Necromancing Diels: computerising the phonological analysis of early Slavonic texts using existing treebank data and a Late Common Slavonic computerised inflectional morphology

0. Introduction

Much progress has been made in the last twenty years in early Slavonic corpus linguistics as a result of the Old Church Slavonic part of the PROIEL project (REFERENCE) and its subsequent expansion as the TOROT treebank (Eckhoff & Berdičevskis 2015), such that currently just over 240,000 words of canonical OCS have been manually lemmatised, part-of-speech and morphologically-tagged, and syntactically parsed. The focus of these projects, however, has been exclusively on the higher-level linguistic domains of syntax, semantics, and pragmatics: surface-morphology has been of only incidental concern, for example in investigations into differential-object marking (Eckhoff 2015, 2022). No inflection-class data is included in these corpora, and phonology has been totally ignored to the extent that some of the texts (esp. Kiev Folia, Codex Suprasliensis, and partially Codex Zographensis) contain quite severe typographical inconsistencies and errors that make them dangerous to use without reference to the manuscripts.

That being said, enough information is included in the lemmatisation and morphology-tagging that, with a few exceptions (e.g. comparatives), the morphological shape of the inflected text-forms can be predicted from just the tag-information, provided that inflection-class annotations are added to the lemmas. This means that the immediate Late Common Slavonic ancestors of surface-text forms can be generated by using a database of LCS inflectional-endings, reconstructing and inflection-class-marking the LCS stems of the lemmas, and then applying inflectional-endings to the stems according to the word's morphology-tag annotation[1]. Such LCS reconstructions are an extremely useful form of 'phonological annotation', since theoretically all the information required to give rise to an attested form must be present in any correct reconstructed proto-form, and the complete regularity of the idealised LCS forms makes texts predictably searchable regardless of orthographic variability, abbreviations, or other irregularities in the surface-texts. When applied to whole texts, they make the exhaustive investigation of almost any phonological or orthographic question trivially easy compared to manually reading and extracting relevant forms, or using TOROT's existing lemmatisation and morphology-tagging to try to gather morphological categories which might contain the sound-groups one is interested in.

In the next section I will describe my computerised LCS inflectional-morphology in more detail, show how it can be used to "autoreconstruct" different OCS texts, and explain how difficulties caused by things like morphological innovations, badly-integrated foreign loanwords, or insufficiently-precise tagging-data can be overcome. (Possibly include here some demonstration of 'exhaustive investigation' of the autoreconstructed Marianus, since that is the highest-quality TOROT text and the only one virtually 100% covered by my lemmas?)

Since morphology-tagging and lemmatisation are a prerequisite for my method of automatic reconstruction, Section 2 will survey recent work on automating these tasks for early Slavonic texts. Thanks to modern deep-learning techniques and the large and growing amount of manually-produced training-data in Eckhoff's corpus, accuracies of 90%+ can easily be reached (depending on the target-text), and I will see how far up this can be pushed by better neural-network design and more careful and informed pre-processing of training and target-data.

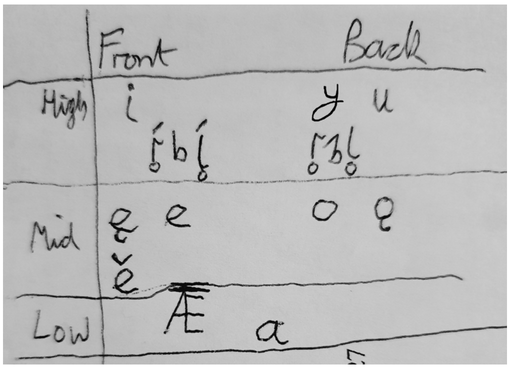
As a test-case of "wholly automatic" phonological annotation, Section 3 will apply such methods to the Codex Assemanianus, an OCS lectionary containing most of the gospels which has been digitised in an ASCII-encoded format by Jouko Lindstedt but is not included in Eckhoff's corpus. Accuracy will be evaluated by comparing both the automatic tagging and lemmatisation, and the resulting LCS reconstructions, to 10 randomly-selected manually-annotated shorter sections.

[1] Morphological innovations and variations are detected by inspecting the text-forms and then applying 'alternative' endings as specified in the inflectional-endings database; see Section 1 for more detail.

Section 4 will then use the wholly-automatically-reconstructed Assemanianus as the basis for a short investigation into aspects of its phonological and orthographic system, which will be compared against existing treatments of this text in the literature, to see to what extent useful insights can be extracted even without any form of manual-annotation.

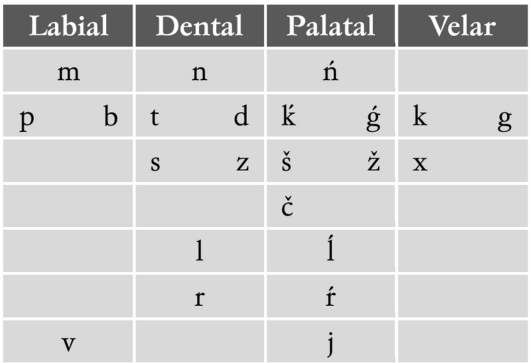
1. Auto-reconstructing texts using a computerised Late Common Slavic inflectional morphology

The premise of my chosen form of "phonological annotation" is that the earliest Slavic texts reflect languages which are **structurally** close enough to the broadly-agreed-upon system of Late Common Slavonic that the forms underlying the manuscript-spellings are more or less trivially derivable (by the application of sound-change rules) from their theoretical LCS ancestors. In order to account for as much of the subsequently attested Slavic as possible, a point after the monophthongisation of diphthongs, but before the Second and Third Velar Palatalisations (PV2 and PV3) is chosen as the point of departure, since differences between the West Slavic /š/ and South/East /ś/ reflex of these two palatalisations of *x (Cz. loc. pl. *duších* <*duxěxъ, Prague Framents вш҃ѣχ' <*vьxěxъ) as well as the probable complete absence of both in the far North of East Slavic (Old Novgorodian nt. acc. sg. воχо *<vьxo, masc. loc. sg. сѣньнıкѣ <*sěnьnikě), prevent us from calling these changes "Common Slavonic".
To be explicit, the native phonemes in my LCS system are given in the tables below:



	Front	Back
High	i ŕ̥ ь ĺ̥	y u r̥ ъ l̥
Mid	ę e ě	o ǫ
Low	Æ a	

Figure 1: LCS vowels after monophthongisation of diphthongs



Labial		Dental		Palatal		Velar	
m		n		ń			
p	b	t	d	ḱ	ǵ	k	g
		s	z	š	ž	x	
				č			
		l		ĺ			
		r		ŕ			
v				j			

Figure 2: LCS consonants before PV2 (from Winslow 2022:304)

In addition, the following symbols are used to represent phonemes of wholly foreign origin in order to represent badly-integrated foreign borrowings, whose level of integration into the native system we deliberately do not take a position on: /ḱ ǵ x́ f ü/, e.g. in respectively к҄итъ <*ḱitъ ıꙉемонъ <*iǵemonъ χитонъ <*x́itonъ иосифъ <*ijosifъ and мѵро <*müro. Almost none of the words containing these symbols would actually have existed in the language during Common Slavonic times, but they need to be included in the indexing-system because they often contain native Slavic elements (f.ex. inflectional endings). Normally they represent specific sounds in the source-language (usually Greek), so including them is useful for investigating the process of these sounds' integration into the native systems. For instance, the variation between adapting Greek /ü/ to native /i/ or /u/ can be seen in variations in the OCS spellings of the word for 'Egypt': еꙉюпта vs егıпта vs егѵптъ vs егоѵптъ vs е҃ꙉѵпт҃а[2], etc.. One might also ask whether the replacement of a separate <ꙉ> letter for /ǵ/ (and the writing of <к҄> with the palatalisation-diacritic) could be linked to the inadmissibility in the native systems of soft [k^j, g^j] sounds, and their replacement with regular <г, к> more likely in systems with some level of native [k^j, g^j] (for instance, in Rus' after the so-

[2] NB apart from the form from Suprasliensis these should all be in Glagolitic, but currently everything in my database is Cyrillic. Cyrillic transcription of Glagolitic is and always has been a bad idea for which Jagic' should be punished

called Fourth Velar Palatalisation, or in Novgorod due to the retention of native velars before front-vowels because of the non-action of PV2, etc.); such questions are far easier to investigate if all relevant forms can be reliably retrieved by giving them even a consciously artificial LCS representation.

I have neglected to include accentual information in my reconstruction of vowels, even though such information is in fact required to explain certain differing manuscript-reflexes, e.g. Russkaja Pravda fem. acc. sg. робꙋ < *orb-ǫ vs Uspenskij Sbornik nt. acc. sg. рало < *ordl-o, the second of which is listed in Derksen's (2007) Etymological Dictionary as *òrdlo and accent-paradigm (a) (no accentual information is given for the first one). For too large a proportion of the vocabulary this information is not securely or uncontroversially reconstructed enough to justify its inclusion, and anyway the (often post-LCS) derivational processes which are responsible for most of the actual words in an OCS or Old Russian texts complicate things even further (that робꙋ example is actually not even listed in Derksen's dictionary, since he actually lists only the masculine *orbъ, and I don't even know where I would go to find the accentuation of the feminine-derivative *orba, let alone a more complicated derivative of this root like in Russkaja Pravda 3rd pres. pl. verb робѡтѧть.)

The syllabic liquids /ŕ̥ ĺ̥ r̥ l̥/ are included as unitary vocalic phonemes, rather than as combinations of /ь ъ/ + /ŕ ĺ r l/, because these groups descend from PIE syllabic liquids and many descendant Slavic dialects which retain syllabic liquids in this position do not show any evidence of an intervening oral-vowel + liquid stage (such a view is shared by Bethin 1991: 71-72; cf. also Bulgarian dialectal evidence in Stojkov 1954: 130-131, where hard consonants precede reflexes of the LCS /ĺ ŕ/ even in dialects with secondarily-palatalised consonants before fallen weak LCS /ь/).

//there needs to be some more bullshit about my LCS system here

The below is all just unstructured crap and random notes related to what I want eventually to talk about
By 'structurally' I am referring to the structure at the phonological level; structural changes at higher levels of analysis (inflectional morphology -> derivational morphology) are of no concern unless they are **only made possible only by intervening phonological changes**.

For example, whether or not there actually existed at the LCS stage a mechanism for deriving secondary-imperfective verbs like OCS разарꙗти <*orzaŕǢti (e.g. Mar. Mark 15 nom. masc. sg. def. pres. act. part. разарѣꙗи) from the prefixed разорити *orzoriti is irrelevant, because LCS /orzaŕǢti/ does not violate the rule of LCS phonotactics: palatal /ŕ/ can be followed by /Ǣ/

An obvious example of morphological change contingent upon structural phonological change, leading to manuscript forms which preclude any construction of their direct LCS-stage ancestors, is the replacement of i-stem endings with those of the corresponding jo- or jā-stems, in nouns whose stems end on labials or the subset of LCS dental consonants which lack palatal counterparts, viz. /d t s z/. Such deviances depend on the development of secondary phonemic palatalisation before LCS front-vowels

Smolensk Treaty татѧ

In the case of Supr. Gsg. masc. ꙁвѣрѣ, for an original i-stem (звѣри <*zvěri), a direct LCS ancestor for the attested form can still be given (*zvěŕǢ), because palatal /ŕ/ already exists in our LCS system, and one plausible explanation for this form is that the Eastern Bulgarian dialect underlying Suprasliensis developed secondary palatalisation of LCS plain *r before front-vowels, which was then phonemicised after the fall of word-final front-jers, and that newly-palatalised /r'/ fell together with original LCS palatal /ŕ/, so that the nom. sg. *zvěrь became /zvěr'/, and its stem now ended on the same consonant /r'/ as original ja- and jo-stems ending on LCS *ŕ like мор҄є and боурꙗ, so it began to be inflected as a jo-stem masculine instead of an i-stem.

Most clearly in Russian, but also in Eastern Bulgarian dialects, the development of and (on the evidence of forms of the word господь) in the dialects of at least some of the antigraphs of some OCS texts,

It should be emphasised that the historical reality of our reconstructions is only of concern at the phonological level, that is, phonemes and phonotactics; the plausibility of higher-level structures built out of these units,