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

Ilia Afanasev
MTS AI LLC
ilia.afanasev.1997@gmail.com

Olga Lyashevskaya
HSE University
Vinogradov Russian Language Institute RAS
olesar@gmail.com

Abstract

The paper presents a computer-assisted exploration of a set of texts, where qualitative analysis complements the linguistically aware vector-based language distance measurements, interpreting them through close reading and thus proving or disproving their conclusions. It proposes using a method designed for small raw corpora to explore the individual, chronological, and gender-based differences within an extinct single territorial lect, known only by a scarce collection of documents.

The material under consideration is the Novgorodian birchbark letters, a set of rather small manuscripts (not a single one is more than 1000 tokens) that are witnesses of the Old Novgorodian lect, spoken on the territories of modern Novgorod and Staraya Russa at the first half of the first millennium CE.

The study shows the existence of chronological variation, a mild degree of individual variation, and almost absent gender-based differences. Possible prospects of the study include its application to the newly discovered birchbark letters and using an outgroup for more precise measurements.

1 Introduction

This article discusses the complexities of studying the variation with the low-resourced data on the material of the corpus of the East Slavic birchbark letters, dated from 1020 to 1500 CE, found in the territories of modern Russia (among others, Staraya Russa, Pskov, Moscow, Smolensk) and Belarus (Viciebsk[1], Mscislaŭ[2]); the most well-known site with the most manuscripts, where they were originally discovered, is Novgorod (Zaliznjak, 2004).

The research investigates three distinct types of variation: variation within a collection of doculects from the same place and time, time variation, and gender variation. The latter two are impossible to study without the first one (with a high level of individual variety, it is not possible to produce more effective research with more approximation) and are crucial for the study of Old Novgorodian, allowing one to understand the social dynamics within the society and the reflection of its development on its language. They are also irreplaceable for the building of the Old Novgorodian lect resources as they help to capture its ever-changing state until its extermination in the XV - XVI centuries CE.

Birchbark letters are usually small fragmented texts, so there is no way to use a more traditional lectometry (Shim and Nerbonne, 2022) or a corpus-based (Gamallo et al., 2020) methodology. The study requires a method designed for small raw corpora. The method relies on the combination of frequency-based metrics, string similarity measures, and a set similarity coefficient and their application to the subtoken-level units.

The research is based on three hypotheses:

H1. The differences detected by the proposed method among the individual documents are insignificant.

H2. The differences among the chronological periods of Old Novgorodian are significant.

H3. Genderlects are present in Old Novgorodian, there were significant differences between

[1]Most commonly called Vitebsk after the Russian variant; this article gives the official Belarusian transliteration for it and the other mentioned Belarusian cities.

[2]Most commonly called Mstsislaw after the Russian variant.

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 153–164
March 2, 2025 ©2025 Association for Computational Linguistics

the style of writing between men and women.

To test the hypotheses, the article uses a combination of quantitative and qualitative analysis, aimed at differentiating between random distributional skewings and regular significant differences. The important constraint is that the proposed method is intended to be preliminary, its results are not set in stone and require subsequent exploration by a human scholar, which this article is going to perform. However, it is necessary to state that a thorough qualitative analysis will require a detailed close reading of hundreds of texts (Zaliznjak, 2004), so the study concentrates on the method application and the exploration of its results.

The structure of the study is as follows. Section 2 expounds on the history of the Old Novgorodian studies and defines the present research gap. Section 3 provides detailed information on the utilised data. Section 4 explains the method and the means of analysis. Section 5 reports on the results of the experiments. Section 6 is a conclusion that outlines the future research prospects.

2 Related Work

The East Slavic birchbark letters have been known in the field of Slavic studies since the second half of the XX century (Zhukovskaja, 1959), however, for a long time they failed to gain recognition, as scholars perceived them as erroneous and illiterate, thus having little to contribute to the language study (Isačenko et al., 1980), which is a common misconception in traditional and generative studies (Otheguy and Stern, 2011). Only during the last two decades have the linguistic features of birchbark letters received acceptance as a full-fledged resource of information on lects spoken at the corresponding territory (Krys'ko, 1998; Zaliznjak, 2004). Since then, a significant body of work has been produced, with topics ranging from the language of these manuscripts(Andersen, 2006; Kwon, 2016; Gippius and Schaeken, 2011; Dekker, 2018), including the genderlect variation (Zaliznjak, 1993) and sociolinguistics (Lebedeva, 2003), to the creation of a network of linguistic databases that includes Birchbark Letter Database[3] (BLD), and Russian National Corpus [4] (RNC).

Old Novgorodian is part of a large group of historical and contemporary lects, generally called fragmented languages, which are attested only partially and by rather low-resourced corpora (in the best-case scenario, less than 100 000 tokens, in the worst-case scenario, less than 100 tokens) (Baglioni and Rigobianco, 2024). These lects present a significant challenge to the NLP methods due to their low-resourcedness and heterogeneity (Swaelens et al., 2023; Doyle and McCrae, 2024; Lyashevskaya and Afanasev, 2021). Old Novgorodian and the cases akin to it (Verhelst, 2020–2021) add a new layer to the complexity of the task, as the texts themselves frequently lack significant parts due to the damage to the original manuscript.

Despite the relative well-studiedness of the Old Novgorodian (Zaliznjak, 2004) and a high awareness of the low-resourcedness problem in NLP (Dione, 2019), there are crucial lacunae in the current research. Some types of language variation in the birchbark letters gained attention (Zaliznjak, 1993), but not all of them: for example, the chronological division remains understudied (Zaliznjak, 2004). The 2010s advancements in computational methodology (Nerbonne et al., 2013) were not applied to the language variation within Old Novgorodian. At the same time, low-resourced NLP rarely problematises the features of the analysed lects from the linguistic perspective (de Graaf et al., 2022), but rather declares these features as obstacles to be overcome via strictly mathematical algorithm enhancement (Nehrdich and Hellwig, 2022) and only rarely with language-aware methods (Prokić and Moran, 2013). The current study aims to become the first step in the direction of a language-aware computer-assisted study.

3 Data

The research corpus consists of 1249 documents available in the BLD as of February 2025. The distribution is heavily skewed in favour of the Novgorod letters which comprise most of the dataset. It complicates the comparison between different regions. At the same time, some of the non-Novgorod charters are still going to influence the results of comparison by any other criterion (gender or time frame), especially given the number of tokens in some of them. For instance, *Mosk_3*, the third of the charters found in Moscow, has 470 tokens. As one of the biggest manuscripts

[3] http://gramoty.ru/birchbark

[4] https://ruscorpora.ru/en/corpus/birchbark

in the dataset, it may quantitatively outweigh a hundred other charters. To eliminate this noise in measurements that use other criteria, the study data is restricted to charters from Staraya Russa and Novgorod that represent Old Novgorodian in the strict sense (Zaliznjak, 2004).

For the study, these letters undergo several stages of preprocessing [5].

The birchbark letters suffer from being very fragmented, and it is barely possible to use them either in their raw form preserving only the fully visible characters (there is not enough information), or in the processed form containing all the reconstructed characters (which may lead to the researcher bias interfering with the existing variation). Thus, preprocessing starts with creating the middle ground.

The initial step is to exclude all completely non-reconstructible tokens, marked with .... The next stage is the deletion of string breaks, marked with ±±. Following this, each of the charters is joined into a single string. After this, the non-reconstructible parts of the existing tokens (... joined to the tokens in the existing forms) undergo replacement with ѧ signs. The same applies to the parts of the tokens that may be inferred from the context but are not present in the charter in any shape or form, originally surrounded by *()*. If such reconstruction spans between two or more tokens, both the end of the first token before the reconstruction and the beginning of the second token after the reconstruction receive the ѧ sign. The present misspellings, originally designated by {}, are excluded from the texts. The parts of the tokens that are not fully visible but reconstructible with a high degree of certainty, surrounded by *[]*, are taken as is; only the designating signs *[]* are excluded from the resulting text. The final step is to merge the consequent break signs ѧ that appear before the token in cases when the break and/or unrecognisable symbols go before the token that contains a non-reconstructible part. Table 1 shows examples of the transformations that the texts undergo.

The further modifications to the dataset have the purpose of adjusting it for the clusterisation: the letters from Novgorod and Staraya Russa still suffer from an imbalance between the size of different charters: some are too small, consisting only of one token, and some are too big, containing hundreds of tokens. Figure 1 shows the distribution.

Original text	**Transformed text**
рж(и)	ржѧ
[с]	с
·к· {блъ} блъ	·к· блъ
–ружиного шло с... ... по	ѧружиного шло сѧ по
дар(у с о)[с]ипова	дарѧ ѧсипова
сел=<lb/>а	села

Table 1: The preprocessed parts of the Novgorod birchbark letter 1, compared to the fully preprocessed version, are present in the BLD database.

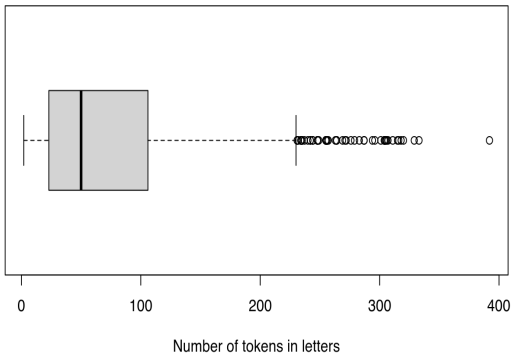

Figure 1: Boxplot for the distribution of the number of tokens in Novgorod and Staraya Russa birchbark letters.

For a more straightforward comparison, the next preprocessing step excludes each letter that consists of less than two tokens and five symbols. The letters that consist of more than 60 tokens (an approximate value of *Q3*[6] + *1.5 * (Q3 - Q1)*, with *Q3 = 27.00* and *Q1 = 6.00*) are shortened to the first 45 largest and the first 15 smallest tokens to preserve their features in the set while partially eliminating imbalance. Figure

[5] https://zenodo.org/records/14808682

[6] Q denotes quartiles, the cut-off points for the range of numbers that split this range into four more or less equal parts. Q1 is the first quartile, below which lie the first 25% of the range values, for example, 25% of the least frequent words in the language. Q3 is the second quartile, below which lie the final 25% of the range values, such as 25% of the most frequent words in the language.

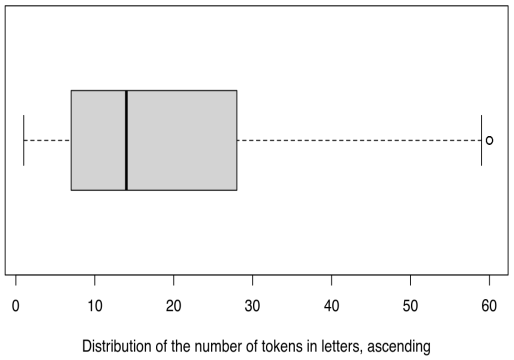

Figure 2: Boxplot for the distribution of the number of tokens in the letters after normalisation.

2 shows the final distribution: here, the only remaining letters are the ones that are more than two tokens or five symbols and less than 60 tokens in length. This contrasts the original distribution, where there was a significant number of letters containing one short token, which are not very useful for comparison purposes, and a dozen of letters that contain several hundred tokens, which significantly overweight the other letters, rendering comparison meaningless.

During the following step, each remaining letter receives two metadata tags, based on the existing analysis: time period and author gender.

Three periods are in the focus of the research: 1020 - 1140 (the early stage (Zaliznjak, 2004)), 1180 - 1240 (marked by the intensive contacts within the Circumbaltic region (Wiemer and Seržant, 2014; Podtergera, 2021)), and 1300 - 1360 (one of the latter stages of the Old Novgorodian development, also marked by the dissolution of the East Slavic area (Stankievič et al., 2007)). The texts assigned to these periods in RNC acquire the corresponding tag, the others get tag *X*.

For most of the texts, there is no possibility to deduce the gender of the author. In such cases, they receive the tag *UNK*. Otherwise, *m* (for authors referred to with masculine gender), and *f* (for authors referred to with feminine gender).

The resulting data frame containing 1158 letters consists of 7 columns, excluding index: *text* (the processed text of the charter), *charter_number* (the index of the letter in the database), *num_token* (number of tokens, excluding punctuation marks), *text_len* (length in symbols, excluding punctuation marks and breaks symbols ѣ), *author_gender* - the gender of the author, *date* - the estimated period of text creation, *place* - the estimated place of text creation. Further preprocessing, required for specific language distance measurement methods, will be discussed in the corresponding section.

Experiment	**Number of letters**	**Number of compared lects**
1020 - 1140 CE period, internal clusterisation	118	118
1180 - 1240 CE period, internal clusterisation	231	231
1300 - 1360 CE period, internal clusterisation	140	140
Chronological clusterisation	489	3
Gender-based clusterisation	397	2

Table 2: The quantity of the letters, used in the experiments, and their internal grouping.

4 Method

This section consists of three parts that elaborate on the preprocessing of the data for the experiments, applied quantitative methodology, and qualitative analysis. The implementation is available via GitHub[7].

4.1 Preprocessing

The first step of the preprocessing stage is to select the required combination of the letters and their grouping for each of the experiments. The latter includes a document-to-document comparison within the three selected periods, a comparison of the three time periods between themselves, and a comparison of the letters by authors of different genders. Table 2 shows the final numbers for each of the experiments.

The first three experiments in Table 2 elaborate on the internal variation within the given period, so the unit of the analysis is a doculect (a lect of the individual letter). The fourth experiment takes a

[7]https://zenodo.org/records/14808716

more distant look at the differences between periods and groups letters from each period together in a single list. The fifth experiment deals with gender-based classification, so the split is between two genderlects.

After the split, the letters are prepared for distance measurement. Tokens within each lect are split into overlapping character 3-grams, further called *3-shingles* (Zelenkov and Segalovich, 2007), as the extremely fragmentary texts make it impossible to use whole tokens as the main unit of comparison. This way of analysing the texts is akin to byte-pair encoding (BPE) (Gage, 1994), which also utilises subtoken units. 3-shingles are a fixed unit, which, in contrast to BPE, complicates semantic comparison but enables a formal one, especially on the phonetical and morphological levels (Lyashevskaya and Afanasev, 2021), better suited for onomaseological lectometry purposes (Shim and Nerbonne, 2022).

The beginning and end of each token receive special marks, ˆ and *$* respectively. Then algorithm removes each 3-shingle containing the ѣ sign as there is no way to deduce the symbols that lie behind it, and, subsequently, it may generate a lot of noise and skew the distributions in a way that does not accurately reflect the linguistic behaviour of the speakers. Thus, the token ѣостерѣ becomes a collection of 3-shingles ост, сте, тер, while token дару becomes a collection of 3-shingles ˆда, дар, ару, ру$. If the letter consists only of fragments of the size of two or fewer symbols, it gets completely removed from the dataset. Note, however, that the intact short tokens remain in place (for instance, ˆa$), as their deletion would significantly skew the distribution, deleting crucial linguistic information (Kestemont, 2014).

The next step is adding symbol embeddings: as the main unit of the analysis is a 3-shingle, its only possible subtoken is a single symbol, so the vector-space representation should be built for it. For embedding producing, the study employs the FastText (Bojanowski et al., 2017) model, which does not possess the inherent bias of large transformers (Devlin et al., 2019), namely, the information on the other languages, used for pre-training, which can add noise. The hyperparameters for the FastText model are in Appendix A.

The following step is to score the alphabet entropy (Shannon, 1948) for each of the analysed lect groupings, which can be approximated as the average value of the probability of the symbols appearing in their respective positions.

The last part of the preprocessing includes merging 3-shingles for each of the lect groupings and scoring their frequency ranks (the most frequent gets 0, the least frequent - *N* - 1, where *N* is the total number of 3-shingles). Frequency ranks are then normalised into the interval of [0;1], as the method requires.

4.2 Distance measurement and clusterisation

As the preliminary experiments have shown, the study employs the most efficient possible setup of the method utilised, which includes multiplying mean DistRank (Gamallo et al., 2017) between the coinciding 3-shingles by a hybrid string similarity measure for the non-coinciding 3-shingles, and dividing by Sørensen-Dice (Sørensen, 1948) coefficient[8] between two lects.

The employed string similarity measure for hybridisation is vector-weighted Jaro distance normalised (VWJDN), a product of Euclidean distance of the sums of symbol embeddings between two 3-shingles, and the Jaro distance between them (Jaro, 1989). The main idea is to emulate the phonetic differences between the sounds that the symbols represent and the distributional differences between the symbols themselves. Jaro distance accounts for transpositions, and thus, for the symbol order. The result of VWJDN undergoes multiplication by alphabet entropy differences between the given lects to account for potential distributional skewings, caused by dissimilarities in the utilisation of the graphic system (Zaliznjak, 2004).

The Sørensen-Dice coefficient between sets (in this case, sets of 3-shingles within the particular lects) *A* and *B* is:

$$\frac{2 * |X \cap Y|}{|X| + |Y|}$$

.

The algorithm is provided below.

The results of the combined metric form the distance matrix between all of the present lects. There are two ways to utilise this metric afterwards.

The first one is to use it for creating a clusterisation as is. Here, the unit of analysis is a

[8] In natural language processing evaluation more frequently referred to as F-score(Derczynski, 2016)

Algorithm 1

Separate 3-shingles that coincide between lects A and B ($A \cap B$) from 3-shingles that do not coincide between A and B (A XOR B)
Calculate mean *DistRank($A \cap B$)* (Gamallo et al., 2017) between coinciding 3-shingles of A and B
for each 3-shingle *a* of A that is in (A XOR B) **do**
    for each 3-shingle *b* of B that is in (A XOR B) **do**
        VWJDN(a, b)
    end for
    Select the pair with minimal *VWJDN(a, b)*
    Calculate *VWJDN(a, b) * DistRank(a, b)*
end for
for each 3-shingle *b* of A that is in (A XOR B) **do**
    for each 3-shingle *a* of B that is in (A XOR B) **do**
        VWJDN(b, a)
    end for
    Select the pair with minimal *VWJDN(b, a)*
    Calculate *VWJDN(b, a) * DistRank(b, a)*
end for
Score mean between all acquired values for non-coinciding 3-shingles (*VWJDND(A, B)*)
*VWJDND(A, B) * DistRank($A \cap B$)/Sørensen-Dice(A, B)*

single lect and the values with which the clusterisation algorithm runs are the distances between this lect and all other lects in the dataset. There are two possible ways to do it: perform a hierarchical bootstrap clusterisation with *pvclust* (Suzuki and Shimodaira, 2006) or perform HDBSCAN (Hahsler et al., 2019) over t-distributed Stochastic Neighbour Embedding (t-SNE) (van der Maaten and Hinton, 2008) over Principal Component Analysis (PCA) results (Jolliffe and Cadima, 2016). These are going to be used for inner clusterisation within the chronological periods.

The second one is to transform it into a lower triangular matrix and build a tree-like clusterisation with UPGMA (Sokal and Michener, 1958). This clusterisation algorithm is more effective for the lesser number of closely related lects and the study applies it to group chronological periods.

4.3 Qualitative analysis

The qualitative analysis is the most crucial research step. It takes the resulting clusterisations and attempts to explain the linguistic reasoning (or lack thereof) behind the decisions of the similarity metrics (whether they are correct or not). It uses the information that the utilised software provides, namely, the tables of comparison between all the 3-shingles, to discover the linguistic patterns in the data. As 3-shingles appear across the different tokens, the detection of a pattern goes through two steps. The first includes going through the generated table of correspondences between lects to check for possibly meaningful, based on the pre-existing body of work, similarities and dissimilarities, the second – going through the texts of the letters to prove the meaningfulness of the discovered distributional skewings. Table 3 provides the example of the generated table of correspondences.

The aim of qualitative analysis is to either state that the dissimilarities between the groups detected by method are not significant, or to explain them on three key levels: individual (on the level of doculects), chronological (on the level of chronolects), and gender-based (on the level of genderlects).

5 Experiments and Analysis

This part provides the summary of the experiments and the subsequent discussion of the linguistic differences detected by the method.

5.1 Inner variation within the time periods

The experiments that investigate the linguistic variation of the individual letters within chronological periods show significant homogeneity in each one (see Figure 3). However, on the individual level, PCA does not demonstrate significant explanatory power, the differences are initially too small and too scattered across the analysed letters.

The next step includes an attempt to dense the data and provide more power to the final analysis on the first period sample, 1020 - 1140 CE. This stage starts with performing bootstrap clusterisation (hyperparameters are in Appendix B), the results of which become the new 13 groups of lects. These new groupings consist of 2 (an outgroup, letters 431 and 557 from Novgorod) to 30 items, and represent higher-level, more reliable, according to the bootstrap clusterisation (*AU* > *85%*),

1180–1240	1020–1140	Metric	Distance
ˆпо	ˆпо	Novgorod birchbark leters by period-1-False-DistRank-True-True-False-weighted_jaro_winkler_wrapper-True - DistRank	0.0012717253073336043
рвн	рва	Novgorod birchbark leters by period-1-False-DistRank-True-True-False-weighted_jaro_winkler_wrapper-True - hybrid	0.4689305328575265

Table 3: A sample of correspondences established by VWJDN.

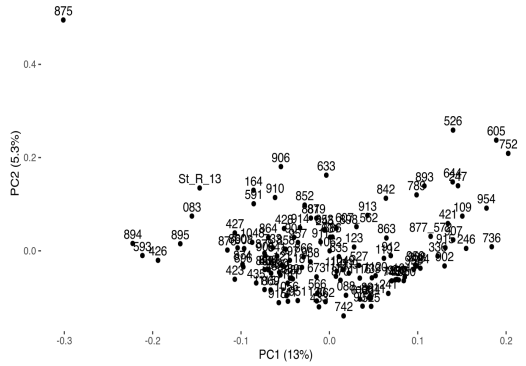

Figure 3: PCA of the distance matrix between the letters, written in the 1020 - 1140 CE.

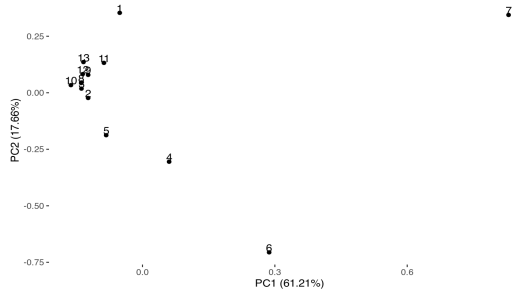

Figure 4: PCA of the distance matrix between the letters, written in the 1020 - 1140 CE, clustered into the higher-level groupings.

groupings. This time, it is easier for PCA to represent the key differences (Figure 4).

It is possible to run t-SNE with HDBSCAN over this result (Figure 5), showing the degree of certainty in cluster grouping. These figures include the same data points, with the first providing the information on the exact data point, and the second - on the reliability of clusterisation.

This shows two distinct bigger clusters, with groups 8 and 6 opposed to 12 and 9 as the centres of the clusters, and other groups joining them with a lesser degree of certainty. Group 7, a small

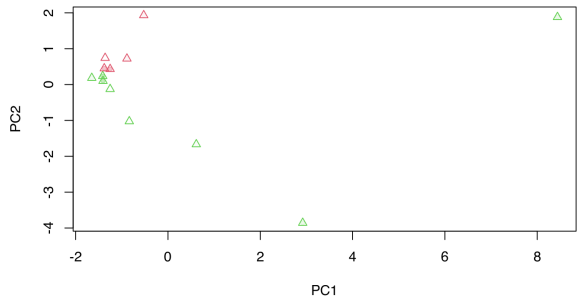

Figure 5: HDBSCAN, run over t-SNE results on PCA of the distance matrix between the letters, written in the 1020 - 1140 CE, clustered into the higher-level groupings.

higher-level outgroup, is an outlier here as well; PC1 is likely to represent the dissimilarities in the size of the cluster, detected by the Sørencen-Dice coefficient.

Interestingly, the consequences of the phonetic processes, such as the reduced vowel fall, help in joining some higher-level groupings together and not in splitting them. Thus, group 12 contains 3-shingle ьло, while group 9 contains 3-shingle ьлъ, with о and ъ known to become interchangeable symbols (Zaliznjak, 2004), as the first had denoted full vowel and the second - its reduced counterpart, before the reduced vowel fall occured. The distance between these 3-shingles is 0.42. At the same time, group 8 contains completely different 3-shingles, which, together with the 3-shingles of group 12, forms such pairs as еть – окь with a distance of 0.47. These dissimilarities in differences are the main cause of the split between two bigger clusters. Yet, overall the letters of a given time period are homogeneous, and it is safe to treat them further as a uniform entity.

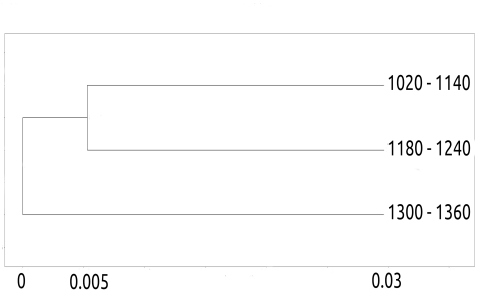

Figure 6: UPGMA (Sokal and Michener, 1958) clusterisation of chronolects, present in the dataset.

5.2 Analysis of chronological clusters

Figure 6 shows the grouping of three chronolects, representing three stages of Old Novgorodian evolution: 1020 - 1140 CE, 1180 - 1240 CE, and 1300 - 1360 CE.

The picture clearly demonstrates the differences between the chronolects, especially between two earlier groups and the later one. It seems that the Old Novgorodian changed between 1240 and 1300 more significantly than between 1140 and 1180, which is likely due to the inner processes as well as to the intensive language contact (Wiemer and Seržant, 2014).

Mostly, however, this is the same lect: the branch length is not exactly large (compare the differences between modern East Slavic territorial lects in (Afanasev and Lyashevskaya, 2024), acquired with the similar methodology, where the branch length is 0.175, and ingroup splits at 0.03). The found pairs of non-coinciding 3-shingles are mostly random (тев of 1300 - 1360 and тет of 1020 - 1140).

Still, some pairs can provide a scholar with a closer look into the ongoing phonetic processes. In the earlier periods, the 3-shingle въх is present in such tokens as въхъ 'entire'. In the later period, the other form for the meaning 'entire' prevailed: вьсь. At the same time, there are graphical differences: the later letters use 3-shingle оду, while the earlier prefer одоу.

Distributions of the coinciding 3-shingles also give a hint into the nature of differences between the stages of the Old Novgorodian development. While sequences with ъ$ and ь that earlier denoted reduced vowels, almost do not change their rank (лъ$ has a value of 0.002), the ones that denoted their full-fledged counterparts changed the distributions significantly (ло$ has a value of 0.15), becoming more frequent.

From the material given, it is possible to conclude the following: the utilised method allows insight into language variation and change which would not be possible on the token level. This becomes crucial in the case of DistRank-based analysis, which uniquely illustrates the dynamics of the reduced vowel fall process, highlighting the complexity of its written dimension.

5.3 Gender-based differences

The genderlects present a significantly more difficult challenge. The distance itself is not big, only 0.12 (for reference, the metric returns the same value between two letter clusters within the same time period). The non-coinciding 3-shingles here demonstrate the absence of any kind of meaningful correspondence, mostly consisting of pairs, akin to саю/ьса.

However, the DistRank behaviour for the symbols that denoted reduced and full vowels is once again suspicious. ло$ has a value of 0.002, while лъ$ has the value of 0.17. Similar occurs with мо$ (0.01) and мъ$ (0.05), то$ (0.07) and тъ$ (0.03), ^въ (0.01) and ^во (0.03). It seems that there were certain dissimilarities in preferences of female and male writers in relation to the ъ$ and о, but even these were restricted (cf. 0.02 for both но$ and нъ$).

The genderlect differences (or, rather, their lack thereof) show the limit of the method utilised. It can pick on the distribution differences, providing a distant reading, based on fixed-size subtoken units, but it inevitably fails when differences are either completely absent (and it seems that Old Novgorodian indeed did not have genderlect differences) or too subtle to pick without the close reading of documents.

6 Conclusion

The paper employed a new method to study individual, chronological, and gender variation within Old Novgorodian. It supported the hypothesis **H2** of chronological variation, showing the similarity between earlier and later periods of the Old Novgorodian development. At the same time, no signs of gender-based variation are present (hypothesis **H3** is thus rejected): from the existing material only it is impossible to claim that Old

Novgorodian had genderlects, which supports the primary qualitative work on the topic, Zaliznjak (1993). Yet the amount of the available material may be misleading: it is possible that there is not enough data. The method statement on the variation within the different time periods highly depends on the letter size, supporting the idea of balancing the corpus before the method application (Afanasev and Lyashevskaya, 2024); hypothesis **H1** is thus supported only partially.

One of the key elements that helped the method to distinguish between different chronological periods and played an important role in other tasks is the contrast between the symbols that denote reduced and full vowels. This is not the only found contrast, as the method was able to find other factors, such as lexical differences. It is also paramount to note that all the components of the combined metric were analysed, and partly proved, during the final qualitative analysis. This affirms the necessity of using lectometry methods for computer-assisted and not computer-driven research.

The acquired classification and the method itself, especially 3-shingle-based representation, will aid the analysis of the newly discovered documents and the exploration of how they fit the existing picture. It will facilitate expert judgment about the period of their creation, aiding theoretical paleographic analysis (Janin and Zaliznjak, 2000). The found similarities and dissimilarities may be included as linguistic features in the existing network of Old Novgorodian databases. The results require further attention and exploration, especially the ones that did not provide any satisfactory conclusions, such as the ъ$ and о distribution differences between the letters authored by men and women. The study shows that the quality of the resources is of the utmost importance for computational methods, especially for language distance measurement. One possible further research direction is using an outgroup (for example, Old East Slavic legal charters) to provide additional linguistic context to the clusterisation trees (Kassian et al., 2021).

Limitations

The research is based on the corpus of fragmented documents that contains all the known data about the Old Novgorodian lect, but definitely not all the data about the lect, which means that the comparison is corpus-driven and may not cover all the spectre of similarities and differences between the sublects (chronolects and genderlects) of Old Novgorodian (Davis, 2017). Furthermore, the dates of the letters creation are approximate, which may have influenced the chronolect comparison results. It is not possible to establish the author's gender for all the letters, therefore the material for the gender-based similarities and differences study is even less than it could have been, which too may have influenced the final comparison.

The applied method uses 3-shingles, the units of sub-token level (Afanasev and Lyashevskaya, 2024), as the main objective of its application is to find the differences between small raw corpora. This means that it captures the variation on the phonological, morphonological, and morphological levels, occasionally being able to account for the lexical differences, thus mostly resembling the character-based comparisons of morphological features and basic vocabulary lists (Kassian et al., 2021; Auderset et al., 2023). The syntactic and pragmatic differences are generally out of scope of this class of methods in general, due to the complications of diachronic syntax studies (Campbell, 2013). And, given the quantity of the material, any kind of the automatic quantitative analysis that does not utilise rigorous manual preprocessing, will not be suitable here as well. These features of Old Novgorodian require further study with other methods.

Acknowledgements

The authors are grateful to the anonymous reviewers for their insightful comments. The remaining errata are ours.

References

Ilia Afanasev and Olga Lyashevskaya. 2024. Measuring language distance based on small raw corpora. In N. Saramandu, M. Nevaci, I. Floarea, l.-M. Farcaș, A. Bojoga, F. R. Constantin, A. Loizo, M. Manta, M. Morcov, and O. Niculescu, editors, *Proceedings of the Xth Congress of the International Society for Dialectology and Geolinguistics*, pages 11–18. Edizioni dell'Orso, Alessandria, Italia.

Henning Andersen. 2006. Future and Future Perfect in the Old Novgorod Dialect. *Russian Linguistics*, 30(1):71–88.

Sandra Auderset, Simon J Greenhill, Christian T DiCanio, and Eric W Campbell. 2023. Subgrouping in a

'dialect continuum': A bayesian phylogenetic analysis of the mixtecan language family. *Journal of Language Evolution*, 8(1):33–63.

Daniele Baglioni and Luca Rigobianco. 2024. *Fragments of Languages: From 'Restsprachen' to Contemporary Endangered Languages*. Brill, Leiden, The Netherlands.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Lyle Campbell. 2013. *Historical Linguistics, third edition: An Introduction*. MIT Press.

Joseph Davis. 2017. The semantic difference between italian vi and ci. *Lingua*, 200:107–121.

Simeon Dekker. 2018. *Old Russian Birchbark Letters: A Pragmatic Approach*. Brill, Leiden, The Netherlands.

Leon Derczynski. 2016. Complementarity, F-score, and NLP evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 261–266, Portorož, Slovenia. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Cheikh Bamba Dione. 2019. Developing Universal Dependencies for Wolof. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 12–23, Paris, France. Association for Computational Linguistics.

Adrian Doyle and John P. McCrae. 2024. Developing a part-of-speech tagger for diplomatically edited Old Irish text. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 11–21, Torino, Italia. ELRA and ICCL.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Pablo Gamallo, Jose Ramom Pichel Campos, and Iñaki Alegria. 2017. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications*, 484:152–162.

Pablo Gamallo, Jose Ramom Pichel Campos, and Iñaki Alegria. 2020. Measuring Language Distance of Isolated European Languages. *Information*, 11(4):181–193.

Alexey A. Gippius and Jos Schaeken. 2011. On direct speech and referential perspective in birchbark letters no. 5 from Tver' and no. 286 from Novgorod. *Russian Linguistics*, 35(1):13–32.

Evelien de Graaf, Silvia Stopponi, Jasper K. Bos, Saskia Peels-Matthey, and Malvina Nissim. 2022. AGILe: The first lemmatizer for Ancient Greek inscriptions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5334–5344, Marseille, France. European Language Resources Association.

Michael Hahsler, Matthew Piekenbrock, and Derek Doran. 2019. dbscan: Fast density-based clustering with r. *Journal of Statistical Software*, 91(1):1–30.

Alexander V. Isačenko, Henrik Birnbaum, L'ubomír Ďurovič, and Eva Salnikow-Ritter. 1980. *Geschichte der russischen Sprache: Bd. Von den Anfängen bis zum Ende des 17. Jahrhunderts [History of the Russian language: the volume from the beginning to the end of the 17th century]*. Geschichte der russischen Sprache [History of the Russian language]. Winter.

Valentin L. Janin and Andrej A. Zaliznjak. 2000. *Novgorodskie gramoty na bereste (iz raskopok 1990–1996 gg.). Paleografija berestjanyh gramot i ih vnestratigraficheskoe datirovanie. [Novgorod birchbark letters (found in 1990-1996). Birchbark letter paleography ad non-stratigraphic dating]*. Russkije slovari [Russian dictionaries].

Matthew A. Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Ian T. Jolliffe and Jorge Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202. Publisher: Royal Society.

Alexei S. Kassian, Mikhail Zhivlov, George Starostin, A. A. Trofimov, Petr A. Kocharov, Anna Kuritsyna, and Mikhail N. Sayenko. 2021. Rapid radiation of the inner indo-european languages: an advanced approach to indo-european lexicostatistics. *Linguistics*, 59(4):949–979.

Mike Kestemont. 2014. Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66, Gothenburg, Sweden. Association for Computational Linguistics.

Vadim B. Krys'ko. 1998. Drevnij novgorodsko-pskovskij dialekt na obshheslavjanskom fone [old novgorodian/pskovian dialect among the slavic languages]. *Voprosy jazykoznanija [Topics in the study of language]*, 3:74–93.

Kyongjoon Kwon. 2016. Reanimating voices from the past: an alternative reading of Novgorod Birch Bark Letter N370. *Russian Linguistics*, 40(1):79–102.

Elena Je. Lebedeva. 2003. Jelement nasilija v bytovom povedenii novgorodcev XI-XV vv. (po materialam novgorodskih berestjanyh gramot) [element of violence in the everyday behaviour of Novgorodians XI-XV centuries (on the material of the birchbark letters)]. *Novgorod i Novgorodskaja zemlja. Istorija i arheologija [Novgorod and Novgorod land. History and archaeology]*, 17:240–253.

Olga Lyashevskaya and Ilia Afanasev. 2021. An HMM-based PoS Tagger for Old Church Slavonic. *Journal of Linguistics/Jazykovedný casopis*, 72(2):556–567.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Sebastian Nehrdich and Oliver Hellwig. 2022. Accurate dependency parsing and tagging of Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 20–25, Marseille, France. European Language Resources Association.

John Nerbonne, Sandrien van Ommen, Charlotte Gooskens, and Martijn Wieling. 2013. Measuring socially motivated pronunciation differences. In Borin, Lars and Anju Saxena, editors, *Approaches to Measuring Linguistic Differences*, number 265 in Trends in Linguistics. Studies and Monographs, pages 107–140. Walter De Gruyter GmbH.

Ricardo Otheguy and Nancy Stern. 2011. On so-called Spanglish. *International Journal of Bilingualism*, 15(1):85–100.

Irina Podtergera. 2021. German, Latin, and Church Slavonic in the language and text of the Smolensk trade treaty of 1229 [in Russian]. *Russkij jazyk v nauchnom osveshhenii*, 1(41):226–276.

Jelena Prokić and Stephan Moran. 2013. Black box approaches to genealogical classification and their shortcomings. In Lars Borin and Anju Saxena, editors, *Approaches to Measuring Linguistic Differences*, number 265 in Trends in Linguistics. Studies and Monographs, pages 429–446. Walter De Gruyter GmbH.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1948.tb01338.x.

Ryan S.-E. Shim and John Nerbonne. 2022. dialectR: Doing Dialectometry in R. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 20–27.

Robert Sokal and Charles Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas science bulletin*.

Jan Stankievič, Valer Bulhakaŭ (ed.), and Juraś Paciupa (ed.). 2007. *Jazyk i jazykavieda [Language and Language Science]*, 2 edition. Instytut bielarusistyki [Institute of Belarusian Studies], Viĺnia [Viĺnius].

Ryota Suzuki and Hidetoshi Shimodaira. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542.

Colin Swaelens, Ilse De Vos, and Els Lefever. 2023. Evaluating existing lemmatisers on unedited byzantine Greek poetry. In *Proceedings of the Ancient Language Processing Workshop*, pages 111–116, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Thomas Sørensen. 1948. A method of establishing group of equal amplitude in plant sociobiology based on similarity of species content and its application to analyses of the vegetation on danish commons. In *Biologiske Skrifter/Kongelige Danske Videnskabernes Selskab*.

Nicholaas Verhelst. 2020–2021. The Carthaginian *SUFETES*: (re-)assessing the literary, epigraphical, and archaeological sources. *Carthage Studies*, 12:31–80.

Björn Wiemer and Ilja Seržant. 2014. Introduction. In Ilja Seržant, Björn Wiemer, Benedikt Szmrecsanyi, Natalia Levshina, Sofija Pozharickaja, Aksana Erker, Nina Markova, James Lavine, Hakyung Jung, Margje Post, Elena Galinskaja, Mirosław Jankowiak, and Anna Żebrowska, editors, *Contemporary Approaches to Dialectology: the Area of North, Northwest Russian and Belarusian Dialects*, number 12 in Slavica Bergensia, pages 11–80. Department of Foreign Languages, University of Bergen.

Andrej A. Zaliznjak. 1993. Uchastie zhenshhin v drevnerusskoj perepiske na bereste [participation of women in the old russian birchbark correspondence]. In *Russkaja duhovnaja kul'tura [Russian spiritual culture]*, Trento. University of Trento.

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.

Lidija P. Zhukovskaja. 1959. *Novgorodskie berestjanye gramoty [Novgorod birchbark letters]*. Gos. uchebno-pedagog. izd-vo [Scientific-pegagogical state publishing house].

Appendix A

Parameter	Value
vector_size	128
window	15
min_count	1
workers	4
epochs	300
seed	1590
sg	1

Table 4: The parameters for FastText training.

Appendix B

Parameter	Value
nboot	1000
method.dist	euclidean
method.hclust	ward.D2

Table 5: The parameters for bootstrap clusterisation.