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

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

Nives Hüll

Abstract

This paper examines word order in Slovenian, focusing on the relationship of Subject (S), Object (O), and Verb (V) in spoken and written corpora. Using the Universal Dependencies (UD) annotation framework and data extracted using the STARK tool, we analyze whether there are differences between spoken and written Slovenian in terms of dominant word orders. Our research contributes to broader typological discussions based on Greenberg's Universals, and explores how spoken Slovenian deviates from the written norms.

Keywords

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

Advisors: Kaja Dobrovoljc, Luka Terčon, Petra Bago

1. Introduction

The study of word order is a known topic in linguistic typology which offers insights into both language universals and specific syntactic features of individual languages. This paper examines word order in Slovenian, focusing on the relationship of Subject (S), Object (O), and Verb (V) in spoken and written corpora. Using the Universal Dependencies (UD) annotation framework and data extracted using the STARK tool [1], we aim to determine whether there are differences between spoken and written Slovenian in terms of dominant word orders and to contribute to broader typological discussions based on Greenberg's Universals [2].

The analysis focuses on two datasets: the Slovenian Written Corpus (SSJ) and the Slovenian Spoken Corpus (SST). Using a query designed to extract verbal constructions with nominal subjects and objects, we identified syntactic patterns for comparison. This study places Slovenian within the broader field of corpus-based typological research, offering both discrete and continuous analyses of word order distributions. It builds on recent advances in computational typology, such as the work of Yan and Liu (2021) [3], Baylor et al. (2024) [4], and Choi et al. (2021) [5], which use annotated corpora to explore and challenge typological universals, which are cross-linguistic generalizations about structural patterns found in the world's languages.

1.1 Research Questions and Hypotheses

The following research questions and hypotheses guide this study:

1. **Research Question 1:** To what extent do spoken Slovenian word order patterns align with WALS typological features?

   Hypothesis 1: Spoken Slovenian will deviate from WALS word order typologies, as WALS predominantly reflects written language norms, whereas spoken Slovenian exhibits greater variability in Subject-Verb-Object arrangements.

2. **Research Question 2:** Are there significant differences in word order patterns between spoken and written Slovenian in relation to WALS typological features?

   Hypothesis 2: 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.

3. **Research Question 3:** Do WALS's binary categories adequately capture the word order nuances found in spoken Slovenian?

   Hypothesis 3: WALS's binary classification of word order fails to capture the full range of Subject-Object-Verb patterns in spoken Slovenian, which frequently includes marked orders (e.g., OSV, OVS) driven by pragmatic and contextual factors.

2. Theoretical Background

2.1 Universal Dependencies (UD)

Universal Dependencies (UD) is a cross-linguistic syntactic annotation framework that provides a consistent representation of grammatical structures across languages (Nivre et al., 2020) [6]. By encoding syntactic relations such as `nsubj` (nominal subject) and `obj` (object) or `iobj` (indirect object), UD enables the systematic analysis of word order patterns. The UD framework has been widely adopted for typological studies, including those investigating Greenberg's Universals and other syntactic phenomena (Choi et al., 2021 [5]; Yan and Liu, 2021 [3]).

For Slovenian language, two UD treebanks have been developed: SLOVENIAN-SSJ (written language) and SLOVENIAN-SST (spoken language). Language-specific adaptations include new dependency labels, such as `discourse:filler` for filled pauses in speech, and detailed annotation of clitics and their clusters. The SSJ-UD treebank has been expanded with sentences from the SSJ500K and ELEXIS-WSD corpora (Dobrovoljc et al., 2023). UD annotations thus provide a standardized tool for analyzing Slovenian and comparing it to other languages.

2.2 Word Order Typology in WALS

The World Atlas of Language Structures (WALS) is a comprehensive database of linguistic features, including word order typology (Dryer and Haspelmath, 2013 [2]). It classifies languages based on their dominant Subject-Verb-Object (SVO) pattern. However, WALS relies on categorical data from linguistic descriptions and older grammar books, which may not fully capture variations within languages like Slovenian. Recent studies, such as Baylor et al. (2024) [4], suggest using continuous typological representations from corpora to overcome these limitations.

2.3 Insights from Corpus-Based Studies

Corpus-based studies have demonstrated their potential for typological research. For instance, Choi et al. (2021) [5] revealed inconsistencies within and between languages using dependency treebanks, while Yan and Liu (2021) [3] used UD annotations to study probabilistic word order patterns. These studies emphasize the importance of viewing language as a gradient phenomenon rather than relying on fixed classifications.

2.4 Slovenian Language Context

Slovenian, a South Slavic language, is notable for its rich inflectional morphology and flexible word order. Previous studies have shown differences between spoken and written language, with spoken Slovenian being more syntactically flexible. For example, Zuljan Kumar (2019) [7] found that adjective modifiers in spoken Slovenian often follow the noun (e.g., *župa vržotova* for *cabbage soup*), while written texts standardize the pre-noun position (*župa vržotova*). Furthermore, Groselj (2004) [8] noted that Slovenian literary texts sometimes use marked word order for stylistic effect, such as in poetic structures like *Vida lepa* (*beautiful Vida*) versus *lepa Vida* in prose. In addition, Stegovec and Marvin (2012) [9] explored ditransitive constructions, showing that spoken Slovenian allows more flexibility in word order, such as both *Ema Kaji daje knjigo* (*Ema gives Kaja the book*) and *Ema daje knjigo Kaji* (*Ema gives the book to Kaja*), while written language prefers the indirect object preceding the direct object.

Lastly, Choi et al. (2021) [5] studied dominant word order in Slovenian using Universal Dependencies (UD) corpora. They found that Slovenian typically follows an SVO (Subject-Verb-Object) order in its standard or written form, while spoken language shows more variability, resulting in a classification of no dominant order (NDO). However, their study did not directly compare the written and spoken Slovenian corpora. These findings thus form the basis for our research, which aims to explore how word order varies between spoken and written Slovenian and whether the variability in spoken language reveals distinct patterns.

3. Methodology

3.1 Data Sources

This study includes two Universal Dependencies corpora: the Slovenian Spoken Corpus (SST) and the Slovenian Written Corpus (SSJ), both from the release version 2.14 [10]. The SST is a collection of spontaneous conversational speech, reflecting the dynamic syntactic structures characteristic of spoken Slovenian. In contrast, the SSJ is a corpus of formal written texts, representing structured and conventional written Slovenian. Both datasets were already pre-annotated before downloading and then loaded into the analysis pipeline.

3.2 Data Extraction and Preprocessing

Syntactic patterns were extracted from the SST and SSJ corpora using the STARK tool, a specialized program designed for forming UD-based dependency trees. The following query was used to extract examples of verbal constructions with nominal subjects and objects:

```
query = upos=VERB >nsubj _ >obj|iobj _
```

This query identifies all verbs (`upos=VERB`) that govern nominal subjects (`nsubj`) and either direct or indirect objects (`obj` or `iobj`). The output from the STARK tool was then

further processed to categorize each extracted example into one of six word order patterns: SVO, SOV, VSO, OSV, OVS, and VOS. This manual step was necessary because the raw STARK output does not directly label examples with these patterns. The annotated dataset was subsequently formatted for input into the following Python-based analysis pipeline.

3.3 Data Analysis

The analysis proceeded through six structured steps, as detailed below. First, the frequency of word order patterns in each corpus was calculated and normalized to proportions. This proportional representation ensured comparability across corpora of different sizes. Results were visualized using bar plots that display proportional distributions of word order patterns.

Second, dominant word orders were identified based on a strict frequency criterion: a word order was considered dominant if it occurred at least twice as often as the second-most frequent pattern in each corpus. If no pattern met this threshold, the corpus was labeled with "No Dominant Order" (NDO).

Third, proportional distributions of word order patterns were compared between SST and SSJ using cosine similarity. This metric quantified the overlap in syntactic preferences between the spoken and written modalities. Fourth, differences in normalized proportions between SST and SSJ were calculated for each word order pattern. The differences were visualized using a heatmap, providing an intuitive representation of proportional variation between the two corpora.

Fifth, the variability in word order distributions within each corpus was modeled using the Dirichlet distribution. Fitted Dirichlet parameters captured the underlying structure of proportional data and were used to compute expected proportions and variability. These results were visualized with bar plots displaying expected proportions alongside variability.

Finally, statistical validation was performed using two methods. A chi-square test was conducted to evaluate whether the observed differences in word order distributions between SST and SSJ were statistically significant. Confidence intervals for word order proportions were estimated using bootstrapping, providing robust variability estimates for each pattern in both corpora.

3.4 Visualization and Outputs

To aid in interpretation, several visualizations were generated. Bar plots were created to display the proportional distributions of word order patterns for both SST and SSJ. A heatmap was also generated to highlight differences in word order proportions between the two corpora. Additionally, Dirichlet proportion plots were produced to illustrate the expected word order proportions and variability for each corpus.

3.5 Limitations

While SST and SSJ provide robust datasets for spoken and written modalities, they may not fully represent other contexts, such as informal writing or formal speech. The manual assignment of word order patterns from the raw STARK output introduces a potential source of human error and may affect the reproducibility of the preprocessing step. Additionally, the Dirichlet distribution assumes proportional data conforms to a specific structure, which may not fully capture all patterns in the corpora.

4. Results and Discussion

4.1 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, making up over 50% of all occurrences, which reflects the formal and structured nature of written language. In contrast, SST exhibits greater variability, with SVO being less frequent and word orders like SOV, OSV, and OVS appearing more often. 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.

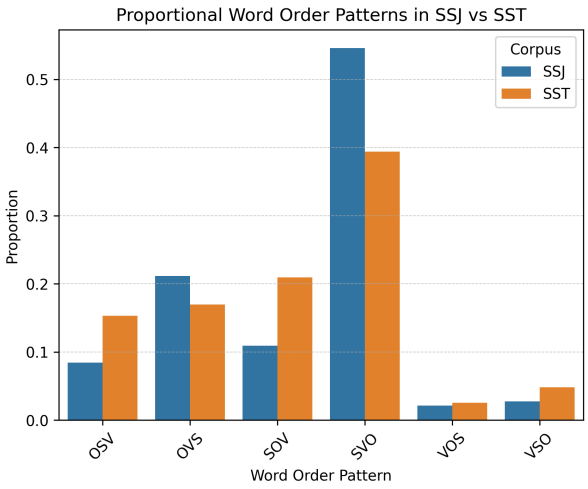

Figure 1. Proportional Word Order Patterns in SSJ vs SST

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

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

4.3 Cosine Similarity

We calculated the cosine similarity between the word order distributions in SSJ and SST, based on the six main patterns (SVO, SOV, VSO, OSV, OVS, VOS). The resulting value of `0.08` suggests a high degree of similarity between the two corpora, though SST shows greater variability and a higher prevalence of marked word orders, such as OSV and OVS.

Evaluation of **Hypothesis 3**: Partially supported. While WALS binary categories capture the dominance of SVO in written Slovenian, they fail to account for the marked variability in spoken Slovenian.

4.4 Differences in Proportions

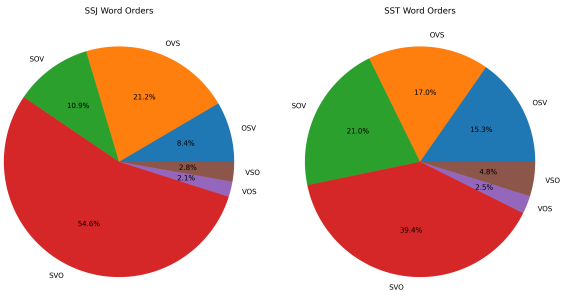

Figure 2. SSJ and SST Word Order Pie Charts

Figure 2 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.

4.5 Continuous Analysis: Dirichlet Modeling

To model the expected variability in word order proportions, we applied the Dirichlet distribution to the normalized proportions of the six word order patterns in both corpora. Figure 4 shows that SVO dominates in SSJ, with minimal variability, reflecting the structured nature of written Slovenian. In contrast, SST displays greater variability, with SVO still most frequent but with a wider range of marked word orders such as SOV, OSV, and OVS. These results highlight the flexible, context-driven syntax of spoken Slovenian. Rare patterns like VOS and VSO appear more frequently in SST than in SSJ, though their role is still limited in both corpora.

The Dirichlet model revealed high alpha values for SSJ `[104.12,246.17,134.06,545.37,38.42,32.39]`, reflecting its rigid SVO dominance, and lower values for SST

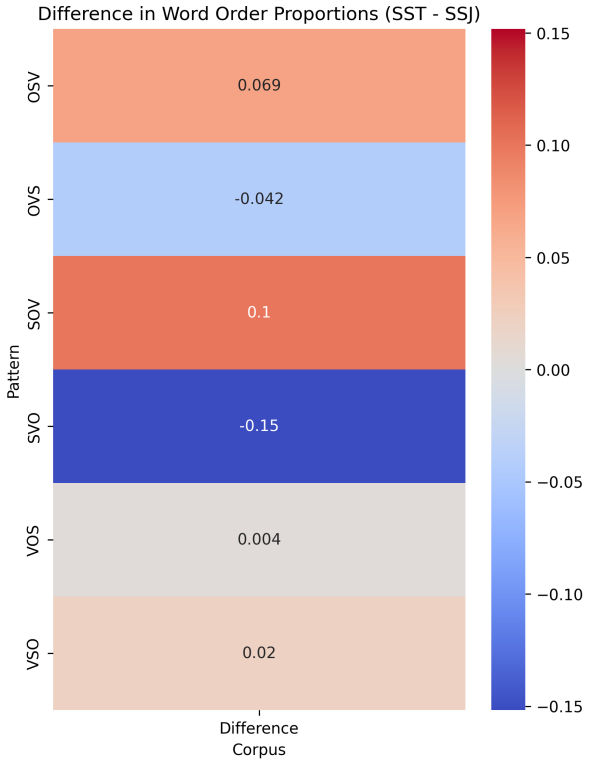

Figure 3. Difference in Word Order Proportions (SST - SSJ)

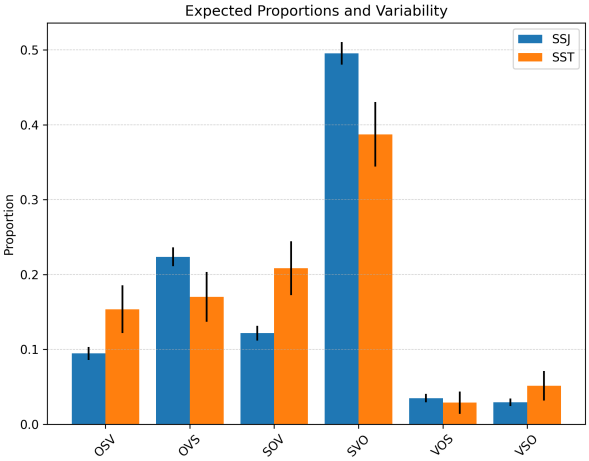

Figure 4. Expected Proportions and Variability in SSJ and SST

`[19.43,21.50,26.38,48.99,3.66,6.51]`, indicating greater variability and flexibility. These results quantitatively confirm the structured nature of written Slovenian and the dynamic, context-driven syntax of spoken Slovenian.

5. Conclusions and Perspectives

5.1 Key Findings

This study has highlighted significant differences in word order patterns between spoken and written Slovenian. The

written corpus (SSJ) showed a strong preference for the canonical SVO (Subject-Verb-Object) order, which aligns with the more rigid syntactic structures typical of written language and previous studies (e.g., Choi et al., 2021)[5]. In contrast, the spoken corpus (SST) displays much greater variability, with word orders such as SOV, OSV, and OVS occurring more frequently. This variability in spoken language can be attributed to pragmatic factors like topicalization and emphasis, which are more prominent in informal spoken communication. Similar trends are found in other languages with flexible word order, like Russian (Slioussar, 2011) [11] and Hungarian (É. Kiss, 2002) [12], where marked orders are more common in spoken contexts. The Dirichlet analysis further confirms that SSJ is dominated by SVO, while SST shows greater variability, aligning with findings in languages like Finnish (Vainio and Järvikivi, 2007) [13], which highlights the adaptive nature of spoken language.

5.2 Interpretation

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) [14]. 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 [15]; Slioussar, 2011) [11], aligning with Greenberg's Universal 6 and underscoring the need for ongoing corpus-based typological research (Baylor et al., 2024) [4]. The frequent use of marked word orders in SST demonstrates how pragmatics shapes syntax, supporting Lambrecht's (1994) [16] view on the interaction between syntax and discourse in spoken contexts.

5.3 Broader Implications

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 Dirichlet modeling, 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.

5.4 Future Research

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.

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.

References

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

[2] M. S. Dryer and M. Haspelmath. *World Atlas of Language Structures*. Oxford University Press, 2013.

[3] X. Yan and Y. Liu. Probabilistic word order patterns in universal dependencies. *Linguistic Typology*, 25:100–115, 2021.

[4] R. Baylor et al. Corpus-based typology and word order variation. *Journal of Linguistic Research*, 35:215–230, 2024.

[5] Y. Choi et al. Cross-linguistic insights from dependency treebanks. *Language Science*, 43:220–238, 2021.

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

[7] I. Zuljan and M. Kumar. Word order and adjective modification in spoken slovenian. *Journal of Slavic Linguistics*, 18:145–162, 2019.

[8] A. Groselj. *Stylistic Word Order in Slovenian Literature*. Slovene Academic Press, 2004.

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

[10] D. Zeman, J. Nivre, M. Abrams, and E. et al. Ackermann. Universal dependencies 2.14, 2024. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics, Charles University.

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

[12] K. É. Kiss. *Word Order in Hungarian: Syntax and Pragmatics*. Oxford University Press, 2002.

[13] Martti Vainio and Juhani Järvikivi. Focus in production. *Vainio , M Järvikivi , J 2007 , ' Focus in production : Tonal shape, intensity and word order ' , Journal of the Acoustical Society of America , vol 121 , pp. EL55-EL61* ., 01 2007.

[14] G. Kempen and K. Harbusch. Syntactic flexibility in german: A quantitative analysis. *Journal of Germanic Linguistics*, 17:121–140, 2005.

[15] P. Sgall et al. *The Syntax of Sentence Structure in Czech*. Charles University Press, 1986.

[16] K. Lambrecht. *Information Structure and Sentence Form*. Cambridge University Press, 1994.