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 (Haug & Jøhndal 2008) 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.
By 'structurally' I am referring to the structure at the phonological level; structural changes at higher levels of analysis (i.e. inflectional morphology, derivational morphology) are of no concern unless they are **only made possible only by intervening phonological changes**.
My contention is that before about 1100 not enough of these structural changes are in evidence in any Slavic text, and thus they can be relatively straightforwardly indexed using a well-chosen LCS system. Before giving examples of structural changes that are problematic for such an indexing-system, it's necessary to first lay out my LCS system in full:

1.1 Late Common Slavonic as a "phonological index"

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, because of the difference between the West Slavic /š/ and South/East /ś/ reflex of these two palatalisations of *x (Cz. loc. pl. *duších* vs Suprasliensis. доусѣхъ <*duxěxъ; Polish *wszak* vs Supr. вьсакъ , Ru. *всяк[ий]* <*vьx-akъ), as well as the probable complete absence of PV2[2] in northern East Slavic (Old Novgorodian, see Zaliznjak 2004: 42-45 for the evidence).
To be explicit, the native phonemes in my LCS system are given in the tables below:

2 The evidence regarding the possible absence of PV3 from Novgorodian is far less convincing: the Birchbark letters abound with examples of the PV3 reflex of *k (e.g. letter №439 from around 1200 has свинеце <*svinьkь and полотенеца <*poltьnьka), and those of *g are not unknown: Zaliznjak (2004: 47) admits that palatalised forms of the Germanic loan кънѧз- <*kъnęg- are the rule, but considers this to be a "supradialectal" word originating outside of the Novgorodian dialect-area; Galinskaja (2014: 10) is less convinced and adduces the form оусьрѧзи 'earrings' from letter №429 as a word of "вполне бытового характера" which thus supposedly shows a native Novgorodian reflex of PV3 of *g. (This is commonly assumed to be a Turkic loan, cognate with e.g. Kazakh *сырға*, but the fact that it appears in Slavic with front-vowels (Ru. *серьга*), unlike its back-voweled Common Turkic cognates, and the fact that it was borrowed early enough to undergo PV3 at all, suggests that Vasmer's derivation of it from "Old Chuvash" (i.e. some form of Oghur or Bulgar Turkic) is correct, and it thus belongs to an earlier layer of Turkic loans than those borrowed from the Kipchak dialects of the Polovtsians (e.g. Ru. *камыш* < *qamış (> Kaz. *қамыс*).)
More importantly, as Galinskaja (op. cit.) points out, in all of the well-known Novgorodian forms of the pronoun *vьxъ 'all' which supposedly show a lack of PV3 by retaining both /x/ and back/hard desinences (e.g. fem. gen. sg. въхоѣ <*vьxoję̌ from letter №850), and which come from letters which otherwise correctly convey the jers (by writing <ь,е> for *ь and <о,ъ> for *ъ), the weak-jer is always written with <ъ,о>, unambiguously suggesting a /ъ/ pronunciation. These forms therefore more likely point to a LCS doublet-form *vъxъ which would never contain the conditioning environment for PV3 anyway, and thus you can't use them as evidence of a lack of PV3 in Novgorodian (on the plausibility of such a doublet see Galinskaja (2014: 14), though cf. Zaliznjak's (2004: 54) less convincing explanation of the /ъ/ in these words as an assimilation of original /ь/ to the back-vowels of the following syllable).

Table 2: LCS Vowels after the monophthongisation of diphthongs

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

Table 1: LCS consonants before PV2/PV3 (adapted from Winslow 2022: 304)

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

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 ки҃тъ <*ḱitъ, ꙇꙉемонъ <*iǵemonъ, хитонъ <*x́itonъ, иосифъ <*ijosifъ[3], 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': ⰵⰼⱛⱂⱅⱏ vs ⰵⰳⱛⱂⱅⱏ vs ⰵⰳⱆⱂⱅⱐ vs егѵптъ vs ⰵⰼⱆⱂⱅⱏ[4], etc.. One might also ask whether a separate <ⰼ> letter for /ǵ/ (and the writing of <ⰽ> with the palatalisation-diacritic) could be linked to the inadmissibility in the native systems of soft [kʲ, gʲ] sounds, and whether their replacement with regular <ⰳ,ⰽ> or <г, к> was more likely in systems with some level of native [kʲ, gʲ] (for instance, in Rus' after the so-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 deliberately not included 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, because 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 the attested texts (and the inevitable accentual levelling processes likely to have occurred in the course of these derivations) complicate things even further.

The two extra nasal-vowels /y̨/ and /ę̌/ are required to account for the split between North (East and West) and South Slavic forms of certain inflectional-endings: *y̨ for the nom. sg. masc./nt. pres. act. participle of certain verb-classes whose present-stem ends on a hard-consonant, which in South Slavic remains high and backed, e.g. Supr. ꙁовꙑ <*zovy̨, Psalterium Sinaiticum ⰸⱁⰲⰴⰳⰵⱘ

3 Of course the sequence /jo/ violates LCS phonotactics as well.

4 Forms are given as they appear in the manuscripts; modern fonts mean that the misleading and unhelpful practice of transcribing Glagolitic into Cyrillic is no longer justified in any context.

<*stergy-jь, Codex Marianus ⱑⰴⱔⰺ <*jǢdy-jь (as is clear from these forms, early OCS retained the nasal character of this vowel and may even have developed the special "hooked" nasal letter <ⱔ> for it), but in North Slavic lowered to /a/: Old Polish (Kazania Świętokrzyskie) *reca* /r'eka/, Ru.Ch.Sl. (Vita Methodii) вьсемогди <*vьxemogy-jь. /ę̌/ is responsible for the NSl. /ě/ vs SSl. /ę/ shapes of jo-stem masc. acc. pl. and the ja-stem nom./acc. pl. and gen. sg. endings, which are reflected in respectively the post- and pre-revolutionary spellings of the Russian nom./acc. pl. long-adjective endings -*ые* < *yjě < *yję̌ vs -*ыя* < *yja < *yję < *yję̌. (I know that Russian adjectives are a bad example but it's also very space-efficient) (See Kortlandt 1979).

The need for the retention of the /Ǣ/ archiphoneme up to this point of LCS, which represents merged Early Common Slavonic *ē *ā in the position after palatal consonants, is explored in detail in Winslow (2022), but the same archiphoneme (along with its short counterpart /Æ/) was explicitly posited by Kortlandt as far back as 1979 (p.266) as part of his ECS system.

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 (including those underlying canonical OCS) 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/). I know I should follow your advice here but frankly the evidence for *ьl, *ьr etc. is so lacking that I don't know why, apart from out of orthographic convention, anyone would ever suggest them at all
//this exploration of my LCS system is still nowhere near finished

An example of morphological change contingent upon structural phonological change, leading to manuscript forms which preclude any valid reconstruction 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/. Evidence for such a change is furnished by the Old Russian masc gen./acc. form тата from the 1229 Treaty between Smolensk, Riga and Gotland (Version A). LCS *tatь is a masc. i-stem noun with genitive *tati, as it still appears in the Codex Suprasliensis translation of John Chrysostom's Homily for Holy Thursday (…то кажетъ владꙑкꙑ чловѣколюбьѥ· ꙗко прѣданника разбоиника тати…), but in the dialect underlying the 1229 Treaty the rise of phonemically palatalised /t'/ after the Jer Shift means that the stem (and the nom. sg. тать /tat'/) of this noun now ends on the same class of "soft" consonants as original jo-stem nouns like *pastyŕь > /pastyr'/, where the original LCS palatal *ŕ has fallen together with secondarily-palatalised /r'/ from plain LCS *r before LCS front-vowels, in e.g. the original i-stem *zvěrь > /zvěr'/. This system thus no longer distinguishes between descendants of the original LCS palatals and the newly secondarily-palatalised consonants like /t'/, both are now together in the set of 'soft' consonants, opposed to their 'plain' or 'hard' counterparts, and so tend towards taking the same set of inflectional enndings (in this case those of the original jo-stems). Consequently, a word like тать has begun to take jo-stem endings, including the Old Russian /a/ reflex of LCS *Ǣ in the gen. sg.
Problem is the LCS /Ǣ/ by definition can only occur after LCS palatal consonants (see below), so a reconstruction *tatǢ is just nonsensical. In the case of the dat. sg. /u/-desinence (which isn't attested in our Treaty but it exists in modern Russian *татю*), we don't even have an LCS archiphoneme available to signal a preceding soft-consonant; there's simply no way of getting from LCS *tatu to Russian /tat'u/, because such a form was only made possible by the rise of phonemic /t'/, so our ability to index it with our LCS system is gone.

For example, if the phonotactic rules of our theoretical LCS system allow the sequence /ŕǢ/ (palatal /ŕ/ < *rj + the archiphoneme /Ǣ/) to occur, then a morphological change which replaces the sequence /ri/ with /ŕǢ/ is of no concern, because both are equally valid LCS. If, however, the same type of morphological change were to

For example, whether or not there actually existed at the LCS stage a mechanism for deriving secondary-imperfective verbs like OCS разарꙗти <*orzaŕǢti 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/ because such a combination exists in the paradigms of wholly securely reconstructable jo-stem nouns, e.g. nt. gen. sg. *moŕǢ (> Pol. morza, OCS морꙗ, Ru. моря, etc.)

In the case of Supr. Gsg. masc. ѕвѣрѣ, 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 морє 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,