

Title: Etymological Markup of the Birchbark Letter Corpus
Author: Andrew Dombrowski, University of Chicago

The digital humanities community has given considerable attention to the use of digital methodologies to make data accessible to a wide audience, but somewhat less attention has been given to the project-internal use of digital methodologies to facilitate the analysis of data bearing on specific research questions. This presentation describes one such project, in which the attested lexicon of the medieval Novgorod birchbark letter corpus (BBL) is marked up with information linking the orthography of each token with information about the Late Common Slavic-era etymology of each segment attested in the token (or etymologically present, but not realized), resulting in data indicating the degree to which orthography and etymology coincide. This information can then be linked to other metadata pertaining to the token, including, but not limited to, the date of the document, the location where it was found, the genre of the document, and information (gender, social status, etc.) about the sender or recipient of the letter, when identifiable. The same approach can be used for inflectional morphology. This allows for a more nuanced, granular, and sociolinguistically-sensitive approach to the historical phonology and morphology of the Novgorod dialect, and is a technique that could potentially be applied to other diachronic investigations of reasonably small corpora. The presentation focuses primarily on the mechanics of the XML markup of the corpora, and in particular on how to translate traditional research objectives into a digital framework. Consideration is also given to how these techniques could be applied in pedagogical contexts.

adombrow@uchicago.edu