools used. It can be accessed here.

5.2 Collaboration

Although I worked independently on most tasks, the collaboration with my colleagues Luka Terčon and Kaja Dobrovoljc was crucial in overcoming technical challenges, especially those related to STARK. This support was essential in making progress with the analysis.

6. Challenges

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.

7. Conclusion and Future Work

This project has provided valuable insights into Slovenian word order and the application of UD in linguistic typology. Although the findings were not groundbreaking, the process of analyzing Slovenian word order using statistical methods and visualizations was highly educational. The results offer a solid foundation for future research, particularly in exploring other WALS features or analyzing additional languages.

The project also contributed to the development of a GitHub repository that documents the research process, making it a useful resource for future work. While some challenges were encountered, they provided opportunities for learning and improvement. Moving forward, I plan to extend the analysis to other languages or linguistic features to expand the scope of the research and potentially publish the findings in a linguistic journal.

Additionally, I plan to proceed with the qualitative analysis and model training for predicting word order patterns, (as proposed here), with a particular focus on the latter. This phase will be instrumental in advancing the project by applying machine learning techniques to predict syntactic structures, enhancing our understanding of word order variation across languages.

The proposal outlines the goal of training a model that predicts sentence structure types (e.g., OSV, SVO) from raw text input. STARK will be used to extract relevant syntactic patterns, which will then be processed into a dataset for model training. The model will be designed to preprocess the text, tokenize it, and predict the sentence order. To increase the accessibility and impact of this research, the dataset will be published on HuggingFace, enabling others to use and build upon it. Additionally, I will create a demo or app using Gradio that allows users to test the model with their own sentences.

These efforts aim to advance research in syntactic analysis and provide a practical tool for linguistic analysis and education.

8. Acknowledgments

This project made use of several generative AI tools to assist with coding, literature review, and content refinement. ChatGPT was used for generating and debugging code, as well as improving the quality of the written content. Perplexity helped with the literature review by identifying and summarizing relevant sources. Codeium provided assistance with coding tasks and guidance on best practices. All ideas and proposed solutions, however, remain my own or are derived from the cited sources.