МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА НА РЕПУБЛИКА БЪЛГАРИЯ
КИРИЛО-МЕТОДИЕВСКИ НАУЧЕН ЦЕНТЪР ПРИ БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. М. Т. КАЛАШНИКОВА
НАУЧНОЕ СООБЩЕСТВО “ПИСЬМЕННОЕ НАСЛЕДИЕ”
DIGITAL MEDIEVALIST SCHOLARLY COMMUNITY
ФОНДАЦИЯ „УСТОЙЧИВО РАЗВИТИЕ НА БЪЛГАРИЯ “

Писменото наследство и информационните технологии

El’Manuscript–2014

Материали от V международна научна конференция
Варна, 15–20 септември 2014 г.

София · Ижевск
2014

Сборникът е издаден с финансовата подкрепа на Министерството на образованието и науката на Република България по процедура за подкрепа на международни научни форуми.

Отговорни редактори: проф. дфн В. А. Баранов
доц. д-р В. Желязкова
д-р А. М. Лаврентьев

Редактори: Нели Ганчева, Веселка Желязкова (български текст)
О. В. Зуга, В. А. Баранов (руски текст)
Кевин Хокинс (Kevin Hawkins) (английски текст)

Писменото наследство и информационните технологии [Текст] : материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. — София ; Ижевск, 2014. — 448 с.

Сборникът съдържа материали от конференция, посветена на разработването и създаването на съвременни средства за съхраняване, описване, обработка, анализ и публикуване на ръкописни и старопечатни книжовни паметници и исторически извори, а също и на въпросите за подготвянето на електронни ресурси в областта на хуманитаристиката и тяхното използване в научните изследвания и преподаването.

ISBN 978–954–9787–25–2

Collating the Rus′ Primary Chronicle (Povest′ vremennyx let)

David J. Birnbaum

Collation, alignment, chronicle, edition

The automated alignment and collation of manuscript variants is complicated by the fact that not all variation is philologically significant. This report describes an approach to preprocessing diplomatic (character-by-character, orthographically detailed) manuscript transcriptions to permit them to serve more effectively as input into an automated collation process[1].

The problem

Alignment and collation of variants in manuscript transmission is a computationally complex problem to which many solutions have been proposed in the theoretical literature (Schmidt, 2009) and implemented in practice (Juxta; CollateX; CTE). The issues are similar to alignment problems in other disciplines (such as biological sequencing), but the philological context imposes its own requirements, including the distinction between philologically *significant* and *insignificant variation*. This presentation reaches back to the early twentieth century to provide a context for understanding those distinctions, which are a crucial preprocessing requirement for real-world text-collation problems. This paper does not discuss the alignment and collation process itself; the focus is on preprocessing the data in a way that makes it more tractable as input to subsequent alignment and collation, and on postprocessing the output of the collation to correct for errors that could not be obviated through preprocessing.

Significant and insignificant variation

Greek and Latin ancient and medieval manuscripts were created at a time when orthography was not regulated by standard dictionaries, which means that the understanding of *correct writing* ('orthography') was different from the way that concept is understood today. Unlike in, for example, the scribal tradition of the Hebrew Torah, where even known textual errors must be copied and reproduced exactly, ancient and medieval scribes sought to write *correctly*, which means that they felt equally free to reproduce or amend their sources as they produced new copies of works (Lunt, 1949). *Text-critical scholarship* (Maas, 1960), which seeks to reconstruct the transmission of a text through copying (a "descent with modification" analogous in some ways to biological evolution, but with important structural differences), is concerned with identifying *significant* patterns of variation, which means that the modern editor must distinguish orthographic variation that matters for collation purposes from variation that does not matter. This requirement means that raw string matching is overly crude because it would respond to differences that the philologist must ignore. A generic approach to fuzzy string matching

[1] Acknowledgement: Minas Abovyan is a co-developer of this project.

would fail to recognize that the closeness of the match in terms of edit distance is insufficiently nuanced for at least two reasons. First, edit distance of the Levenshtein variety (or similar models) may fail to take into consideration that manuscript transmission has social properties that impose their own requirements for quantifying and evaluating similarity (Birnbaum and Dubin, 2004). Second, matches that are equally close in edit distance may have different philological properties, so that, for example, the replacement of X with Y may be insignificant variation in one position and significant variation in another. In practice, the Classical Greek and Latin traditions (outside such subdisciplines as epigraphy) tend to rely on heavily normalized texts, where the editor is responsible for neutralizing philologically insignificant orthographic variation before undertaking the collation, alignment, and analysis of significant variation. This normalization is common editorial practice whether the collation is then to be performed manually or with computational assistance.

The Slavistic medievalist tradition, on the other hand, relies extensively on non-normalized texts, where transcriptions may retain orthographic variation of the sort that would complicate, or even frustrate, automated collation. For example, in the recent paper text-critical edition of the Rus′ Primary Chronicle (Ostrowski, 2004), parallel diplomatic transcriptions from manuscript witnesses are printed in interlinear collation, and the evaluation of variation in order to construct a hypothetical *alpha* text was based on the editor's mental process of ignoring insignificant variation and evaluating patterns of significant variation. If computational methods are to be used to support the collation, alignment, and analysis of variation using orthographically precise diplomatic transcriptions, which preserve variation that is unneeded from a text-critical perspective (and therefore a practical impediment to it), insignificant variants must be neutralized in a preprocessing stage in a way that does not sacrifice the ability to render the original, diplomatic transcription at the reporting stage.

Soundex

Soundex is an algorithm first developed in the early twentieth century to facilitate locating records of English-language surnames that might be spelled variously (Odell and Strong, 1947). Soundex is thus a specialized form of fuzzy matching, neutralizing orthographic distinctions selectively according to their significance for determining pronunciation. The algorithm has been modified and refined several times, and has been adapted to different languages, but core features include the following: 1) the first letter of the name is retained exactly; 2) non-word-initial vowels and a few other letters are ignored; 3) remaining consonants are conflated according to phonetic features (e.g., all nasal consonants are given the same representation); 4) geminate representations are simplified; and 5) representations are padded or truncated to a uniform length of four characters.

Adapting Soundex to manuscript collation

Early Cyrillic writing turns out to have properties that pose challenges comparable to those that Soundex was designed to address. For example, non-significant orthographic

variation affects vowel letters more often than consonant letters (cf. property #2, above); consonant variation most often affects classes of letters that have phonetic features in common (property #3); and gemination is not significant (property #4). The average length of a word (as spelled in the manuscript) in Old Church Slavonic (OCS) is approximately 5.37 characters (calculated from the Codex Suprasliensis, the longest of the OCS manuscripts), and more variation occurs at the end of the word than at the beginning (largely because the substantially agglutinative structure of Church Slavonic morphology means that lexical information tends to be located toward the beginning of the word and grammatical information toward the end). These correspondences between the properties of early Cyrillic writing and the Soundex algorithm mean that applying simplification of the sort performed by Soundex to early Cyrillic manuscript transcriptions usually leaves enough distinguishing information to enable effective recognition of what philologists would regard as matching and non-matching strings. It thus outperforms raw string matching (which is, effectively, hopeless because of the extreme prevalence of non-significant orthographic variation) as well as linguistically naïve fuzzy matching.

Postprocessing

The collation process itself in our system is delegated to CollateX, which is capable of accepting input structures that associate a raw manuscript word token (without normalization) with a normalized representation (created by preprocessing according to our Soundex-inflected algorithm), performing the collation on the latter, and returning both, so that the eventual output will retain the original, orthographically precise tokens. The CollateX alignment algorithm is not error-free; in particular, it can find exact matches accurately and efficiently, but in the absence of an exact match it does not evaluate and select the closest match (according to, for example, some measure of edit distance or other pairwise comparison of the match candidates), and defers instead to the first in a sequence of non-matching tokens (without regard to edit distance or other measure of similarity or difference). This is a necessary limitation imposed by the computational complexity of the alignment and collation task, where pairwise comparison of all tokens in all witnesses has exponential complexity, and therefore quickly scales up to become computationally intractable. Known alignment algorithms that avoid that complexity, including those incorporated into CollateX, rely on exact matching, and cannot perform comprehensive edit-distance comparisons of all tokens to find the closest non-exact match.

For this reason, we implement a postprocessing routine designed to correct misalignments in the CollateX output. Performing this correction in postprocessing reduces the complexity to a tractable level by limiting the comparisons to a small number of candidates, where the fact that the comparison process is potentially exponentially complex has no adverse practical consequences because the number of comparisons is guaranteed to be small.

Our postprocessing algorithm examines the output of CollateX in situations only where both of the following conditions are met: 1) the Soundex values in an aligned col-

umn vary and 2) there is a gap in an adjacent column. If the Soundex values correspond, we assume that the collation was performed correctly and that no adjustment is required. If the values do not match but there are no adjacent gaps, we assume that we have a *forced match*, a situation where, for example, the alignment of ABC and ADC lets us infer, from the perfect matches of A and C, that B and D are corresponding tokens and should be aligned. In that situation, as well, no adjustment is required.

In situations were both conditions are met, we run two subsequent comparisons of the Soundex normalizations of the tokens, which we use to adjust the alignment. First, we maintain a thesaurus, seeded by collecting forced matches, which we then edited manually, and thesaurus matches are assumed to represent synonyms or other types of variation that may occur across tokens that should be aligned. Second, we perform an edit-distance comparison of the tokens we are examining, and we move a token from the position assigned by CollateX to an adjacent gap if it matches one of the tokens in the gap column more closely than any token in the column to which it was originally assigned. This adjustment is recursive, so that a gap newly created by moving a token is then examined according to this same process.

Preliminary results

Our goal is to take a corpus of manuscript variants and employ computational tools to collate and align the variants. The actual collation and alignment process is not the focus of the present report, which concentrates instead on developing a mechanism for preprocessing the data to prepare it to serve as input into the collation and alignment process, and a mechanism for postprocessing the results to adjust for limitations in the CollateX alignment algorithm. Preliminary results are available at http://pvl. obdurodon.org (the collation itself at http://pvl.obdurodon.org/browser.xhtml and a description of the preprocessing routine at http://pvl.obdurodon.org/doc/manual.html). Development is open source under a Create Commons BY-NC-SA license, with materials available at http://github.com/obdurodon/collateos.

References

Birnbaum, David and David Dubin. “Measuring similarity in the contents of medieval miscellany manuscripts.” Presented at the ninth biennial meeting of the International Federation of Classification Societies, Chicago, IL, July 2004.

CollateX. http://collatex.sourceforge.net/

CTE. Classical text editor. http://cte.oeaw.ac.at/

Juxta. http://www.juxtasoftware.org/

Maas, Paul. *Textkritik.* Fourth edition. B. G. Teubner, Leipzig 1960. (First edition: 1927)

Lunt, Horace G. 1949. “The orthography of eleventh century Russian manuscripts.” Unpublished Columbia University PhD dissertation.

Ostrowski, Donald, ed. 2004. *The* Povest' vremennykh let. *An Interlinear Collation and Paradosis.* Cambridge, MA: Harvard UP.

Schmidt, Desmond. 2009. “Merging Multi-Version Texts: a Generic Solution to the Overlap Problem.” Presented at Balisage: The Markup Conference 2009, Montréal, Canada, August 11–14, 2009. In *Proceedings of Balisage: The Markup Conference 2009. Balisage series on markup technologies*, vol. 3.

Strong, Margaret K. and Earl P. Odell. 1947. *Records management and filing operations.* New York and London: McGraw-Hill.