A panchronic corpus of Old East Slavic and Russian : bringing together Slavic historical and modern corpus resources

A panchronic corpus is a resource representing texts of multiple different historical periods of a given language or a branch of a language group. Typically, most large-scale diachronic corpora capture a given language only within a single historical period (“Old”, “Middle” or “Modern” lect), which is the case, for example, with the COHA language of American English and the GRAC corpus of Ukraininan (both, roughly, encompassing the period of 1820s-2020s). Families of historical corpora may be further divided by centuries, which is the case with historical corpora of Polish (see eg. https://spxvi.edu.pl/ for the 16th century, https://sxvii.pl/ for the 17th century, https://korba.edu.pl/ for 1600-1772).
On the other hand, panchronic corpora are essentially large-scale diachronic corpora encompassing the bulk of the known history of the lect in question. A good example of a panchronic corpus is the *Frantext* database (https://www.frantext.fr/) that includes the texts from the whole written history of French starting from the 9th to the 21st century. A useful tool, it allows for building queries for Old French, Middle French and Modern French alike, but its lemmatization and annotation heavily depends on the modern orthographic and grammatical standard and is far from being accurate even with high-frequency tokens. Another panchronic resource is *Corpus Corporum* of the Zurich University (Roelli 2014), representing different stages of Latin and built as merger of different pre-existing Latin databases.
Panchronic corpora can be used for statistical study of lingustic *phénomènes de longue durée* on different levels, including orthography, morphosyntax, grammaticalized constructions, and semantics. It is of particular use in studying the so-called submerged phenomena (see e. g. for Latin: Adams, Vincent eds. 2016) that are not reflected in written sources during a large timespan but are shared by earliest and latest attestations of the language.
The paper presents the experience in bringing together the existing corpus resources within the Russian National corpus for Old East Slavic (a common ancestor of Russian, Ukrainian and Belarusian), Middle Russian, and Modern Russian, as well as a separate corpus of Old East Slavic birchbark letters. This unified resource is now searchable as the Panchronic corpus within the Russian national corpus. The source corpora had been annotated using different morphological tagsets and lemma standards stemming to different historical dictionaries. Main issues in bringing together these resources are related to mapping correspondences between the Old East Slavic phonetic rendering of lemmas prior to the loss of the short vowels known as yers (ъ and ь), and, further, between Middle and Modern Russian, using rule-based and neural network algorithms with manual post-correction. General phenomena of historical linguistics such as split and merger of different lemmas due to phonetic changes and semantic divergence are discussed within this context. The issue of unified annotation of changing and emerging grammar is also to be addressed, particularly within the context of grammaticalization of East Slavic aspect and animacy.
The panchronic corpus within the RNC is also annotated by semantic classes, using Modern Russian cognates; parallels between this solution and the approach in the Historical thesaurus of English (Kay et al. eds. 2009) are discussed in the talk. As many words changed their semantics drastically this approach has inherent setbacks and should be used with caution but they can be compensated by gains in research availability for the majority of lexicon. This is illustrated in the talk by an example of searching within the panchronic corpus of “lists of sins” (a literary tradition known both in literature and vernacular birchbark writing) using a simple semantic query of three abstract nouns with negative connotation in a row, that yields relevant results in Old East Slavic, Middle Russian, and Early Modern Russian tiers alike.
The technology can be further applied to both Ukrainian and Belarusian (building of a comprehensively annotated Old Ukrainan / Old Belarusian / Ruthenian corpus being a prerequisite) as well, and also to other Slavic languages.

References

J. N. Adams, N. Vincent (eds.) 2016. *Early and Late Latin Continuity or Change?* Cambridge, Cambridge University press
C. Kay et al. (eds.) 2009. *Historical Thesaurus of English.* Glasgow : Glasgow University.
Ph. Roelli, 2014. The Corpus Corporum, a new open Latin text repository and tool. *Archivum Latinitatis Medii Aevi* 72, 2014