Polesian Noun Project:

# Part 1: Descriptive Statistics: [R Markdown](https://docs.google.com/document/d/19lXvR3N9SAlMKFLq6VGj8-g1_FAhi9O8/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)

*Decided with Kristian on one table per variation/form*

## Summary Table for each Cell:

The columns in the resulting tables are:

* Variable: Describes the category or group.
* Total: The total occurrences of the variation for each group.
* Percentage\_of\_Total: The proportion of the total occurrences that each group's total represents. (Calculated based on how many times the variation was used)
* Percentage\_of\_Variation\_Users: The proportion of unique users of the variant that belong to each group. (Calculated based on how the presence of the variation, binary)

1. [Person Accusative Plural](https://docs.google.com/spreadsheets/d/14GnO7_qa9v2S4tX1P476eskj5ed9sdKA/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
2. [Person ADMN](https://docs.google.com/spreadsheets/d/17J4p8qzROkBpSH4C1eWX10p2DOWrI8Rf/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
3. [Person Dative Plural](https://docs.google.com/spreadsheets/d/1DIIqCzOHdRzb1MSu97wLm7H0L2XHW3Ps/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
4. [Person Dative Singular](https://docs.google.com/spreadsheets/d/19MeJ1tCMZV7TLCP5NtH5NtJuJrdg9w0m/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
5. [Person Genitive Plural](https://docs.google.com/spreadsheets/d/1an8LpGhMWStviLqn-OzjOT_-kI4GDAs0/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
6. [Person Nominative Plural](https://docs.google.com/spreadsheets/d/19nJurhdo36YvF2YGvCnmwmXI-SFzYCgn/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
7. [Person Nominative Singular](https://docs.google.com/spreadsheets/d/1Nfl52f07Su2-usbQRzsrrQFKyT2p0F5X/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
8. [Year Accusative Plural](https://docs.google.com/spreadsheets/d/14Y_aC0yzddXJYEuic6AI_kemI3PcYn5s/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
9. [Year Accusative Singular](https://docs.google.com/spreadsheets/d/1HYYcBOvlFCPICDVc2NvYAZRxfZgGOiJj/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
10. [Year ADMN](https://docs.google.com/spreadsheets/d/1-OwKrXk0-2PC0BPXle9JRsH1VPkwYe3A/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
11. [Year Genitive Plural](https://docs.google.com/spreadsheets/d/1FNZ11f4CctXWXijmcbauvIsG_3kP2y94/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
12. [Year Genitive Singular](https://docs.google.com/spreadsheets/d/1LWwRefY9r4b3O6a5s8Nsel_X1KMZpTyx/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
13. [Year Ins Plural](https://docs.google.com/spreadsheets/d/1Jd-lOm7wROXIg8l91b7ZvYwrFaCdlbEv/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
14. [Year Locative Plural](https://docs.google.com/spreadsheets/d/1A-dzX2B7UvkrSGyJUrbR4kwFynpm5oHu/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
15. [Year Locative Singular](https://docs.google.com/spreadsheets/d/14zW0-x0CKLIzUPfQOGKPJHjxu0U_Sbxj/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
16. [Year Nominative Plural](https://docs.google.com/spreadsheets/d/1F_RduDaVw52dJkuwP2rruUAmQJDtd9D8/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)
17. [Year Nominative Singular](https://docs.google.com/spreadsheets/d/14VJ8FZwZecEdT4gyZakB_7AxN7kFvDMr/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)

## 

## Response Table:

The columns in the resulting table 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.
8. [Response Table](https://docs.google.com/spreadsheets/d/1ds-YIntktMGGepWsW3VMY4dazhe0QJAv/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)

# Part 2: Analysis

Started by doing Multilevel Logistic Regression on each cell for each form as Jean suggested, but since there are very few instances where a variation has 1s and because we are fixing (gender.age, length, and Area) the results are uninterpretable.

## Research question:

Is there a specific characteristic or group or people that produce multiple forms when speaking?

## Part 1 Visualise using [Cluster analysis](https://drive.google.com/file/d/1dvwLG8qqkiPsLNAR2c2Y25V2baN4AYFY/view?usp=drive_link): [R Markdown](https://docs.google.com/document/d/1aA5rRDAg52yBHN5Eas7s-6DyARn5KZ07/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true) [Explanation](https://docs.google.com/document/d/1K-oLlXhOkW7uytR179HJd7foBdVLhEBWXERosbNyaHY/edit?usp=drive_link)

* Used to discover patterns and structures
* Helps in revealing hidden patterns and structures in a dataset that might not be apparent through other analytical methods

1. Person Accusative Plural:

Cluster 4 seems to be the group of interest since it has the highest proportion of individuals who produce more than one form while speaking. The predominant characteristics of this cluster are “female 75 plus” from the “East” area.

1. Person Genitive Plural:

Cluster 2 stands out as the group with the highest proportion of individuals producing more than one form while speaking. This cluster predominantly comprises older females (75 plus) from the "East" area.

Cluster 1, with males from Poland, and Cluster 3, with males from East, also have notable proportions of individuals producing more than one form.

Cluster 4, despite being the largest cluster, has the lowest proportion of individuals producing more than one form.

1. Person Nominative Singular:

Cluster 1 is the only group where a fraction of individuals produce more than one form while speaking. Specifically, males from Poland seem to be the group with this tendency, albeit at a relatively low percentage of 6.25%. The other clusters (2-4) don't seem to have any individuals with tot.form equal to 1. This indicates that neither older females from the East nor males from the East show a tendency to produce more than one form in this dataset. Given that Cluster 4 is the largest cluster and none of its members produce more than one form, it suggests that the "female 75 plus" group from the "East" area predominantly uses a single form when speaking in this dataset.

1. Year ADMN:

Since none of the individuals used more than one form in this cell, we cant do cluster analysis.

1. Year Genitive Plural:

Cluster 4 stands out as the group with the highest proportion of individuals producing more than one form while speaking. Specifically, older females from the West seem to be the group with this tendency, with half of them producing more than one form. Cluster 2, comprising older females from the East, also shows a relatively high tendency, with approximately 30.4% of them producing more than one form. Both Cluster 1 and Cluster 3 have the same proportion of individuals producing more than one form (20%), but they differ in the predominant area. Given this clustering result, the data suggests that older females, whether from the East or West, have a higher likelihood of producing more than one form while speaking compared to males in the dataset

1. Year Genitive Singular:

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. The other clusters (2-4) show no such tendency. No individual in these clusters produces more than one form

1. Year Locative Singular:

Cluster 3 stands out as the only cluster where individuals, on average, use forms. This cluster is smaller in size and is characterised by males from the West.

1. Year Nominative Singular:

Clusters 1 and 3 are dominated by males, but they differ in their primary regions (Poland vs. East). Clusters 2 and 4 are dominated by females aged 75 and above from the East. Cluster 4 is notably larger than Cluster 2. The characteristic or centroid value seems to be relatively consistent across the clusters, with only Cluster 3 showing a minor difference.

## 

## Part 2 Chi-Squared Test: [R Markdown](https://docs.google.com/document/d/1BgEc5Z35JCdKhL0-VFfBP0td19Nr7uGR/edit?usp=drive_link&ouid=109271868152726343587&rtpof=true&sd=true)

Primary assumptions of Chi-Squared Test:

1. Independence of Observations: Each observation used in the calculation of the Chi-squared statistic should be independent of the others. This means, for example, that each individual in a sample contributes to only one cell in the contingency table.
2. Categorical Data: Both variables under consideration in the Chi-squared test should be categorical.
3. Sample Size: The test requires that the dataset be a random sample and sufficiently large. A commonly used rule of thumb is that the test is applicable when the smallest expected count is 5 or greater for at least 80% of the cells, and no cell has an expected count less than 1.

### Person

#### 1. Association between being a multiform speaker for the "person" noun and Area

Null Hypothesis (H0 ): There is no association between the area of residence and being a multiform speaker for the "person" noun.

Alternative Hypothesis (Ha ): There is an association between the area of residence and being a multiform speaker for the "person" noun.

**East Poland West**

**0 21 10 15**

**1 4 6 6**

The Monte Carlo simulation was employed for the Chi-squared test because some cells in the contingency table had expected counts below the recommended threshold, potentially violating test assumptions; this simulation method estimates the p-value by repeatedly sampling from the observed data to better approximate the true distribution of the test statistic.

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

**replicates)**

**data: table(person\_data$multiform, person\_data$Area)**

**X-squared = 2.4822, df = NA, p-value = 0.3158**

Given that the p-value is 0.3158, which is greater than the typical significance level of 0.05, you would fail to reject the null hypothesis. This suggests that, based on the data and at the 0.05 significance level, there's no statistically significant association between being a multiform speaker for the "person" noun and the area of residence.

In simpler terms: The data does not provide sufficient evidence to say that the area of residence affects whether or not someone is a multiform speaker for the "person" noun.

#### 2. Association between being a multiform speaker for the "person" noun and Length

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

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

**1 2 3 4**

**0 10 20 11 5**

**1 0 4 8 4**

Monte Carlo simulation employed.

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

**replicates)**

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

**X-squared = 8.7944, df = NA, p-value = 0.02799**

The p-value of 0.02799 is less than the conventional significance level of 0.05. Therefore, we would reject the null hypothesis, suggesting that there is a statistically significant association between the multiform status (being a multiform speaker for the "person" noun) and the Length of the interview.

In simpler terms: The data provides evidence, at the 0.05 significance level, to suggest that the length of the interview has some relationship with whether or not someone is a multiform speaker for the "person" noun.

#### 3. Association between being a multiform speaker for the "person" noun and Gender.age

Null Hypothesis (H0): There is no association between the gender and age group of the individual and being a multiform speaker for the "person" noun.

Alternative Hypothesis (Ha): There is an association between the gender and age group of the individual and being a multiform speaker for the "person" noun.

**female 75 plus female under 75 male**

**0 18 14 14**

**1 6 4 6**

Monte Carlo simulation employed.

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

**replicates)**

**data: table(person\_data$multiform, person\_data$Gender.age)**

**X-squared = 0.31262, df = NA, p-value = 0.8621**

The p-value of 0.8621 is much greater than the typical significance level of 0.05. Thus, you would fail to reject the null hypothesis, implying that there's no statistically significant association between being a multiform speaker for the "person" noun (multiform status) and the Gender.age group of the individual.

In simpler terms: Based on the data and at the 0.05 significance level, there's no evidence to suggest that the gender and age group of an individual affects whether or not they are a multiform speaker for the "person" noun

### Year

#### 1. Association between being a multiform speaker for the "year" noun and Area

Null Hypothesis (H0 ): There is no association between the area of residence and being a multiform speaker for the "year" noun.

Alternative Hypothesis (Ha ): There is an association between the area of residence and being a multiform speaker for the "year" noun.

**East Poland West**

**0 17 10 10**

**1 8 6 11**

**Pearson's Chi-squared test**

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

**X-squared = 2.0415, df = 2, p-value = 0.3603**

The p-value of 0.3603 is greater than the standard significance level of 0.05. Therefore, you would fail to reject the null hypothesis. This suggests that, based on the data and at the 0.05 significance level, there's no statistically significant association between being a multiform speaker for the "year" noun (multiform status) and the Area of residence.

In simpler terms: The data doesn't provide enough evidence to suggest that the area of residence has an influence on whether or not someone is a multiform speaker for the "year" noun.

#### 2. Association between being a multiform speaker for the "year" noun and Length

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

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

**1 2 3 4**

**0 9 18 8 2**

**1 1 6 11 7**

Monte Carlo simulation employed.

**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.

In simpler terms: The data provides strong evidence, at the 0.05 significance level, to suggest that the length of the interview affects whether or not someone is a multiform speaker for the "year" noun.

#### 3. Association between being a multiform speaker for the "year" noun and Gender.age

Null Hypothesis (H0): There is no association between the gender and age group of the individual and being a multiform speaker for the "year" noun.

Alternative Hypothesis (Ha): There is an association between the gender and age group of the individual and being a multiform speaker for the "year" noun.

**female 75 plus female under 75 male**

**0 13 12 12**

**1 11 6 8**

**Pearson's Chi-squared test**

**data: table(year\_data$multiform, year\_data$Gender.age)**

**X-squared = 0.66915, df = 2, p-value = 0.7156**

The p-value of 0.7156 is greater than the standard significance level of 0.05. Therefore, you would fail to reject the null hypothesis. This suggests that, based on the data and at the 0.05 significance level, there's no statistically significant association between being a multiform speaker for the "year" noun (multiform status) and the Gender.age group of the individual.

In simpler terms: The data doesn't provide enough evidence to suggest that the gender and age group of an individual has an impact on whether or not they are a multiform speaker for the "year" noun.

# Part 3: Overall Analysis

### Summary:

In our investigation into linguistic variations, we sought to determine whether there exists any specific demographic characteristic, such as area of residence, length of the interview, or gender and age group, that predisposes individuals to using multiple forms when speaking.

Our primary analytical tool was the Chi-squared test, which is adept at discerning associations between categorical variables. However, it is essential to remember that a critical assumption of this test is that the expected frequency of each cell in the contingency table should be at least five. Violating this assumption may render the test results unreliable. Hence, whenever this assumption was threatened, we employed the [Monte Carlo](https://drive.google.com/file/d/1uahkFGze34D0W4gXFBxgPPUNZu5G8bgc/view?usp=drive_link) simulation method with the Chi-squared test. This method repeatedly samples from the observed data to obtain a more accurate p-value, especially beneficial when expected counts are low.

### Interpretation & Future Directions:

While certain factors like the length of the interview showed association with multiform speaking for specific nouns, others like area, gender, and age group were less influential. This might be due to cultural, regional, or individual differences in linguistic patterns. However, it's also possible that other unexplored factors or deeper linguistic nuances might explain the observed patterns better.

For future research, it might be fruitful to delve deeper into contextual factors, like the topic of conversation or the relationship between the interviewer and interviewee. Moreover, a more granular age categorization or exploring other demographic factors could provide more insights. Additionally, employing qualitative analysis alongside quantitative could elucidate the subtle nuances and reasons behind multiform speaking.

### 

### 

### 

### Limitations:

Obtaining more data could potentially reveal an association between gender.age/Area and multiform that might not be evident with a smaller sample size. Here's why:

1. Increased Power: The statistical power of a test refers to its ability to correctly detect a true effect. With more data, the power of the Chi-squared test (or other statistical tests) increases, which means it becomes more sensitive to detecting even small associations that might be missed with fewer data.
2. Smoothing Out Random Variation: A larger dataset can help average out random noise or variations. In smaller samples, random fluctuations might mask true underlying patterns.
3. Mitigating Biases: If the initial data collection had any biases (e.g., over-representing a particular group or under-representing another), collecting more data, especially if done systematically and inclusively, can help mitigate these biases.
4. Class Imbalance: Class imbalance refers to a situation in classification problems where the classes are not represented equally. In this instance, with a binary classification problem, having very few 1s compared to 0s would be an example of class imbalance. Many machine learning algorithms assume an equal number of instances for each class. With class imbalance, the model may be overly biassed towards the majority class, leading to poor performance on the minority class.

However, it's also essential to note:

Even with more data, there's no guarantee of finding a significant association if none truly exists. Gathering more data can be resource-intensive, so it's crucial to weigh the potential benefits against the costs. In conclusion, while obtaining more data can improve the chances of uncovering associations, it's also essential to ensure that the data is collected systematically and represents the population of interest accurately.