

# Under the magnifying glass. Dimensions of variation in the contemporary Timok variety Documentation

Teodora Vuković

08.07.2020

## Introduction

The present document is appendix to the manuscript Under the magnifying glass. Dimensions of variation in the contemporary Timok variety.

The manuscript deals with morphosyntactic and socio-geographic variation in a South Slavic Timok variety spoken in Southeast Serbia. Four linguistic features are analysed in the context of variation between East South Slavic/Standard Serbian on the one side, and Balkan Slavic/non-standard on the other. The features selected for the analysis are:

- marking of indirect object and possessor
- post-positive demonstratives
- dative reflexive *si* as a particle
- auxiliary omission in the perfect tense

The present document follows the analysis presented in the paper and provides data and methodological processes used. It thus orderly refers to the sections and subsections from the manuscript.

For the purposes of the present paper, corpus files were searched using Python. The published online version of the corpus will provide different search options.

Note that in this document, some pieces of code have been hidden to make it more readable. The entire code is available in the source script with the .Rmd extension.

## 3. Facets of variation

### 3.1 The analysis of morphosyntactic factors

#### 3.1.1 Marking of indirect object and possessor

The analysis is based on the following variables:

- Dependent variable: type of marking (na + general oblique case vs. inflectional dative)
- Independent variables: function (indirect object, possessor), part-of-speech (nouns, pronouns, ‘other’), nominal categories (proper/common nouns, grammatical number, grammatical gender, animacy)

The data used in the analysis is stored in the file `1_data.xlsx`. The data was extracted from the corpus semi-automatically. Firstly Python script was used to extract all the instances of dative or “NA” + noun/pronoun patterns.

`00_IO_na_search.py`

`00_IO_dative_search.py`

Noun forms were approximated using word endings for inflected and non-inflected forms. For pronouns, a list of all pronominal forms was used (see in scripts). Context where IO is expected to appear is approximated with a list of verbs requiring an IO (see in scripts).

This data was then filtered manually example, by example. The final list of examples was labelled manually for the parametres included in the analysis. The filtered data was further segmented by focusing on particular criteria for each analysis.

Frequencies of na 'on' + general oblique case and synthetic dative are normalized with regard to the overall number of relevant parts of speech and nominal categories retrieved from the corpus and multiplied with 10.000 in case of the PoS, gender and number, but with 1.000 in case of type of noun and animacy.

The file 1\_marking\_examples.xlsx is organized in sheets as follows:

1. Case, PoS, Function - rows contain examples extracted from the corpus. Columns contain information about Case, Function, PoS for each example (manually annotated)
2. IO PoS RAW - data from Case, PoS, Function, only for IO. It contains also a summary table with absolute frequencies regarding PoS.
3. POSS PoS RAW - data from Case, PoS, Function, only for POSS. It contains also a summary table with absolute frequencies regarding PoS.
4. Freq PoS tabls - repeated summary tables from zbirne table 2. IO PoS RAW and 3. POSS PoS RAW, with calculated percentages, normalized per total number of the respective category.
5. Nominal categories RAW data - (for nouns only!) rows contain examples extracted from the corpus. Columns contain information about nominal categories: Type of Noun (proper, common), Gender (masculine, feminine, neuter), Number (singular, plural), Animacy (animate, inanimate).
6. % for Nominal categories - Summary table based on data from 5. Nominal categories RAW data, with percentages and normalized frequencies per total number of nouns of each type/gender/number/animacy. The data for Type of Nouns is marked in yellow. The final table used for Figure 3 is highlighted in red.
7. corpus\_PoS\_frequencies - frequencies extracted from the corpus for each PoS and nominal categories. The last row shows total frequency for each column.

In what follows analyses are presented as they appear in the paper.

Chi square test is used to compare analysed observations of analytic vs. synthetic marking in the whole sample. The test is performed using the data in the file 1\_analytic\_synthetic\_marking.csv' which contains all examples of IO and POSS extracted from the corpus, labelled for the type of marking: analytic=0, synthetic=1 (from the file 1\_data.xlsx, sheet 1. Case, PoS, Function, column Case). The values were relabelled below 0="NA+OBL", 1="DAT" here for clearer representation.

```
head(analytic_synthetic_marking)
```

```
##           Informant Case
## 1 TOR_C_0001_tagged.txt    0
## 2 TOR_C_0001_tagged.txt    0
## 3 TOR_C_0001_tagged.txt    0
## 4 TOR_C_0001_tagged.txt    0
## 5 TOR_C_0001_tagged.txt    1
## 6 TOR_C_0001_tagged.txt    0
```

The sum of each category is used as input for Chi-square test.

```
head(analytic_synthetic_marking_chisq)
```

```
##
##    0    1
## 763 132
```

```
chisq.test(analytic_synthetic_marking_chisq)
```

```
##
## Chi-squared test for given probabilities
##
## data:  analytic_synthetic_marking_chisq
```

```
## X-squared = 444.87, df = 1, p-value < 2.2e-16
```

Chi-square test is used to compare frequencies of analytic and synthetic type of marking with regard to their function (indirect object, possessive).

```
head(marking_function_chisq)
```

```
##      analytic synthetic
## IO      480      112
## POSS    283      20
```

```
chisq.test(marking_function_chisq, simulate.p.value = TRUE)
```

```
##
## Pearson's Chi-squared test with simulated p-value (based on 2000
## replicates)
##
## data:  marking_function_chisq
## X-squared = 24.187, df = NA, p-value = 0.0004998
```

The percentage of each category is visualised in Figure 1, based on the data from the file 1\_marking\_type\_function.csv. The data was obtained by categorizing each example based on the type of marking and function (see 1\_data.xlsx, 1. Case, PoS, Function, columns Case and function).

```
marking_type_function
```

```
##   X marking_type marking_function marking_count marking_percent X.1
## 1 1      NA+OBL      IO (66.14%)         480         53.63   NA
## 2 2      NA+OBL      POSS (33.86%)        283         31.62   NA
## 3 3      DATIVE      IO (66.14%)         112         12.51   NA
## 4 4      DATIVE      POSS (33.86%)         20          2.23   NA
```

```
Figure1
```

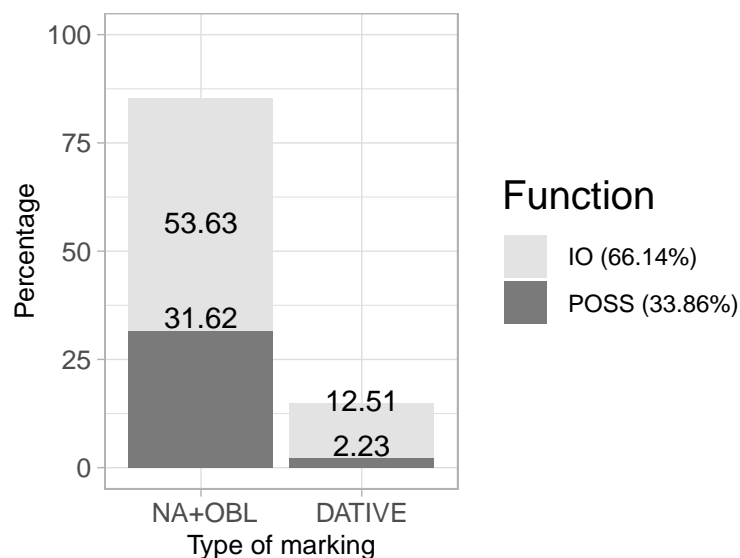


Figure 1: Figure 1: Type of marking: overall and per functions

The data for analytic and syntetic marking was sorted based on part-of-speech categories. Frequencies were extracted from the 1\_data.xlsx file and presented in the file 1\_marking\_function\_pos.csv.

```
head(marking_function_pos)
```

```
##   X Function Type_of_marking   POS Values
## 1 1      IO      NA+OBL      Noun 43.61
## 2 2      IO      NA+OBL Pronoun 14.83
## 3 3      IO      NA+OBL  Other  8.57
## 4 4      IO      DAT      Noun  4.10
## 5 5      IO      DAT Pronoun  9.01
## 6 6      IO      DAT      Other  0.86
```

The values for indirect object and possessive function with respect to PoS categories are presented in Figure 2.

Figure2

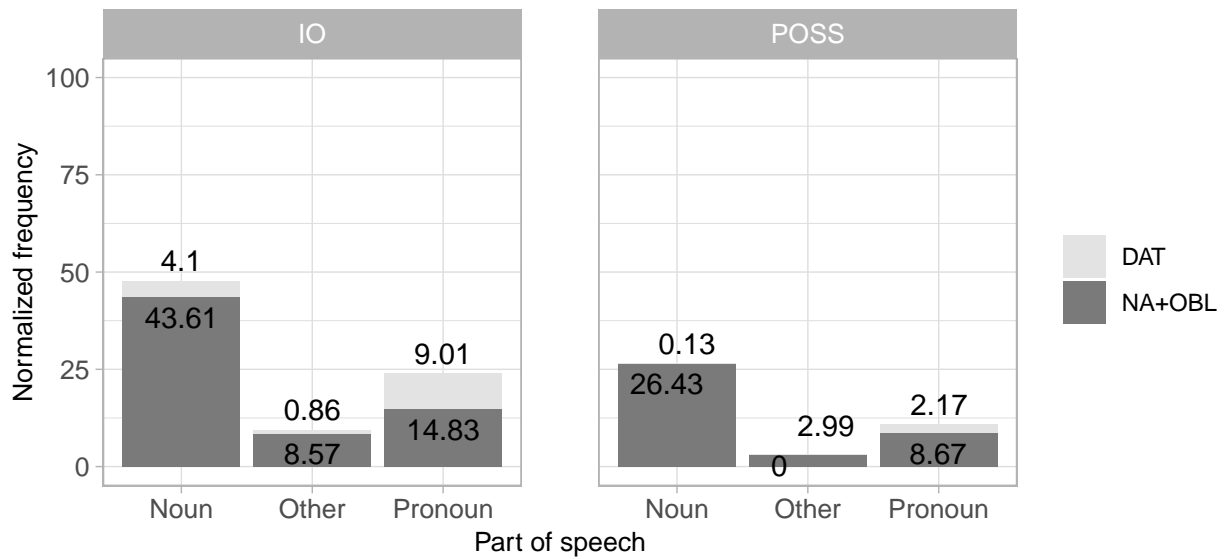


Figure 2: Figure 2: Marking of case in IO and POSS function with respect to PoS

Data for IO was categorized based on nominal categories (type of noun, gender, number animacy) and stored in the file 1\_marking\_nominal\_category.csv. It is visualised in Figure 3.

Figure3

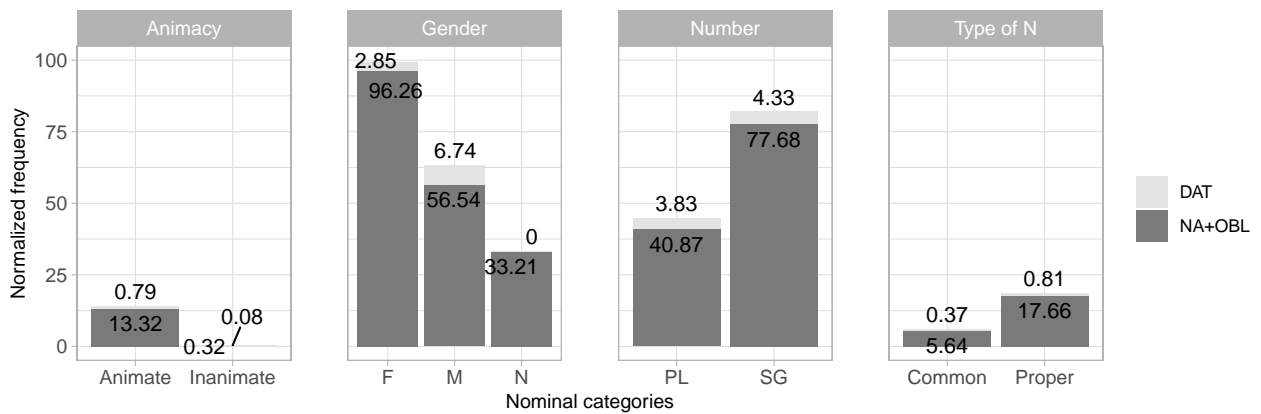


Figure 3: Figure 3: Case marking with regard to nominal categories

### 3.1.2 Post-positive demonstratives

In order to identify the distribution of different forms of PPD (nominative/unmarked vs. accusative/oblique, as well based on gender), nouns containing PPD were compared against bare nouns. The comparison regarding gender includes all nouns, while the comparison concerning case takes into account only nouns of the grammatical feminine gender ending in -a and masculine animate nouns ending in a consonant (regardless of the syntactic position). The following variables were used:

- Dependent variable: frequency of the nouns containing PPD and bare nouns (absolute and normalized per 10000 nouns)
- Independent variables: gender of nouns (masculine ending in consonant, feminine ending in -a, neuter), case of nouns (nominative/unmarked and oblique/accusative singular)

Words with articles were extracted from the corpus based on their form. The resulting list contains 817 types of all PoS categories (1313 tokens). These words were manually annotated for PoS categories for the purposes of the analysis, because some PoS labels had been initially wrong. The examples of words containing PPD are stored in the file 2\_PPD\_examples.xlsx.

The analysis in the present study involved nouns only, as explained in the manuscript. For the analysis of nouns of all three genders, the data was extracted and categorized using PoS tags. The extraction of nouns of grammatical feminine gender (feminine and masculine nouns ending in -a) and animate masculine nouns ending in consonant was based on manually selected lists of lemmas of each category. The lists were created by extracting all feminine and masculine nouns ending in -a and removing the incorrect instances. The feminine group includes the first 1337 correct lemmas, because the proportion of unwanted results became much bigger afterwards. Both masculine groups contain all lemmas retrieved from the corpus fitting the criteria. Results regarding case were obtained using the lists of lemmas and the morphological form. The lists of lemmas are available in files 2\_PPD\_masculine\_nouns\_in\_a.txt, 2\_PPD\_masculine\_animate\_nouns\_in\_consonant.txt, 2\_PPD\_feminine\_nouns\_in\_a.txt. The number of elements in each list is shown below (not included in the manuscript).

lists\_of\_lemmas\_gender

##	Category	List_size
## 1	Masculine animate in consonant	336
## 2	Feminine in -a	1337
## 3	Masculine in -a	109

All nouns were compared for gender, categorized based on gender and the presence of PPD. The total number of bare nouns of all genders is 74769. The total number of nouns with PPD is 1182. The data used in the analysis is presented in the file 2\_PPD\_gender\_absfreq.csv.

Absolute frequencies of each gender in bare nouns and nouns containing a PPD are presented in the file 2\_PPD\_gender\_prop.csv.

PPD\_gen\_all

##	Bare_nouns	Nouns_with_PPD
## F	31549	612
## M	34100	413
## N	9120	157

Proportions of each gender in bare nouns and nouns containing a PPD is shown in Figure 4.

Figure4

Chi-square test shows that there is a significant difference in distribution of gender in bare nouns and nouns carrying a PPD.

chisq.test(PPD\_gen\_all)

##

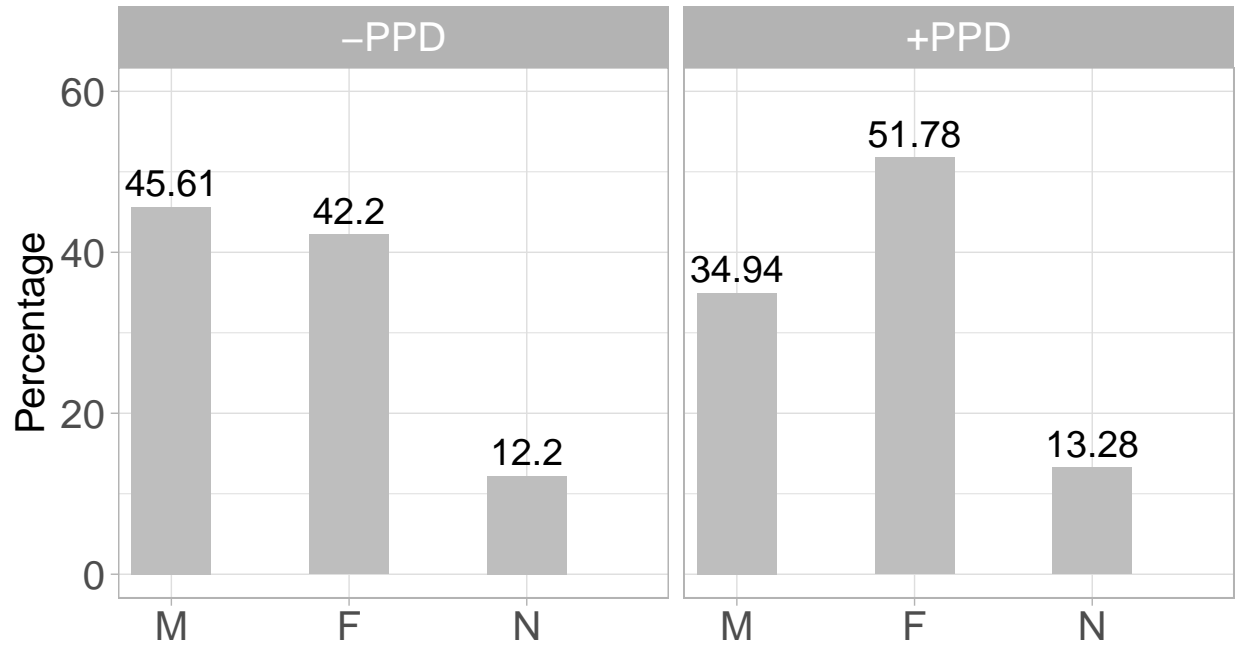


Figure 4: Figure 4: PPD and gender of nouns

```
## Pearson's Chi-squared test
##
## data: PPD_gen_all
## X-squared = 55.482, df = 2, p-value = 8.96e-13
```

Data used for the analysis of the distribution of case marking has been categorized based on the presence of PPD (bare vs. with PPD) and case inflections. The same cateorization was was performed for masculine and feminine nouns separately.

```
ppd_case_gender
```

```
## X Case PPD All_nouns Masculine_animate Feminine
## 1 1 NOM -PPD 50.29 68.51 46.24
## 2 2 OBL -PPD 49.71 31.49 53.76
## 3 3 NOM +PPD 59.31 79.27 56.63
## 4 4 OBL +PPD 40.69 20.73 43.37
```

Mosaic plots presenting the proportion of nouns marked and unmarked for case (all nouns, masculine, feminine nouns) is displayed in Figure 5. Figure 5: Proportions of nominative/unmarked and oblique/accusative case forms in nouns with and without PPD

```
Figure5 = grid.arrange(ppd_mosaic_all, ppd_mosaic_masc, ppd_mosaic_fem, nrow = 1)
```

### 3.1.3 Particle SI

The analysis is based on the following variables:

- Dependent variable: absolute and normalized frequency of the clitic si used non-pronominally (per 1,000 verbs)
- Independent variables: properties of the verb (person and number, animacy, reflexivity, lexical type), variation in the syntactic patterns in the contact position between si and the verb

The search was done semiautomatically. A python script was used to search for all the occurrences of the word 'si' and some unwanted results were excluded (such as the forms of the 2nd person auxiliary, e.g. Ti si

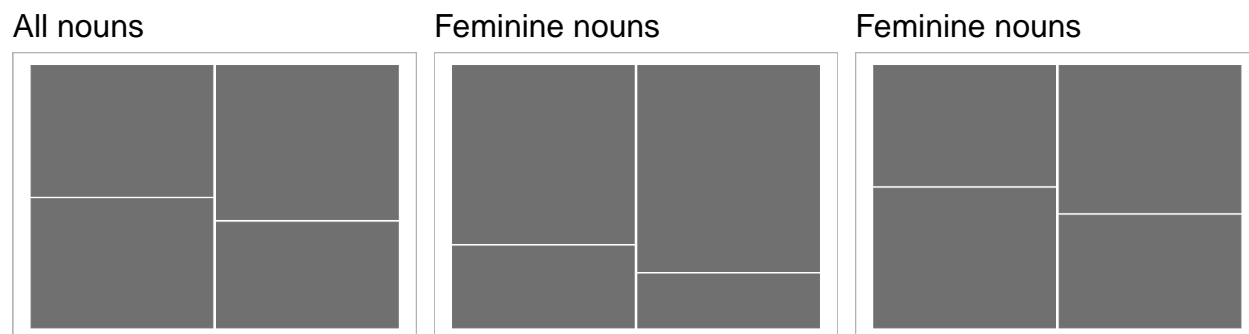


Figure 5: Figure 5: Proportions of nominative/unmarked and oblique/accusative case forms in nouns with and without PPD

gledal. ‘You were watching.’). The rest was removed manually, by checking each example. Each examples was annotated manually for the criteria described in the manuscript.

Manually annotated data used in the analysis is shown in the file XX.xlsx

The frequency of particle SI categorized based on person and number is shown below (see file 3\_si\_person.csv).

si_person					
##	X	si_person_pers	si_person_labels	si_person_value	
##	1	1	SG	1SG	19.13
##	2	2	SG	2SG	0.80
##	3	3	SG	3SG	44.22
##	4	4	PL	1PL	16.15
##	5	5	PL	2PL	3.56
##	6	6	PL	3PL	16.15

Figure6

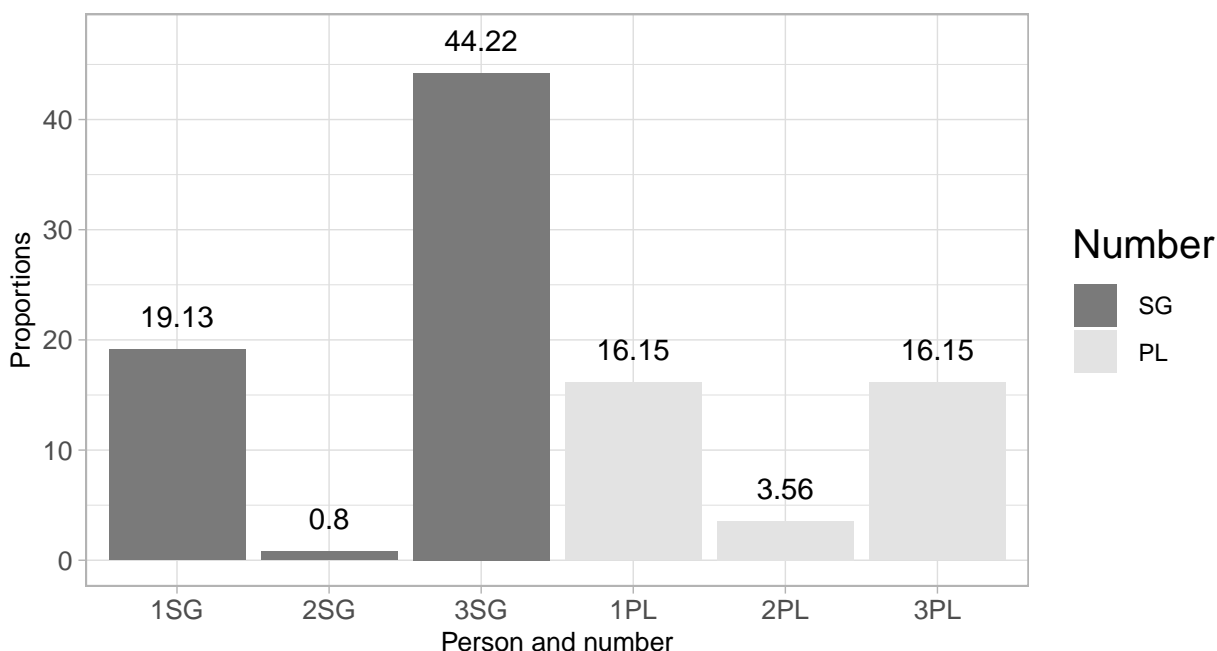


Figure 6: Figure 6: Si particle frequency: Person and number of the verb

Frequency is compared on the basis of grammatical categories: animacy of the subject, reflexivity of the predicate, voice of the predicate.

Animacy (see file 4\_si\_animacy.csv):

```
si_animacy

##   X si_animacy_label si_animacy_value
## 1 1      Animate      83.35
## 2 2    Inanimate     16.65
```

Reflexivity (see file 4\_si\_refl.csv):

```
si_refl

##   X si_refl_label si_refl_value
## 1 1 Non-reflexive   91.78
## 2 2   Reflexive     8.22
```

Voice (see file 4\_si\_voice.csv):

```
si_voice

##   X si_voice_label si_voice_value
## 1 1      Active    96.15
## 2 2     Passive     3.85
```

Figure 7 shows the frequencies of the occurrences of the particle 'si' categorized based on the three linguistic features: animacy, reflexivity, voice.

```
Figure7 = grid.arrange(si_animacy_plot, si_refl_plot, si_voice_plot, nrow = 1)
```

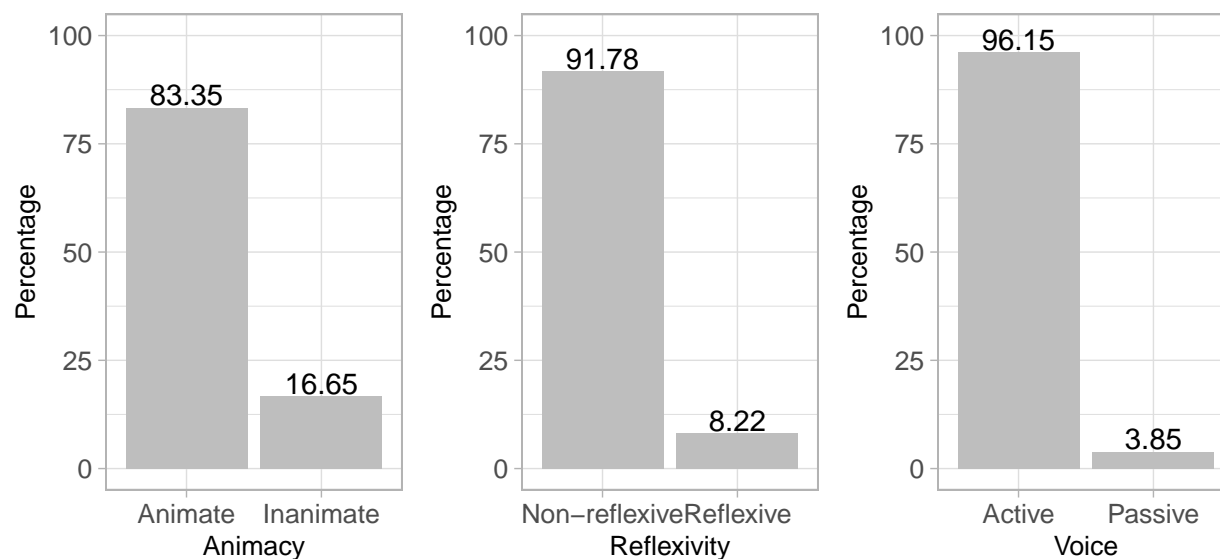


Figure 7: Figure 7: Si particle frequency: Animacy, reflexivity, voice

The data presenting the analysis of the order of particle 'si' and the verb is shown in Figure 8 (see file 4\_si\_order\_csv).

Figure8



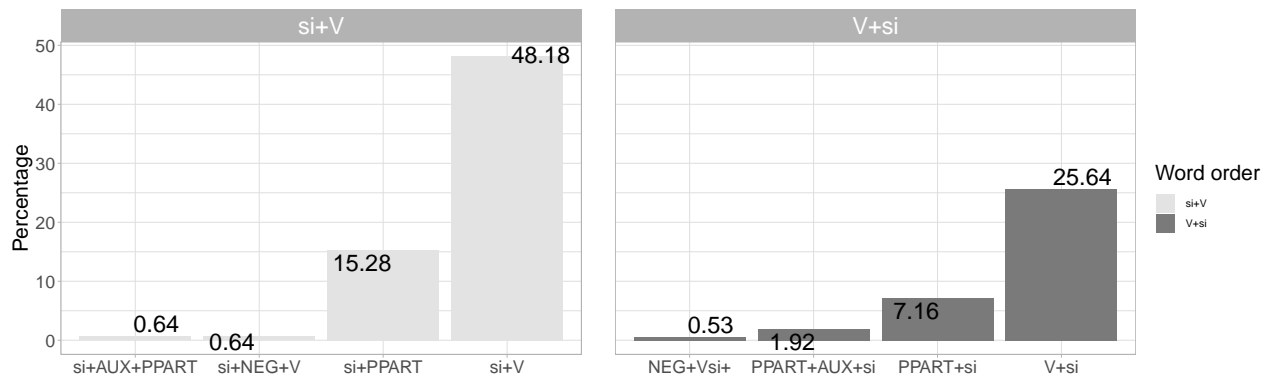


Figure 8: Figure 8: The proportions of contact patterns between si and the verb

### 3.1.4 Auxiliary omission in perfect tense

The quantitative analysis of the use of the -AUX forms is based on the following variables:

- The dependant variable: normalized (to the total number of the examples of the use of the perfect tense) frequency of the -AUX and +AUX forms per location.
- The independant variables: gender, age + several categorical linguistic variables: aspect, transitivity, lexical group.

The automatic search for relevant examples in the Timok corpus made with a user Python script required all the clauses where perfect participle tense is used. These examples were automatically divided into three groups: clauses with -AUX perfect forms, clauses with +AUX perfect forms and clauses with potential mood (the latter group was subsequently excluded from the analysis). The data are presented in two tables. In the table “verb\_timok.csv”, each observation represents texts from a single location. In cases, where there were several recordings from one location, the scores were merged in one observation, assuming that every speaker from a single location represents the same local variety. The table contains the following columns (not all of the information was used in the present study, see in more detail in (Makarova 2021)): ID (recording ID), LOCATION (name of the village), LATITUDE, LONGITUDE, total (total number of examples), total\_aux (total number of +AUX forms), dist\_boarder (distance to the bulgarian border), no\_aux (total number of -AUX forms), total\_aux\_prop (proportion of + AUX forms). In the table “verb\_tim\_soc2.csv”, every observation is a text from a single recording (i.e. the data from same locations are not merged in one observation) contains additional data on age and gender of every informant. The table “gramm.csv” contains contingency table used for the chi-squared test of the categorical linguistic variables.

The file 4\_overall\_freq.csv shows the frequency of analysed examples of the perfect tense that display +AUX (total\_aux) and -AUX (no\_aux) pattern per transcript (normalized per 1,000 occurrences of the perfect tense).

```
aux_overall = read.table('4_aux_overall_freq.csv', sep = '\t', header = TRUE, row.names = "ID")
head(aux_overall)
```

```
##          total_aux no_aux
## TOR_C_0001      547   453
## TOR_C_0002      382   608
## TOR_C_0003      342   658
## TOR_C_0004      483   517
## TOR_C_0005      523   471
## TOR_C_0006      526   474
```

The distribution of +AUX/-AUX patterns in the overall sample is shown in Figure 9.

Figure 9: +AUX and -AUX frequencies in the overall sample

Figure9

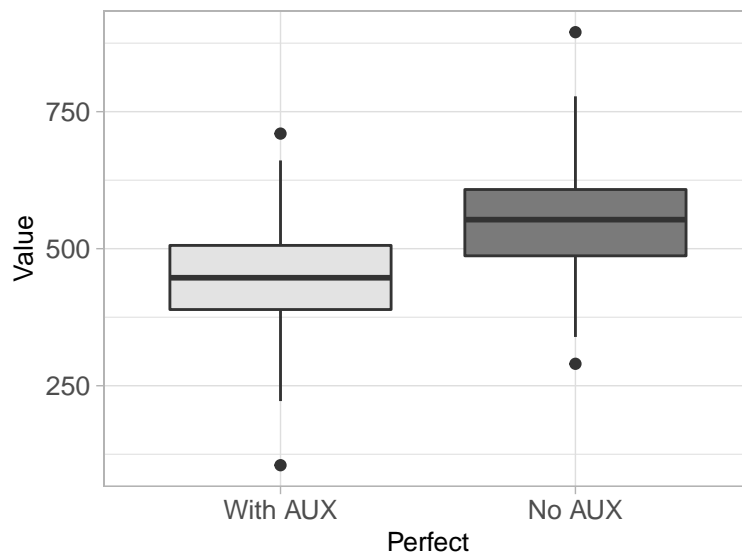


Figure 9: Figure 9: +AUX and -AUX frequencies in the overall sample

The total frequency of +AUX and -AUX pattern are presented below (see 4\_aux\_overall\_chisq.csv):

```
head(aux_overall_chisq)
```

```
## total_no_aux total_with_aux
## 1          35844          28849
```

Chi-squared test is used to compare the total frequencies of +AUX and -AUX.

```
chisq.test(aux_overall_chisq)
```

```
##
## Chi-squared test for given probabilities
##
## data: aux_overall_chisq
## X-squared = 756.34, df = 1, p-value < 2.2e-16
```

The data used in the analysis of verb categories on the use of AUX is kept in the file 4\_gramm.csv.

```
aux_gramm
```

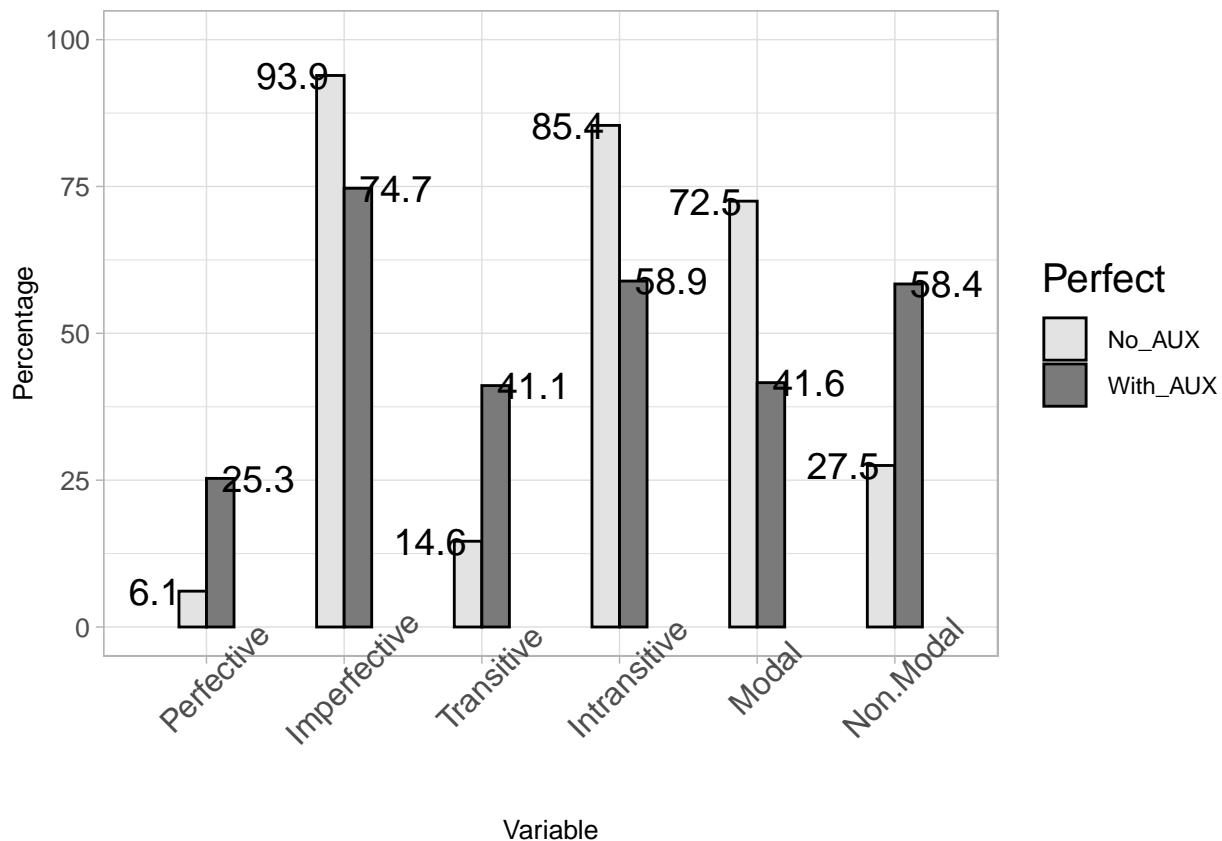
```
## Perfect Perfective Imperfective Transitive Intransitive Modal Non.Modal
## 1 No_AUX          84          1290          201          1173    996          378
## 2 With_AUX        332           979          539          772    546          765
```

The proportions of the linguistic properties through the -AUX and +AUX forms are displayed in Figure 10.

Figure 10: Linguistic properties of -AUX and +AUX forms in Timok corpus (proportions)

Figure10

```
## Warning: position_dodge requires non-overlapping x intervals
```



Chi-squared tests are performed for each verb category separately: aspect, transitivity, lexical group (+/- modal).

Aspect:

```
gramm_table_aspect
```

```
##           Perfective imperfective
## No_AUX           84           1290
## With_AUX          332           979
```

```
chisq.test(gramm_table_aspect)
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  gramm_table_aspect
## X-squared = 187.63, df = 1, p-value < 2.2e-16
```

Transitivity:

```
gramm_table_trans
```

```
##           Transitive Intransitive
## No_AUX           201           1173
## With_AUX          539           772
```

```
chisq.test(gramm_table_trans)
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
```

```
##
## data: gramm_table_trans
## X-squared = 234.38, df = 1, p-value < 2.2e-16

Lexical group (+/-modal):
gramm_table_lex

##           Modal Not modal
## No_AUX    996      378
## With_AUX   546      765

chisq.test(gramm_table_lex)

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: gramm_table_lex
## X-squared = 259.76, df = 1, p-value < 2.2e-16
```

### 3.2 Analysis of the socio-geographic factors

Analysis of social and geographic factors involved the dependent variables:

- proportion of the analytic marking of the indirect object and the possessive per total examples analysed per location
- normalized frequency of PPD per 1000 nouns per location
- normalized frequency of particle SI per 1000 verbs
- normalized frequency of AUX omission per 1000 cases of perfect tense

The independent variables regarding geographic distribution are:

- geographic longitude
- geographic latitude
- altitude
- distance from the city of Knjaževac

The independent variables regarding socio-demographic distribution are:

- age
- gender

#### Analysis of the geographic factors

We firstly present the comparison of the linguistic frequencies with geographic variables (longitude, latitude, altitude, distance from the city). For the analysis of the geographic variables, frequency values have been aggregated for each location. The dependant variables and the geographic variables are continuous. The dependant variable in all 4 analyses does not have normal distribution, so Kendall's correlation test was used. Geographic distribution of frequencies of each feature is presented on maps. (not included in the manuscript)

#### Marking of indirect object and possessor:

```
head(marking_geo)

##           LOCATION N.of.NA.Oblq N.of.DAT ALL..IO.POSS. Freq.NA...ALL
## 1           Žukovac           3         0           3           100
## 2             Žlne           3         0           3           100
## 3  Gornja Bela Reka           1         0           1           100
## 4  Gornja Sokolovica          15         0          15           100
```

## 5	Drvnik	27	0	27	100		
## 6	Mali Izvor	20	0	20	100		
##	Freq.DAT...ALL	LATITUDE	LONGITUDE	Location	Altitude	DIST_Bul	DIST_city
## 1	0	43.53035	22.28190	4	274	15.57	5.89
## 2	0	43.52175	22.23101	5	320	20.28	5.10
## 3	0	43.76383	22.16492	7	235	38.61	14.82
## 4	0	43.52082	22.31761	9	305	13.23	8.34
## 5	0	43.53809	22.37374	10	597	7.92	11.96
## 6	0	43.73677	22.33321	11	205	5.99	18.24

Kendall's rank correlation between analytic case marking frequencies and geographic variables.

```
cor.test(marking_geo$Freq.NA...ALL, marking_geo$LONGITUDE, method = c("kendall"))
```

```
##
## Kendall's rank correlation tau
##
## data: marking_geo$Freq.NA...ALL and marking_geo$LONGITUDE
## z = 1.0804, p-value = 0.28
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.1017915
```

```
cor.test(marking_geo$Freq.NA...ALL, marking_geo$LATITUDE, method = c("kendall"))
```

```
##
## Kendall's rank correlation tau
##
## data: marking_geo$Freq.NA...ALL and marking_geo$LATITUDE
## z = 0.41866, p-value = 0.6755
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.03944419
```

```
cor.test(marking_geo$Freq.NA...ALL, marking_geo$Altitude, method = c("kendall"))
```

```
##
## Kendall's rank correlation tau
##
## data: marking_geo$Freq.NA...ALL and marking_geo$Altitude
## z = 0.013506, p-value = 0.9892
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.00127351
```

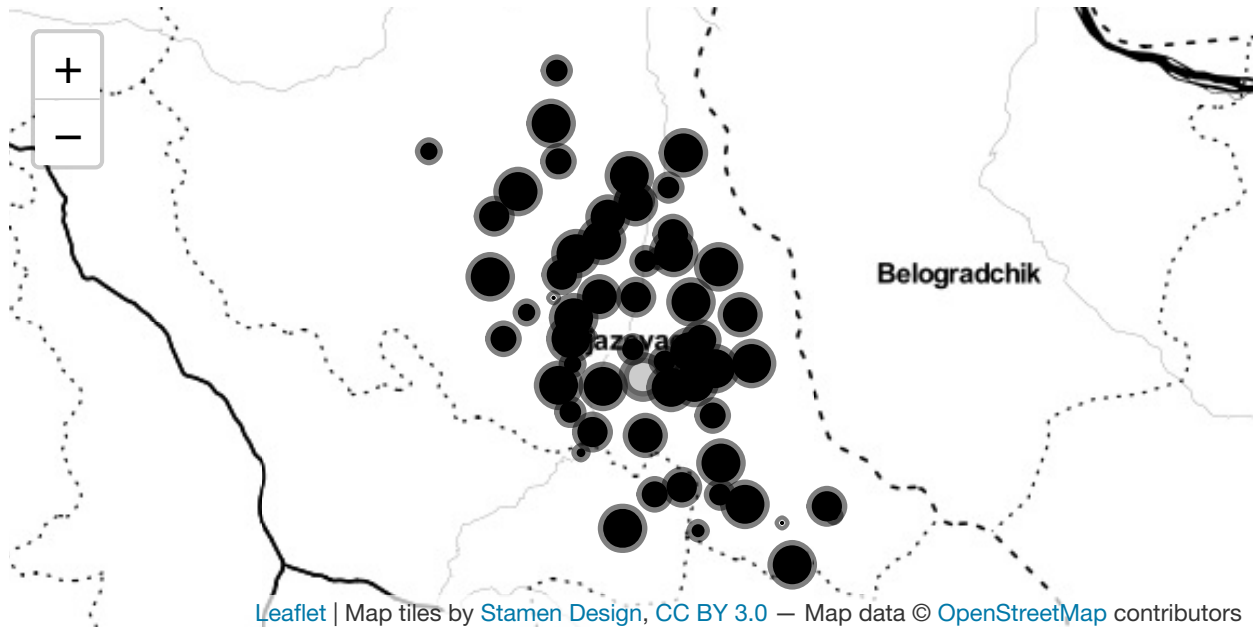
```
cor.test(marking_geo$Freq.NA...ALL, marking_geo$DIST_city, method = c("kendall"))
```

```
##
## Kendall's rank correlation tau
##
## data: marking_geo$Freq.NA...ALL and marking_geo$DIST_city
## z = -0.29037, p-value = 0.7715
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
```

```
##      tau
## -0.02736446
```

The map presenting the areal distribution of the analytic case marking in IO and POSS:

marking\_map



Post-positive demonstratives:

`head(ppd_geo)`

```
##      X      LOCATION art_freq LATITUDE LONGITUDE Altitude DIST_city X.1 X.2 X.3
## 1 1      Aldinac      10 43.54287  22.41992     623    16.44  NA  NA <NA>
## 2 2      Balanovac     12 43.58993  22.13367     327     7.04  NA  NA <NA>
## 3 3      Balinac      70 43.56462  22.35576     605    11.58  NA  NA <NA>
## 4 4 Balta Berilovac    20 43.39568  22.45872     419    27.00  NA  NA <NA>
## 5 5      Borovac       2 43.73822  22.00940     199    18.68  NA  NA <NA>
## 6 6      Bučje       38 43.67853  22.09256     514    16.05  NA  NA <NA>
```

Kendall's rank correlation between post-positive demonstratives frequencies and geographic variables.

```
cor.test(ppd_geo$art_freq, ppd_geo$LONGITUDE, method = c("kendall"))
```

```
##
## Kendall's rank correlation tau
##
## data:  ppd_geo$art_freq and ppd_geo$LONGITUDE
## z = 3.7682, p-value = 0.0001644
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.3320316
```

```
cor.test(ppd_geo$art_freq, ppd_geo$LATITUDE, method = c("kendall"))
```

```
##
## Kendall's rank correlation tau
```

```
##
## data:  ppd_geo$art_freq and ppd_geo$LATITUDE
## z = -2.3157, p-value = 0.02058
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## -0.2040447

cor.test(ppd_geo$art_freq, ppd_geo$Altitude, method = c("kendall"))
```

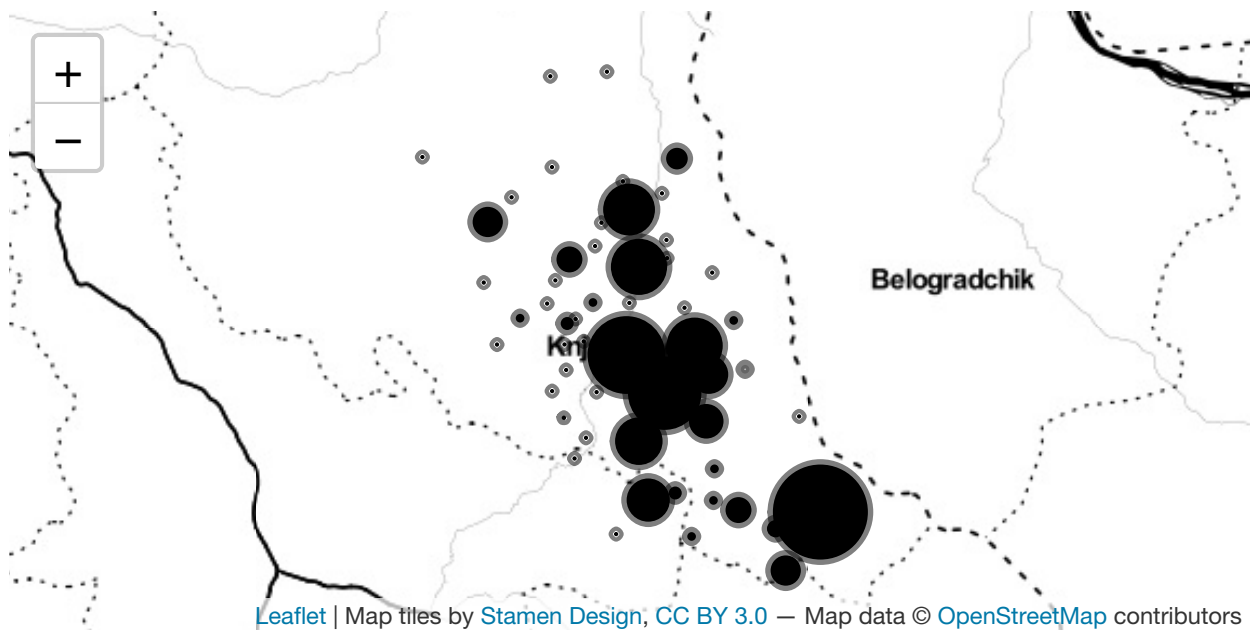
```
##
## Kendall's rank correlation tau
##
## data:  ppd_geo$art_freq and ppd_geo$Altitude
## z = 1.649, p-value = 0.09915
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.1453711
```

```
cor.test(ppd_geo$art_freq, ppd_geo$DIST_city, method = c("kendall"))
```

```
##
## Kendall's rank correlation tau
##
## data:  ppd_geo$art_freq and ppd_geo$DIST_city
## z = 1.774, p-value = 0.07606
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.1563519
```

The map presenting the areal distribution of the post-positive demonstratives:

```
ppd_map
```



## Particle SI:

```
head(si_geo)
```

```
##      LOCATION Number.of..si. Number.of.verbs Normalized.FREQ.of..SI. LATITUDE
## 1   Ošljane      48             996             4.8192771 43.66194
## 2   Lepena      91            5005             1.8181818 43.58023
## 3 Trgovište     28            1938             1.4447884 43.55598
## 4   Žukovac     31            1717             1.8054747 43.53035
## 5     Žlne       7             734             0.9536785 43.52175
## 6   Vasilj      16            2648             0.6042296 43.56564
##      LONGITUDE Altitude DIST_Bul DIST_city
## 1  22.31988      520      3.06      16.11
## 2  22.16977      315     17.35       9.05
## 3  22.26894      230     16.56       2.62
## 4  22.28190      274     15.57       5.89
## 5  22.23101      320     20.28       5.10
## 6  22.10432      415     26.75       7.51
```

Kendall's rank correlation between particle 'si' frequencies and geographic variables.

```
cor.test(si_geo$Normalized.FREQ.of..SI., si_geo$LONGITUDE, method = c("kendall"))
```

```
##
## Kendall's rank correlation tau
##
## data: si_geo$Normalized.FREQ.of..SI. and si_geo$LONGITUDE
## z = 0.2482, p-value = 0.804
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.02238355
```

```
cor.test(si_geo$Normalized.FREQ.of..SI., si_geo$LATITUDE, method = c("kendall"))
```

```
##
## Kendall's rank correlation tau
##
## data: si_geo$Normalized.FREQ.of..SI. and si_geo$LATITUDE
## z = -0.32869, p-value = 0.7424
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## -0.02964307
```

```
cor.test(si_geo$Normalized.FREQ.of..SI., si_geo$Altitude, method = c("kendall"))
```

```
##
## Kendall's rank correlation tau
##
## data: si_geo$Normalized.FREQ.of..SI. and si_geo$Altitude
## z = 0.98612, p-value = 0.3241
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.08898307
```

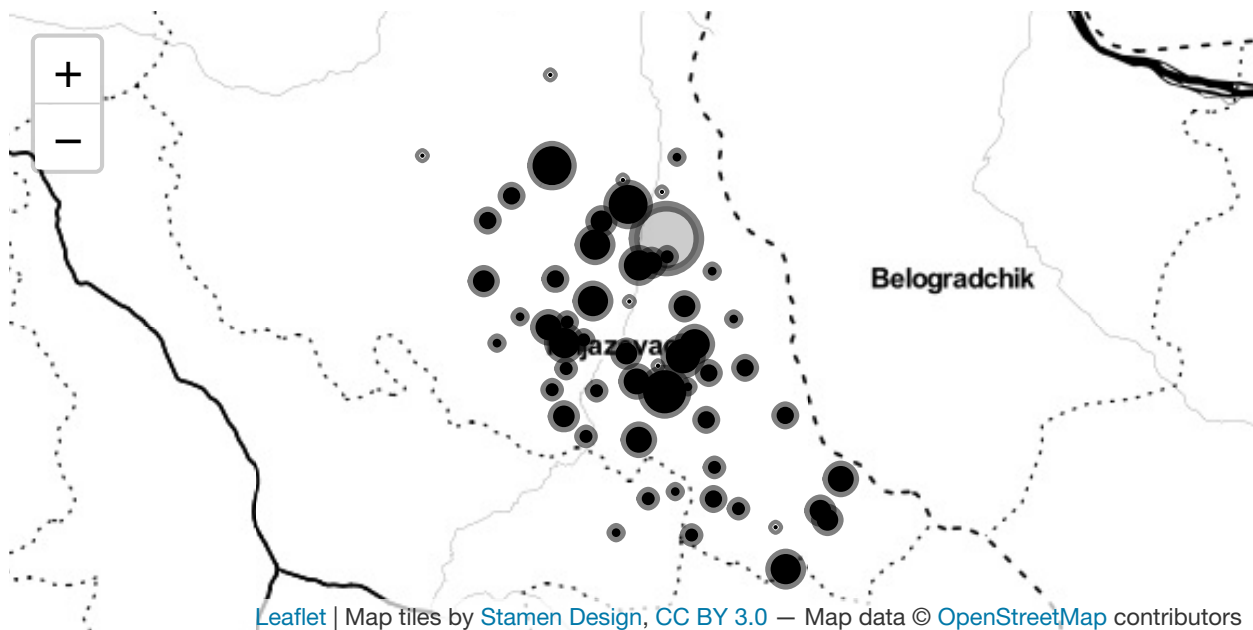


```
cor.test(si_geo$Normalized.FREQ.of..SI., si_geo$DIST_city, method = c("kendall"))
```

```
##
## Kendall's rank correlation tau
##
## data: si_geo$Normalized.FREQ.of..SI. and si_geo$DIST_city
## z = -0.17441, p-value = 0.8615
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## -0.01573374
```

The map presenting the areal distribution of the particle 'si':

```
si_map
```



Auxiliary omission in the perfect tense:

```
head(aux_geo)
```

```
##      X      LOCATION total total_aux no_aux LATITUDE LONGITUDE Altitude DIST_city
## 1 1      Ošljane     95      52      43 43.66194  22.31988      520      16.11
## 2 2      Drvnik    204      78     124 43.53809  22.37374      597      11.96
## 3 3      Balinac   184      63     121 43.56462  22.35576      605      11.58
## 4 4      Čuštica    89      43      46 43.35698  22.47159      794      33.74
## 5 5 Gornje Zuniče  155      81      73 43.60401  22.27268      235       4.13
## 6 6      Novo Korito 19      10       9 43.63191  22.37807      423      17.68
```

Kendall's rank correlation between Auxiliary omission in the perfect tense frequencies and geographic variables.

```
cor.test(aux_geo$no_aux, aux_geo$LONGITUDE, method = c("kendall"))
```

```
##
## Kendall's rank correlation tau
##
## data: aux_geo$no_aux and aux_geo$LONGITUDE
```

```
## z = -0.046358, p-value = 0.963
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## -0.00397912
```

```
cor.test(aux_geo$no_aux, aux_geo$LATITUDE, method = c("kendall"))
```

```
##
## Kendall's rank correlation tau
##
## data: aux_geo$no_aux and aux_geo$LATITUDE
## z = 0.16805, p-value = 0.8665
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.01442789
```

```
cor.test(aux_geo$no_aux, aux_geo$Altitude, method = c("kendall"))
```

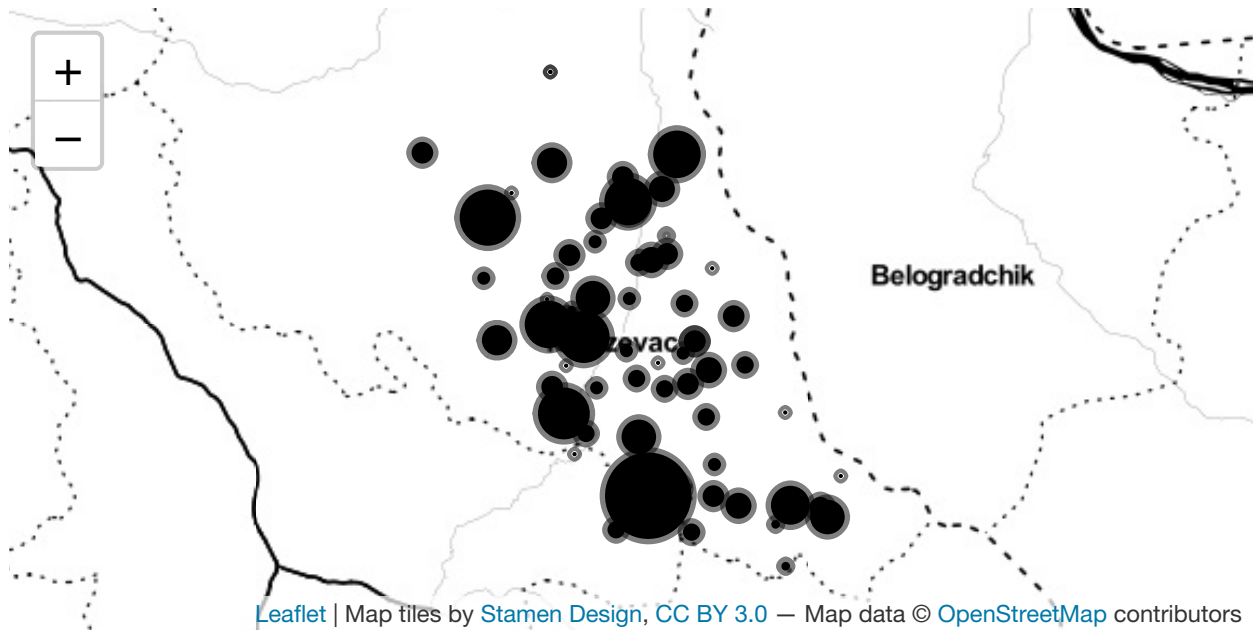
```
##
## Kendall's rank correlation tau
##
## data: aux_geo$no_aux and aux_geo$Altitude
## z = -0.81136, p-value = 0.4172
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## -0.0697385
```

```
cor.test(aux_geo$no_aux, aux_geo$DIST_city, method = c("kendall"))
```

```
##
## Kendall's rank correlation tau
##
## data: aux_geo$no_aux and aux_geo$DIST_city
## z = 0.41725, p-value = 0.6765
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.03584768
```

The map presenting the areal distribution of the auxiliary omission in the perfect tense:

```
aux_map
```



### Analysis of the socio-demographic factors

What follows is the correlation of the linguistic frequencies with socio-demographic variables (age, gender). For the analysis of the geographic variables, frequency values have been aggregated for each location. The dependant variable is continuous, while the geographic variables are binary. The dependant variable in all analyses except PPD do not have normal distribution, so Wilcoxon Rank Sum test was used, while for PPD, we used Pearson's rank correlation.

### Marking of indirect object and possessor:

(see file 1\_marking\_socio\_all.csv)

Analytic marking and age:

```
head(marking_age)
```

```
##          Informant N.of.NA.Oblq N.of.DAT ALL..IO.POSS. Freq.NA...ALL
## 1 TOR_C_0001_tagged.txt          6          1          7      85.71429
## 2 TOR_C_00010_tagged.txt        17          0         17     100.00000
## 3 TOR_C_00011_tagged.txt          0          1          1       0.00000
## 4 TOR_C_00013_tagged.txt          3          0          3     100.00000
## 7 TOR_C_00017_tagged.txt          1          0          1     100.00000
## 9 TOR_C_00019_tagged.txt        15          0         15     100.00000
##   Freq.DAT...ALL AGE
## 1      14.28571 OLD
## 2       0.00000 OLD
## 3     100.00000 OLD
## 4       0.00000 OLD
## 7       0.00000 OLD
## 9       0.00000 OLD
```

Mann-Whitney test used to compare the distribution across OLD and YOUNG speakers.

```
wilcox.test(Freq.NA...ALL ~ AGE, data = marking_age)
```

```
##
```

```
## Wilcoxon rank sum test with continuity correction
##
## data: Freq.NA...ALL by AGE
## W = 145, p-value = 0.0006728
## alternative hypothesis: true location shift is not equal to 0
```

Analytic marking and gender:

```
head(marking_gender)
```

```
##           Informant N.of.NA.Oblq N.of.DAT ALL..IO.POSS. Freq.NA...ALL
## 1 TOR_C_0001_tagged.txt          6          1          7      85.71429
## 2 TOR_C_00010_tagged.txt        17          0         17     100.00000
## 3 TOR_C_00011_tagged.txt          0          1          1      0.00000
## 4 TOR_C_00013_tagged.txt          3          0          3     100.00000
## 5 TOR_C_00015_tagged.txt          3          0          3     100.00000
## 6 TOR_C_00016_tagged.txt        24          5         29     82.75862
## Freq.DAT...ALL Gender
## 1      14.28571 FEMALE
## 2       0.00000 FEMALE
## 3     100.00000 FEMALE
## 4       0.00000 FEMALE
## 5       0.00000 FEMALE
## 6     17.24138 FEMALE
```

Mann-Whitney test used to compare the distribution accross MALE and FEMALE speakers.

```
wilcox.test(Freq.NA...ALL ~ Gender, data = marking_gender)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: Freq.NA...ALL by Gender
## W = 734.5, p-value = 0.0003797
## alternative hypothesis: true location shift is not equal to 0
```

**Post-positive demonstratives:**

(see files 2\_PPD\_age.csv and 2\_PPD\_gender.csv)

Post-positive demonstratives and age:

```
head(ppd_age)
```

```
##           ID NORM_ART YEAR_OF_BIRTH AGE OLD
## 1 TIM_SPK_0001     59.64         1925 OLD  1
## 2 TIM_SPK_0003    147.87         1930 OLD  1
## 3 TIM_SPK_0004     0.00         1954 OLD  1
## 4 TIM_SPK_0005    587.30         1957 OLD  1
## 5 TIM_SPK_0006     0.00         1957 OLD  1
## 6 TIM_SPK_0007    61.16         1927 OLD  1
```

Mann-Whitney test used to compare the distribution accross OLD and YOUNG speakers.

```
wilcox.test(NORM_ART ~ OLD, data = ppd_age)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
```

```
## data:  NORM_ART by OLD
## W = 185, p-value = 0.002198
## alternative hypothesis: true location shift is not equal to 0
```

Post-positive demonstratives and gender:

```
head(ppd_gender)
```

```
##           ID ART NOUN TOKEN FEMALE GENDER NORM_ART
## 1 TIM_SPK_0161  0  149  1062      0  MALE      0.00
## 2 TIM_SPK_0164  0  136  1111      0  MALE      0.00
## 3 TIM_SPK_0014  6  191  1155      0  MALE    314.14
## 4 TIM_SPK_0163  0  210  1312      0  MALE      0.00
## 5 TIM_SPK_0162  0  200  1357      0  MALE      0.00
## 6 TIM_SPK_0134  0  149  1366      0  MALE      0.00
```

Mann-Whitney test used to compare the distribution accross MALE and FEMALE speakers.

```
wilcox.test(NORM_ART ~ FEMALE, data = ppd_gender)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  NORM_ART by FEMALE
## W = 340, p-value = 0.003727
## alternative hypothesis: true location shift is not equal to 0
```

## Particle SI:

(see file 3\_si\_socio.csv)

Particle 'si' and age:

```
head(si_age)
```

```
##           Informant Age Normalized.FREQ.of..SI.
## 1 TOR_C_0001_tagged.txt OLD              4.819277
## 2 TOR_C_00019_tagged.txt OLD             3.082395
## 3 TOR_C_0046_tagged.txt OLD             2.949062
## 4 TOR_C_00033_tagged.txt OLD             2.944444
## 5 TOR_C_00013_tagged.txt OLD             2.815534
## 8 TOR_C_0050_tagged.txt OLD             2.367628
```

Mann-Whitney test used to compare the distribution accross OLD and YOUNG speakers.

```
wilcox.test(Normalized.FREQ.of..SI. ~ Age, data = si_age)
```

```
##
##  Wilcoxon rank sum test
##
## data:  Normalized.FREQ.of..SI. by Age
## W = 74, p-value = 0.7236
## alternative hypothesis: true location shift is not equal to 0
```

Particle 'si' and gender:

```
head(si_gender)
```

```
##           Informant Gender Normalized.FREQ.of..SI.
## 1 TOR_C_0001_tagged.txt FEMALE              4.819277
```

```
## 2 TOR_C_00019_tagged.txt FEMALE 3.082395
## 3 TOR_C_0046_tagged.txt FEMALE 2.949062
## 4 TOR_C_00033_tagged.txt FEMALE 2.944444
## 5 TOR_C_00013_tagged.txt FEMALE 2.815534
## 6 TOR_C_0038_tagged.txt FEMALE 2.647059
```

Mann–Whitney test used to compare the distribution accross MALE and FEMALE speakers.

```
wilcox.test(Normalized.FREQ.of..SI. ~ Gender, data = si_gender)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: Normalized.FREQ.of..SI. by Gender
## W = 475, p-value = 0.001399
## alternative hypothesis: true location shift is not equal to 0
```

### Auxiliary omission in the perfect tense:

(see files 4\_aux\_age.csv and 4\_aux\_gender.csv)

Auxiliary omission in the perfect tense and age:

```
head(aux_age)
```

##	ID	LOCATION	LONGITUDE	LATITUDE	total	total_aux	no_aux	Year	AGE
## 1	TOR_C_0001	Oöljane	43.66194	22.31988	95	52	43	1925	OLD
## 3	TOR_C_0003	Balinac	43.56462	22.35576	184	63	121	1952	OLD
## 4	TOR_C_0004	?uötica	43.35698	22.47159	89	43	46	1955	OLD
## 5	TOR_C_0005	Gornje Zuni?e	43.60401	22.27268	155	81	73	1934	OLD
## 6	TOR_C_0006	Novo Korito	43.63191	22.37807	19	10	9	2005	OLD
## 7	TOR_C_0007	Trnovac	43.67783	22.23714	123	57	65	1941	OLD

Mann–Whitney test used to compare the distribution accross OLD and YOUNF speakers.

```
wilcox.test(no_aux ~ AGE, data = aux_age)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: no_aux by AGE
## W = 222, p-value = 0.0332
## alternative hypothesis: true location shift is not equal to 0
```

Auxiliary omission in the perfect tense and gender:

```
head(aux_gender)
```

##	ID	LOCATION	LONGITUDE	LATITUDE	total	total_aux	no_aux	GEN
## 1	TOR_C_0001	Oöljane	43.66194	22.31988	95	52	43	FEMALE
## 2	TOR_C_0002	Drvnik	43.53809	22.37374	204	78	124	FEMALE
## 3	TOR_C_0003	Balinac	43.56462	22.35576	184	63	121	FEMALE
## 4	TOR_C_0004	?uötica	43.35698	22.47159	89	43	46	FEMALE
## 5	TOR_C_0005	Gornje Zuni?e	43.60401	22.27268	155	81	73	FEMALE
## 6	TOR_C_0006	Novo Korito	43.63191	22.37807	19	10	9	FEMALE

Mann–Whitney test used to compare the distribution accross MALE and FEMALE speakers.

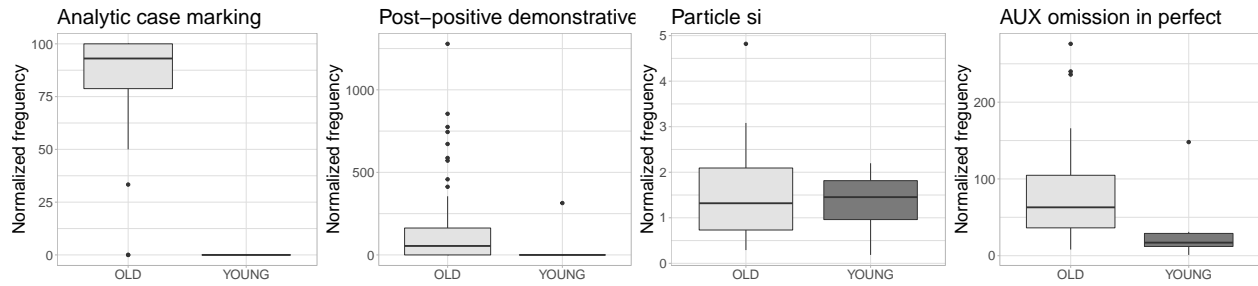
```
wilcox.test(no_aux ~ GEN, data = aux_gender)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: no_aux by GEN
## W = 541.5, p-value = 0.02942
## alternative hypothesis: true location shift is not equal to 0
```

The ranges of values of the linguistic frequencies categorized according to age are shown in Figure 11.

Figure 11: Age

```
Figure11 = grid.arrange(marking_age_plot, ppd_age_plot, si_age_plot, aux_age_plot, nrow = 1)
```



The ranges of values of the linguistic frequencies categorized according to gender are shown in Figure 12.

Figure 12: Gender

```
Figure12 = grid.arrange(marking_gender_plot, ppd_gender_plot, si_gender_plot, aux_gender_plot, nrow = 1)
```

