Thesis Plan

Title: Computer-aided phonological investigation of early Slavonic manuscripts

There are two main reasons why applying computers to the study of early Slavic phonology is a good idea: one is the nature of Slavic itself at the time of its earliest attestations, i.e. a system which sociologically at least is still seen as unitary, but whose unity is rapidly being undermined by very important structural differences in the various dialects' phonological systems. These deep structural differences should manifest themselves across multiple classes of phonological feature, as well as in morphological innovations, so rather than study narrow features of the texts individually, we should aim to account for their systems as a whole, which requires the exhaustive investigation of multiple phonological and morphological features simultaneously.
The second reason is that we now have lots of computer-readable data about many of the relevant texts, in corpora like Dr. Eckhoff's TOROT, and this can be used to massively speed up the kind of annotation that is needed for the kinds of holistic and multi-vector investigations mentioned above.

Because the project includes a significant methodological aspect, an article-based thesis is more appropriate than a monolithic study of one narrow question, so I envisage a suite of four or five articles, one or two on computational and methodological aspects, and then three or four case-studies where I use those methods on actual texts.

The submitted piece is the Introduction and Section 1 of the first of those articles, the aim of which is to provide theoretical justifications for my method of “phonologically annotating” early Slavic texts using Late Common Slavonic reconstructions, as well as to outline the principles behind my Autoreconstructor program, which uses already-existing TOROT annotations to help produce such reconstructions automatically.
To repeat what is said in the Introduction, Section 2 of that article will look at modern deep-learning-based morphological-tagging and lemmatisation as a way to expand the amount of annotated-text consumable by the Autoreconstructor, and then those techniques will be applied in Section 3 to the Codex Assemanianus[1].
In Section 4 I will to extract as much information about Assemanianus' phonological and morphological system as possible with the least amount of effort, by Autoreconstructing it based on these wholly-automatic taggings, and then I'll compare the conclusions thus drawn to what's already in the literature, as a way of testing just how far you can now get with no manual annotations at all.

The case-study articles will be more linguistically serious[2] :

---

1 I have already Cyrillicised Jouko Lindstedt's ASCII-encoded version of this text (far more accurately than the woeful TITUS conversion), and it is accessible on the https://ocstexts.co.uk website, but with an automatic lemmatisation and tagging based on (an improved version of) the outdated method described in Berdičeviskis, Eckhoff & Gavrilova (2016).

2 I would like to apply modern phonological theory to my analysis of texts, but my experience of attending the Graduate Foundations Phonology course last year suggests that generativists don't understand the difference between phonology and morphology (i.e. they think that words being derivationally or morphologically related has an effect on the phonemic identity of the forms; that a word like Ru. моряк contains an /o/ despite never being pronounced with one, or even that нёс contains an /e/, or that Arabic broken-plurals is a synchronically phonological phenomena rather than an obviously morphological one). Things like Optimality Theory simply restate facts and explain nothing at all (e.g. Padgett 2001, the main idea of which, contrast-maximisation, is utterly independent of the Optimality-Theory framework surrounding it), and even some of the more traditional structuralist tenets, according to which there should be no opposition between e.g. OCS /ję/ and /ę/ or /je/ and /e/ (since they are in total complementary distribution with each other), appear to be contradicted by the very history of OCS writing, where devices to disambiguate such pairs are literally spontaneously invented by the scribes (see pg. 9 of the submitted piece).

Topic 1: Consonant-palatalisation, vowel-length, and syllabic-liquids in the earliest East Slavic: to what extent are the extreme structural difference seen today between Russian and southwestern Ukrainian detectable in pre-Mongol Rus'ian texts?
For this I would need to lemmatise, tag, and autoreconstruct (then correct) the 11th century East Slavic works like Ostromir's Gospel, the Archangel'sk Gospel, and the Izborniki 1073 and 1076 (digitised in the Trondheim-Sofia corpus), and ideally also the Galician Gospel of 1144 (possibly necessitating AI transcription-tools, as explored in e.g. Rabus 2019).

Topic 2: The same but for South-Slavic, focusing on classifying the OCS texts along the spectrum from East Bulgarian to Serbian[3]

Topic 3: Comparative study of the development of the nominal declensions in various texts, where the Autoreconstructor's ability to find morphological deviances will be used to comprehensively study the innovations in e.g. i-stem and jo-stem nouns, to see whether systems where we expect more Russian-style secondary consonant palatalisation also expand their jo- and ja-stem classes. (For this some of the later non-canonical texts like the Psalterium Bononiense might be useful.)

Timeline:
The first introductory article should be finished by the end of Hilary 2026, and the section on neural-network tagging/lemmatisation means I will be in a position to tag/lemmatise the early East Slavic texts needed for the first case-study much more quickly. Eckhoff already has some early Old Russian texts (e.g. Uspenskij Sbornik) in TOROT with manual tagging.
I would expect to get the new texts fully processed by mid-summer 2026, and an article written in Michaelmas 2026.

3 My chief interest with Topics 1 and 2 will be in how system-level structural differences arising in the disintegrating Slavic dialects might find expression in spellings, given that we have very early texts which purport to share a literary-medium but are from dialect-areas as diverse as northern Macedonian (verging on Serbian) and northern Russian, which occupy two ends of a spectrum along which Slavic phonological systems can be classified, viz. the extent of the development of phonemic secondary consonant-palatalisation before LCS front-vowels and thus number of hard/soft consonant-pairs which are opposed to each other only by the distinctive-feature of 'tonality'. It can be seen from the phonological systems of modern Ukrainian and Russian that the more hard/soft consonant pairs in the system the greater the importance of these consonantal tonality-distinctions for the organisation of the system as a whole: in Russian all consonants, even unpaired palatals like /č ž j š c/, are 'paradigmatically' hard or soft, meaning they determine the pronunication of surrounding vowel-phonemes and can never themselves be hardened or softened by surrounding vowels, whereas in Ukrainian, where phonemically soft labials and (dialectally) /r'/ have gone, it's the vowels which hold the whip-hand: /c/ is usually soft but hardens before /e/ and /ÿ/, whereas /č ž š/ are usually hard but soften before /i/ and when geminated. Precisely *when* such differences arose in East Slavic is very difficult to determine, because of both the nature of the Cyrillic alphabet and the fact that pre-Mongol Rus', where we expect such differences to have developed, had basically a single written culture.
In Townsend & Janda (1996: 107-8) our attention is brought to a fascinating inverse-relationship in the Slavic dialects between the extent of phonemic tonality-distinctions in consonants on the one hand, and the longer maintenance of pitch-distinctions and distinctive vowel-length on the other. This is linked to earlier loss of jers in central vs. peripheral areas, with far southwestern Ukrainian being the most central and thus most likely to have either lost or not developed secondary consonant palatalisation and to have longer maintained distinctive vowel-length (the best précis of such intra-East Slavic dialect differences remains Trubetzkoy 1924).
My comprehensive phonemic indexing of texts would allow us to look for signs of such correlations, because not only could I easily quantify the extent of the Jer Shift, but I could also quickly look for signs of vowel-length (like loss of inter-vocalic /j/ and contraction in e.g. *aje or *oje groups, in e.g. long-adjectives or the 3sg. pres. verb ending, cf. OCS -ѧѧтъ spellings), and differences in the paradigmatisation of hard/soft consonant oppositions could be probed by studying the letters used after <ш ч ц ж> (do we expect e.g. <ю ѣ а> to be *more* likely as devices for conveying the softness of such palatal-phonemes in systems like Russian or Eastern Bulgarian where this softness is more systematically important than it is in systems like Ukrainian or northern Macedonian/Serbian?). Even things like use of a palatalisation-diacritic to distinguish LCS *ń and *Í before *e might hint at the lack of secondary palatalisation of LCS plain *l *n before *e, as Shevelov (1964: 490) mentions in relation to the Archangel'sk Gospel and the Izbornik 1073. Giving people (me) the ability to quickly verify the philological facts adduced in such arguments in the literature (by searching texts like those for *ńe, *Íe etc. groups) was one of the main motivations behind my whole idea of using LCS reconstructions as phonological annotation.

The Topic 2 article would be much faster because most of the texts are already digitised and lemmatised/tagged, so I think that could be finished by Hilary 2027.

Topic 3 is almost not difficult enough to deserve its own article, given that a side-effect of the Autoreconstructor is to inflection-class-mark every single word of a text, so that one should be doable in the remainder of the 2027 academic year.

I would then spend the summer of 2027 writing the introduction and conclusion of the thesis have something submittable by Michaelmas 2027.

Simultaneously with the more focused article-writing work, I would continue reconstructing and inflection-class-marking OCS and Old Russian[4] lemmas. The recent digitisation of the two OCS dictionaries (the *Slovník nejstarších staroslověnských památek* for canonical OCS and the *Slovník jazyka staroslověnského* which also includes later manuscripts) at http://gorazd.org/gulliver/ means that a full list of canonical OCS lemmas can be extracted (via web-scraping; cf. the dictionary-widget on ocstexts.co.uk which just steals the data of individual entries) and added to my lemma-spreadsheet even before they are encountered in annotated texts, so even as-yet out-of-vocabulary (i.e. not-in-the-training-data) lemmas should be available to automatic lemmatisation methods.

References

Brants, T. (2000). TnT – a statistical part-of-speech tagger. *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000.* Seattle, WA.
Berdičevskis, A., Eckhoff, H., & Gavrilova, T. (2016). The beginning of a beautiful friendship: Rulebased and statistical analysis of Middle Russian. In Computational linguistics and intellectual technologies. Proceedings of Dialogue 16. Moscow.
Rabus, Achim. 2019. Recognizing Handwritten Text in Slavic Manuscripts: a Neural-Network Approach Using Transkribus. Scripta & e-Scripta 19, 9–32.
Rabus, Achim; Besters-Dilger, Juliane. 2021. Neural Morphological Tagging for Slavic: Strengths and Weaknesses. Scripta & e-Scripta 21, 79–92.
Shevelov, George Y. 1956. Konsonanten vor e, i in den protoukrainischen Dialekten. In Bräuer, Herbert & Woltner, Margarete (eds.), Festschrift für Max Vasmer zum 70. Geburtstag, 482–494. Wiesbaden.
Townsend, Charles E. & Janda, Laura A. 1996. *Common and comparative Slavic: Phonology and inflection. With special attention to Russian, Polish, Czech, Serbo-Croatian, Bulgarian*. Columbus (Ohio).
Trubetzkoy, N. 1924. Einiges über die russische Lautentwicklung und die Auflösung der gemeinrussischen Spracheinheit. Zeitschrift für slavische Philologie 1(3–4). 287–319.

4 Most of the texts in the Old Russian part of TOROT are too late to be sensibly autoreconstructed (the main exceptions being Uspenskij Sbornik, the First Novgorod Chronicle, and Russkaja Pravda). The quality of the manual-tagging is also noticeably worse than that of the OCS texts.