The Application of Corpus-Based Language Distance Measurement to the Diatopic Variation Study (on the Material of the Old Novgorodian Birchbark Letters)

Ilia Afanasev (ilia.afanasev.1997@gmail.com)
Olga Lyashevskaya (olesar@gmail.com)

Overview

- Introduction
- Research problem
- Method
- Results and interpretation
- Conclusion

Birchbark letters: an overview

- Old Novgorodian (but not only)
- Short (most less than 100 tokens) documents
- approx. 1000 – 1500 CE

Current state of study

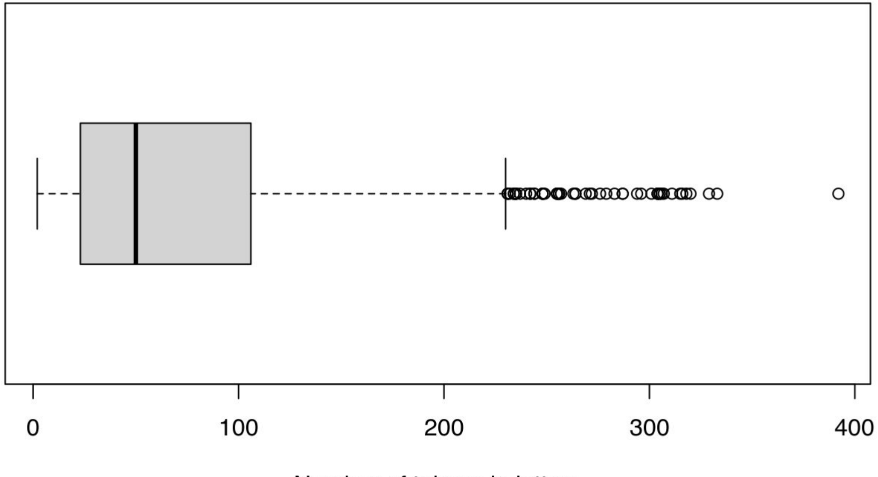
- Theoretical description (Zaliznjak 2004)
- Digitization in progress
- Lack of computational studies and inner classification

Task

- Discover individual variation
- Discover chronological variation
- Discover gender-based variation

Dataset issues

- Non-reconstructable tokens (and general fragmentedness of the letters)
- Researcher-imposed reconstructions
- Disbalanced dataset



Preprocessing I

- Eliminating non-reconstructable tokens
- Eliminating some of the researcher-imposed reconstructions
- Eliminating letters that are too long, or too short
- Split the rest to clusters:
  - individual
  - chronological
  - gender-based

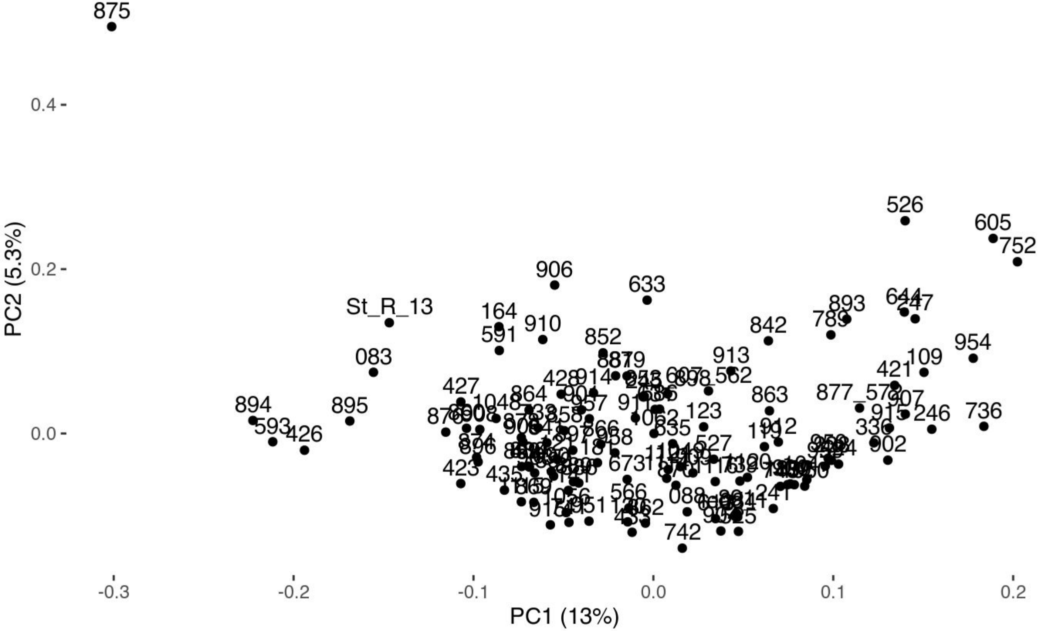
Preprocessing II

- Splitting into 3-shingles (*дару* > *ˆда, дар, ару, ру$*)
- Symbol embeddings with FastText
- Scoring alphabet entropy
- Scoring frequency ranks for 3-shingles

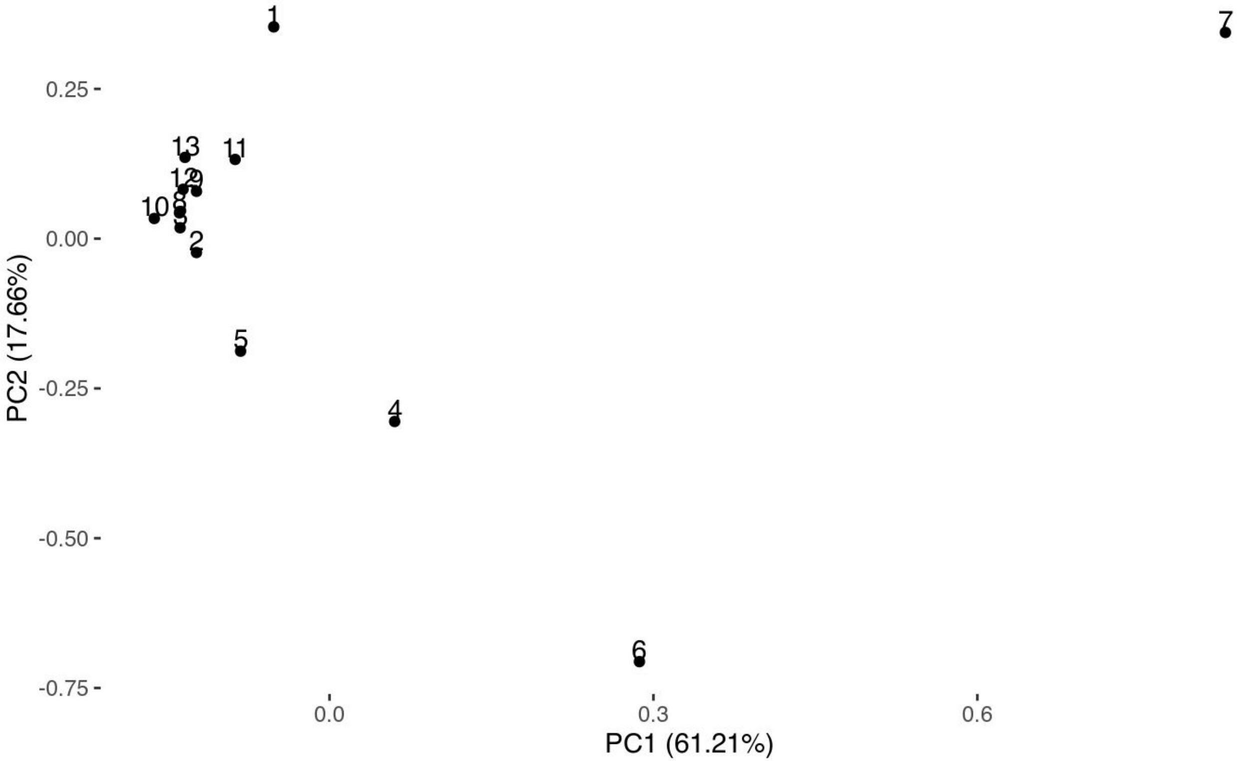
Method (Afanasev and Lyashevskaya 2024)

- The combination of metrics:
  - Mean DistRank for coinciding 3-shingles
  - Mean DistRank/string similarity measure hybrid for non-coinciding 3-shingles
  - Sørensen-Dice coefficient for lects
  - Split of the first by the third, multiplied by the second
- Vector-based string similarity measure
  - Euclidean distance between sums of symbol vectors of 3-shingles
  - Jaro distance between 3-shingles
  - Multiplication by alphabet entropy
- UPGMA classification
- Statistical analysis through PCA and HDBSCAN
- Qualitative analysis

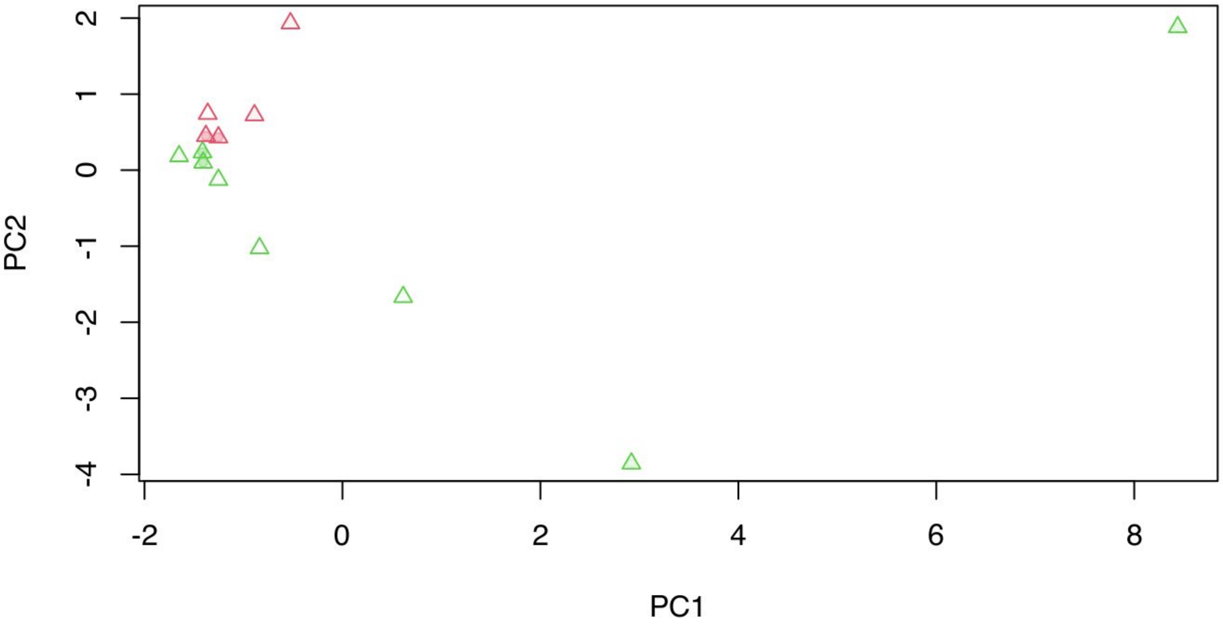
Individual variation – individual letters



Individual variation – clusters



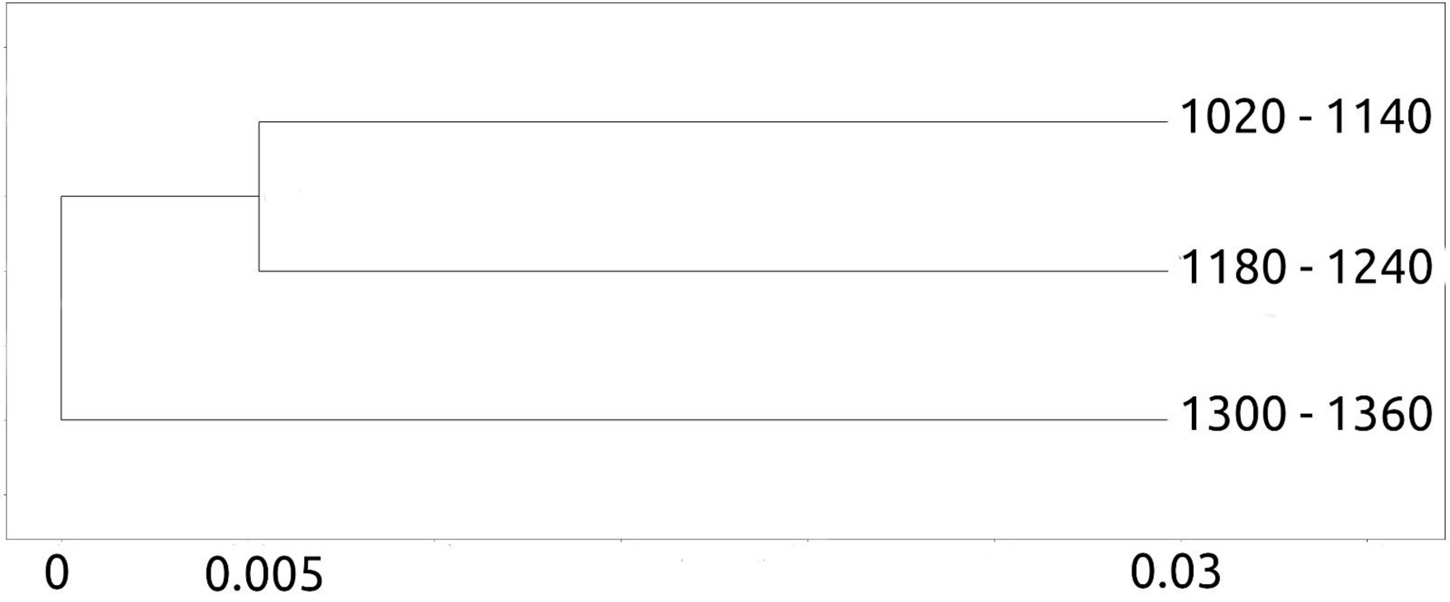
Individual variation – HDBSCAN



Individual variation – qualitative analysis

- Shared innovations (*ьло* vs. *ьлъ*)
- Noise in data (*еть* vs. *окь*)

Chronological division – classification



Chronological division – qualitative analysis

- Linguistic change aftermath: *въх* in *въхъ* ’entire’ in earlier periods vs. *вьсь* ’entire’ in the later period
- Linguistic change signal: increase in frequencies of symbols, denoting full vowels (e.g. *ло$*)
- Overall, mostly connected with the reduced vowel fall

Were there genderlects in Old Novgorodian?

- Distance between masculine and feminine authors is 0.12 (approximately equal to the mean distance of the letters within the same time period)
- Lack of differences in the usage of symbols that denote full vowels (0.002 for *ло$*)
- The absence of differences in the usage of symbols that denote reduced vowels (*лъ$* has the value of 0.17)
- Not universal (0.02 for both *но$* and *нъ$*)

Conclusion

- There was a significant chronological variation in Old Novgorodian
- There was no significant gender-based variation in Old Novgorodian
- Individual variation seems to be insignificant, however, letters within the same time period may form clusters

Prospects

- Inclusion of results as linguistic features into the databases
- Further exploration of found differences in distributions
- Using an outgroup for the additional linguistic context

Thank you!

References

Afanasev, I., Lyashevskaya, O. Measuring language distance based on small raw corpora // Proceedings of the 10th Congress of the International Society for Dialectology and Geolinguistics, Bucharest, September 4 – 8, 2023. / Eds. M. Nevaci [et al.]. 2024.

Andrej A. Zaliznjak. 2004. Drevnenovgorodskij dialekt [Old Novgorodian dialect]. Jazyki slavjanskoj kul’tury [Languages of the Slavic culture], Moscow.

Yuri G. Zelenkov and Ilya V. Segalovich. 2007. Comparative analysis of near-duplicate detection methods of web documents. In Digital Libraries: Advanced Methods and Technologies, Digital Collections, 9th All-Russian Scientific Conference RCDL’2007 Proceedings, Pereslavl’-Zalessky.