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 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.

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 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 **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 texts 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. Ⱄⱁⱆⱍⱏⱈⱏ <*duxěxъ; Polish *wszak* vs Supr. ⰲⱐⱍⱄⰰⰽⱏ , Ru. *вс̱як[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), and the blocking of PV2 by an intervening *v in West Slavic (Pol. *gwiazda*, Cz. *květ* <*gvězda, *květъ, etc.).
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 ⱄⰲⱁⱀⱑⱌⱑ <*svinьkь and ⱂⱁⰾⱁⱅⱑⱀⱑⱌⰰ <*poltьnьka), and those of *g are not unknown: Zaliznjak (2004: 47) admits that palatalised forms of the Germanic loan �best- <*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 ⱁⱛⱄⱐⱃⰰⰵⱁⱀ 'earrings' from letter №429 as a word of "вполне бытового характера" which thus supposedly shows a native Novgorodian reflex of PV3 of *g.
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. ⰲⱏⰅⱁⱏ <*vьxojě from letter №850), and which come from letters which otherwise correctly convey the jers (by writing <ⱐ,ⱑ> for *ь and <ⱁ,ⱏ> for *ъ), the weak-jer is always written with <ⱏ,ⱁ>, 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 к҄итъ <*ḱitъ, ıꙉемонъ <*iǵemonъ, χитонъ <*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 extent to which Greek /ü/ is integrated into either native /i/ or /u/ can be seen in variations in the OCS spellings of the word for 'Egypt' (*eǵüpьtъ): ⰵⰼⱛⱂⱅⱏ vs ⰵⰳⱛⱂⱅⱏ vs ⰵⰳⰹⱂⱅⱐ vs єгоуптъ vs ⰵ̄ⰼⰹⱂⱅ̄ⱏ[4], etc.. One might also ask whether a separate <ⰼ> letter for /ǵ/ (and the writing of <ⰳ҄> 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 *ky,*gy,*xy > [kʲi, gʲi, xʲi], or in Novgorod due to the retention of native velars before front-vowels because of the non-action of PV2, etc.). In any case, such questions are far easier to investigate if all relevant forms can be reliably retrieved by giving them even a consciously artificial LCS representation.

Vowels

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. робѹ < *orb-ǫ vs Uspenskij Sbornik nt. acc. sg. рало < *ordl-o, because for too large a proportion of the vocabulary this information is not sufficiently securely and uncontroversially reconstructed 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

3 Of course the sequence /jo/ violates LCS phonotactics as well.
4 Forms are given as they appear in the manuscripts; modern fonts and Unicode symbols mean that the misleading and unhelpful practice of transcribing Glagolitic into Cyrillic is no longer defensible.

Slavic remains high and backed, e.g. Supr. ⰸⱁⰲⱏⰹ <*zovy̧, Psalterium Sinaiticum ⱄⱅⱐⰳⱁⱗⱖ <*stergy̧-jь, Codex Marianus ⰰⰱⰵⰴⱛ <*jǢdy̧-jь (these forms lead Kortlandt (1979:260) to posit that some dialects of early OCS retained some kind of nasal character in this vowel and may even have developed the special "hooked" nasal letter <ⱗ> for it), but which in most of North Slavic lowered to /a/: Old Polish (Kazania Świętokrzyskie) has both *reca̱* and *recǫ* (with the special Old Polish letter for the merged reflex of *ę and *ǫ) <*reky̧, and in other texts also *biorǫ* <*bery̧; Russkaja Pravda рєка, Uspenskij Sbornik донда <*dojьdy̧, Ru.Ch.Sl. (Vita Methodii) вьсемогди <*vьxemogy̧-jь. Kortlandt's positing of a CS *y̧ (which he writes as *aN) is far from universally accepted, and others consider these forms the result of various dialect-specific analogical process; see references and discussion in Olander (2015: 88-92). Whatever the truth of the matter, our *y̧ is a convenient placeholder which allows all the relevant evidence to be retrieved.
/ę̆/ 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 < *yję̆ vs *-ыя* < *yja < *yję < *yję̆.[5]

The need for the retention of the /Ǣ/ archiphoneme, which represents merged Early Common Slavonic *ē *ā in the position after palatal consonants, up to this point of LCS, 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. In short, a combination of:

1.) the lack of any device in the Glagolitic alphabet to render /ja ńa ŕa ĺa/ sequences (for which Glagolitic texts must use the jat' <ⰰ> letter whose base-value is /ě/);

2.) overwhelming spellings of palatal-letter (<ⱍⱎⰶⰴ>) + jat' in the Kiev Folia (the oldest and therefore least distant ms. from the 'original' OCS, as first codified by Cyrill and Methodius and for which the Glagolitic alphabet was devised) for the reflexes of LCS *č/š/ž/ħ + *Ǣ (e.g. ⱁⰱⰰⰲⰰⰾⱏ <*ob-věħǢlъ, ⰴⱆⱎⰰⰿⰹ <*dušǢmi), as well as occasional traces of such spellings in later Glagolitic OCS (e.g. Psal. ⱍⰰⱎⰵ <*čǢšę); and

3.) the evidence of certain modern Bulgarian dialects, which have reflexes of LCS *ě in words like *ж'еба* <*žǢba 'toad' (Stojkov 1954: 74–78),

all together point very strongly towards there having occurred a split on the Southeastern periphery of Slavic between LCS dialects which have /ě/ < *Ǣ and the majority of the rest which got /a/, and that original OCS ('Urkirchenslavisch') was an *Ǣ > /ě/ dialect. I posit that *Ǣ remained until the opposition /a/:/ě/ after palatal consonants was reintroduced when PV2 and PV3 brought new soft-consonants /c ś dź/ into the system, which could be followed by both /a/ and /ě/: fem. nom. sg. /stьdźa/ < *stьga (PV3) vs fem. loc. sg. /nodźě/ <*nogě (PV2) (Winslow 2022: 304-305). Thus since my LCS system is based on a point just *before* PV2 and PV3, I must also retain the *Ǣ archiphoneme.[6]

---

5 Other troublesome pre-LCS morphological isoglosses reflected in the texts include the masc./nt. instr. sg. *o- and *jo-stem endings *-ъmь/*-ьmь (N.Sl., e.g. KF ⱁⰱⱃⰰⰸⱏⰿⱐ, Uspensk. Sbor. кназьмь) and *-omь/*-emь (S.Sl., e.g. Supr. ѻбраꙁомъ, кнаꙁемь), which are most commonly (e.g. Olander 2015:168) thought to be analogical replacements of the original instr. sg. ending ECS *-ā which is preserved in the adverb *vъčera 'yesterday', and the *-tь (N.Sl.) vs *-tъ (S.Sl.) verbal endings of 3rd sg. and pl. present (plus its extension to 2nd and 3rd sgl. aorists like OCS начатъ, OR (Uspensk. Sbor.) бысть, начать). Here I have no choice but to index them with dummy-symbols in the database: *-Omь/*-Emь for the instr. sg. ending and *-tQ for the verb-endings.

6 It's possible to argue that the short *Æ counterpart to *Ǣ persisted in East Slavic until after the Fall of the Jers, and that the ESl. so-called e > o shift before hard-consonants / back-vowelled syllables is actually just the resolution of this archiphoneme as /o/ (where palatalisation of the preceding consonant remained, in e.g. Ukr. *бджола* <*pъčÆla, or was newly phonemicised, in e.g. Ru. *вёсла* <*v'Æsla <*vesla), and that there was never a stage when these words had /e/ (based among other things on <o> spellings regardless of stress after palatal-letters in very early texts, and even after the letters for secondarily-soft LCS plain consonants in the Birchbark documents (Le Feuvre 1993, Nakonečnyj 1962), but there isn't space to elaborate on the issue here (see Winslow 2022: 304 fn.16). Unlike the situation with long *Ǣ, OCS shows no sign of anything but an /e/ reflex of short *Æ (and indeed the fact that

The syllabic liquids /ŕ̥ ĺ̥ r̥ l̥/ are included as unitary vocalic phonemes, following Schenker (1995: 94), rather than as combinations of /ь ъ/ + /ŕ ĺ r l/, because these groups descend from PIE syllabic liquids and many descendant South Slavic dialects which retain syllabic liquids in this position (including most of those underlying canonical OCS) do not show any evidence of an intervening oral-vowel + liquid stage (such a view is shared by Bethin 1998: 71-72; cf. also Bulgarian dialectal evidence in Stojkov 1954: 130-131, where hard consonants precede reflexes of the LCS /ĺ̥ ŕ̥/ even in dialects with secondarily-palatalised consonants before fallen weak LCS /ь/).
The need for both front and back *ŕ̥ *r̥ is unambiguously shown by the East Slavic reflexes /er/ and /or/ (Ru. *смерть*, *морковь*), but *ĺ̥ vs *l̥ is more complicated: PIE *pl̥nos, *wl̥kʷos > Lithuanian *pilnas* 'full', *wilkas* 'wolf' (LCS *pĺ̥nъ, *vĺ̥kъ) vs Lith. *stulpas* (LCS *stl̥pъ 'pillar') suggests that Balto-Slavic had differentiated front/back variants of the PIE syllabic *l̥ (Bethin 1998: 69), but the ancestor to East Slavic backed all vowels preceding tautosyllabic /l/ (Ru. молоко < Proto-ESl. *molko < LCS *melko > OCS млѣко), and thus only has /ol/ reflexes here: Ru. *волк*, *столб*, *полный*. It's true that Polish has *wilk* and *milczeć* (<*mĺ̥čǢti), but the Polish reflexes are complicated and likely have more to do with the surrounding consonants: *pĺ̥nъ by contrast gives *pełny* with hardened /l/ and the Polish non-palatalising-/e/ reflex of *ъ, and the differing reflexes in *wierzch* <*vŕ̥xъ, *śmierć* <*sъmŕ̥tь and *martwy* <*mŕ̥tvъjь rule out any explanation based on the nature of the LCS syllabic-liquid alone (for more discussion see Bethin op. cit.: 73-75).
While most OCS shows no sign at all of a front-back distinction in the syllabic-liquids and writes the reflexes of these groups overwhelmingly with <ръ> and <лъ>, the Kiev Folia, which is the only OCS text that reflects a pre-Jer Shift stage and is very nearly flawless in its etymologically correct rendering of the jers, also spells *ŕ̥ *r̥ and *ĺ̥ as one would expect: ⱄⱏⰲⱃⱐⱂ- ⱅⰲⱃⱐⰴ- ⱄⱃⱐⰴⱐⱌ- ⰴⱃⱐⰶ- <*ŕ̥, ⱄⱏⱃⱏⱎⱏⱂⱅ <*r̥, and ⱂⰰⰴⰾⱐⱂⰵⱂⱑ <*ĺ̥ (Winslow 2022: 313), and even Zographensis spells all 5 occurences of *vĺ̥k- 'wolf' with ⰲⰾⱏⰽ-/ⰲⰾⱏⱌ- and all 15 instances of its *-mĺ̥č- root with -ⰿⰾⱏⱍ- (e.g. ⰻⰿⰾⱏⱍⰰⱎⰵ). Therefore, taken as a whole the Slavic evidence pretty securely points to front and back variants of both syllabic liquids, and for searching purposes it's far preferable to denote them with separate symbols[7] rather than as the sequences /ьr ъr ьl ъl/[8].

Consonants

Dejotation

the East Slavs inherited their writing system ultimately from the Urkirchenslavisch system designed for such a dialect, rather than one which had a clear way of writing <soft consonant> + <o>, is likely the reason that /o/ reflexes are so rarely detectable in the early texts, since <e> had to be used for both /e/ and /'o/, cf. the spelling ȇвшанъ of the Kipchak word /jovşan/ 'wormwood' in the Hypatian Codex, whose modern cognates (Turkmen *ýowşan* /jowşan/, Kazakh *жусан* /žuwsan/, Azeri *yovşan*) unambiguously point to a Kipchak /o/), and the history of the East Slavic /o/ reflexes remains the subject of much disagreement, so it's simpler for everyone if I continue the traditional practice of writing LCS *e after palatals, even if that strictly speaking is inconsistent with my use of *Ǣ.

7 In the database I will have to use the single Unicode characters <ṛ ṝ ḷ ḹ>, rather than what's shown in my table, since the latter cannot actually be rendered without using the letters for /r ŕ l ĺ/ plus the 'combining ring below' U+0325 symbol, which means searches for the consonantal liquids on their own will also return results containing syllabic liquids. The same problem affects /ẹ̌ y̨/, which I will have to replace with <ẽ̌ ỹ>.

8 To my mind the only evidence in support of a genuine jer + liquid stage comes from the paradigms of verbs like OCS сътрьти < *sъtŕ̥ti, where the syllabic /ŕ̥/ in the stem alternates with /ьr/ depending on the vocality of the following morpheme: the e.g. 3sg. pres. *sъtьretъ (Zogr., Supr. сътьретъ) or (one possibility of the) 3rd sg. aorist *sъtьre (Supr. сътьре) must have /ьre/, while the 3rd pl. aorist *sъtŕ̥šę (Supr. сътрьшѧ) and the other possibility for the 3rd sg. aorist *sъtŕ̥ (Psal. Sin. ⱄⱏⱅⱃⱐ, or with a different prefix Mar. ⱁⱅⱃⱐ <*otŕ̥), being word-final or pre-consonantal, must be syllabic /ŕ̥/. The same alternation occurs in the zero-grade forms of verbs like *umerti, as is clear from the Polish reflexes *umarł* <*umŕ̥lъ vs *umrę* <*umьrǫ. The argument could be that at some stage, before the LCS tendency towards Open Syllables became dominant, the stems in these paradigms were surely unitary /tĭr/, /mĭr/, i.e. 3sg. aor. /sŭ.tĭ.re/ vs 3rd. pl. aor /sŭ.tĭr.šę/, and that the latter's closed /tĭr/ syllable was only forced to open itself up by changing to /tŕ̥/ because of the Law of Open Syllables. Thus at least one source of the syllabic-liquids could be shown to have developed from a vowel + liquid stage, but that still doesn't prove that they all did, or that the change of /ĭr/ to /ŕ̥/ in these verb-forms was not merely a move to an already-existing syllabic-liquid phoneme.

Reflexes of the so-called jot-palatalisation are all written either as unitary palatal phonemes, or in the case of jot-palatalised labials as /vĺ mĺ bĺ pĺ/, rather than as sequences of consonant + /j/, hence /ń ĺ ŕ/ for *nj *lj *rj. The ‘dejotated’ reflexes of *tj (and *kt+front-vowel) and *dj are denoted using the modern Serbian Cyrillic letters /ћ/ and /ђ/ respectively, because the commonly used alternatives, i.e. /ť ď/ (as used in e.g. Olander 2015) or /ḱ ǵ/ (as used by me in Winslow 2022), or variations thereof, are visually too close to symbols used elsewhere in the system. /ḱ, ǵ/ are anyway already used in my system for foreign /k, g/ before front-vowels, and /ť ď/ look too similar to the common denotations of secondarily-palatalised post-Jer Shift /t' d'/, as used in discussions of systems like Russian or Eastern Bulgarian where they arise.
The compelling hypothesis, first proposed by Durnovo (1929: 55-58) but most recently elaborated by Vermeer (2014: 209-214), and accepted by Mathiesen (2014: 197 fn. 22) and Winslow (2022: 310 fn.25), according to which the Urkirchenslavisch reflexes of *ђ,ћ were close enough to foreign /g k/ before front-vowels that the original Glagolitic system used <ⰼ ⱋ> for both sets (i.e. alongside attested ⰻⰼⰵⰿⱁⱀⱏ < ἡγεμών would have been **ⱁⱄⱘⰼⰵⱀⰻ <*osǫђeni, and alongside attested ⰴⱏⱋⰵⱃⱐ <*dъћerь would have been **ⱋⰻⱀⱏⱄⱏ < κῆνσος[9]), does not prevent us from keeping the foreign sounds separate for our LCS stage, since clearly they differed enough in all the dialects underlying actually attested OCS to be written separately.
Pre-dejotation *stj and *zdj are differentiated from the PV1 reflexes of *sk and *zg by writing the former as *šћ and *žђ and the latter as *šč and *žǯ, even though their modern reflexes do not differ from each other anywhere and so must've fallen together in the CS period, because they often alternate with their respective un-palatalised counterparts morphologically and derivationally, e.g. očistiti:očišћenьje vs. jьskati:jьščǫ, jǢzditi:jǢžђǫ vs jьzgъnati:jьžǯenǫ.

There are convincing arguments for PV2/3 having preceded dejotation, at least in more central areas, most recently presented in e.g. Vermeer (2014: 197) and Wandl & Kavitskaya (2023: 244-247), and therefore it could be objected that my system, which contains the dejotation reflexes /ћђńĺŕ/ but not the PV2/3 reflexes /c ś dź/, is ahistorical. However it should be reemphasised that the primary goal of my LCS reconstructions is to act as an index which allows reflexes in texts to be found, not to be a historically realistic description of some actually-existing LCS dialect. The absence of PV2 in Novgorodian shows that it can’t have preceded dejotation everywhere in Slavic, and in any case the replacement of the sequences /tj dj nj lj rj/ by articulatorily distinct combined units, no longer associated by speakers with their /t/ and /j/ phonemes, is structurally completely irrelevant unless and until these new units merge with existing phonemes (or new sequences of dental + /j/ are introduced), as e.g. in the KF dialect where /tj/ merged with /c/ from PV2/3, or in ESl. where it merged with /č/ from PV1. A language which had distinct Czech-like palatal [c, ɟ] reflexes of *tj and *dj, and also no new sequences of [tj, dj], could not convicingly be argued to have undergone dejotation at the phonemic level, as these new units would just be phonetic realisations of /tj, dj/. Analysed like that, the symbols /ћђńĺŕ/ in my system strictly speaking would really just be cover-symbols for the pre-jotation sequences, but such notation is preferable since it prevents searches for groups containing /j/ alone from returning results polluted by all the dejotation-groups. As I explored in my previous article (Winslow 2022), the status of /j/ as a phoneme in the earliest OCS texts is an intricate problem, so the ability to investigate the reflexes of *j in isolation from the dejotation-reflexes is important.

/j/

Word-initial *jǢ-/*a-
The tendency for ECS *ā- to have taken prothetic /j/ by LCS times (in accordance with the drive towards open syllables) can make it difficult to distinguish this group from *jǢ- in the absence of

9 Interestingly, this aspect of the hypothesised Urksl. orthographic system has rearisen in the modern Macedonian standard due to Turkish loanwords: *ќемер* < Tk. *kemer* ‘belt’, *ќе* < *[хъ]ћe[тъ]; *ѓон* < Tk. *gön* ‘leather’, *меѓу* <*meђu.

wider Indo-European evidence. Normally I've followed Derksen (2009), or the ESSJA, but for certain lexemes, e.g. *ama 'pit', which in OCS is spelt overwhelmingly with ⱑⰿ- or ꙗм-, the single Greek cognate ἄμη adduced by ESSJa I p.70 in favour of jot-less *am- is not enough to categorically exclude the alternative *jǢma. In particular the 1sg. nom. pronoun *azъ/jǢzъ is especially problematic: I follow ESSJA I p.100 which ultimately plumps for *azъ, but Derksen doesn't discuss it at all. (A lengthy discussion of the evidence can be found in Teneva's (2012) article on the subject.)
Forms with insecure etymologies can't under any methodology be used as good evidence in phonological investigations, so in difficult cases like the above I simply mark the lemma in the database and provide some short discussion, so that eventually the web-interface can flag such forms in some way and inform users of the specific difficulties.

Like Derksen, I assume that roots going back to PIE jot-less long *ē or dipthongal *oi-, e.g. the root for 'to eat', PIE *h_1ēd-, all took prothetic *j and merged with *jǢ- from other sources, unlike Durnovo (1929: 54), who seems to think that such a development was limited to Bulgarian and Macedonian dialects, including those underlying OCS (where in the Cyrillic mss. we get regular ꙗсти etc.). Isolated nominal forms like Ru. *язва* (which Derksen derives from a Balto-Slavic *oi-based on Lith. *aiža* and Old Prussian *eyswo*) suggest that *ě reflexes in the modern forms of verbs like Ru. *ехать*, Pol. *jeść* are later generalisations from prefixed forms like OR сънѣсти, where no jot-prothesis could take place (cf. Schenker 1995: 88, Winslow 2022: 302 fn.14).

<u>Jers before *j</u>
As explored more fully in Winslow (2022: 313-315), OCS spellings seem to suggest that free-variation between <ь,ъ> and <и,ꙑ> was a feature of the pre-Jer Shift Urkirchenslavisch orthographic system for conveying the reflexes of the sound-groups *ьj and *ъj (so-called 'tense jers'), regardless of whether they were in strong or weak position. The examples given were: Zogr. ⰸⱀⰰⰿⰵⱀⱐⱑ vs ⰸⱀⰰⰿⰵⱀⰻⱑ <*znamenьjǢ, ⱆⰴⰰⱃⱐⰺ <*udaŕь-jь vs ⱁⰿⱁⱍⰻⰹ <*omočь-jь; Mar. ⱁⱄⱘⰴⱔⱅⱐ ⰹ vs ⱁⱄⱘⰴⱔⱅⱐⰹ ⰻ <*osǫdętъ ‿jь; KF ⰿⰻⰾⱁⱄⱅⱐⱙ vs ⰿⰻⰾⱁⱄⱅⰻⱙ <*milostьjǫ, ⰲⱐⱈⱁⰿⱁⰳⱐⰹ vs ⰲⱐⱈⱁⰿⱁⰳⱐⰻⰹ < *vьxomogy-jь). Of importance here is the fact that the same orthographic system characterises both pre- (i.e. KF) and post-Jer Shift texts; that even in strong position in a text like Zographensis, which shows pretty clear signs of having undergone the Jer Shift, spellings like ⱆⰴⰰⱃⱐⰺ, ⰱⱁⰾ̑ⱐⰹ, ⰲⱔⱋⱅⱐⰹ for what in the live dialect underlying Zogr. must surely have been /udaŕij/, /boĺij/, /vęštij/, are not infrequent[10]. The fact that the same alternation occurs in the pre-Jer Shift KF (i-stem gen. plurals ⰸⰰⱂⱁⰲⱑⰴⱐⰹ <*zapovědьjь vs ⰾⱓⰴⰻⰺ <*ĺudьjь) suggests that it is a common inheritance from the Urkirchenslavisch spelling system, and thus that in pre-Jer Shift Slavic the difference between /ь ъ/ and /i y/ was neutralised before /j/, and we should perhaps posit archiphonemes (which I call /Î/ and /Ŷ/) in this position. These archiphonemes are, in slightly different terms, effectively posited by Trubetzkoy (1954: 70) in his analysis of the Urkirchenslavisch phoneme-system[11].
However, for simplicity and accessibility's sake it's better to avoid overburdening the indexing-system with unfamiliar and controversial archiphoneme-symbols, so I keep *ьj/*ъj as the denotations for these groups.

Difficulties arise though when deciding how to denote foreign sources of /ij/[12] which may or may

10 Marianus and Psalterium Sinaiticum, on the other hand, frequently show a Russian-style /ej/ reflex of strong tense *ьj: Psal. ⰲⱁⱃⰱⱐⰹ <*vorbьjь, ⱂⰾⱏⱅⱐⰹ <*plъtьjь, ⰿⱐⱀⱐⰹ <*mьńьjь; Mar. ⰸⰰⱂⱁⰲⱑⰴⱐⰻ <*zapovědьjь, ⱆⰴⰰⱃⱐⰻ <*udaŕь-jь, ⰳⰲⱁⰸⰴⱐⰻⱀⱐⰹⱑ <*gvozdьjьny-ję̌. Psal. (Psalm 21) even has definite past. act. part. Nsg. masc. form ⰻⱄⱅⱃⱐⰳⱐⰹ <*jьstrgъ-jь that suggests an /oj/ reflex of *ъjь.

11 Though Trubetzkoy, like me, believes Urkirchenslavisch to have been based on a /j/-less dialect, so in that particular system the archiphonemes would be conditioned by the position before *vowels*, rather than before /j/.

12 The sequence /ij/ is not totally banned from native words, since it appears to be preserved across morpheme-boundaries, such as in prefixed-verbs like приꙗти <*prijęti or long-form adjectives like masc. nom. pl. дроугыи <*drugi-ji, but within roots it does seem restricted to these post-LCS loanwords.

not have been integrated into the native system as reflexes of /Îj/: words like марига < Μαρία, стадии < στάδιον, which are well-integrated into the morphological system as a fem. ja-stem and masc. jo-stem respectively, could either be reconstructed as consciously-foreign *marijǢ, *stadijь, or as nativised *marьjǢ, *stadьjь, but there are no occurrences of jer-spellings in these words in the OCS texts in TOROT. Other similarly-Greek words like дигаволъ (< διάβολος), however, do show up in OCS with jer-spellings: Supr. дьга́вола, Zogr. Luke 8 and Psal. Psalm 108 ⰴⱐⱑⰲⱁⰾⰰ, which (alongside the modern Macedonian *ѓавол* with the reflex of *ђ produced by the Macedonian so-called 'new jotation' of /d/ after the fallen jer brought it into contact with /j/) clearly suggest an early adaption of this foreign /ij/-group to native /Îj/. Old Russian texts even show spellings of марига suggestive of full nativisation: Laurentian Primary Chronicle м҃рьга, м҃рьєю, Zadonshchina марга, марьга, as well as First Novgorod Chronicle gen. sg. васильга (jo-stem василии < Βασίλειος).
Since we can't ever be sure of the precise timing or route by which these late borrowings entered the various Slavic dialects, or of the extent of their adoption by Slavs beyond a tiny and often Greek-knowing scribal-class, the best solution is to set all such foreign /ij/ groups apart from the native vocabulary by using an *ij reconstruction, even where we can be pretty sure that early nativisation to reflexes of *ьj occurred: *dijǢvolъ, *vasilijь, *marijǢ etc.

Word-initial *jь-/*ji-/*i-
With native Slavic word-initial *ji-/*jь-, I follow Derksen's (2009: 16) practice of writing *jь-, even though Derksen himself (2003) has argued for a split between *ji- and *jь- conditioned partly by accentological factors (which, as stated above, I have chosen not to consider). Most of the modern languages reflect these groups as just /i-/, except for Czech and Ukrainian: forms like Cz. *jdou* and Ukr. (after vowels) *йдуть* appear to have dropped the weak-jer in *jьdǫtь just like any other and retained the /j/, and Ukr. *ськати* <*jьskati (with the restricted meaning 'look for nits/fleas in someone's hair' after the base-meaning 'seek' was taken over by the Polonism *шукати*) shows the expected Ukr. softening of the /s/ after fallen weak-jer in *ьsk groups (cf. *польський*).
I make an exception for certain forms of the personal-pronoun *jь, however, and write *jimь, *jima, *jixъ *jimъ and *jimi for the masc/nt. instr. sg. and dat./instr. dual/pl., because Czech here has *jim jich jimi*.
In badly-integrated clearly post-LCS foreign words, such as Biblical names like игаковъ (borrowed via Gk. *Ἰακώβ*), or ѩемонъ (< *ἡγεμών*), I keep a bare initial *i-, though this is rather an arbitrary choice and done partly as a way of marking such words as non-native[13] (cf. my treatment of foreign initial *e- below). An exception is made for исоусъ < Gk. *Ἰησοῦς*, which I have as *jisusъ, because of the greater likelihood that Slavs will have heard of Jesus even before the first biblical translations, and because spellings like Zogr. ⰺⱐⰻ ⰻ̄ⱄⱏ suggest that it causes the same /Ŷ/ archiphoneme reflex of *ъ before *j as you get in e.g. native Mar. ⰲⱐⰻ ⰻⱄⱅⰻⱀⱘ < *vъ ‿ jьstinǫ (see above).
Prefixed forms like *do-jьti 'to come, arrive' for morphological reasons have to be distinguished from the class 4 verb *dojiti/dojiši/dojimъ etc. 'to breastfeed' (and its derived noun *dojidlika), a difference which is reflected in the modern Ukrainian *дійти* (<*dojьti with compensatorily-lengthened /o/ > /i/) vs *доїти*. Thus /i/ can follow /j/ when the former is part of a morpheme which just happens to be stuck onto a /j/-ending stem: I similarly allow words like *šujika (шюица) and *vojinъ 'warrior' (воинъ, as opposed to *vojьnъ, the gen. pl. of *vojьna), or the loc. sg/pl. desinences of any jo-stem noun whose stem ends on /j/, e.g. Psal. ⰶⱃⱑⰱⱐⰻ <*žerbьji.

Word-initial *je-/*e-

13 Spellings like Zogr. Mark 13:3 "ⱂⰵⱅⱃⰰ. ⰻ ⱑⰽⱁⰲⰰ. ⰻ ⰻⱁⰰⱀⰰ." "Peter and Jacob and John" would suggest that this initial *i- can get dropped after an /i/ of a preceding word, but whether this points to a dropping of the non-native *i-, simple deletion of a double /i i/ (haplology), or a native-like reflex of a weak-jer /*i *jьjǢkovъ/ > /i jakov/, is not really knowable, so indexing such words with a markedly foreign initial *ij- group is again the best way of allowing such difficult cases to be investigated.

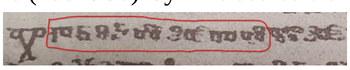
No Glagolitic text makes any effort to distinguish /je/ (after vowels or word-initially) from post-consonantal /e/, writing both with <ⰵ>, unlike the situation with the reflexes of *ję vs *ę, where in Zogr. and Mar. and partially in Assem. (Velcheva 1981: p.168) the full front-nasal digraph <ⱗ> is reserved for *ję, while just the second 'nasalising component' <ⱔ> is used for post-consontal *ę, e.g. Mar. 3rd pl. aorist ⱗⱄⱔ <*jęsę, as opposed to KF ⱂⱃⰹⱗⱅⰹ <*prijęti vs ⰲⱏⰸⱗⰾⰹ <*vъzęli[14]. Glagolitic evidence alone therefore would suggest that foreign borrowings with word-initial /e-/ were simply adapted to whatever the reflex of native LCS *je was. Suprasliensis, though, which uses the jotated <ѥ> letter, does in fact make an extremely consistent spelling distinction between foreign borrowings and native Slavic words: of the 157 occurrences of the 13 foreign lemmas I have so far reconstructed with word-initial *e/*je- which appear in Supr. (*episkupъ, evanǵelьje, eǵüpьtъ, elisavetь, elinъ, evanǵelistъ, eǵüpьtьskъ, elinьskъ, episkupьstvo, evrejьskъ, elisejь, emъmausъ, etijopьskъ*), the only spellings with <ѥ> are ѥлисеи, Ѥ҃ппъ, Ѥ́лини, and ѥ́лина, i.e. 4/157 or 2.5%. By contrast, of the 3172 native Slavic words in Supraslienisis which I Autoreconstruct as starting with *je- (not all of whose *lemmas* start with *je-, e.g. forms of **byti*), just 88 are written with initial <є>, vs 3070 with <ѥ>[15]. Thus 97.2% of native word-initial *je- in Supraslienisis is spelt with <ѥ>, while 97.5% of the occurrences of the clearly post-LCS Greek-mediated foreign borrowings listed above instead use plain <є>, suggesting that *some* sort of difference was felt, at least by the scribes of Supraslienisis, and that we probably shouldn't index these with the same *je- as used for native forms. I therefore use non-jotated *e- for such foreign borrowings, and the extent to which they take prothetic *j- and fall together with the native vocabularly is left as something for investigators to determine based on the evidence of each manuscript.

Prefixes

The last particularity of my LCS indexing-system worth mentioning relates to the handling of consonant-clusters in prefixes: as exhaustively exemplified by Diels (1963: 121-125), Common Slavic permitted only a restricted set of consonant-combinations in the syllable onset, generally either combinations of the continuants *s/*z plus obstruent or sonorant (except *r, see below), or of obstruents plus sonorant (with some curiosities such as the seeming dialectal diversity in the tolerance of *bn but not *pn: OCS гъıбнѫти <*gyb-nǫti > Ukr. *гинути*, vs OCS оусъпнѫти <*usъp-nǫti (cf. 3sg. aor. оусъпѥ), Ru. *тонуть* <*top-nǫti, though see Meillet (1965: 142)). Geminate consonants were banned and either simplified (исѣшти <*jьs-sekti) or dissimilated (процвисти <*prokvit-ti).

The ban on *sr/*zr is dealt with by insertion of *t and *d respectively, but the commonly-cited examples of *str <*sr (сестра, строуꙗ, остръ) all concern root-internal *sr where insertion of *t is common also to the Germanic and sometimes Baltic cognates. The examples given by Meillet (1965: 136) include: (for строуꙗ) Lith. dial. *srauja* next to Latvian *strauja*, then Germanic *straum- (> Eng. *stream*, Old Norse *straumr* etc.); (for остръ) Lith. *aštrus*, Gk. *ἄκρος* (here the *s is from PIE *ḱ). As Meillet says, "*ce n'est pas un developpement germano-balto-slave ; d'une part, le developpement d'un* -t- *dans le groupe* sr *est chose naturelle et se retrouve ailleurs (fr. pop.* castrole *de* casserole*) et, d'autre part, le developpement de* t *en ces conditions n'est pas general en baltique:* str *est regulier en lette, mais* sr *subsiste couramment en lituanien.*", so we can't really be sure when the Slavic change took place or whether it was still active during our LCS stage. The only indication of its activity in OCS is the single Psal. ⱄⱅⱃⱏⰿⱁⰿⱐ <*sorm-omь spelling cited by Diels (p. 122); otherwise new /sr/ from metathesised *sErC groups is tolerated unchanged.

---

14 Psalterium Sinaiticum contains at least five occurrences of non-digraph <ⱔ> for *ę: Eckhoff's digitisation suggests ⱎⱃⰰⰿⱔ, ⰲⱁⰸⰾⱏⰺⱔ, ⱄⱅⱁⱑⱔ, ⱅⱔ, ⰳⱔ, and ⱂⱃⱁⰺⰾⱔⱅⱁ, but the last one ⱂⱃⱁⰺⰾⱔⱅⱁ (which is from Psalm 151 in the part discovered in 1975) comes from a mistake in Mareš's (1997: 50) Cyrillicised edition: the manuscript-photograph in Tarnanidis (1988: 260) clearly shows ⱂⱃⱁⰺⰾⱗⱅⱁ: 

15 The leftover 14 are things like 1st. pres. dual. ѝмавѣ which Eckhoff's corpus wrongly lemmatises as *jьmati instead of *jьměti, and which thus get reconstructed as *jemĺevě instead of *jьmavě. At the time of writing only 3227/6862 Supraslienisis lemmas have been reconstructed, but those 3227 cover 89713/99194, or 90.4%, of the words.

New occurences of *zr, on the other hand, are regularly generated in the language right up to OCS times, not only in the derivational-morphology because of the verb-prefixes *orz-, *vъz-, *jьz- (e.g. Supr. 3sg. aor. въ<u>зд</u>роу ‘roared’, from *vъz-ruti), but also because of the clitic prepositions *jьz and *bez, which form one phonological word with whatever follows them and thus cause OCS spellings like Mar. Luke 1 ⰹ<u>ⰸⰴⱃ</u>ⱘⰽⱏ <*jь.z ‿rǫ.kъ. Meillet (p. 136) also cites the Old Polish adverb *zdręki* <*jьz ‿rǫ.ky, which proves that the phenomenon is not limited to SSl. or OCS. Curiously, though, despite this overwhelming evidence of a synchronic /zr/ > /zdr/ rule in OCS, /zr/ from the metathesised *zork- root is never spelt <здрак> and so seems to be tolerated, even though Diels cites prepositional forms like Supr. бе<u>ѕдр</u>аѕоума <*bez ‿*orzuma, бе<u>здр</u>ала <*bez ‿*ordla, which come from metathesised *orT- groups but *do* show inserted /d/. Such inconsistency is hard to explain unless the addition of /d/ has been partly morphologised as a variant of specifically the prepositions before /r/.
With such a sound-change that appears most often at morpheme or straight-up word-boundaries, there is a strong drive to restore the underlying shape of the constituent parts, hence the modern languages have mostly restored /zr/ groups in e.g. Russian *разрешить*, and there are traces of this even in Psalterium Sinaiticum: Psalm 48 ⰹ<u>ⰸⱃ</u>ⱘⰽⱏⱇ (Diels 1963: 122). In Old Russian, the Uspenskij Sbornik is pretty consistent in keeping prefixed verb-forms like раздрушить <*orzrušitь, but by the time of the Laurentian Codex we get forms like възрадуем and неизреченное.
Therefore even though *sr > *str and *zr > *zdr appear to be simply voiced and unvoiced variants of the same sound change, the practical effects are very different because the former is, from the LCS perspective, totally ‘opaque’, since it only occurs in roots and thus is not analysable by speakers into constituent morphemes without the inserted stop, in the way that /bez ‿drǫky/ can be identified with separate /bez/ and /rǫky/.
For this reason I don’t include /zdr/ <*zr at prefix or preposition-boundaries in my LCS system, so that investigators can see for themselves the extent of each text’s adherence to the expected phonological development vs restoration of /zr/ under morphological pressure.

Following the same logic I also retain illegal *ss and *sš groups in prefixed-verbs like Psal. ⰹ<u>ⱄⱄ</u>ⱔⱍⰵ <*jьs-sęče, Mar. ⰹ<u>ⱎ</u>ⱏⰴⱏ <*jьs-šьdъ, Assemanianus ⱃⰰ<u>ⱎ</u>ⰹⱃⱑⱓⱅⱏ <*ors-širÆjǫtъ, because that same drive towards restoration of the underlying shapes of the prefixes *jьs-/*ors- etc. can be seen in modern Russian *и<u>сс</u>якнуть*, *ра<u>сш</u>ирять*, and Laurentian Codex ра<u>сш</u>едше.[16] This treatment is also more consistent with my handling of verbs like *jьs-kěliti (> OCS и<u>ц</u>ѣлити/и<u>сц</u>ѣлити) where simplification *must* have occurred posterior to our pre-PV2 LCS stage (since /sk/ is always totally permissible), and where manuscripts show great diversity, e.g. Zogr. and Assem. consistently have ⰹ<u>ⱌ</u>ⱑⰾ- while Mar. and Psal. keep ⰹ<u>ⱄⱌ</u>ⱑⰾ- (more discussion of the wider Slavic reflexes of this group, including the OCS <єт> spellings, can be found in *replace with just ‘see also’?* Meillet 1965: 133).

<u>Morphological innovations that scupper LCS reconstruction</u>

The units of the phoneme-system sketched above serve as the building-blocks for all higher-level linguistic systems, most immediately the inflectional-morphology and derivational-morphology systems, whose features are thus constrained by said phoneme-system and the distributional-

16 Conversely, sequences of *sk, *zg at prefix-boundaries which show PV1 reflexes, like Mar., Zogr. ⱃⰰ<u>ⱎⱍ</u>ⱐⱅⰵⱅⱏ <*orš-čьtetъ (ECS *skĭt- > *ščĭt-), Psal. ⱃⰰ<u>ⰶⰴ</u>ⰹⰳⰰⰵⱅⱏ <*orž-ǯigajetъ (ECS *zgīg- > *žǯīg-) are kept as *šč, *žǯ. Such forms may well not go all the way back to the time of PV1, and instead be just the result of a synchronic rule prohibiting /zž/ and /sč/ (> /žǯ/ and /šč/) that remained active until much more recently, especially given prepositional-phrase forms like Psal. ⰹ<u>ⱎⱍ</u>ⱃⱐⰲⰰ <*jьs‿*červa, so this is arguably inconsistent with my treatment of *ss, *sš etc. My justification is firstly that *sk, *zg > *šč, *žǯ are *conspicuously* PV1-changes, which we *know* originated well before our target LCS point, whereas the precise timing of de-gemination or simplification of *sš is less clear-cut; and secondly that even in languages like Russian which *orthographically* have restored <сч> and <зж> spellings in compounds like *и<u>сч</u>езнуть* and *ра<u>зж</u>ечь*, the pronunciations are still arguably direct reflexes of LCS *šč and *žǯ, viz. [ɕː] and [ʑː] (or, in the conservative Moscow-dialect, the palatalised [ʑː] found also in до<u>ждь</u> <*dъžǯь).

restrictions of its units (i.e. *phonotactics*). Changes which occur in the phoneme-system between the time of our theoretical LCS and the time of our texts can therefore trigger (or allow) restructuring of these morphological systems, which in turn can produce forms containing phoneme-sequences with no direct LCS ancestor-sequences.

An example of such morphological change contingent upon structural phonological change, leading to forms which preclude any direct LCS-stage reconstruction, 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/[17]. 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 владъікъі чловѣколюбьѥ· ꙗко прѣданника разбойника 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 *końь > /kon'/, where the original LCS palatal *ń has fallen together with secondarily-palatalised /n'/ from plain LCS *n before LCS front-vowels, in e.g. the original i-stem *bornь > /boron'/. 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 endings (in this case those of the original jo-stems)[18]. Consequently, a word like тать has begun to take jo-stem endings, including the Old Russian /a/ reflex of LCS *Ǣ in the genitive/accusative singular.
LCS /Ǣ/, though, by definition can only occur after LCS palatal consonants (see above), so a reconstruction *tatǢ is just nonsensical. In the case of the dat. sg. /u/-desinence (which isn't attested in our Treaty but 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.

Were the same shift from i-stem to jo-stem to occur in a word like *zvěrь, then the structural change would not be so catastrophic, because our LCS system *does* contain a palatal *ŕ which any allophonically-softened LCS hard *r could easily be subsumed into. Indeed, interestingly Suprasliensis does in fact contain 3X gen. sg. звѣрѣ, with what looks like a jo-stem reflex of *ŕǢ (spelt with jat' as an overhang of the Glagolitic tradition, cf. 2X морѣ vs 1X морꙗ spellings), suggesting that Russian-style secondary palatalisation of *r > /r'/ may have occured in the Bulgarian dialect underlying it[19]. You don't, though, get anything like тата[20], so the system-wide development of secondary-palatalisation does not seem to have advanced enough to have

17 In some dialects (notably East Slavic) the PV3 reflexes *ś and *ʒ became patalised counterparts to plain /s z/, i.e. /s' z'/, and merged with the /s' z'/ that developed from LCS *s,z before front-vowels, but in most OCS they seem to have just hardened to /s, z/: searching my database for the sequence *ьxǫ, for example, turns up exclusively <ⱄⱘ> spellings in Marianus, just one <ⱄⱘ> in Zogr., and exclusively <сѫ> in Suprasliensis, with only Assem. and Psal. containing a significant minority of <ⱄⱘ> spellings.

18 Russian feminine i-stems like *весь* (<*vьsь, 'village') do not fall together with ja-stems in the way the masculines like *zvěrь fall together with jo-stems, but they do all still take the /am, ax, ami/ endings in the dat., loc. and instr. pl., e.g. *ве́сям*, which contain the same LCS ja-stem *-Ǣ- vocalism which can only occur after LCS palatals, meaning they too end up totally unreconstructable due to an illegal **sǢ sequence.

19 Numerous spellings in Supr. like боура <*buŕǢ, оукараѥтъ <*ukaŕǢjetъ, мороу <*moŕu etc., however, point to a hardening of LCS palatal *ŕ to plain /r/, so it's difficult to know whether звѣрѣ spellings stem from a genuine /zvěŕa/ form in the history of the language, or if they instead represent synchronic /zvěra/, i.e. with a hard o-stem ending, but with confusion by the scribe between <ра> and <рꙗ>/<рѣ> spellings for what in his/her dialect would've all been /ra/.

20 Except the numerous gen. sg. господа and dat. sg. господоу for the i-stem *gospodь, but this word seems to be an isolated special case, because it bafflingly turns up even in early Glagolitic OCS with endings like ⰳ҃ⰵⰲⰺ (ju-stem dat. sg. *-evi, cf. Supr. господеви), ⰳ҃ⱆ (jo-stem dat. sg. *-u), and ⰳ҃ⱑ (jo-stem gen. sg. *-Ǣ). See Van Wijk (1929).

caused the sort of fundamental structural reorganising which shifted *tatь into the jo-stems in Old Russian.

Forms like татѧ, then, though they frustrate our goal of reconstructing entire texts, do provide us some objective measure of 'linguistic distance' between stages of a language, because their existence presupposes at least one intervening stage where the structure of the phonological system has changed enough from our LCS stage to have caused/allowed restructuring of the morphological system.

1.2 LCS Morphology and the Autoreconstructor

The ten-place morphology-tags included as part of the word-level annotations in Eckhoff's TOROT corpus constitute a veritable goldmine of linguistic data, because, based as they are on the *form* of a word rather than the *function*, they bridge the gap between the higher (syntax, semantics etc.) and lower (phonology, orthography, morphology) levels of linguistics analysis. An example TOROT annotation for the word ⰲⱏⰸⰲⱑⱋⱘ is given below:

```
<token id="3589172" form="възвѣштѫ" citation-part="70.17"
lemma="възвѣстити" part-of-speech="V-" morphology="1spia----i"
relation="pred" presentation-after=" "/>
```

Figure 1: TOROT annotation for Psal. Sin. Psalm 70 ⰲⱏⰸⰲⱑⱋⱘ in XML format

TOROT token XML-tags include various attributes, but for the Autoreconstructor all that's needed are *form*, *lemma*, *part-of-speech*, and *morphology.* The *form* attribute is used to check for morphological variations/innovations, to ensure that what gets produced is the direct phonological ancestor of the actually-occurring word (see section ??? for more about this deviance-detection). The *lemma* and *part-of-speech* attributes, when concatenated, serve as a unique key linking each word to its lemma[21] and thus to its LCS reconstruction and inflexion-class information. Finally the *morphology* attribute consists of a 10-character string to hold values for the 10 morphological-features used by TOROT (and the wider PROIEL corpora). Not all features are relevant for all words, in which case a dash '-' is used as a placeholder.
A detailed explanation of each feature can be found in Section 6 of Eckhoff et al. (2018: p.?), but here it suffices to say that in this example the tag "1spia----i" is telling us that ⰲⱏⰸⰲⱑⱋⱘ is **1**st person, **s**ingular, **p**resent-tense, **i**ndicative-mood, **a**ctive-voice, has no gender, case, degree, or strength features, and is **i**nflectable rather than non-inflecting.

Of importance here is the *present* tense tagging, even though възвѣстити (even in OCS) can be taken as a perfective verb, opposed to its imperfective counterpart възвѣщати, and thus has future-tense meaning in its non-past indicative forms (and the Greek Septuagint here has ***ἀναγγελῶ*** *τὰ θαυμάσιά σου*, with a morphologically future-formed *ἀναγγελῶ* 'I will proclaim' from *ἀναγγέλλω*). The tagging thus follows the *inflectional*-morphology of ⰲⱏⰸⰲⱑⱋⱘ, rather than the future-meaning which is carried by the *derivational*-morphology. This is important because the Autoreconstructor works by *inflecting* already-reconstructed lemmas; //a lot more crap needs to be added

21 Identical lemmas with the *same* part-of-speech tag, such as вести 'to lead' <*ved-ti and вести 'to drive' <*vez-ti, both of which have 'V-' for verb, are differentiated by appending #2 etc. to the extra homomorphs, i.e. вести vs. вести#2.

Deviance detection

217066	приведошѧ	3paia----i	privedošę	privedǫ
217101	въздъхнѫвъ	-supamn-si	vъzdъxnǫvъ	vъzdъxъ
217112	развръзосте	3daia----i	orzvŕ̥zoste	orzvŕ̥zete
217261	б҃гвивъ	-supamn-si	bolgoslovivъ	bolgoslovĺь
217266	ѣшѧ	3paia----i	jǣšę	jǣsę
217272	вьзѧшѧ	3paia----i	vъzęšę	vъzęsę
217302	начѧшѧ	3paia----i	načęšę	načęsę
217316	въздъхнѫвъ	-supamn-si	vъzdъxnǫvъ	vъzdъxъ
217455	приведошѧ	3paia----i	privedošę	privedǫ
217600	начѧтъ	3saia----i	načętъ	načę
217605	с҃ноу	-s---md--i	synu	synovi
217635	начѧтъ	3saia----i	načętъ	načę
217787	поѩтъ	3saia----i	pojętъ	poję
217832	мосѣомь	-s---mi--i	mosijomь	mosijemь
217856	мосѣови	-s---md--i	mosijovi	mosiju
217869	быс	3saia----i	bystъ	by
217960	с҃нѣ	-s---ml--i	syně	synu
218067	невѣрьнъ	-s---mvpsi	nevěrьnъ	nevěrьne
218108	быс	3saia----i	bystъ	by
218173	нѣмы	-s---mvpwi	němъjь	němejь
218175	глоухы	-s---mvpwi	gluxъjь	glušejь
218198	бысть	3saia----i	bystъ	by
218206	оумрѣтъ	3saia----i	umertъ	umer
218388	ꙇмени	-s---nl--i	jьmeni	jьmene
218419	ꙇмени	-s---nl--i	jьmeni	jьmene
218561	окомь	-s---ni--i	okomь	očesьmь
218648	мѫжю	-s---md--i	mǫžu	mǫževi
218688	мѫжа	-s---mg--i	mǫžǣ	mǫžu
218755	прѣл҃юбы	-s---fa--i	perĺuby	perĺubъvь
218762	поустивъши	-supafn-si	pustivъši	pušťьši

Figure 2: Auto-detected and -reconstructed morphological deviances from a small part of the Book of Mark in Codex Zographensis

The screenshot above shows some raw data from my autoreconstructed SQLite database of the TOROT OCS texts; in this case it's forms from Zographensis (around Mark 7 to Mark 10) where the Autoreconstructor has detected morphological innovations. The fourth column shows what the Autoreconstructor thinks is the direct phonological ancestor to the text-form, but the ancestor of the 'original', 'correct', or 'default' morphological form is also generated and stored in the fifth column, so that such cases of innovation can be easily searched-for and counted (since non-innovated forms have NULL values in this column).

Types of innovation detected here include:

- extended S-aorists (*-ox- aorists) of class 1 verbs: 3rd pl. приведошѧ vs. приведѫ, 3rd dual развръзосте vs *развръзете[22]

22 Koch (1990: 293) lists only sigmatic aorists as possibilities for the *-verz-/*-vŕ̥z- stem verbs, and it seems that outside of the 3rd sg. (e.g. Psal. ⱃⰰⰸⰲⱃⱏⰸⰵ, Zogr. ⰲⱏⰸⰲⱃⱏⰸⰵ) no root-aorists are attested in any Slavic text, so maybe I am wrong to set up asigmatic root-aorists like 3rd dual *-vŕ̥zete as a possibility alongisde primary sigmatic *-verste (e.g. Mar. Mark 7 ⱃⰰⰸⰲⱃⱑⱄⱅⰵ). My justification is that the *-verg-/*-vŕ̥g- stem verbs *do* attest such root-aorists, e.g. 3rd pl. ⰲⱏⰸⰲⱃⱏⰳⱘ in Psal. Psalm 77, and I don't see what, apart from the nature of the final stem-consonant (obstruent vs. continuant), could be grounds for classifying these two verbs differently.

- unetymological extension of the RUKI-rule-produced *š in 3rd pl. primary sigmatic aorists: ѣшѧ vs. ѣсѧ <*jǢd-s-ę, вьзѧшѧ vs. възѧсѧ <*vъzьm-s-ę, начѧшѧ vs. начѧсѧ <*načьn-s-ę (neither *d, *m, nor *n have ever been RUKI sounds)
- extension of the *-nǫ- suffix to the past. act. part. of class 2 verbs like въздъхнѫти: въздъхнѫвъ (cf. Mar. ⱆⱄⱏⱈⱏⱎⰵ from оусъхнѫти)
- addition of the *-tQ suffix from the 3rd sg. pres. (see fn. 5 above) to 3rd sg. aorist forms: начѧтъ, поѩтъ, оумрѣтъ
- original u/ju-stem nouns taking o/jo-stem endings: dat. sg. с҃ноу, мѫжю; loc. sg. с҃нѣ, gen. sg. мѫжа
- past act. part. of class 4 verbs using the suffix *-ivъ rather than *-jь: б҃гвивъ, поустивъши (cf. Mar. Mark 10 ⱂⱆⱎⱅⱏⱎⰻ <*pust-jьši)

- Consonant-stems – with the *teĺ- suffix agent-nouns, I mostly follow people like Meillet (1965: 426) in taking consonant-stem endings in most of the plural, but the nom. pl. it's difficult to agree with his positing of a plain /le/, as opposed to palatalised /ĺe/ desinence (i.e. with the consonant-stem vowel on the jo-stem stem), because Zographensis and Suprasliensis are consistent in marking such forms with their palatalisation-diacritic.
- related is derivation-morphology difficulties such as whether adjective *voĺьnъ should have a palatal /ĺ/, or whether the *voĺǢ is specifically differentiated from the root *vol- by a *-jǢ noun-forming suffix. Spellings are similarly suggestive of *voĺ-

-Talk about the pres. forms of *telħi and link back to the discussion about the difficulties with syllabic *ĺ̥, saying that Derksen and the two Czech dictionaries cite *tl̥kǫ forms, and that Zogr. mostly spells this group with <лъ> as well, which would suggest an switch from e- to o-grade ablaut between the full-grade and zero-grade stems, but also that the issue is confounded by the existence of an o-grade form of the verb *tolħi suggested by the PPP form ⱂⱃⱁⱅⱁⰾⱍⰵⱀⱘⱙ <*protolčenǫjǫ in Psal. Psalm 138

-The seeming impossibility of reconstructing aberrations like Supr. жласти, жладьба, from what can only be an original *geld- root and likely a Germanic loan (cf. 1X жлѣдеть in the same text, or OR желести) are real barriers to

-Could use нарицати as example of too coarse-grained lemmatisation in TOROT

-Psalm 84 сърѣтете tagged as present, which means it gets reconstructed as sъręħete, when it's clearly aorist 3rd dual

-Mention the problems with my class "16" verbs in the morphology-section – i.e. , PAPs in /v/ aren't very realistic for *žьńǫ, bastard Suprasliensis has PPP заклатъ (a noisome foulness), etc.
-Could talk about the impossibility of dealing with sъmęšę deviances of S-aorist sъmęsę (vs. sъmętǫ sъmętošę), since unlike with nasal-stems, the deviance-slot here is taken up with the -ox-aorists, leaving no room for deviantly-RUKI'd S-aorists

-Could use the овьсяныи OR adjective as another example of derivational-morphology made possible only be the rise of soft /s'/ (if it can be confirmed that the *ěnъ adjectival suffix in e.g. OCS оловѣнъ 'leaden' is LCS)

-възлакати would be a good one to use to talk about the unmetathesised groups like *old-, *olk- etc., because the “corpus-forms” table of my thing shows many examples of metathesised and unmetathesised forms