Thesis Plan

The submitted piece is the Introduction and Section 1 of an envisaged larger article, 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 alread-existing TOROT annotations to help produce such reconstructions automatically. While I think that the programming and data-processing challenges involved in making the Autoreconstructor are methodologically interesting in themselves, the real goal is the correct and complete phonological indexing of the texts and then the carrying out of comprehensive investigations into their orthographic and underlying synchronic phonological systems.

Since TOROT is missing a good chunk of both canonical OCS and the earliest East Slavic copies of South Slavic originals, my immediate aim is to expand the amount of text that the Autoreconstructor can work on by:

1.) applying modern statistical / machine-learning methods of automatic tagging and lemmatisation (using the manually-tagged corpus as training-data) to already-digitised texts, such as the Codex Assemanianus and the Savvina Kniga (digitised in ASCII format by Jouko Lindstedt as part of the *Corpus Cyrillo-Methodianum Helsingiense*, https://www.kielipankki.fi/download/ccmh-src/www/index.html), and 11th century East Slavic works like the Ostromir’s Gospel, Archangel’sk Gospel, and the Izborniki 1073 and 1076, digitised as part of the Trondheim-Sofia corpus, https://www.hf.ntnu.no/SofiaTrondheimCorpus/index2.html). The Assemanianus (freshly Cyrillicised from the ASCII file far more accurately than the woeful TITUS conversion) is already included on the ocstexts.co.uk website with automatic-tagging and an Autoreconstruction based on said tagging, but for that I used (with some fixes and improvements) the outdated method described in Berdičeviskis, Eckhoff & Gavrilova (2016), based on the Trigrams’n’Tags statistical-tagger (Brants 2000). I would like to instead apply modern neural-network methods, similar to those used in e.g. Besters-Dilger & Rabus (2021), to get the automatic tagging and lemmatisation, and thus autoreconstructions, as accurate as possible and to lessen the amount of manual-correction needed. Separately I think it'd be an interesting standalone-project to see just how much of such a text can be accurately autoreconstructed wholly automatically in this way (i.e. what I envisage for Sections 3 and and 4 of the article).
2.) for interesting texts which haven’t even been digitised, such as (as far as I know) the Galician Gospel of 1144, I’d like to use modern neural-network-based handwritten-text-recognition tools like Transkribus (as explored in e.g. Rabus 2019) either on the manuscript or on printed editions, to speed up the production of accurate digisations that can then be fed into the automatic tagging and reconstruction tools mentioned in 1.).
   I think automatic transcription would also be useful for aligning already-digitised texts with their manuscript-layout, not only so that the links to manuscript-page images can be provided for texts in the web-interface, but so that the digital text itself can optionally be laid out like it is in the manuscript, similar to this recent online edition of Psalterium Sinaiticum: https://www.punco.slavistik.lmu.de/psalter_view.php?single=true&cyr=1&verse=150-1.

Alongside work to expand the annotated corpus of text, I would continue reconstructing and inflection-class-marking OCS and Old Russian 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 dictionarywidget 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.

I would expect the corpus-expanding and -tagging work outlined above to take up most of this year, and then in the final year I would turn to actual investigations of the texts:

My chief interest here 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 early East Slavic 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 -аатъ spellings), and differences in the paradigmatisation of hard/soft consonant oppositions could be probed by studying the letters used after <ш ч ц ж> graphemes (do we expect e.g. <ю ѭ ѧ> letters to be *more* likely as devices for conveying the softness of such palatal-phonemes in systems like Russian where this softness is more systematically important than it is in systems like Ukrainian?). Even things like use of the palatalisation-diacritic to distinguish LCS *ń and *ĺ before *e can 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, *ĺe etc. groups) was one of the main motivations behind my whole idea of phonological-annotations for texts.

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/); 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).

References

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.