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

Nives Hüll

January 20, 2025

Introduction

This project analyzes Slovenian word order using Universal Dependencies (UD) and the World Atlas of Linguistic Structures (WALS).

- **Universal Dependencies (UD):** A framework for annotating syntactic relations to analyze sentence structure across languages.
- **World Atlas of Linguistic Structures (WALS):** A database providing typological information, including common word order patterns for cross-linguistic comparison.

Main goal: To explore linguistic patterns, focusing on the subject-verb-object (SVO) structure in Slovenian.
The project integrates linguistic theory with computational methods for analysis.

Introduction 2: UD

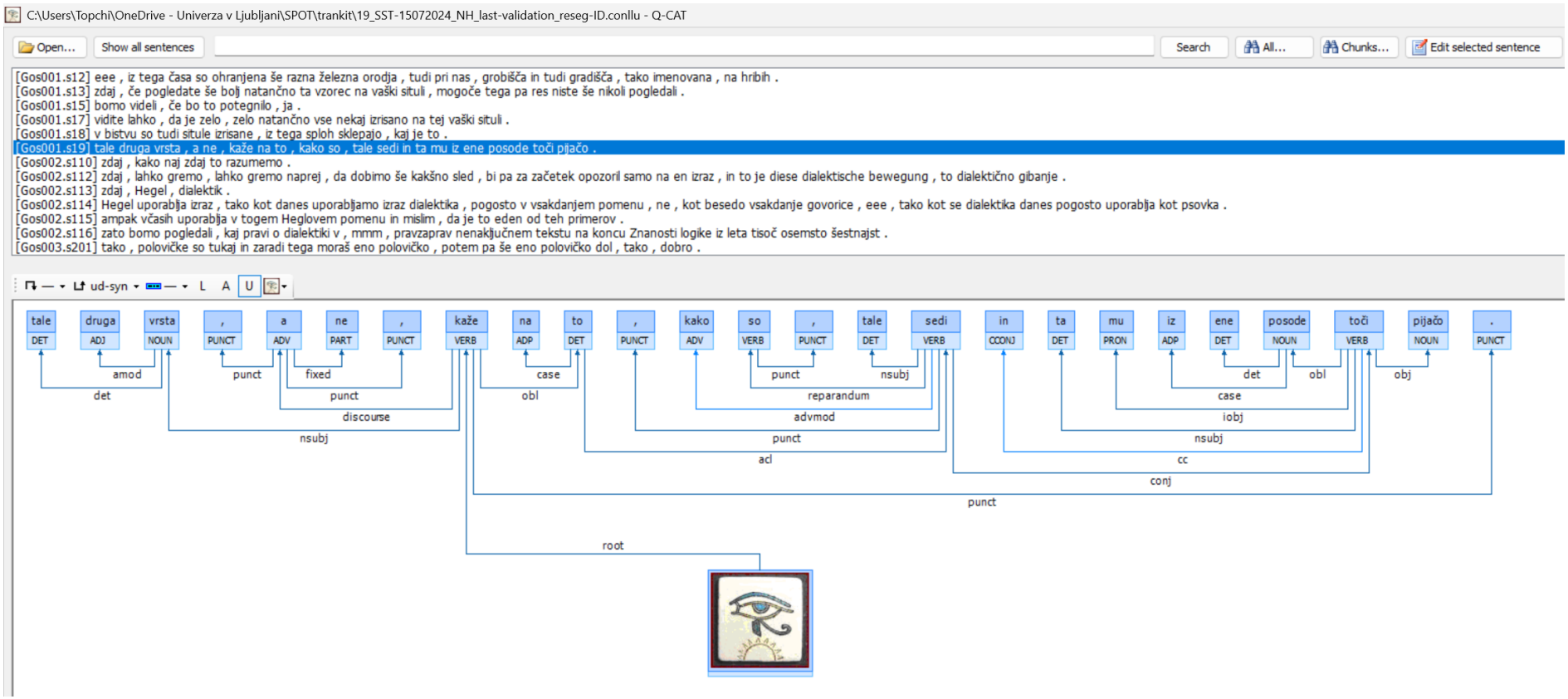

Figure: Example of UD annotated sentence in Q-CAT

Introduction 3: WALS

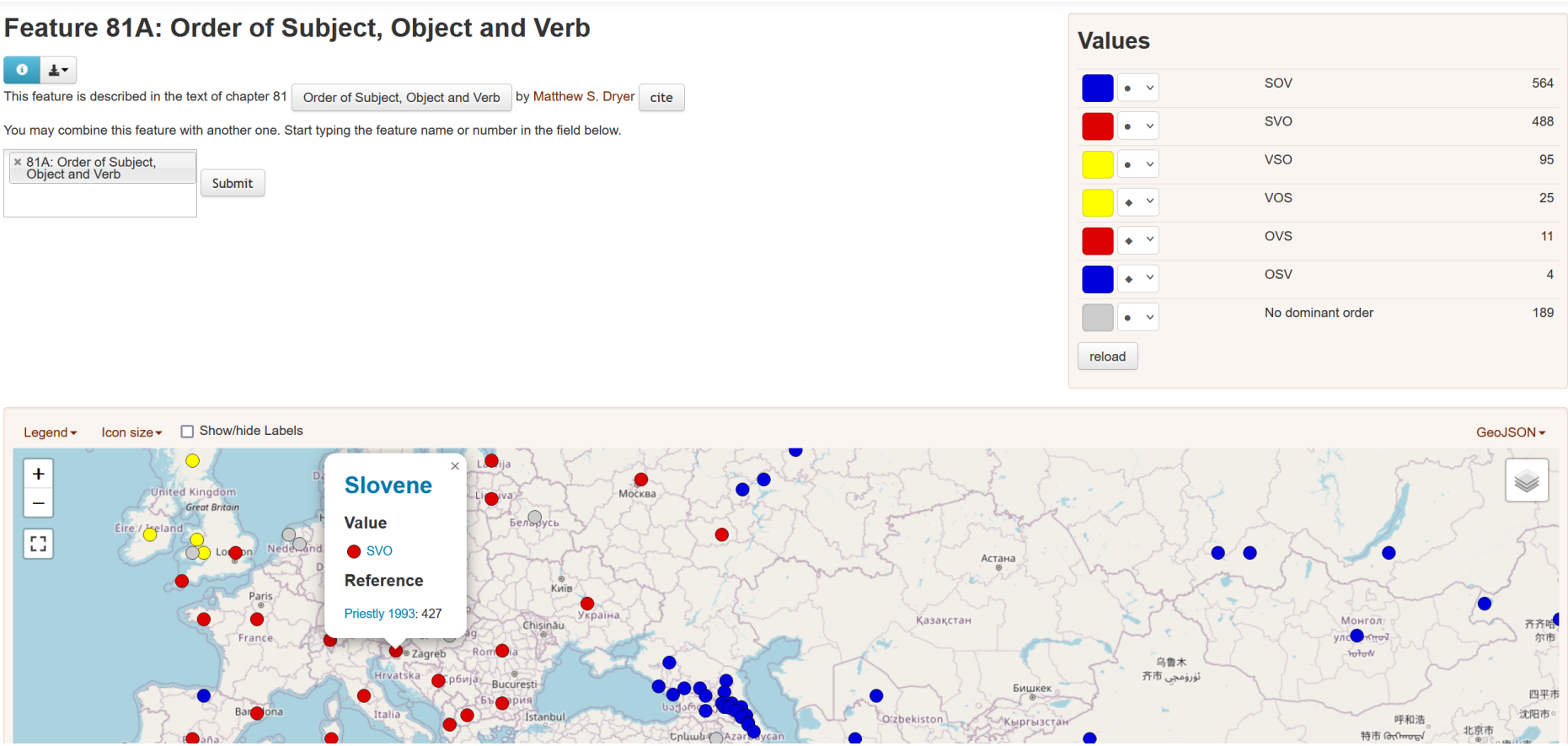

Figure: Example of WALS feature

Research Questions and Hypotheses

1. **RQ1:** To what extent do spoken Slovenian word order patterns align with WALS typological features?
   H1: 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. **RQ2:** Are there significant differences in word order patterns between spoken and written Slovenian in relation to WALS typological features?
   H2: 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. **RQ3:** Do WALS's binary categories adequately capture the word order nuances found in spoken Slovenian?
   H3: 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.

Feature Table

In the first phase, I extracted linguistic features from the WALS database, focusing on those relevant to Slovenian, with Word Order having the most representation.
Initially focused on Slovenian, the table was later expanded to include features from all languages in WALS.
This work also contributed to the development of LLMs in Slovenia, simplifying UD queries for researchers and the public.
The final feature table and interactive app are available here:

- Feature Table
- Interactive Feature App

Feature Extraction

A major challenge was extracting feature data from the WALS database.
Web scraping was difficult due to dynamic JavaScript elements, so I manually copied and formatted the data into a ‘.txt‘ file.
Initially focused on Slovenian, I later expanded the table to include features from all languages in WALS for broader comparisons.
The final table contains 192 features, with 43 (22.4%) relevant to Slovenian.

Feature Distribution for Slovenian

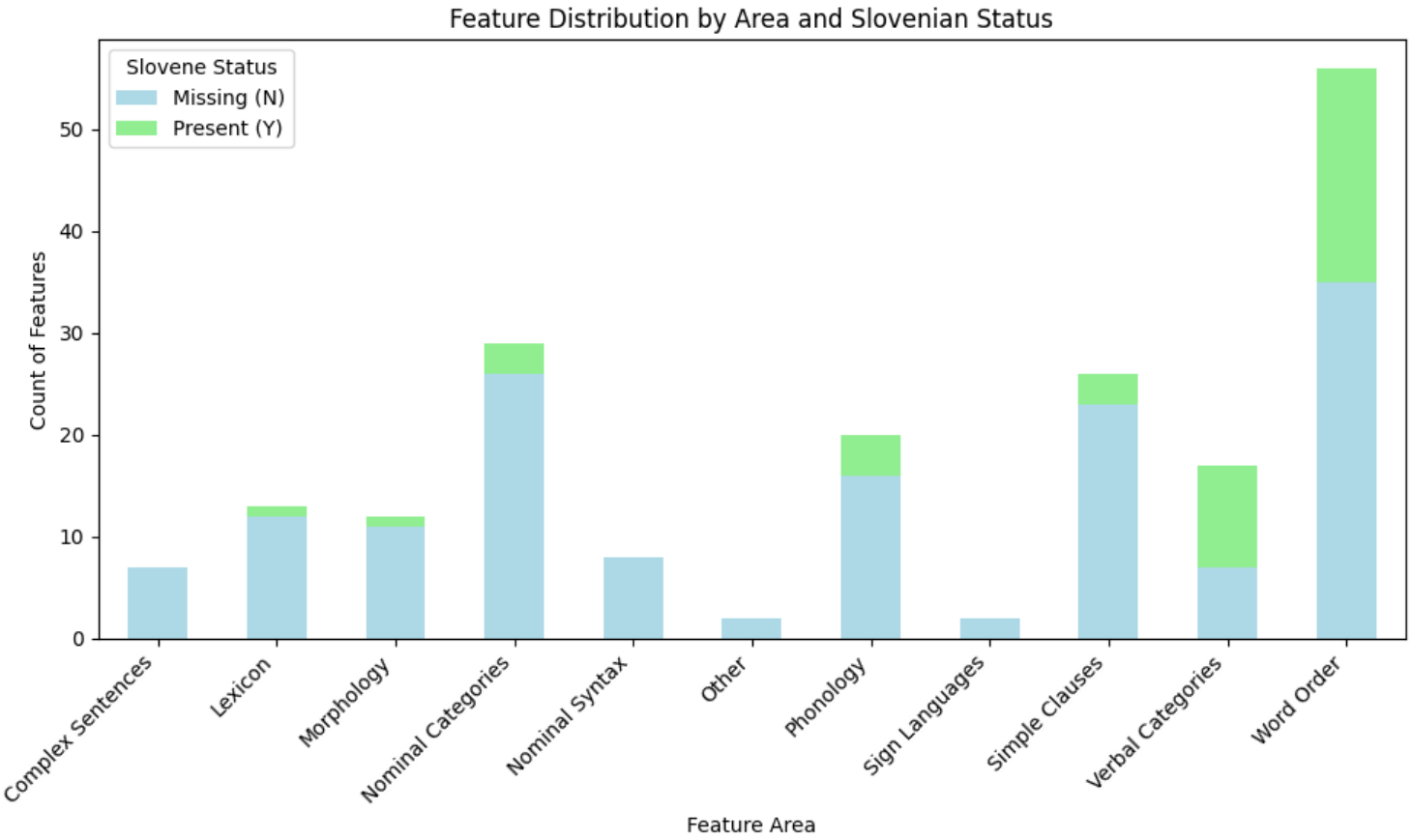

Figure: Distribution of WALS features in Slovenian

Strong representation in Word Order and Verbal Categories, minimal presence in other categories.

Slovenian Word Order: Written vs. Spoken Language

While WALS classifies languages by typology, it doesn't fully capture variations within languages, like those in Slovenian. Slovenian shows flexible word order, especially in spoken language, where variability often leads to No Dominant Order (NDO), unlike the fixed SVO pattern in written language.

Analysis of WALS Feature 82A

I focused on analyzing **WALS Feature 82A: Order of Subject and Verb**, which identifies the order of subject and verb in basic declarative clauses. To explore this in Slovenian, I used two pre-annotated corpora (SST, SSJ) and cleaned up punctuation in SSJ to ensure comparability with SST. STARK was employed to extract dependency trees, but challenges appeared when handling indirect relations between words. These required more complex processing to accurately capture the syntactic structure.

STARK Tool 1

.codegpt	7. 01. 2025 19:44	File folder	
.git	8. 01. 2025 11:23	File folder	
logos	19. 11. 2024 12:22	File folder	
sample	7. 01. 2025 19:41	File folder	
scripts	19. 11. 2024 12:22	File folder	
scripts_n	19. 11. 2024 12:26	File folder	
stark	19. 11. 2024 12:30	File folder	
tests	19. 11. 2024 12:22	File folder	
.gitignore	7. 01. 2025 19:44	Git Ignore Source File	1 KB
advanced.md	19. 11. 2024 12:22	Markdown Source File	7 KB
config.ini	15. 01. 2025 12:19	Configuration settings	2 KB
install.bat	19. 11. 2024 12:22	Windows Batch File	1 KB
LICENSE.txt	19. 11. 2024 12:22	Text Document	12 KB
MANIFEST.in	19. 11. 2024 12:22	IN File	1 KB
README.md	19. 11. 2024 12:22	Markdown Source File	9 KB
requirements.txt	19. 11. 2024 12:22	Text Document	1 KB
run.bat	19. 11. 2024 12:22	Windows Batch File	1 KB
run.sh	19. 11. 2024 12:22	sh_auto_file	1 KB
settings.md	19. 11. 2024 12:22	Markdown Source File	15 KB
setup.py	19. 11. 2024 12:22	Python Source File	1 KB
stark.py	19. 11. 2024 12:22	Python Source File	1 KB
stark-multiresult.py	19. 11. 2024 12:22	Python Source File	2 KB

Figure: Files of STARK tool

STARK Tool 2

```
;___GENERAL SETTINGS___
input = sample/wals/input/sl_sst-ud-merged.conllu
output = sample/wals/output/sst_patterns_NOVO.tsv

;___TREE SPECIFICATIONS___
node_type = form
labeled = yes
label_subtypes = yes
fixed = yes

;___TREE RESTRICTIONS___
size = 2-10000
;head = upos=VERB|upos=NOUN
ignored_labels = punct|reparandum
;allowed_labels = nsubj|obj|obl

;___SEARCH BY QUERY___
query = upos=VERB >nsubj _ >obj|iobj _

;___ADDITIONAL STATISTICS___
node_info = yes
association_measures = no
;compare = sample/fr_gsd-ud-dev.conllu

;___VISUALISATION___
example = yes
grew_match = yes
depsearch = no

;___OUTPUT THRESHOLD___
;frequency_threshold = 5
;max_lines = 100


; ************** ADVANCED SETTINGS (see advanced.md) **************
;internal_saves = ./internal_saves
;cpu_cores = 12
;continuation_processing = no
greedy_counter = yes
complete = no
;processing_size = 1-7
;sentence_count_file = number_of_matched_trees_per_sentence.txt
;detailed_results_file = list_of_all_sentences_with_matched_trees.txt
```

Figure: Example of STARK configuration

STARK Output

Tree	Node A-fo	Node B-fo	Node C-fo	Number o	Head nod	Example
kaj <obj jaz <nsubj vem	kaj	jaz	vem	3	vem	in pač ne vem zdaj kaj ali imeli svoj lasten piknik ali A[kaj] B[jaz] C[vem]
kaj <obj to <nsubj pomeni	kaj	to	pomeni	3	pomeni	kaj A[kaj] zdaj B[to] C[pomeni]
mi <obj je >nsubj poklic	mi	je	poklic	3	je	in pravzaprav A[mi] B[je] zato moj C[poklic] všeč
to <nsubj mi <obj zdi	to	mi	zdi	3	zdi	in A[to] se B[mi] C[zdi] neki tak proces učenja e kaj se je pa zdaj zgodilo v
aktiv <nsubj dobi >obj rezultat	aktiv	dobi	rezultat	3	dobi	ker osmi šesti tukaj piše A[aktiv] B[dobi] skupni C[rezultat] vseh
jaz <nsubj ga <obj spomnim	jaz	ga	spomnim	3	spomnim	A[jaz] se B[ga] ne C[spomnim]
kaj <obj ti <nsubj govoriš	kaj	ti	govoriš	3	govoriš	a no daj veš kdo bo šel na dvoboj A[kaj] B[ti] meni zdaj C[govoriš]
kaj <obj vi <nsubj pravite	kaj	vi	pravite	3	pravite	mhm kaj pa vi pravite na to gos- ja k- A[kaj] B[vi] C[pravite] gospo-
kar <nsubj mi <obj zdi	kar	mi	zdi	3	zdi	zdaj mi se srečujemo tulele s predlogi vlade ne ker prinese izračune na d
kar <nsubj tega <obj tiče	kar	tega	tiče	3	tiče	ja eee ker svet ni več tako zelo prijazen vsaj A[kar] se B[tega] C[tiče]
kdo <nsubj ga <obj zagleda	kdo	ga	zagleda	3	zagleda	A[kdo] B[ga] od daleč že C[zagleda] [name:personal] prosim
kraj <nsubj vam <obj je	kraj	vam	je	3	je	e kateri izletniški A[kraj] B[vam] C[je] všeč in zakaj
mi <nsubj tega <obj predlagali	mi	tega	predlagali	3	predlagali	A[mi] B[tega] nismo C[predlagali]
mi <obj bil >nsubj film	mi	bil	film	3	bil	najbolj všeč A[mi] je B[bil] C[film] e Hitri in drzni drugi del zaradi tega ker

Figure: STARK output

Search query: `query = upos=VERB >nsubj` $_{>}$*obj*|*iobj*

Data Cleaning and Processing

After extracting the relevant data, I used Python to clean it by removing unnecessary columns, assigning correct word order patterns, and merging data from both corpora. A key challenge was tagging parts of speech (POS), which was resolved after adjustments. I also attempted to automate sentence extraction but reverted to the original method for better results.

CONLLU File

```
# text = ravno danes sva se slišali, je rekla, da imata ob pol dvanajstih vaje, tako da vaje imajo normalno, ne.
# sent_id = iriss.942
1   ravno   ravno   PART    Q   _   2   advmod  _   seg_id=Artur-N-G5020-P600012.s36|tei_tok_id=Artur-N-G5020-P600012.to
2   danes   danes   ADV Rgp Degree=Pos  5   advmod  _   seg_id=Artur-N-G5020-P600012.s36|tei_tok_id=Artur-N-G5020-P60001
3   sva biti    AUX Va-r1d-n    Mood=Ind|Number=Dual|Person=1|Polarity=Pos|Tense=Pres|VerbForm=Fin  5   aux _   seg_id=A
4   se  se  PRON    Px------y   PronType=Prs|Reflex=Yes|Variant=Short   5   expl    _   seg_id=Artur-N-G5020-P600012.s36
5   slišali slišati VERB    Vmbp-df Gender=Masc|Number=Dual|VerbForm=Part   0   root    _   seg_id=Artur-N-G5020-P600012
6   ,   ,   PUNCT   Z   _   8   punct   _   seg_id=Artur-N-G5020-P600012.s36|tei_tok_id=Artur-N-G5020-P600012.tok247
7   je  biti    AUX Va-r3s-n    Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|VerbForm=Fin  8   aux _   seg_id=A
8   rekla   reči    VERB    Vmep-sf Aspect=Perf|Gender=Fem|Number=Sing|VerbForm=Part    5   parataxis   _   seg_id=Artur
9   ,   ,   PUNCT   Z   _   11  punct   _   seg_id=Artur-N-G5020-P600012.s37|tei_tok_id=Artur-N-G5020-P600012.tok250
10  da  da  SCONJ   Cs  _   11  mark    _   seg_id=Artur-N-G5020-P600012.s37|tei_tok_id=Artur-N-G5020-P600012.tok251
11  imata   imeti   VERB    Vmpr3d-n    Aspect=Imp|Mood=Ind|Number=Dual|Person=3|Polarity=Pos|Tense=Pres|VerbForm=Fin
12  ob  ob  ADP Sl  Case=Loc    14  case    _   seg_id=Artur-N-G5020-P600012.s37|tei_tok_id=Artur-N-G5020-P600012.tok253
13  pol pol DET Rgp PronType=Ind    14  advmod  _   seg_id=Artur-N-G5020-P600012.s37|tei_tok_id=Artur-N-G5020-P600012.to
14  dvanajstih  dvanajst    NUM Mlc-pl  Case=Loc|Number=Plur|NumForm=Word|NumType=Card  11  obl _   seg_id=Artur-N-G5020
15  vaje    vaje    NOUN    Ncfpg   Case=Gen|Gender=Fem|Number=Plur 11  obj _   seg_id=Artur-N-G5020-P600012.s37|SpaceAf
16  ,   ,   PUNCT   Z   _   20  punct   _   seg_id=Artur-N-G5020-P600012.s37|tei_tok_id=Artur-N-G5020-P600012.tok257
17  tako    tako    CCONJ   Cc  _   20  cc  _   seg_id=Artur-N-G5020-P600012.s38|tei_tok_id=Artur-N-G5020-P600012.tok258
18  da  da  SCONJ   Cs  _   17  fixed   _   seg_id=Artur-N-G5020-P600012.s38|tei_tok_id=Artur-N-G5020-P600012.tok259
19  vaje    vaja    NOUN    Ncfpn   Case=Nom|Gender=Fem|Number=Plur 20  obj _   seg_id=Artur-N-G5020-P600012.s38|tei_tok
20  imajo   imeti   VERB    Vmpr3p-n    Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Polarity=Pos|Tense=Pres|VerbForm=Fin
21  normalno    normalno    ADV Rgp Degree=Pos  20  advmod  _   seg_id=Artur-N-G5020-P600012.s38|SpaceAfter=No|tei_tok_i
22  ,   ,   PUNCT   Z   _   23  punct   _   seg_id=Artur-N-G5020-P600012.s38|tei_tok_id=Artur-N-G5020-P600012.tok263
23  ne  ne  PART    Q   Polarity=Neg    20  discourse   _   seg_id=Artur-N-G5020-P600012.s38|SpaceAfter=No|tei_tok_id=Ar
24  .   .   PUNCT   Z   _   5   punct   _   sentence_ending=True|seg_id=Artur-N-G5020-P600012.s38|tei_tok_id=Artur-N-G50
```

Figure: Example of CONLLU file

Statistical Analysis

Summary: Statistical analysis revealed significant differences between written (SSJ) and spoken (SST) Slovenian.

- In SSJ, SVO was dominant (54.6%).
- In SST, SVO occurred less (39.4%) with more variability:
  - SOV: 20.9%
  - OSV: 15.3%
- Written Slovenian had a clear dominant order (SVO), while spoken Slovenian showed No Dominant Order (NDO).

Statistical Analysis: Frequency and Proportional Distribution 1

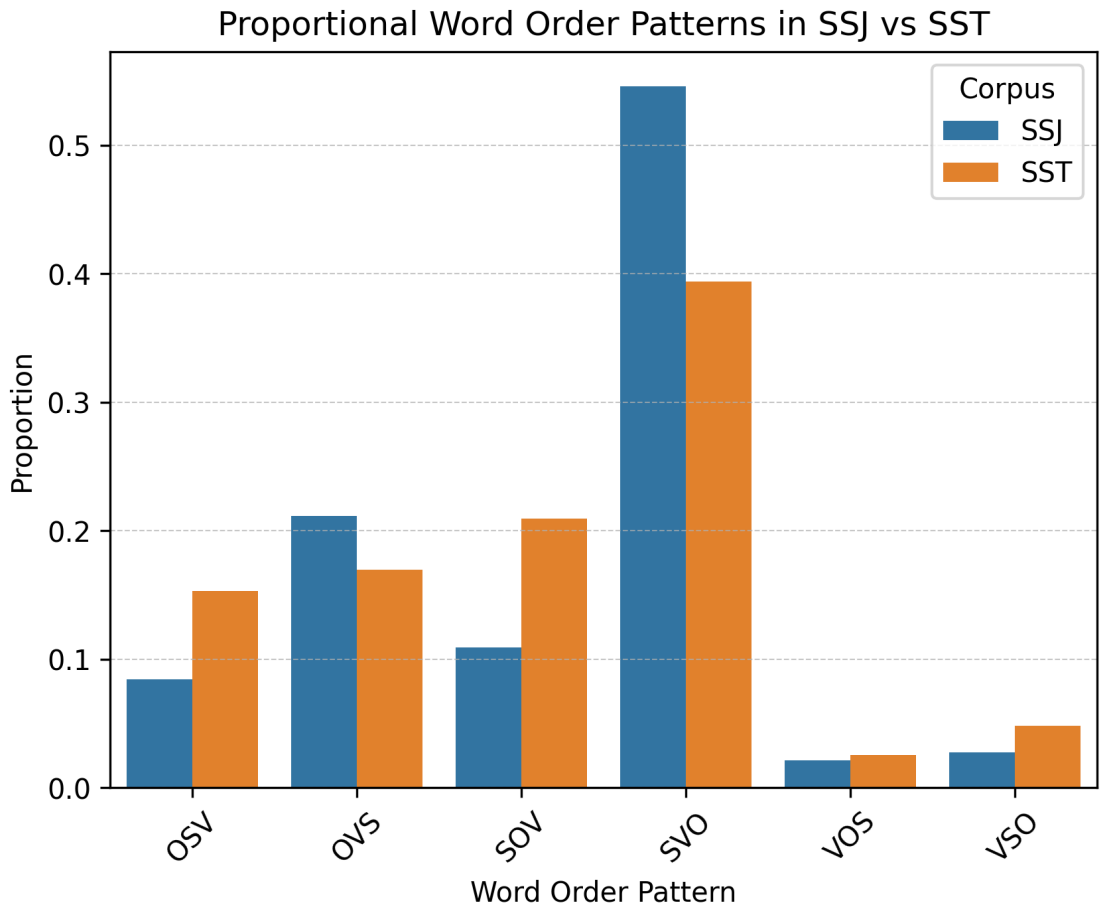

Figure: Distribution of Word Order Patterns in Written and Spoken Slovenian

Statistical Analysis: Frequency and Proportional Distribution 2

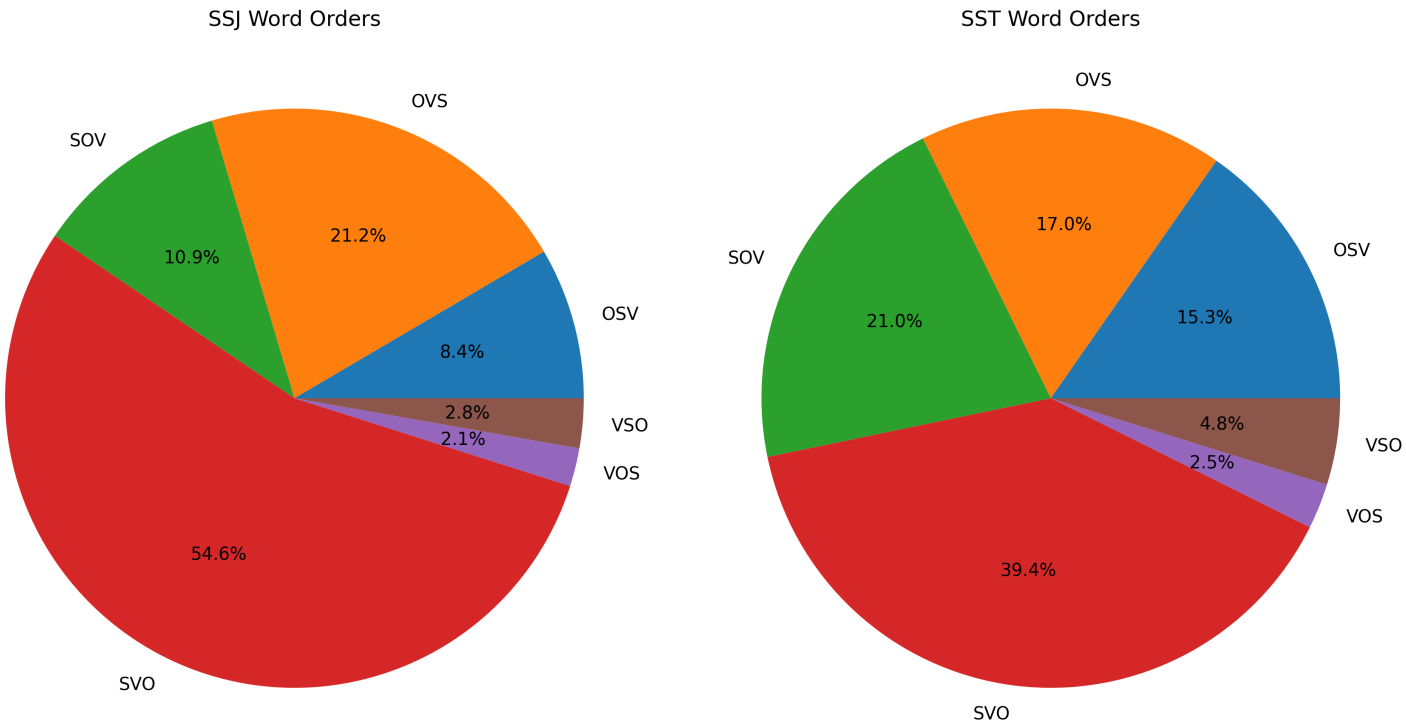

Figure: Frequency and Proportional Distribution in Written and Spoken Slovenian

Statistical Analysis: Frequency and Proportional Distribution 3

Word order patterns in written Slovenian (SSJ) and spoken Slovenian (SST):

- In SSJ, the SVO pattern is dominant (54.6%).
- In SST, SVO is less frequent (39.4%), with more variability:
  - SOV: 20.9%
  - OSV: 15.3%

This shows that spoken Slovenian is more flexible, with word order influenced by context. Statistical analysis confirmed significant differences ($\chi^2 = 133.59, p < 0.001$) between the two corpora.

Hypothesis 1: Supported – Spoken Slovenian shows more variability than written Slovenian.

Chi-Square and p-value:

- The Chi-Square value ($\chi^2 = 133.59$) indicates the size of the difference between expected and observed frequencies.
- The p-value ($p < 0.001$) indicates that these differences are statistically significant and not due to chance.

Statistical Analysis: Dominant Word Orders

Dominant word order is the word order that appears most frequently in a corpus, typically more than twice as often as the next most frequent pattern.

- In SSJ, SVO is the dominant word order, appearing more than twice as often as OVS.
- In SST, there is no clear dominant order, classified as No Dominant Order (NDO).

Hypothesis 2: Supported – SVO dominates in SSJ, while SST shows more variability.

Statistical Analysis: Distributional Comparisons

Word order patterns across SST and SSJ according to distributional metrics:

- **Jensen-Shannon Divergence:** 0.148 (moderate divergence)
- **Entropy:** SST (1.52) has more variability than SSJ (1.29)
- **Euclidean Distance:** 0.20 (notable but not overwhelming difference)
- **Pearson Correlation:** 0.93, **Spearman Rank Correlation:** 0.94 (strong alignment)

These values indicate that spoken Slovenian (SST) shows more variability in word order than written Slovenian (SSJ), but both corpora have similar overall ranking of patterns.

Hypothesis 3: Partially supported – SST shows greater variability and markedness, diverging from the typological norm.

Statistical Analysis: Differences in Proportions 1

The heatmap shows the proportional differences between SST and SSJ:

- Positive values: Patterns more frequent in SST (spoken Slovenian)
- Negative values: Patterns more frequent in SSJ (written Slovenian)
- OSV, SOV are more common in SST, reflecting its flexibility.
- SVO dominates in SSJ, showing the structured nature of written language.
- Less common patterns (VOS, VSO) show minor differences, indicating their limited role.

Statistical Analysis: Differences in Proportions 2

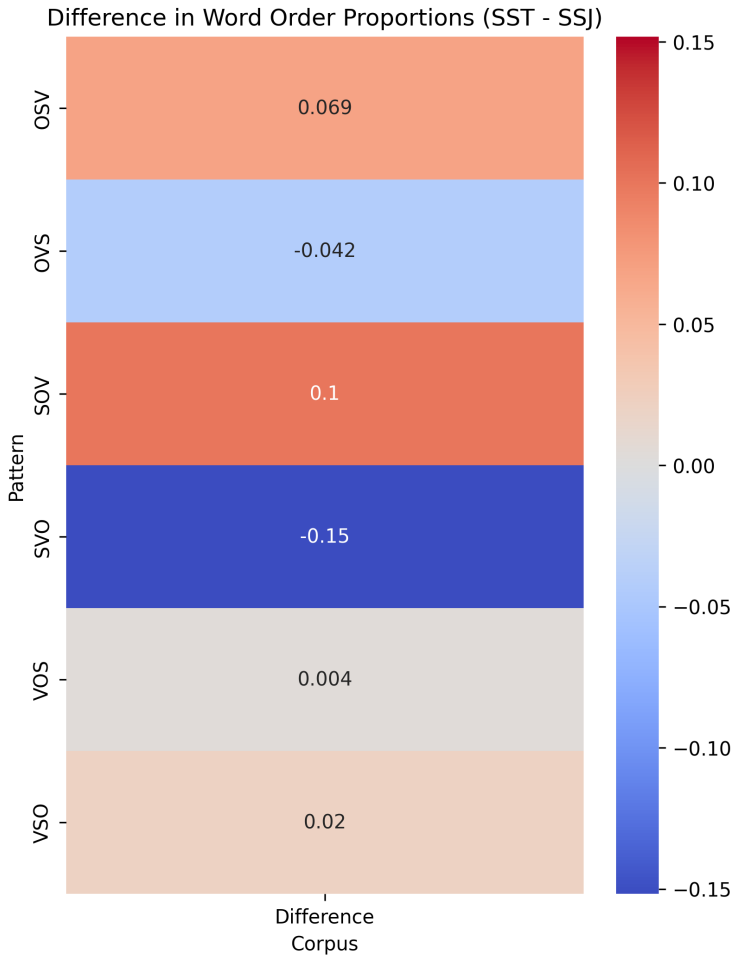

Figure: Comparison of Word Order Variability in Written and Spoken Slovenian

Drafting the Paper

The final phase involved drafting the research paper, available here. Key points from this project include:

- Comparison of spoken and written Slovenian word orders.
- Statistical analysis and the metrics used to evaluate differences.
- Implications of these findings for linguistic theory and computational applications.
- Providing a foundation for future research on other patterns.

Since this field is so specific and there is limited research on it, these findings may be valuable for others interested in the area.

Challenges

Key challenges included:

- Data extraction from WALS was difficult due to dynamic website elements, requiring a manual process.
- Tool limitations with STARK, especially in handling indirect relations.
- Probabilistic modeling was complex, indicating the need for more advanced techniques.
- Deployment of the Streamlit app faced additional difficulties.

Conclusion and Future Work

This project provided insights into Slovenian word order and UD application in typology. While the findings were not groundbreaking, it laid a foundation for future research.
The project also contributed to a GitHub repository.
Future plans:

- Extend analysis to other languages.
- Train models to predict word order patterns.
- Publish the dataset on HuggingFace.
- Create a Gradio app for users to test the model with their own sentences.