**Overabundance, inter-and intra-speaker variation and suppletion**

# Abstract

Previous studies on morphological typology have primarily focused on language-internal processes as the origin of overabundance (e.g. heteroclisis or suppletion) (Thornton 2011; 2019). In this paper, we highlight the relevance of language-external factors for explaining the origin and preservation of overabundance, based on fieldwork data and inferential statistics.

Overabundance is prominent in West Polesian/Podlasian (East Slavic) at intra- and inter-speaker level, even within the same household. The use of West Polesian is stigmatised, and speakers belong to the same socio-economic class, which according to Dorian (2010), creates an ecosystem that favours the conservation of overabundance. Concomitantly, they are exposed to four standardised and closely related Slavic languages (extended diglossia) (cf. Fishman 1967). Hence, unsurprisingly, many of the overabundant forms in the paradigms have cognates in the more prestigious neighbouring languages. In this paper, we tackle the following question: is this extensive variation (and thus, overabundance) related to any sociolinguistic variables or is it mostly the result of language contact and the lack of a standard?

We use Chi-Squared Test to determine whether there is any association between sociolinguistic variables (age, gender and village of origin; as indicators of exposure to more prestigious related varieties) and measures of overabundance in the language (i.e. how many different forms speakers use for a single cell). We focus on the suppletive nouns for ‘person’ and ‘year’, which in many Slavic languages display two stems (e.g. [Polish] *jeden* ***rok***; [Russian] *odin* ***god*** ‘one year’; vs. [Polish] *pięć* ***lat***; [Russian] *pjat’* ***ljet*** ‘five years’). In West Polesian such nouns have up to three different stems in some idiolects and are also heterogeneous concerning suffixation and stress (e.g. for the gen. pl of ‘year’ either *ɾɪk, ˈɾok-ɪʋ* or *ɾoˈk-ɪʋ*). Data are drawn from a corpus of free texts by c. 60 speakers, collected through fieldwork in Belarus and Poland and containing 960 tokens of the nouns. Some paradigm cells only show evidence of one or two possible combinations (e.g. acc.sg, ins.sg); in part due to their lower frequency in the corpus. Conversely, we have recorded up to eight different forms for other cells e.g. ‘year’[loc.sg] *ˈɦodu ~ ɦoˈdu ~ ɦod ~ ˈɦoʣʲe ~ ɾoˈkovɪ ~ ˈɾoʦɪ ~ 'ɾoku* ~ ˈvjekovɪ. Note that we have excluded dialectal differences in phonology; e.g. *ʧoloˈvɪk ~ ʧɪlaˈʔek ~ ʧuloˈvjek* ‘person’ [nom.sg].

The results of our analysis suggest that this may be a canonical instance of sociolinguistically-neutral variation, combined with extended diglossia, resulting in high overabundance. Concerning the noun ‘person’, there is a statistically significant association between age and the suppletive stems available (i.e. the older the person, the more stems they use). However, this applies neither to other forms of variation (stress position or suffix type) nor to the noun ‘year’; i.e. speakers of all ages, genders and origins have shown high levels of overabundance in free speech. This leaves the sociolinguistic diglossic landscape as the most plausible relevant factor triggering and preserving overabundance, which calls for an expansion of current theories on morphological typology.

**Keywords: *intra-speaker variation*; *multifactor statistical analysis; overabundance; suppletion; West Polesian***

# Introduction

## Variation / overabundance hypotheses & research question

In this paper, we want to show the relevance of language-external factors for explaining the origin and preservation of overabundance, based on fieldwork data and inferential statistics. Previous studies on morphological typology have primarily focused on language-internal processes as the origin of overabundance and defectivity (e.g. Baerman 2011; Guzmán Naranjo & Bonami 2021; Sims 2015; Thornton 2019).

However, Dorian (2010) showed that certain sociolinguistic settings could create a climate that favours the preservation of sociolinguistically neutral inter- and intra- speaker variation (resulting in inflectional overabundance). Among the most defining factors would be being a small-speech community, poor (and thus, everyone belongs to the same socio-economic class) and a lack of awareness of a standard form of their variety. Moreover, it is well known that multilingual settings can affect grammar, but only recently have researchers started to pay attention to overabundance in such contexts (e.g. Meakins & Wilmoth 2020), since most previous studies on overabundance and defectivity have been heavily biased towards W.E.I.R.D. (Western, educated, industrialised, rich, democratic) settings.

We try to replicate part of those studies with West Polesian/Podlasian (WP) a minoritised East Slavic variety spoken around three international borders, with the addendum of a multi-side and more pronounced language contact with other closely related and more prestigious Slavic varieties. We try to answer the following question: are there any sociolinguistic variables (including gender, age or origin) which predict speakers’ behaviour with inflection (i.e. using more than one form for a given cell)? Do any of these variables serve to diagnose the speakers’ level of exposure to other standardised varieties?

For this endeavour, we will focus on the use of the nouns ‘person’ and ‘year’ in West Polesian/Podlasian (WP), which are particularly open to overabundance, and which also intersect with very particular instances of suppletion.

## 1.2. Sociolinguistic setting

In order to understand the motivation of this study, it is necessary to give a short overview of the sociolinguistic landscape of Western Polesie. Western Polesie is a region that covers southwestern Belarus (Brest), northwestern Ukraine (Volynja) and the easternmost region of Podlasie (Poland). The variety spoken in Poland is often referred as Podlasian (aka. Podlachian) and is considered to be a closely related but different branch of West Polesian (WP) by some, although we will treat them as one.

Western Polesie is located in a very marshy area. The frequent floods, particularly with the melting of the snow in spring, made the region gain the name *Herodotus’ Sea* for centuries (Klimčuk 1992). Travelling between villages as well as outside of the area was very difficult. Until the draining of swamps and construction of roads from the eighties on, often this was only possible by boat. Notwithstanding its long history of hermeticism, this community has been exposed to several closely related standardised varieties for the past decades. Moreover, there have been multiple rulers in a short period of time, which has been reflected in phenomena like the language for schooling. For example, I have interviewed speakers within the same village (or a few villages around) who had been schooled in Polish, Ukrainian, Russian or Belarusian depending on the year they were born. Therefore, there are two processes that can be observed. On the one hand, the isolation, which has resulted into a distinct culture and linguistic variety; with a different evolution from other East Slavic varieties. On the other hand, being exposed to a very pronounced language contact with several closely related varieties. Thus, the language change and input are not unilineal (as for example with the Belarusian-Russian situation). Yet, speaking West Polesian is highly stigmatised and there is no real standard for it. Most people who use it regularly live in rural areas and have blue-collar jobs (most often farmers). When interacting with outsiders they try to switch to a more prestigious and standardised variety, to the point of even denying any knowledge of it when surrounded by socially higher circles. All this results in a very unique setting, which we hypothesise, opens the doors for overabundance to flourish.

## 1.3. Numeral Phrases & suppletion

There are several conditions determining the distribution of the suppletive stems for ‘year’ and ‘person’ in Slavic.Nevertheless, for the vast majority of Slavic languages the most determining factor is being headed by a numeral (i.e. being in a Numeral Phrase) and the class it belongs to.[[1]](#footnote-1) For those who are not very familiar with Slavic languages, we shall give a short overview of the functioning of Numeral Phrases (NumPs)in Slavic and , in West Polesian (WP).[[2]](#footnote-2) The syntax of Numeral Phrases (NumPs) was already complicated in Common Slavic, and with the erosion of the dual number in most contemporary Slavic varieties, it has become more complex. Very briefly, NPs headed by the cardinal numeral ‘one’ would take nom.sg; those followed by lower numerals (‘two’ to ‘four’ and derived) nom pl or (what looks like) gen sg (or adnumerative, in the case of WP) and higher numerals (‘five’ to ‘twenty’) gen pl. [Should I use sentences/ examples instead?] See (Figure 1):

|  |  |  |  |
| --- | --- | --- | --- |
|  | ‘one’ | ‘two’, ‘three’, ‘four’ | ‘five’ to ‘twenty’ |
| Noun form | sg (any **case**) | gen sg/ nom pl/ adnumerative | gen pl/ greater adnumerative |
| Verbal agreement | sg | **pl** | **pl** |

Figure Most common nominal and verbal agreement with numerals in Slavic [Roncero, Forthcoming]

Now, depending of the heading numeral the nouns ‘year’ and ‘person’ display suppletive stems in many Slavic languages; e.g. [Polish] *jeden* ***rok***; [Russian] *odin* ***god*** ‘one year’; vs. [Polish] *pięć* ***lat*** (\**rok-ów*); [Russian] *pjat’* ***ljet*** (\**god-ov*) ‘five years’. What we observed with West Polesian is that speakers were combining most, if not *all* (for ‘year’) Slavic stems for these nouns and that their distribution was very heterogeneous from speaker to speaker, even within the same household ((Roncero 2019; 2022). Note that the examples (1)-(3) were produced by the same speaker within the span of a few minutes, within the same context (same room, same hearers, and same style).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| (1) | (TL6.5 00:45) | | | |  |  |  |
|  | uʒe | **Sim** | **roˈkɪʋ** | V | Jomu | bu-l-o |  |
|  | already | seven | year.gen.pl | In | 3sg.dat.m | be-pst-n.sg |  |
|  | ‘He was already seven years old.’ | | | | | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| (2) | (TL6.5 02:40) | | |  |  |  |  |
|  | V | **sorok** | **Lit** | vɪn | i | ʋmer |  |
|  | in | forty | year.gen.pl/gradnm | 3sg.nom.m | and | die.pst.m.sg |  |
|  | ‘He died when he was forty years old.’ | | | | | | |

|  |
| --- |
| [But some minutes later] |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| (3) | (TL6.7 02:00) | | |  |  |  |  |
|  | oj, | i | **mnɪɦa** | ʋʒe | mɪnjɪ | **ɦaˈdoʋ** | uʒe, |
|  | oh | and | Many | Already | 1sg.dat | year.gen.pl | already |
|  | **dɪvɪnosta** | | **Dva** | **ˈrokɪ**! |  |  | |
|  | ninety | | two.nom.m | year.adnm | |  | |
|  | ‘Oh, I'm already very old (lit. many of years old); ninety-two years.’ | | | | | | |

A closer look at the data shows that speakers within a village or a household, not only differ on the suppletive stems they may use, but even when they use the same stems, there are other differences on the type of suffix and stress they use for a cell between speakers. Therefore, the overall picture looks more complex than initially thought, but it also makes it more interesting for the study of overabundance.

# Methodology

Unfortunately, the full-scale invasion of Ukraine began a day before the official start of this project. Therefore, doing fieldwork in Ukraine and Belarusian Polesie has been impossible ever since. Furthermore, the region of Podlasie in Poland has had limited access for most of the time, given the tensions because of the migration crisis with Belarus, and most recently, the deployment of NATO tanks. For this study, first, we have gathered data from free texts (i.e. non-directly elicited by tasks or translations) from speakers in the region of Brest (Belarus) (2016-2017) and Podlasie (Poland) (August 2022). More than 60 speakers participated in the study, though some of their contributions were too short and/or did not include any of the nouns under study here. Therefore, we have had to trim the number of speakers included. We gathered informed consent from every speaker and each one was assigned a specific code to protect their identity, which we have used for the rest of our study. We have gathered a total of 960 tokens: 560 for ‘year’ and 400 for ‘person’ from the free texts.

Second, after extracting all the tokens of ‘year’ and ‘person’ from the recordings and transcribing them, we looked for cognates in other Slavic languages to see if the alternating forms were genuine suppletive stems or just synonyms ((Roncero 2022), FORTHCOMING). We coded all the utterances and specific tokens at multiple levels using NVivo. We have considered the different syntactic contexts (e.g. the type of numeral they were bound by), stem, stress-type, suffixation, speaker characteristics for each token.

Third, we generated queries in NVivo which gave us an overview of all the possible combinations for each case/number cell. We condensed all of them manually merging into the same form variants which are only distinguished by well-established dialectal differences in the phonology (*see attachment).*

Fourth,

Fifth, once in RStudio, we first picked the short datasets for each **case/number** combination and filtered out all the speakers uttering a cell less than once using the *%>% filter* function of the "*dplyr*" package and assisted by *Chat GPT3* (e.g. Figure 2). In this fashion, we wanted to avoid adding unnecessary noise to data, since, if a speaker only uttered once a specific form there is no room for variation. This reduced the number of forms we could realistically study, as some cells are so infrequent that there is only a handful of tokens for them. We also used the “perc” and “cat” functions of the *dplyr* package (also, assisted by *Chat GPT3*) to extract the percentages of the people using more than one form (having used that cell more than once). For the noun ‘year’ we focused on the gen sg, gen pl and loc sg, which also correspond to the cells with the higher number of forms (6, 6 and 8, respectively). As for ‘person’, we chose the nom pl, acc pl and gen pl, which also correspond to the cells with the higher number of forms (4, 4 and 8, respectively).

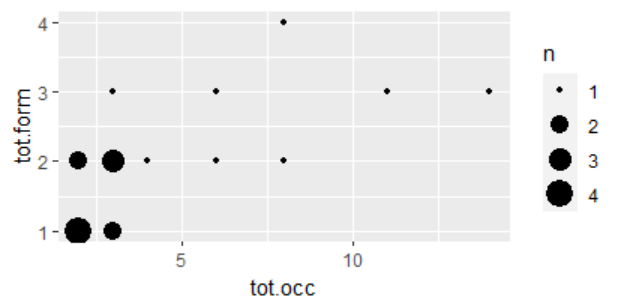


Figure Dot-plot of the gen pl of ‘person’ showing the proportion of gen pl tokens produced by each speaker (=x), and the number of different forms produced (=y).

Sixth, we then moved to a multilevel exploratory analysis for the chosen cells. We used the *ggplot2* and *esquisse* packages to generate plots, covering the frequency of the forms across villages, length of the corpus, age and gender (initially separated, but later combined) (e.g. Figure 3); as well as the frequency of each individual form (e.g. Figure 4).

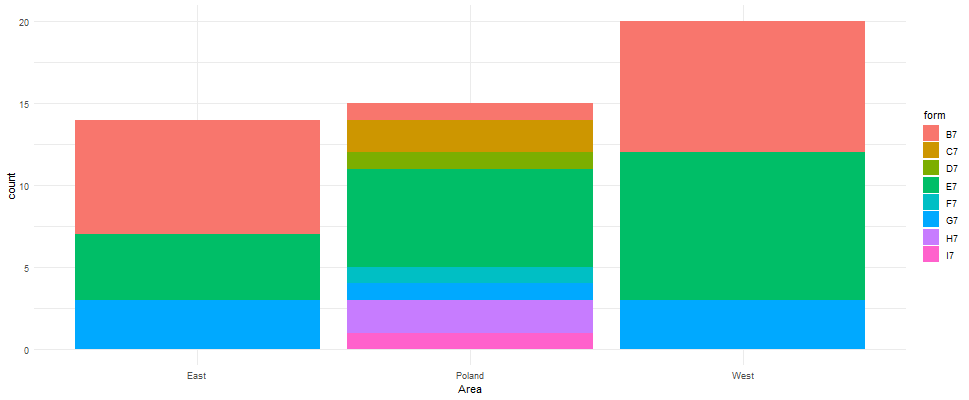


Figure Distribution of gen pl forms of ‘person’ across area of origin

A picture containing screenshot, colorfulness, square, line

Description automatically generated

Figure Frequency of gen pl forms of ‘person’ in the corpus

The initial observations suggested only limited correlations. We had initially thought of doing poisson and multilevel logistic regressions, but they were providing too many zeroes. As a result, we moved to other types of tests described below.

## Distribution of dependent variables

There is a remarkably smaller number of male speakers, especially, from Belarusian Polesie given several demographic factors. Life expectancy is lower for men than women, in part because many of the peers of older women died during WWII. Many men of a working age were often doing seasonal jobs in the cities (most frequently in the construction sector), so they were far less available for the interviews. Furthermore, in general, since most men in Belarus had served in the military, where they did their service in Russian, they tended to switch to Russian when approached, which made them less desirable participants. Finally, women over 75 were often at home and had more free time, besides being the first choice of key community members, when asked about participants who could contribute to the project.

In order to address this issue, we have had to combine age and gender into three categories: men, women under 75 and women over 75 by using variable interaction tests .

Concerning village of origin, we have had participants from over 20 different villages. Taking into account geographic proximity and traditional dialect divisions (Klimčuk 1983), we initially posited six values for this variable. To reach frequency levels enabling statistical analysis, we clustered them into three larger areas: Eastern Brest (Belarus); Western Brest (Belarus) and Podlasie (Poland). In spite of traditional divisions posited for West Polesian dialects, we have observed a high level of similarity in these parameters between varieties within these three created areas.

We have also considered the length of the interventions (i.e. how long the speakers spoke for in the corpus of free texts) for each **case/number** combination. The observations for this variable were not surprising: the longer a participant had spoken for, the higher the number of specific **case/number** forms used and, for most cases, the higher the number of forms for a cell. This has helped to see why some cells are extremely rare (e.g. dat pl), as they only appear in the corpora of those who have spoken for the longest. In any case, we have excluded this dependent variable from our final analysis. We were also going to include religion; however, we did not have enough data for all participants and in most cases, there is a strong correlation between village of origin/residence and religion. Therefore, this possible covariant was excluded from the final analysis.

## Data availability

The corpus of transcribed recordings with all the tokens of ‘year’ and ‘person’, as well as the NVivo codes are available at\*\*\*[Zotero/GitHub]\*\*, following the FAIR data share principles. Note that because of ethical and data protection reasons the audio recordings are not publicly available. Also note that the comments in the corpora which included sensitive data (e.g. gossip, naming people) have been censored, following the common ethical guidelines [FORTHCOMING].

All the R-coding, spreadsheets and plots are also available at [\*\*\*GitHub?\*\*\*]

# 3. Analyses

For both nouns, one of the cells with the highest amount of forms has been the gen pl (8 for ‘person’, 6 for ‘year’), As for ‘year’, also the loc sg (9), and the gen sg (6) provided the highest amount of forms, whereas there is not a single token of loc sg in the corpus of ‘person’, but the nom pl (4) and the gen sg (4) showed to be the most prolific for variation. Some cells, such as the ins sg of ‘year’ did not provide a single token in the corpus, thus, we have excluded them from the table.

We proceeded to test multilevel-logistic regressions on this data, considering all the variables. Most of the explanatory variables are between speaker. i.e. an individual can only be male or female during the conversation and cannot swap genders half way through, they are speaking in a specific area etc. Thus, most of the information is on between person variation. Thus we wanted elaborated tables documenting the presence or absence each form for each person to then perform multi-level logistic regressions. These would allow us to find out if variation in form is related to gender, origin, number of times used, etc. However, such tests were producing many unwanted zeroes given the size limitations of our dataset. Therefore, we decided to move towards creating many simple logistic regressions (one for each cell and form) and explore data further with a k-means cluster analysis.

On (Table 1) we can see a description of the most prolific cells for variation: both because we had enough data of individual speakers using these cells more than once *and* also enough of that subset of speakers using more than one form. Beware that the distribution of the cells for any noun will never be homogeneous in any real life corpora. Some **cases/number** combinations are overall very rare (e.g. dat pl), whilst other have semantically motivated morphosyntactic restrictions; e.g. loc is not available for human animates (thus, ‘person’ is defective in our case). Conversely, the dative (especially of masculine and neuter nouns and pronouns) seems to be being displaced by the genitive with the preposition *do*.[[3]](#footnote-3)

The columns in (Table 1) are:

1-Cell

2-Speakers\_using\_cell: Represents the number of speakers where the total occurrences (tot.occ) is not zero. This essentially counts how many speakers have used the specific cell (i.e. case/number) at least once in their corpus.

3-Max\_usages: The maximum number of times a cell has been used by any individual speaker in the corpus. It takes the highest value from the tot.occ column.

4-Speakers\_using\_cell\_more\_than\_once: The number of speakers who have used the form more than once. It counts the rows where tot.occ is greater than 1.

5-Num\_forms\_produced: The number of different forms (or variations) produced by the speakers; number of different forms present in the dataset.

6-Speakers\_using\_multiple\_forms: Represents the number of speakers who have used multiple forms. It counts the number of rows where tot.form is greater than 1. The tot.form column indicates how many different forms each speaker has used.

7-Max\_forms\_by\_individual: The maximum number of different forms used by any individual speaker in the dataset. It fetches the highest value from the tot.form column.

**Table 1** Results of the most overabundant cells

| Cell | Speakers using cell in corpus | Maximum nº of uses of cell by an individual speaker | Nº of speakers using cell more than once | Nº of forms produced by speakers | Nº of speakers using multiple forms | Maximum nº of forms used by an individual speaker |
| --- | --- | --- | --- | --- | --- | --- |
| Person | | | | | | |
| NOM SG | 23 | 17 | 16 | 3 | 1 | 2 |
| NOM PL | 28 | 22 | 20 | 4 | 1 | 2 |
| ACC PL | 17 | 6 | 3 | 4 | 2 | 1 |
| GEN PL | 30 | 14 | 19 | 8 | 13 | 4 |
| ADNM | 5 | 2 | 4 | 3 | 1 | 2 |
| DAT SG | 7 | 4 | 3 | 2 | 1 | 2 |
| DAT PL | 12 | 2 | 5 | 3 | 1 | 2 |
| Year | | | | | | |
| NOM SG | 25 | 6 | 12 | 2 | 1 | 2 |
| NOM PL | 10 | 3 | 4 | 3 | 0 | 1 |
| ACC SG | 12 | 3 | 2 | 4 | 0 | 1 |
| ACC PL | 3 | 5 | 1 | 2 | 0 | 1 |
| GEN SG | 41 | 13 | 26 | 6 | 7 | 2 |
| GEN PL | 52 | 25 | 42 | 6 | 19 | 5 |
| ADNM | 14 | 3 | 6 | 3 | 0 | 1 |
| INS PL | 2 | 1 | 0 | 3 | 0 | 1 |
| LOC PL | 11 | 3 | 2 | 4 | 0 | 1 |
| LOC SG | 30 | 13 | 13 | 8 | 5 | 3 |

Based on the data for the most prolific cells for intra-speaker variation (overabundance) we created individual summary tables to explore inter-speaker variation with RStudio (*see attachment*). There is a table for each form of each cell, showing the total of times each individual form has been used (=n) for each individual variable; the proportion of the total occurrences of this specific form (face to others) on the corpus of this cluster (variable); and the proportion of speakers of the group using it. See for example (Table 2).

Table 2 Distribution of the *ɦod* form (B7) of the GEN PL of ‘year’

**Variable Total(=n) Occurrences % Total %\_ of users in cluster**

1 Women < 75 31 49.21 33.33

2 Women 75+ 21 33.33 33.33

3 Men 11 17.46 33.33

4 East 51 80.95 76.19

5 West 10 15.87 14.29

6 Poland 2 3.17 9.52

7 Length 1 4 6.35 9.52

8 Length 2 37 58.73 52.38

9 Length 3 5 7.94 23.81

10 Length 4 17 26.98 14.29

We then calculated multiple cluster analyses using the k-means cluster method (see attached data). Cluster analysis, or clustering, is a statistical technique that sorts different objects or points into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. We were looking for inherent structures in data and grouping them in the best possible way. In most datasets the elbow method (optimal number of clusters) was 4. The reason why we get this number may be also the common source of the datasets and the way they were generated or just simply not diverse enough patterns. The axes created by those clusters are the 2-D simplification of multiple dimensions of the data, which would be otherwise impossible to see for the human eye. Effectively, what we were trying to figure out is whether there is any characteristic of the population that is well defined and more or less prone to overabundance. We have observed some tendencies: the two more recurrent distinctive clusters seemed to be men from Poland and women over 75 from Eastern Belarus. For example, in (Figure 5).

**Cluster avg\_tot\_form count predominant\_gender\_age predominant\_area avg\_length**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| <int> | <dbl> | <int> | <chr> | <chr> | <dbl> |
| #1 | 0 | 14 | male | Poland | NA |
| #2 | 0 | 10 | female 75+ | East | NA |
| #3 | 1 | 5 | male | West | NA |
| #4 | 0 | 34 | female <75 | East | NA |

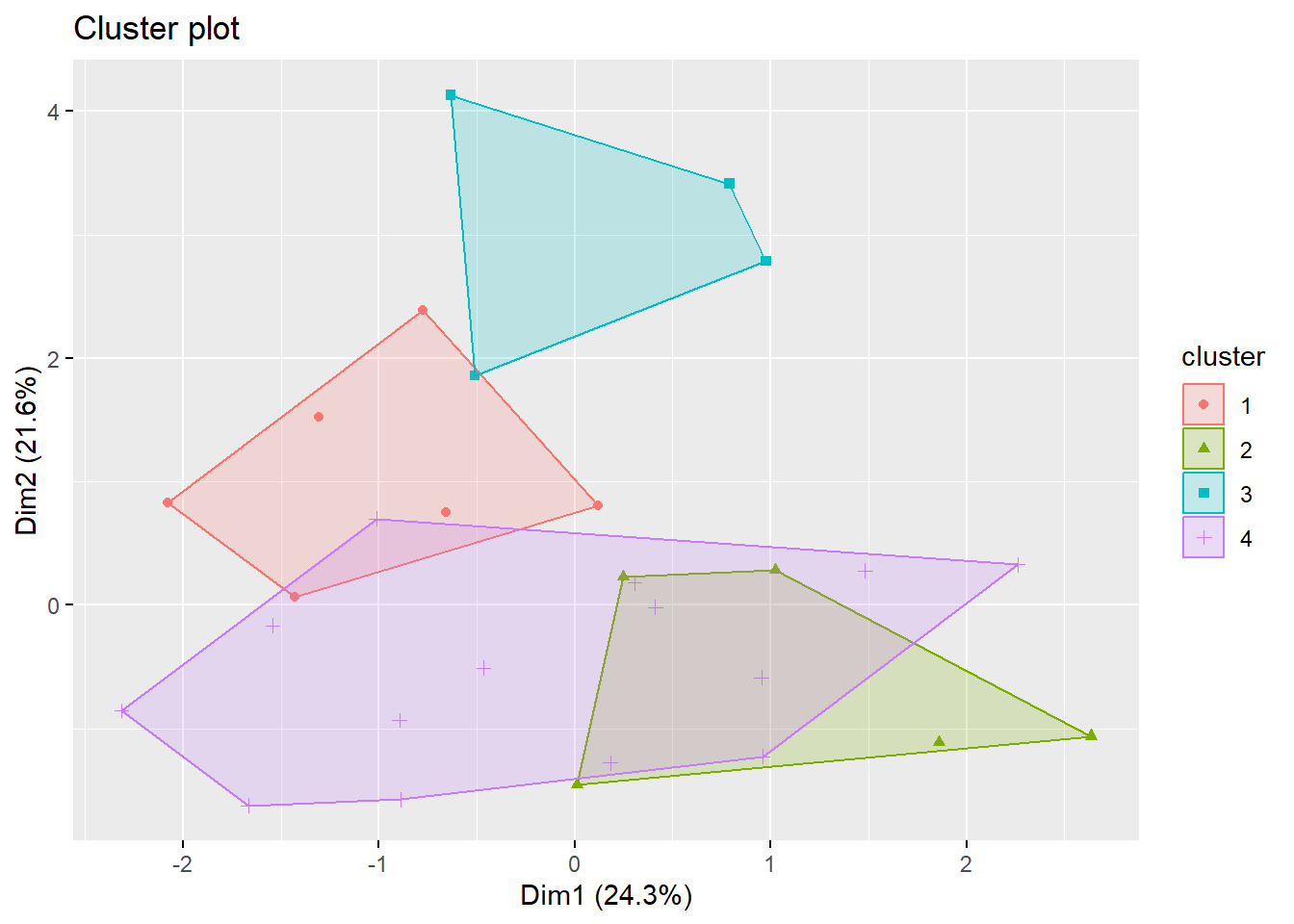


Figure Overabundance patterns for loc sg of the noun ‘year’

In the plot on (Figure 5), each dot symbolises a data point and the different colours and shapes represent the clusters. The spatial distribution on the plot provides a sense of distance or dissimilarity. Points that are closer together on the plot are more alike concerning the features under study; whilst, distance between points means dissimilarity. The most distinctive cluster in (Figure 5) is #3 representing men from Podlasie, followed by #1, which also involves men, though from Western Belarus. The majority of predictors in the logistic regression binomial model are not statistically significant, with the exception of **gender/age:** male. This suggests that being male might have some influence on the outcome, but the evidence is not very strong.

Nevertheless, after an overall look at all the plots for all the cells under study there are many overlaps between clusters as in (Figure 6). Cluster 1 stands out as the only group where every individual produces more than one form while speaking. Specifically, males from the East consistently show this tendency, whereas the remaining clusters (2-4) do not show no such tendency. No individual in these clusters produces more than one form. In any case, this method has only exploratory purposes and cannot be used for validation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | <int> | <dbl> | <Data > | <Gender/age> | <Origin> | <dbl> |
| #1 | #1 | 1 | 7 | Male | East | NA |
| #2 | #2 | 0 | 9 | Female 75+ | East | NA |
| #3 | #3 | 0 | 15 | Male | East | NA |
| #4 | #4 | 0 | 32 | Female 75+ | West | NA |

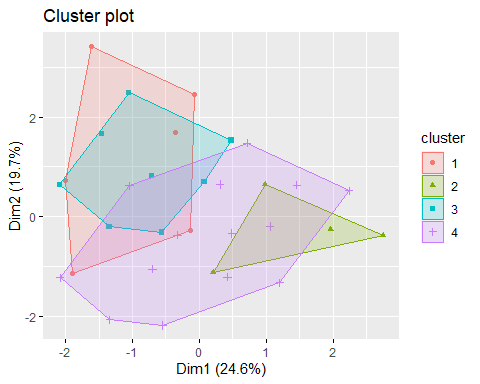


Figure Overabundance patterns for the gen.sg of ‘year’

For our last step, we have elaborated two χ2 tables (one for each noun) to test for our hypotheses with the most recurrent clusters; i.e. whether there is or not a correlation between the amount of different forms speaker produce and specific sociological characteristics (variables). As it is common practice in Linguistics, we have set our *p-value* on 0,05. We have merged all the tables for each cell into two big groups (one for each noun) and created a column with binaries; i.e. every time a speaker had used more than one form (for any cell) they got a 1.

One of the conditions or assumptions for Chi-squared tests is to have a minimum value of 5 in the contingency table. However, given the scarcity of data for some cells, some of did not meet this condition and, thus, we employed the Monte Carlo method. This method helps to get a more accurate p-value by randomly sampling from the observed data to simulate what the distribution of the test statistic would look like and improve the condition (Halton 1970; Johansen & Evers 2007).

For example, we want to ask whether there is an association between the length of the interview and producing more forms:

(H0 ): There is no association between the length of the interview and being a multiform speaker for the "year" noun.

(Ha): There is an association between the length of the interview and being a multiform speaker for the "year" noun.

**CONFUSION MATRIX**

**1  2  3  4**

**0  9 18  8  2**

**1  1  6 11  7**

0= No overabundant found in speakers

1= OA present

For example, 2 speakers that were amongst those who spoke for the longest did not show traces of OA, whereas 7 of them did. In this particular example, this assumption (of being greater than 5) was not met for every combination, therefore, we employed a Monte Carlo simulation:

**Pearson's Chi-squared test with simulated p-value (based on 2000**

**replicates)**

**data:  table(year\_data$multiform, year\_data$Length)**

**X-squared = 13.848, df = NA, p-value = 0.003498**

The p-value of 0.003498 is less than the conventional significance level of 0.05. Therefore, you would reject the null hypothesis, indicating that there is a statistically significant association between being a multiform speaker for the "year" noun (multiform status) and the Length of the interview.

**Year χ2**

The Pearson’s Chi-squared data for ‘year’ show that the only relevant factor for individuals using more or less forms would be the length of their input in the corpus; i.e. the longer they spoke for, the higher the chances for them to use more than one form in a given cell, which is not surprising at all. The p-values of the results are the following: p=0.3603 for area of origin; p= 0.001999 for length of recording and p= 0.7156 for gender/age.

**Person χ2**

In the same vein as the noun ‘year’, the chi-squared data for ‘person’ show that the only statistically significant variable for individuals being prone to overabundance is the length of their input. The p-values of the results are the following: p=0.3108 for area of origin; p= 0.03598for length of recording and p= 0.8741 for gender/age.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **‘Year’ χ2** | |  | **‘Person’ χ2** | |
| **Variable** | **p=** |  | **Variable** | **p=** |
| area | 0.3603 |  | area | 0.3108 |
| length | 0.001999 |  | length | 0.03598 |
| gender/age | 0.7156 |  | gender/age | 0.8741 |

# 4. Summary and conclusions

We have selected 960 tokens of the nouns ‘year’ and ‘person’ in West Polesian/Podlasian from a corpus of free texts. These nouns are particularly open to variation, as they display several suppletive stems and also multiple stress and suffix possible combinations. In this study we ask which sociolinguistic variables could help us predict whether someone will use more than one form for a given cell (overabundance). Given the unequal distribution of speakers we have merged age and gender into a single category, with three possible outcomes. We have applied logistic regressions, which gave us many results lower than the threshold that we needed, due to data scarcity, especially for some cells. Therefore, we have ended up applying Pearson’s Chi Squared tests. None of the studies are sufficient to answer the reasons for variation. There are some associations/implications and effects, but none of them provides a fully satisfactory answer (co-causation?) to reject the null hypothesis. The only control variable, that provided a p-value that rejected the null hypothesis was speakers’ length of input; i.e. the longer a speaker spoke for the corpus, the higher the chances to use more than one form for a specific case/number cell.

Now we would like to discuss some of the limitations of our study. We are aware that obtaining more data may reveal an association between the origin, gender or age of the speakers and their inclination to overabundance (increased power). We are also concerned about our population sample bias and how to mitigate it. A better study should include more men and women under 75 (admitting, there are no many men over 75 alive, especially, in Belarus); which could also smooth out random fluctuations that mask underlying patterns. And obtaining more data in general (even from the same speakers) would also reduce the amount of zeroes for some of the less common case/number values, which we have not been able to explore further. Yet, we also acknowledge the above mentioned difficulties to carry out fieldwork in the area, mostly exacerbated by the full-invasion of Ukraine.

Finally, for we would like to propose a question for further exploration of the data. So far we have looked at overall presence or absence of OA in speakers idiolects. However, for a future, we would like to ask whether there is more or less OA depending on the origin, gender and age of the speakers or not (considering all case/number values). We understand that for some case/number combinations there were more than two possible forms (e.g. gen pl) whilst for others the choice was only between two.

# 5. References

Baerman, Matthew. 2011. Defectiveness and homophony avoidance. *Journal of Linguistics* 7. 1–29.

Bortnik, Ninel´. 1979. Nekotorye osobennosti osnov obrazujuščix dvojnye formy množestvennogo čisla. *Russian Language Journal* 32. 43–58.

Chumakina, Marina & Hippisley, Andrew & Corbett, Greville G. 2004. Istoričeskie izmenenija v russkoj leksike: slučaj čeredujuščegosja suppletivizma. *Russian Linguistics* 28. 281–315.

Dorian, Nancy C. 2010. *Investigating Variation. The Effects of Social Organization and Social Setting*. New York/Oxford: Oxford University Press.

Fishman, Joshua. 1967. Bilingualism with and without diglossia. Diglossia with and without bilingualism. *Journal of Social Issues* 23(2). 29–38.

Guzmán Naranjo, Matías & Bonami, Olivier. 2021. Overabundance and inflectional classification: Quantitative evidence from Czech. *Glossa: a journal of general linguistics* 6(1). (doi:10.5334/gjgl.1626) (https://www.glossa-journal.org/article/id/5468/) (Accessed March 1, 2023.)

Klimčuk, Fëdar D. 1983. *Havorki Zaxodnjaga Palessja: Fanetyčny narys 4-ox havorak*. Minsk: Navuka i tèxnika.

Krasovitsky, Alexander. 2019. Case Loss in Pronominal Systems: Evidence from Bulgarian. In Baerman, Matthew & Bond, Oliver & Hippisley, Andrew (eds.), *Morphological perspectives: papers in honour of Greville G. Corbett*, 193–204. Edinburgh: Edinburgh University Press.

Meakins, Felicity & Wilmoth, Sasha. 2020. Overabundance resulting from language contact: Complex cell-mates in Gurindji Kriol. In Arkadiev, Peter & Gardani, Francesco (eds.), *The Complexities of Morphology*, 81–104. Oxford University Press. (https://doi.org/10.1093/oso/9780198861287.003.0004) (Accessed March 20, 2023.)

Roncero, Kristian. 2019. *A typological approach to West Polesian morphology and syntax.* Guildford (Surrey, UK): University of Surrey. (PhD Thesis.) (Application/pdf.) (https://doi.org/10.15126/thesis.00851715)

Roncero, Kristian. 2021. Revisiting numeral phrases in East Slavic: Insights from West Polesian. *Slavonic and East European Review* 99(4). 601–646.

Roncero, Kristian. 2022. A very unpredictable ‘person’: A corpus‐based approach to suppletion in West Polesian. *Russian Journal of Linguistics* 26(1). 116–138. (doi:10.22363/2687‐0088‐26828)

Sims, Andrea D. 2015. Productivity, defectiveness, and syncretism. 82–132. Cambridge: Cambridge University Press. (doi:10.1017/CBO9781107053854.005) (http://ebooks.cambridge.org/chapter.jsf?bid=CBO9781107053854&cid=CBO9781107053854A036)

Thornton, Anna M. 2011. Overabundance (multiple forms realising the same cell): A non-canonical phenomenon in Italian verb morphology. In Maiden, M. & Charles Smith, J. & Goldbach, M. & Hinzelin, M. (eds.), *Morphological autonomy: Perspectives from Romance inflectional morphology*, 357–381. Oxford: Oxford University Press.

Thornton, Anna M. 2019. Overabundance: A canonical typology. In Rainer, F. & Gardani, F. & Luschützky, H. & Dressler, W. U. (eds.), *Competition in Inflection and Word-Formation*, 189–224. Dodrecht: Springer.

1. For more details about the conditions involved in Russian and West Polesian see (Bortnik 1979; Chumakina et al. 2004; Roncero 2022). [↑](#footnote-ref-1)
2. For a more detailed explanation and discussion see (Roncero 2021). [↑](#footnote-ref-2)
3. Note that this phenomenon is not rare in Slavic and has also been documented in South Slavic (Krasovitsky 2019). [↑](#footnote-ref-3)