Spelling variation and word clusters in the Middle Russian Corpus

Olga Lyashevskaya\textsuperscript{1,2}, Dmitry Sichinava\textsuperscript{2,1}

\textsuperscript{1}National Research University Higher School of Economics,  
\textsuperscript{2}Vinogradov Institute of the Russian Language RAS

In historical corpora, variation in orthography is usually powered by the diversity of regional language variants, varying literacy skills of the authors, different orthographical schools, diversity of genre registers, and sometimes varying impact of a higher standard language (cf. Church Slavic as a high bookish code in Old Russian documents). However, in the Middle Russian Corpus (Gavrilova et al. 2016) there exists yet another source of diversity since the collection includes only published documents rather than archive documents digitalized from scratch. The manuscripts were published by historians, philologists and linguists since the middle of the 19th century until very recently. The editors’ impact on spelling representation varies from more or less exact mirroring of what is seen in the original document (letter-to-letter and line-to-line, the so-called ‘diplomatic’ edition) to graphical and/or orthographical (in terms of Zalizniak 1982) attempts at normalization (what was considered to be ‘correct’ in different periods) and further on to modernization (readers of historical books should be able to understand what is said in old documents).

For example, the earliest Slavic codices, birchbark letters, early parchment documents or inscriptions are typically edited with maximal accuracy with regard to line-final division, choice of graphemes, interpuncts, and superscripts. For the Middle Russian period more advanced stages of this continuum are generally selected. They can vary from almost letter-by-letter rendering only with omitted line-division (like in the editions of Muscovite gramotki, or private letters, of the 17th century) to a heavy modernization eliminating all the letters cancelled by the reform of 1918 (which occurs in some historical and literary editions). Alongside with this, the principles of orthography can vary even within a given series (as this is the case in the Collected Russian Chronicles, published during more than 150 years, or in the Library of the Old Rus’ Literature). A separate problem is whether the text in question is to be published with added punctuation and capitalization (that are used in Slavic medieval manuscripts very rarely). If this option is chosen, the rules of either punctuation or capitalization would naturally follow the lines of the epoch when the edition is prepared and reflect both these rules and the publishers’ linguistic interpretation of the syntax of the text.

We consider such a ‘publishing history’ an important piece of information in the corpus which should not be neglected as it reflects a valuable external context, how the document was introduced to the community and perceived in science and culture during the past two centuries. The data on publishing source is preserved in meta-annotation layer and is made visible to the users of the corpus.

An example can be quoted from the description of the text called Chin svadebnyj ‘The Marriage Rite’ of the 16th century. It is reproduced in the RNC using the edition in the Library of the Old Rus’ Literature, edited by Dmitri Likhachev and his colleagues in 2000. The textological comment to this text from the edition is quoted in the metatextual tagging of the RNC in full. They state that Chin is published by a MS. from the Russian State Library, but the variants and revisions are cited by two previous editions dating back from 1881/1882 and 1916. The text is represented in an orthography with only two additional pre-1918 letters – yat and yer – and with modernized punctuation and capitalization.
At the same time, we do not attempt to restore back the spelling sequences of the original manuscript and work with words just as they were filtered by publishers.

That said, it is evident that spelling variation of different kinds becomes a challenge to corpus search design, creation of word indexes, dictionaries based on the corpus and so on. Integration of the corpus and historical dictionaries or other historical corpora of Russian is also problematic since it is not immediately clear what the lemma is in terms of spelling representation. For example, in the index to Zalizniak’s *Old Novgorod dialect* (Zalizniak 2004) the lemmata in the birchbark letters are cited in the later spelling (without reduced vowels), whereas different historical dictionaries use different decisions regarding this alternative. In order to take the first steps toward developing a linked knowledge platform for Old Russian language resources we constructed a network of word forms represented in the Middle Russian corpus and *Dictionary 11–17* (Slovar 1975–). Our aim is to reveal word clusters which are either different grammatical forms of the same lemma or different spelling representations of the grammatical form. If such a cluster is recognized correctly, we can apply various techniques to define spelling representation of the lemma and lemma alignment.

Firstly, we calculate similarity between word forms in terms of weighted Levenshtein’s distance using publisher as an encoding feature in a noisy channel model (Brill, Moore 2000) that increases edit costs. We argue that the accuracy of word clustering improves on our data compared to the model in which the factor of publisher is excluded. Secondly, we suggest a metric to estimate orthographic consistency (OC) within and between subcorpora of certain publishers. Within a corpus, the OC index is calculated as an aggregated log-frequency divided by the number of words in the cluster and then averaged over top-N frequent clusters. We also discuss two possible approaches to measure the OC difference between corpora.

**References**


*Slovar’ russkogo jazyka XI—XVII vekov*. [A Dictionary of the Russian language of the 11th-17th centuries]. Moscow, 1975–. [In Russian]


Zalizniak A. A. *Drevnenovgorodskii dialect* [Old novgorod dialect]. Moscow, 2004. [In Russian]