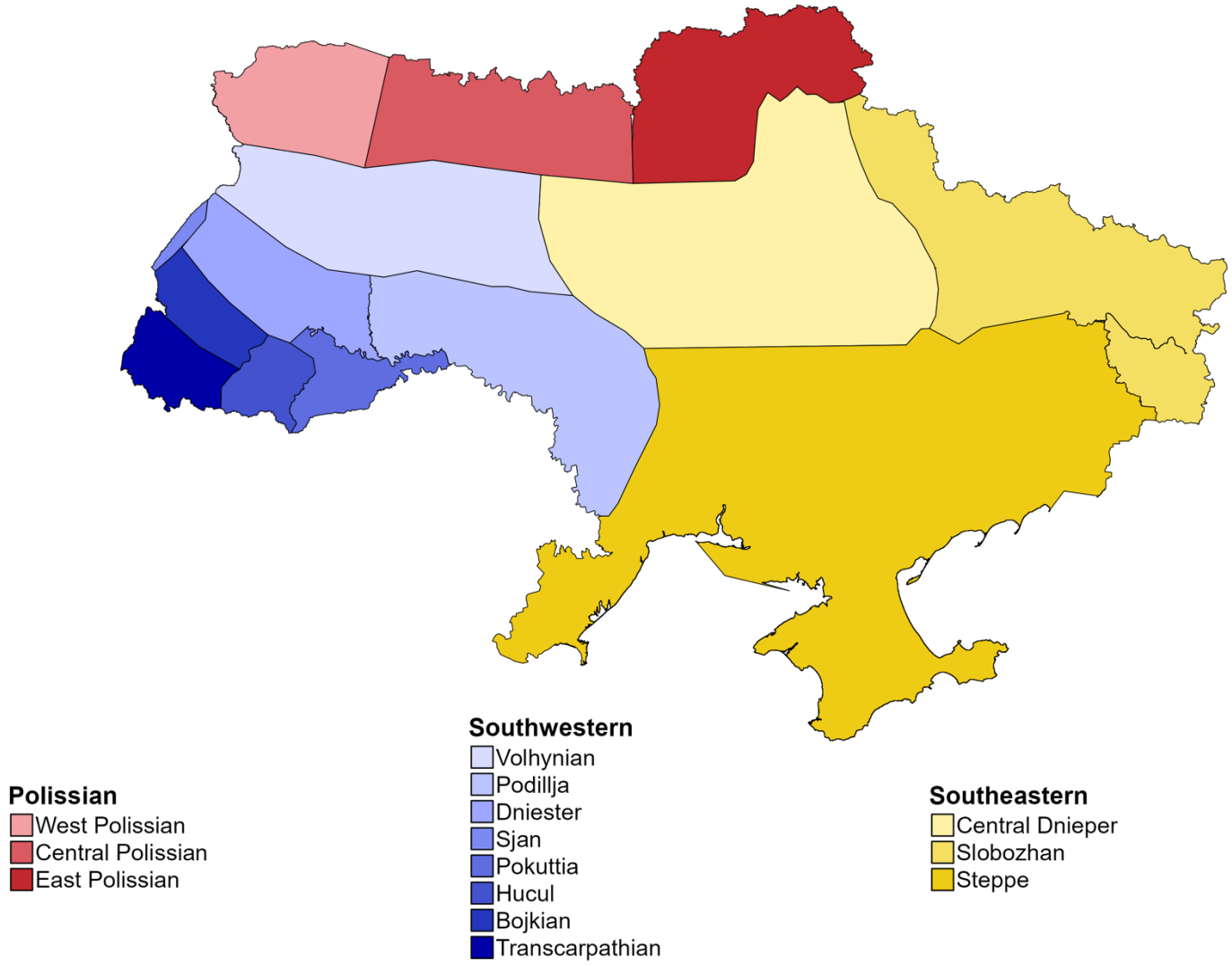
DIALECTOMETRIC ANALYSIS OF UKRAINIAN DIALECTS

Ilja A. Seržant, Kirill Kozhanov, Dmitri Sitchinava
- Universität Potsdam
- 15. Slavistiktag
- 30.09.–3.10.2026, Jena

POTSDAM SLAVIC VARIATION LAB

Distribution of Ukrainian dialects



Classification of Ukrainian dialects

• Dialect differentiation existed already in prehistoric times: the northeastern vs. southwestern areas, possibly reflecting ancient tribal distinctions among the East Slavs.

• The former northeastern dialects evolved into the contemporary Northern group, while the former southwestern dialects correspond to today's Southwestern group.

• The Southeastern dialects emerged in the 16th–18th centuries, following mass relocations of Ukrainian-speaking populations.

Classification of Ukrainian dialects

- It is generally assumed that the classification of contemporary Ukrainian dialects is based on phonetic, grammatical, and lexical criteria (Matvijas 1990: 33; Del Gaudio 2017: 55).

- This assumption has never been tested systematically, and this study represents a first step toward addressing this issue.

- We provide an aggregate analysis of Ukrainian dialectal data using methods typical of dialectometry (for an overview, see Weiling & Nerbonne 2015).

Research questions

- When aggregate analysis is applied, do Ukrainian dialects **pattern** in the same way as in the traditional classification? How different are the dialects from one another?

- How do features from **different linguistic levels** contribute to the observed spatial distribution, and how do distance measures derived from individual linguistic levels correlate with those obtained from the overall analysis?

- Which dialects are the **closest to Standard Ukrainian**, and which are more distant?

- How do Ukrainian dialects differ in terms of **preservation of archaic features**?

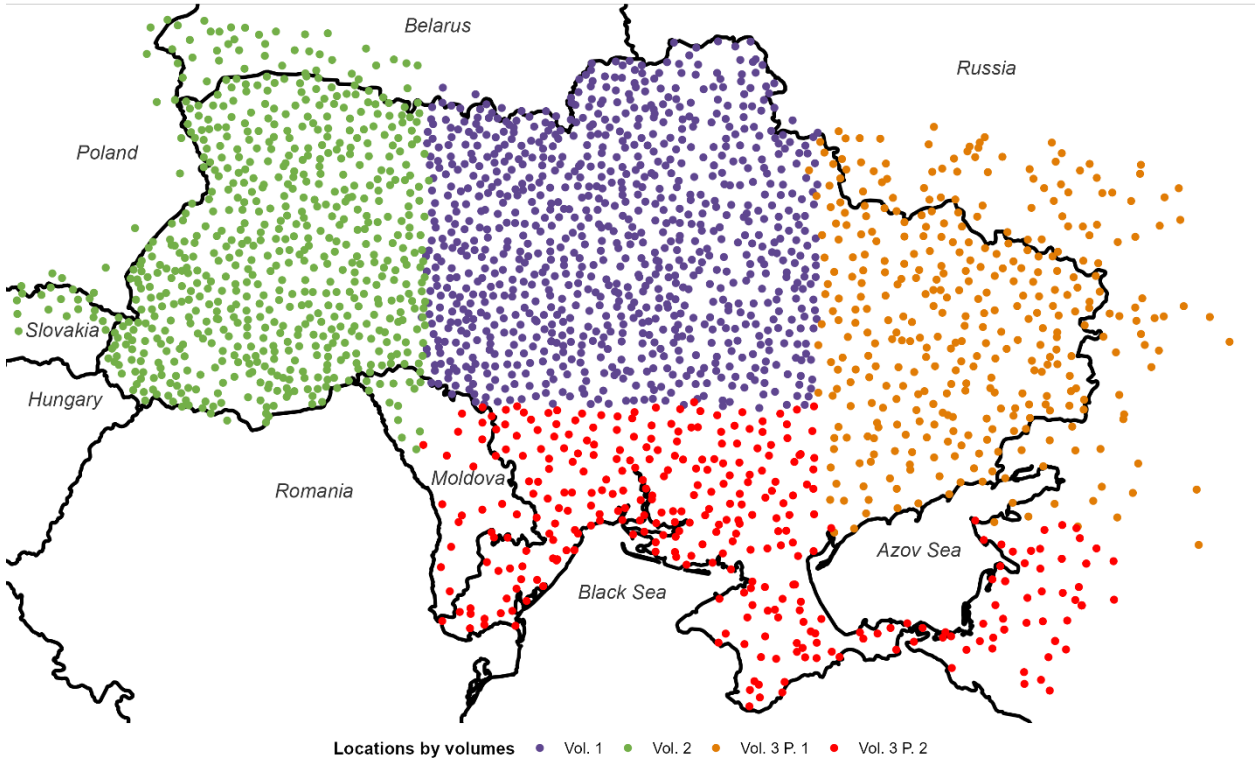
Data and annotation

Atlas of the Ukrainian Language (AUM)

Volume	Locations	Maps
Vol. 1	922	391
Vol. 2	673	407
Vol. 3, part 1	364	144
Vol. 3, part 2	346	151
Total	2305	1093

Data and annotation

Atlas of the Ukrainian Language (AUM)



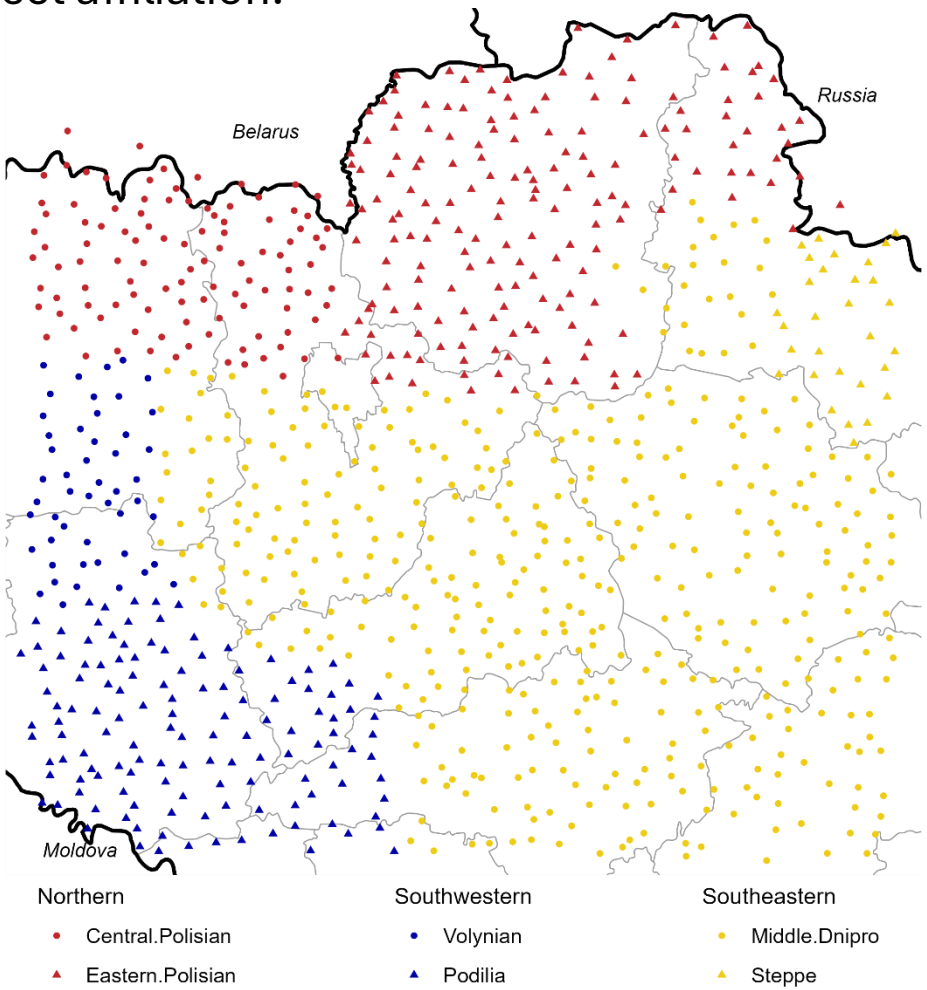
Data and annotation

Our dataset is based on Volume 1 of the AUM.

Domain	Maps	Locations	Rows	Missing data
Phonetics and phonology	157	922	160877	7321
Morphology	107	922	115255	4412
Lexis	69	922	79987	4786
Syntax	8	922	9232	216
Total	341	922	365351	16785 (~4.6%)

Data and annotation

Locations in the dataset by dialect affiliation.



Methodology

• We focus exclusively on **categorical variables** and calculate **linguistic distances** between locations using **the Jaccard distance**.

• Jaccard distance = as $\mathbf{1 - I/U}$, where I is the number of shared patterns (intersection) and U is the total number of attested patterns (union).

• This results in a **distance matrix** representing **dissimilarities** between all surveyed locations.

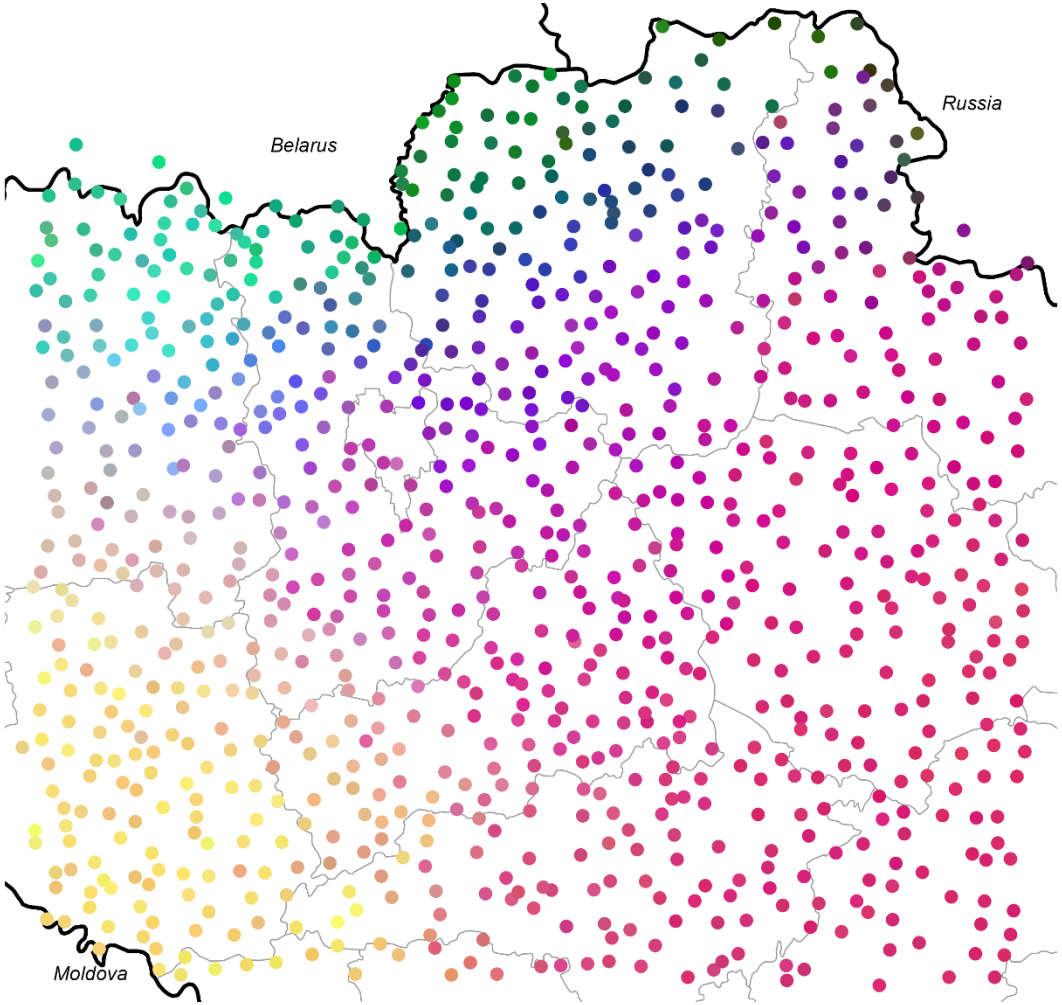
Methodology

- We apply a **multidimensional scaling** (MDS) procedure with three dimensions to visualize the distance matrix.

- The resulting coordinates of the **three MDS dimensions** were converted into **RGB (red, green, blue) color values**, which can be displayed on a geographical map (Heeringa, 2004).

Patterning of Ukrainian Dialects

MDS of Ukrainian dialects (three dimensions as RGB colors)



Patterning of Ukrainian Dialects

Clusterization of Ukrainian dialects (cf. Wichmann et al. 2026)

- We performed hierarchical clustering using five different methods.
  - complete-linkage clustering (= “complete”)
  - unweighted pair-group method with arithmetic mean (UPGMA = “average”)
  - weighted average linkage clustering (= “mcquitty”)
  - Ward’s method (= “Ward.D” and “Ward.D2”)

Patterning of Ukrainian Dialects

Clusterization of Ukrainian dialects (cf. Wichmann et al. 2026)

- The optimal number of clusters and the best clustering method were determined based on three criteria: **silhouette score**, **adjusted Rand index** (ARI), and **stability**.

- The **silhouette score** indicates how well-defined the obtained clusters are: points within a cluster should be close to each other and distant from points in other clusters.

- **ARI** measures the correlation between a given clustering solution and the averaged solution across all methods and cluster numbers.

- The **stability** index assesses the robustness of the clustering under resampling (for each solution, we used 100 bootstrap iterations with a sample ratio of 0.8).

Patterning of Ukrainian Dialects

Clusterization of Ukrainian dialects (best solutions)

Method	*k*	silhouette	ARI_mean	Stability
average	2	0.399	0.491	0.932
ward.D2	2	0.399	0.481	0.930
complete	2	0.398	0.453	0.776
mcquitty	2	0.344	0.574	0.622
ward.D	2	0.323	0.573	0.747
average	5	0.316	0.505	0.875
average	4	0.314	0.505	0.929
average	3	0.313	0.496	0.930
mcquitty	3	0.311	0.613	0.644
ward.D2	3	0.263	0.425	0.881

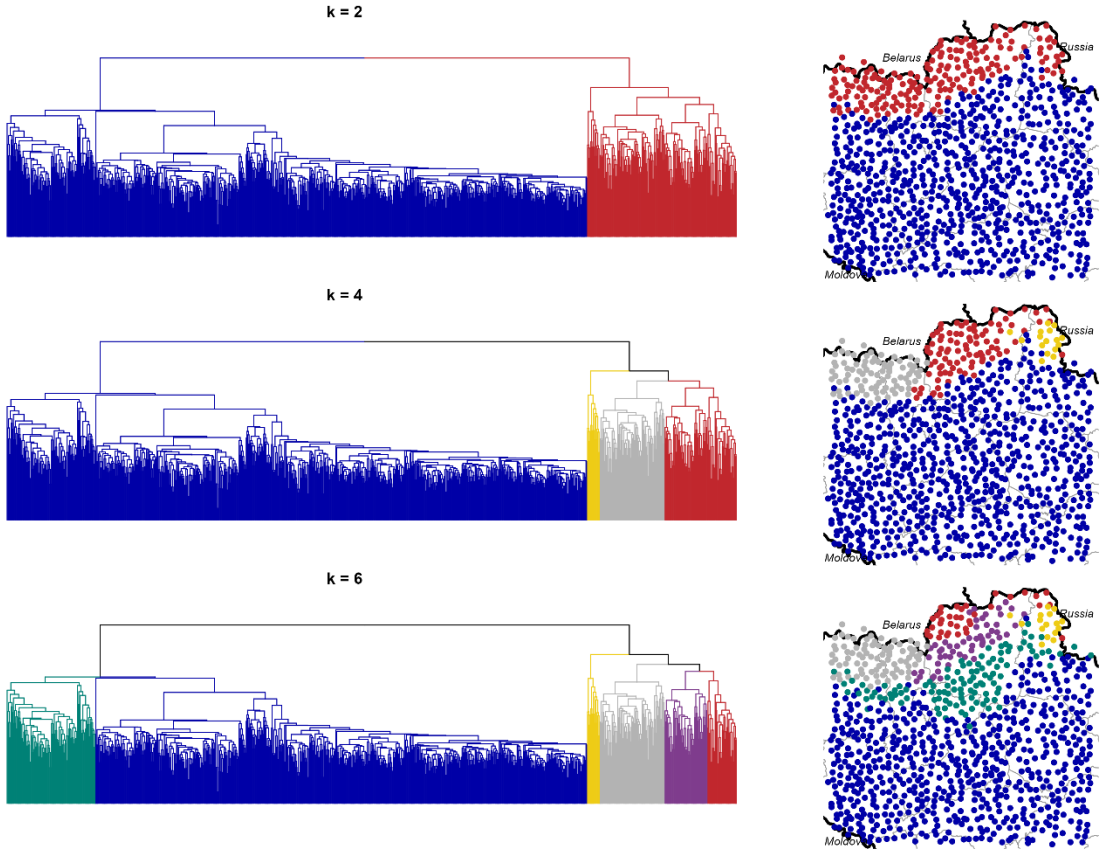
Patterning of Ukrainian Dialects

Clusterization of Ukrainian dialects

- The optimal clustering solution divides the data into **two** groups.
- However, since a more fine-grained partition is linguistically more informative, we also examined the clustering solution with **four** groups.
- Following the traditional dialect classification which distinguishes six subgroups for this area, we additionally cut the tree into **six** clusters.

Patterning of Ukrainian Dialects

Clusterization of Ukrainian dialects (2, 4, 6 cluster solutions)



Patterning of Ukrainian Dialects

Clusterization of Ukrainian dialects

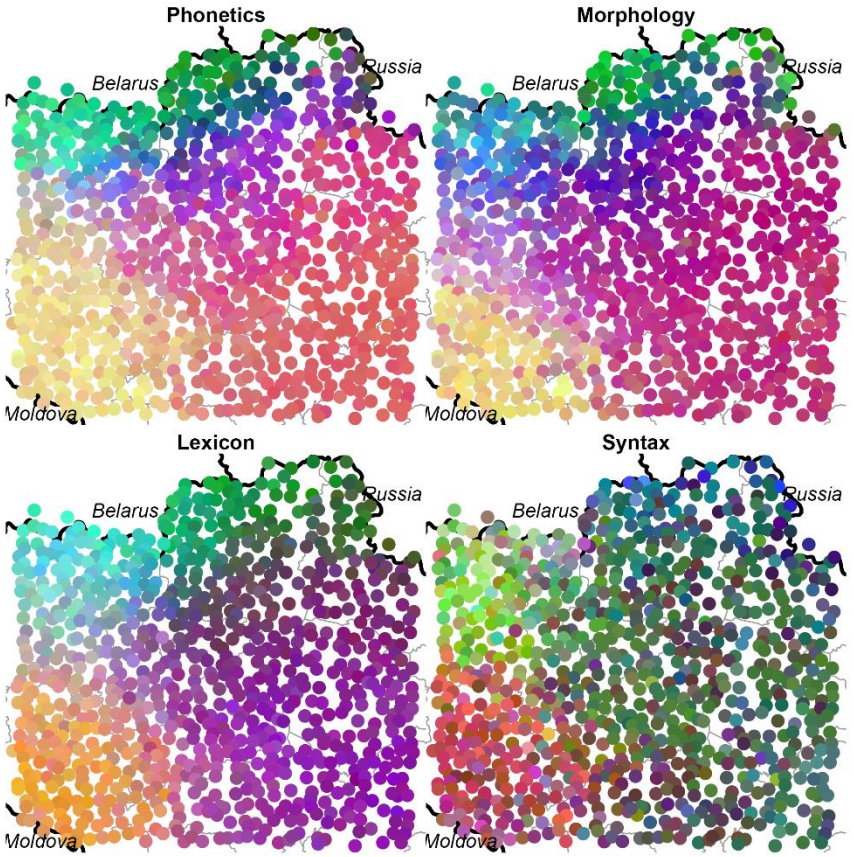
- In the two-group solution, **the primary contrast** in our dataset lies between **the Northern dialects**, on the one hand, and the **South**western and **South**eastern dialects combined, on the other.

- Further partitioning of the data does not reveal the three major dialect groups as traditionally predicted, but, instead, **additional clusters emerge within the Northern group**.

- The six-cluster solution likewise deviates from the traditional model, yielding **two additional horizontally oriented clusters** roughly corresponding to the **boundary** between the Northern and Southern dialects.

Variation among linguistic domains

MDS of the Ukrainian dialects by domains



Variation among linguistic domains

Correlation between distance matrices (Mantel test)

Matrix 1	Matrix 2	Mantel *r*	*p*-value
by all features	by phonetics	0.98	0.001
by all features	by morphology	0.97	0.001
by all features	by lexicon	0.86	0.001
by all features	by syntax	0.43	0.001
by phonetics	by morphology	0.91	0.001
by phonetics	by lexicon	0.78	0.001
by phonetics	by syntax	0.34	0.001
by morphology	by lexicon	0.81	0.001
by morphology	by syntax	0.43	0.001
by lexicon	by syntax	0.51	0.001

Ukrainian dialects vs. Standard Ukrainian

Linguistic distances between dialects and Standard Ukrainian

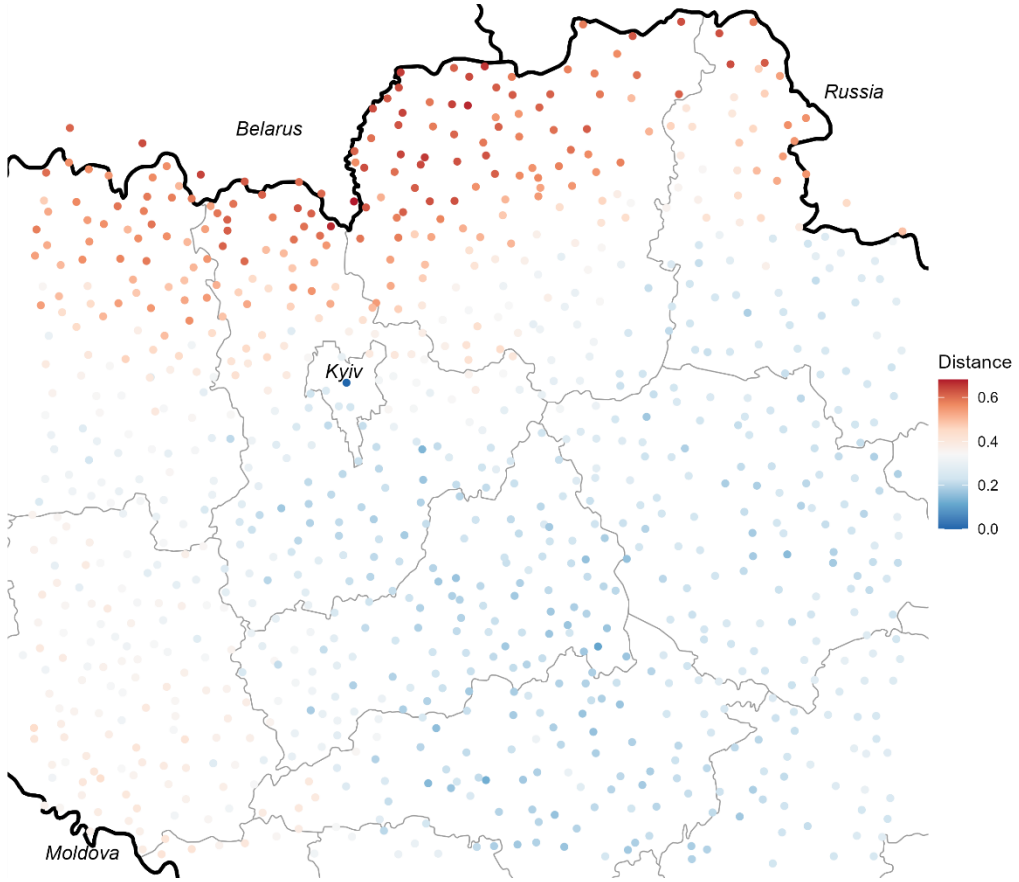
- We created a new location representing Standard Ukrainian, labeled “0,” with the coordinates of the capital city, Kyiv, and annotated it for all features from the dataset.

- We then calculated linguistic distances between the dialects in the dataset and the standard language.

RESULTS

Ukrainian dialects vs. Standard Ukrainian

Linguistic distances between dialects and Standard Ukrainian



Ukrainian dialects vs. Standard Ukrainian

Linguistic distances between dialects and Standard Ukrainian

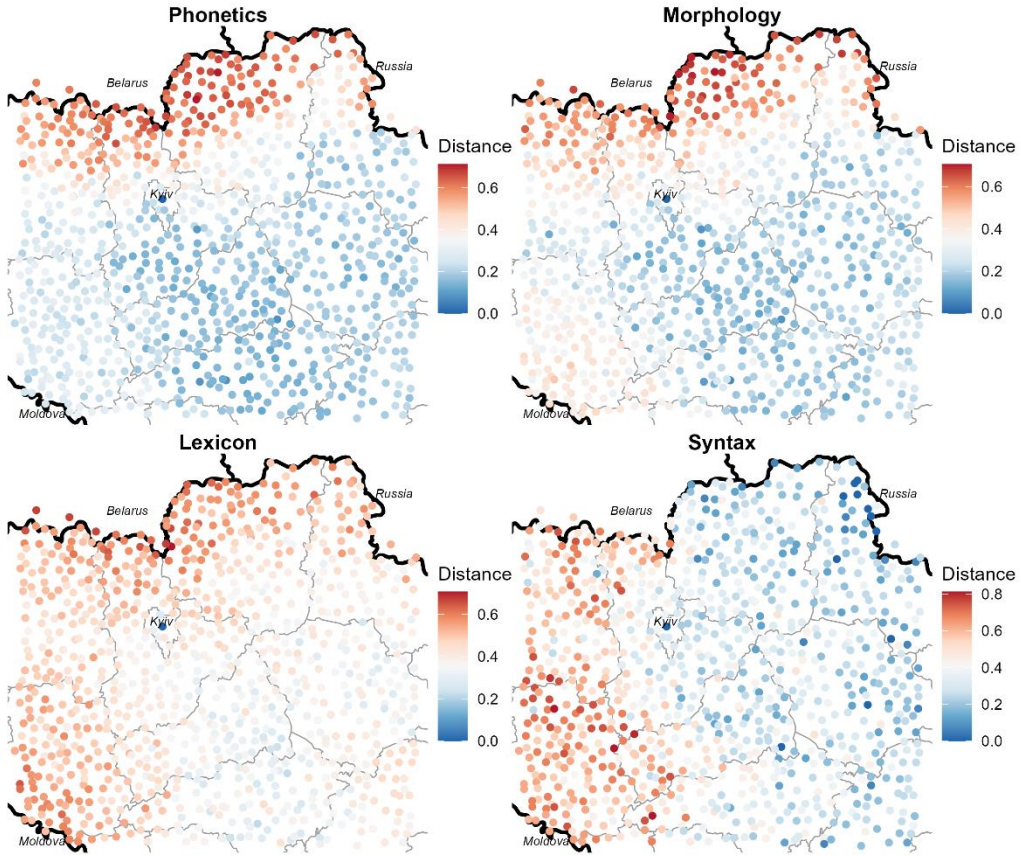
- Our data supports the claim that Standard Ukrainian is most closely related to the **Southeastern** dialects, showing particular similarity to the **Central Dnieper** dialects.

- The dialects most dissimilar to Standard Ukrainian are located in the **north**. A somewhat higher degree of dissimilarity is also observed in the dialects situated in the **southwest** of the area.

RESULTS

Ukrainian dialects vs. Standard Ukrainian

Linguistic distances between dialects and Standard Ukrainian by domains



Ukrainian dialects and linguistic innovations

Archaicity Score

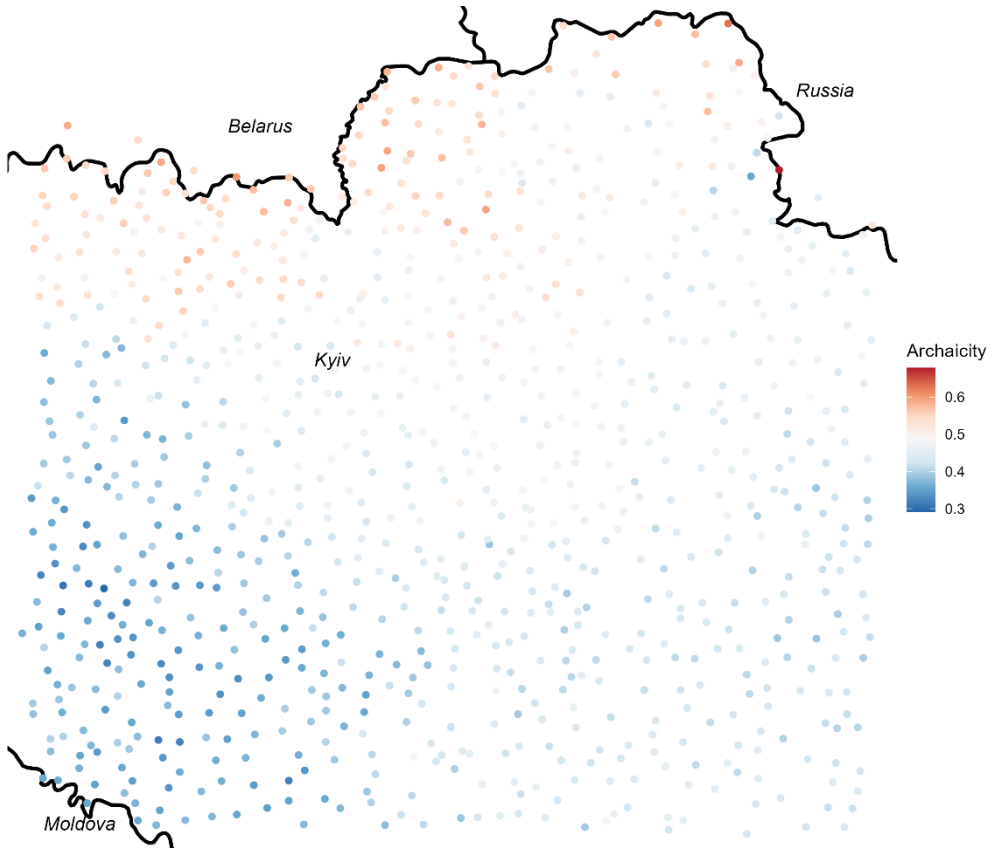
- For many features, original Old East Slavic states are attested. Following this approach, we additionally annotated 256 maps with regard to their degree of **innovativeness**.

- We then calculated an **archaicity score** for each location by dividing the number of archaic values by the total number of values attested for this location (the higher the score, the more archaic the dialect).

RESULTS

Ukrainian dialects and linguistic innovations

Archaicity Score by locations in the dataset



Conclusions

- Our analysis generally **supports** the traditional classification, **but**:
    - clear difference between the Northern dialects and the others
    - no significant difference between Southwestern and Southeastern dialects
    - a more fine-grained clusterization reveals greater internal differentiation within the Northern dialects than within the southern zone

Conclusions

• The traditional dialect classification aligns most closely with phonetic and morphological features, while the syntactic domain exhibits the greatest divergence from the traditional groupings.

• Standard Ukrainian is most similar to the Southeastern dialects, particularly the Central Dnieper varieties, whereas the Northern dialects differ the most.

• In terms of archaicity, the Northern dialects appear to more archaic, while the Southwestern dialects carry larger number of innovations.

References

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*, PhD thesis, Rijksuniversiteit Groningen, Groningen.
Matvijas, I. G. (1990). *Ukrajins'ka mova i jiji hovory*. Naukova dumka.
Wichmann, S., Jonjić, D., & Guzmán Naranjo, M. 2026. Foundations of Russian dialectometry. *Russkij jazyk v nauchnom osveshchenii – Russian Language and Linguistic Theory*.