

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), but the sort of inflection-class-marking which would enable the retrieval of, say, masculine o-stems vs i-stems is lacking. The needs of historical phonologists especially are ill-served, since 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 Eckhoff's 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 reconstructing and inflection-class-marking the LCS lemmas, and then using a computerised LCS inflectional-morphology to inflect each text-word's lemma according to its morphology-tag¹. Such LCS reconstructions are an extremely useful form of 'phonological annotation', because theoretically all the information required to give rise to an attested form must by definition 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.

The goal of this article is to describe just such a computerised LCS inflectional-morphology, and to show how it can be used to "autoreconstruct" different OCS texts from the TOROT corpus, while explaining how difficulties caused by things like morphological innovations, badly-integrated foreign loanwords, or insufficiently-precise tagging-data can be overcome. The resulting 'phonologically annotated' texts should allow investigators of the lower-level linguistic domains (phonology, orthography, and morphology) to benefit from the huge amount of manual annotation work that has gone into the OCS part of TOROT, and the computerised LCS inflectional-morphology by itself is valuable in that it allows linguistically-rigorous and comprehensive inflection and conjugation-tables to be generated (reference to Поливанова 2013????).

A web-interface for autoreconstructed TOROT OCS text is being developed at <https://ocstexts.co.uk>, which aims to demonstrate the benefits of this extra layer of annotation by offering things like LCS phoneme-search of texts, conversion of each autoreconstructed word into canonical "normalised" OCS, for easy comparison with the manuscript-form, and computer-generated inflection and conjugation-tables for every reconstructed lemma. It also includes TOROT's existing morphology, lemmatisation, and (for three of the texts) Greek-alignment data, plus integration of the recently digitised OCS dictionaries hosted at gorazd.org.

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 1.2.

Slavonicists are in the fortunate position of having not only a widely and diversely attested set of daughter-dialects from which to reconstruct the proto-language, but also a historical written record that is so close in time to the Common Slavonic period that the far-southeastern Bulgarian dialect of Thessaloniki was considered a suitable vehicle for proselytising the proto-Czech speakers of Moravia. The hypothetical LCS system which I employ for my reconstructions is therefore mostly uncontroversial and widely agreed-upon (except in trivial matters of notation); nevertheless, a corpus-focused project like mine where the aim is to reconstruct *all* words of a text requires an exhaustiveness and rigour that leaves no room for the dodging of difficult points, so Section 1 of this article is taken up with a comprehensive exposition and justification of my chosen LCS system. Section 2 then looks in more detail at my computerised LCS inflectional-morphology and the structure of the existing TOROT data, and shows that together they can be used to produce the direct phonemic ancestor-forms of text-words, despite various challenges, including the morphological variation and restructuring which is occurring in even the earliest texts.



1. My Late Common Slavonic (LCS) phoneme-system

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:

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. *dušich* vs Suprasliensis *доуѣхъ* <*duxĕxъ; Polish *wszak* vs Supr. *вѣѣкъ*, Ru. *вѣѣк[уй]* <*vĕx-akъ), as well as the probable complete absence of PV2 in northern East Slavic² (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* <*gvĕzda, *kvĕť, 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 *сѣннѣѣ* <*svinĕkъ and *полотѣнѣѣ* <*polĕnĕka), and those of *g are not unknown: Zaliznjak (2004: 47) admits that palatalised forms of the Germanic loan *кѣнѣѣ* <*kĕnĕg- are the rule, but considers this to be a "supradialectal" word originating outside of the Novgorodian dialect-area; Галинская (2014: 10) is less convinced and adduces the form *оуѣѣрѣѣ* (cf. Russian *серьѣѣ*, commonly assumed to be an Oghur, i.e. Bulgar, Turkic loan, cognate with e.g. Kazakh *сырѣѣ*) '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 Галинская (op. cit.) points out, in all of the well-known Novgorodian forms of the pronoun *vĕxъ 'all' which supposedly show a lack of PV3 by retaining both /x/ and back/hard desinences (e.g. fem. gen. sg. *вѣѣхѣѣ* <*vĕxoĕ from letter №850), and which come from letters which otherwise correctly convey the jers (by writing <ѣ, ѣ> for *ъ and <ѣ, ѣ> for *ь), the weak-jer is always written with <ѣ, ѣ>, unambiguously suggesting a /ъ/ pronunciation. These forms therefore more likely point to a LCS doublet-form *vĕxъ 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 Галинская (2014: 14), though cf. Zaliznjak's (2004: 54) less convincing explanation of the /ъ/ in these words as an assimilation of original /ъ/ to the back-vowels of the following syllable).

Table 2: LCS Vowels after the monophthongisation of diphthongs

	Front		Back	
High	i		y	u
	í ь ĭ		y ɣ ɤ ĭ	
Mid	e ɛ		ɔ ɒ	
	ě ě			
Low	Æ			
		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	ḥ	k	g
		s	z	š	ž	x	
				č			
		l		ĺ			
		r		ř			
v				j			

1.1 Foreign sounds

In addition, the following symbols are used for 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: /ḳ ǵ x̣ f̣ ü/, e.g. in respectively кѣтъ <*ḳiṭъ, ѣмѣмонъ <*ǵeṃoṇъ, хитѣонъ <*x̣iṭoṇъ, ѡсиѣѣъ <*ij̣osif̣ъ³, and мѣро <*ṃüro. 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' (*eg̣üpỵṭъ): ѡѣѣѣѣѣ vs ѡѣѣѣѣѣ vs ѡѣѣѣѣѣ vs ѣѣѣѣѣ vs ѡѣѣѣѣѣ⁴, etc.. One might also ask whether a separate <ѣ> letter for /ǵ/ (and the writing of <к̣>/<ѣ> with the palatalisation-diacritic) could be linked to the inadmissibility in the native systems of soft [ḳʲ, g̣ʲ] sounds, and whether their replacement with regular <ѣ, ѣ> or <ѣ, ѣ> was more likely in systems with some level of native [ḳʲ, g̣ʲ] (for instance, in Rus' after the so-called Fourth Velar Palatalisation *ky, *gy, *xy > [ḳʲi, g̣ʲi, x̣ʲi], 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.

1.2 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. ѣѣѣѣѣ <*orb-ɔ vs Uspenskij Sbornik nt. acc. sg. ѣѣѣѣѣ <*ordl-o, because for

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. Where specific forms from Eckhoff's corpus-texts are cited, they are hyperlinked to that place on the ocstexts.co.uk website. Care should be taken with Eckhoff's digitisations, particularly in texts like Psalterium Sinaiticum where certain rusesome decisions from the editors Severjanov (1922) and Mareš (1997) have been compounded by further information-destroying simplifications (for instance, Severjanov transcribes <ѣ> with <ѣ>, but Eckhoff then replaces Severjanov's roundy diacritic with a titlo, leading to nonsense transcriptions like Psalm 8 <ѣѣѣѣѣ> for ms. <ѣѣѣѣѣ>). Cleaned up digitisations, using the Glagolitic originals as the base-text, and a mechanism for displaying manuscript-images, are planned for the near future.



since my LCS system is based on a point just *before* PV2 and PV3, I must also retain the * \bar{A} archiphoneme.⁶

[illegible]

- 6 It's possible to argue that the short **Ā* counterpart to **Ē* 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 <*bъċĒla, or was newly phonemicised, in e.g. Ru. вёсла <*v'Ēsla <*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 **Ē*, OCS shows no sign of anything but an /e/ reflex of short **Ā* (and indeed the fact that 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 ѿвѣнъ of the Kipchak word /jovšan/ 'wormwood' in the Hypatian Codex, whose modern cognates (Turkmen ýowšan /jowšan/, Kazakh жусан /žuwsan/, Azeri yovšan) 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 **Ē*.
- 7 In the database I will have to use the single Unicode characters <ř ř ĩ Ĭ>, rather than what's shown in my table, since the latter cannot actually be rendered without using the letters for /r ř l Ĭ/ 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 /ċ y/, which I will have to replace with <ċ ŷ>.
- 8 To my mind the only evidence in support of a genuine jer + liquid stage comes from the paradigms of verbs like OCS сѣтърѣти <*sъtŕĕti, where the syllabic /t̪/ in the stem alternates with /t̪r/ depending on the vocalicity of the following morpheme: the e.g. 3sg. pres. *sъtŕetŕ (Zogr., Supr. сѣтърѣтъ) or (one possibility of the) 3rd sg. aorist *sъtŕe (Supr. сѣтърѣ) must have /ŕe/, while the 3rd pl. aorist *sъtŕĕŕ (Supr. сѣтърѣша) and the other possibility for the 3rd sg. aorist *sъtŕĕ (Psal. сѣтърѣ, or with a different prefix Mar. сѣтърѣ <*otŕĕ), being word-final or pre-consonantal, must be syllabic /t̪/. The same alternation occurs in the zero-grade forms of verbs like *umerti, as is clear from the Polish reflexes umarł <*umŕĭł vs umrę <*umŕĕ. The argument could be that at some stage, before

1.3 Consonants - Dejotation

Reflexes of the so-called jot-palatalisation are all written either as unitary palatal phonemes, or in the case of jot-palatalised labials as /v́ ḿ b́ ṕ/, rather than as sequences of consonant + /j/, hence /ń í ř/ for *nj *lj *rj. The ‘dejotated’ reflexes of *tj (and *kt+front-vowel) and *dj are denoted using the modern Serbian Cyrillic letters /h/ and /ĥ/ respectively, because the commonly used alternatives, i.e. /t̥ d̥/ (as used in e.g. Olander 2015) or /k̥ g̥/ (as used by me in Winslow 2022), or variations thereof, are visually too close to symbols used elsewhere in the system. /k̥, g̥/ are anyway already used in my system for foreign /k, g/ before front-vowels, and /t̥ d̥/ 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 *ĥ, ĥ were close enough to foreign /g k/ before front-vowels that the original Glagolitic system used <Ѣ Ѣ> for both sets (i.e. alongside attested  < ἡγεμών would have been **ѢѢѢѢѢѢ <*osqĥeni, and alongside attested  <*dĥerĥ would have been **ѢѢѢѢѢѢ < κῆνος⁹), 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 *šh and *žh and the latter as *šč and *žž, 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:očišhenje vs. jъskati:jъščq, jъzditi:jъžhĥ vs jъzgъnati:jъžženq.

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 /h ĥ ń í ř/ but not the PV2/3 reflexes /c ś dž/, 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, j] reflexes of *tj and *dj, and also no new sequences of [tj, dj], could not convincingly 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 /h ĥ ń í ř/ 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

the LCS tendency towards Open Syllables became dominant, the stems in these paradigms were surely unitary /t̥r̥/, /m̥r̥/, i.e. 3sg. aor. /sũ.t̥r̥.re/ vs 3rd. pl. aor. /sũ.t̥r̥.š̥e/, and that the latter's closed /t̥r̥/ syllable was only forced to open itself up by changing to /t̥r̥/ 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 /t̥r̥/ to /r̥/ in these verb-forms was not merely a move to an already-existing syllabic-liquid phoneme.

- 9 Interestingly, this aspect of the hypothesised Urksl. orthographic system has rearisen in the modern Macedonian standard due to Turkish loanwords: *кемер* < Tk. *kemer* ‘belt’, *ке* < *[x̥]he[t̚]; *ѓон* < Tk. *gön* ‘leather’, *меѓу* < *meĥu.

10 Marianus and Psalterium Sinaiticum, on the other hand, frequently show a Russian-style /ej/ reflex of strong tense *bj̥j: Psal. ႁႃႆႠႩႦ <*vorbyj̥j, ႱႃႇႣႩႦ <*rlybyj̥j, ႩႮႦႩႦ <*mnybyj̥j; Mar. ႧႦႲႦႡႩႦႢႩႦ <*zarovēdyj̥j, ႩႮႦႢႩႦ <*udafy-j̥j, ႩႡႦႩႦႢႩႦႢႩႦ <*gvozdyj̥ny-j̥j. Psalm. (Psalm 21) even has a past act. part. Nsg. masc. def. form ႩႮႦႩႦႢႩႦ <*jystrgy-j̥j that suggests an /oj/ reflex of *bj̥j.

are, in slightly different terms, effectively posited by Trubetzkoy (1954: 70) in his analysis of the Urkirchenslavisch phoneme-system¹¹.

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 *ɨj/*ɨ̋j as the denotations for these groups.

Difficulties arise though when deciding how to denote foreign sources of /ij/¹² which may or may not have been integrated into the native system as reflexes of /ĭj/: words like **мариѡ** < Μαρία, **стадиѡ** < 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 *marijĀ, *stadijĀ, or as nativised *marɨjĀ, *stadɨjĀ, but there are no occurrences of jer-spellings in these words in the OCS texts in TOROT. Other similarly-Greek words like **дѡбѡлѡ** (< διάβολος), however, do show up in OCS with jer-spellings: Supr. **дѡдѡлѡ**, Zogr. Luke 8 and Psal. Psalm 108 **дѡдѡлѡ**, which (alongside the modern Macedonian *ѓавол* with the reflex of *ɨ̋j produced by the Macedonian so-called 'new jotation' of /d/ after the fallen jer brought it into contact with /j/) clearly suggest an early adaptation of this foreign /ij/-group to native /ĭj/. Old Russian texts even show spellings of **мариѡ** suggestive of full nativisation: Laurentian Primary Chronicle **мѡрьѡ**, **мѡрьѡ**, Zadonshchina **мѡрьѡ**, **мѡрьѡ**, as well as First Novgorod Chronicle gen. sg. **вѡснѡлѡ** (jo-stem **вѡснѡлѡ** < 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 *ɨj occurred: *diĭjĀvolĀ, *vasilijĀ, *marijĀ etc.

1.4.3 Word-initial *jɨ-/ *ji-/ *i-

With native Slavic word-initial *ji-/ *jɨ-, I follow Derksen's (2008: 16) practice of writing *jɨ-, even though Derksen himself (2003) has argued for a split between *ji- and *jɨ- 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 *jɨdɔtĭ just like any other and retained the /j/, and Ukr. *ськати* < *jɨskati (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 *ɨsk groups (cf. *польський*).

I make an exception for certain forms of the personal-pronoun *jɨ, however, and write *jimĭ, *jima, *jixĭ *jimĭ 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 **ѡѡковѡ** (borrowed via Gk. *Ἰακώβ*), or **ѡѡмонѡ** (< 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-native¹³ (cf. my treatment of foreign initial *e- below). An exception is made for **ѡѡсѡсѡ** < Gk. *Ἰησοῦς*, which I have as *jisusĭ, because of the greater likelihood that Slavs will have heard of Jesus even before the first biblical translations, and because spellings like Zogr. **ѡѡсѡсѡ** suggest that it causes the same /ĭ̋/

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 **прѡѡѡти** < *priĭeti or long-form adjectives like masc. nom. pl. **дрѡѡѡни** < *drugĭ-jĭ, but within roots it does seem restricted to these post-LCS loanwords.

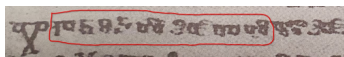

13 Spellings like Zogr. Mark 13:3 "**ѡѡсѡсѡ. ѡ ѡѡѡѡѡ. ѡ ѡѡѡѡѡ.**" "Peter and Jacob and John" would suggest that this initial *i- can get dropped after an /i/ of a preceding word, but whether this points to a dropping of the non-native *i-, simple deletion of a double /i/ (haplology), or a native-like reflex of a weak-jer /*i *jɨjĀkovĭ / > /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.

Prefixed forms like *do-jyti '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 *дої́му* (<*dojyti with compensatorily-lengthened /o/ > /i/) vs *доїму*. 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 (шуйица) and *vojinyъ 'warrior' (воинъ, as opposed to *vojnyъ, the gen. pl. of *vojъna), or the loc. sg./pl. desinences of any jo-stem noun whose stem ends on /j/, e.g. Psal. *жєбєѣѣ* <*žerbjji.

No Glagolitic text makes any effort to distinguish /je/ (after vowels or word-initially) from post-consonantal /e/, writing both with <ѣ>, unlike the situation with the reflexes of *jē vs *ĕ, where in Zogr. and Mar. and partially in Assem. (Велчева 1981: 168) the full front-nasal digraph <ѥ> is reserved for *jē, while just the second 'nasalising component' <є> is used for post-consontal *ĕ, e.g. Mar. 3rd pl. aorist ѡѣсѧ <*jese, as opposed to KF ϣϣѢѤѦѨ <*prijeti vs ϣϣѠѢѤѦѨ <*vъzeli¹⁴. 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 jotted <ѿ> 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*, *evanĝelъje*, *eĝŭрътъ*, *elisavetъ*, *elinъ*, *evanĝelistъ*, *eĝŭрътъskъ*, *elinъskъ*, *episkupъstvo*, *evrejskъ*, *elisejъ*, *етѣтауsъ*, *etiјоръskъ*), the only spellings with <ѿ> are ѿлиси, ѿппъ, ѿлини, and ѿлина, i.e. 4/157 or 2.5%. By contrast, of the 3172 native Slavic words in Suprasliensis 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 <ѿ>, vs 3070 with <ѿ>¹⁵. Thus 97.2% of native word-initial *je- in Suprasliensis is spelt with <ѿ>, while 97.5% of the occurrences of the clearly post-LCS Greek-mediated foreign borrowings listed above instead use plain <ѿ>, suggesting that some sort of difference was felt, at least by the scribes of Suprasliensis, and that we probably shouldn't index these with the same *je- as used for native forms. I therefore use non-jotted *e- for such foreign borrowings, and the extent to which they take prothetic *j- and fall together with the native vocabulary is left as something for investigators to determine based on the evidence of each manuscript.

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 */*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 гъзѣнѣти <*gyb-nqti > Ukr. *гунути*, vs OCS оуѣзѣнѣти <*usъp-nqti (cf. 3sg. aor. *оуѣзѣ*), Ru. *тонуть* <*top-nqti, though see Meillet (1965: 142)). Geminate consonants were banned and either simplified (*нѣшѣти* <*jъs-sěkti) or dissimilated (*процвѣсти* <*prokvit-ti).

14 Psalterium Sinaiticum contains at least five occurrences of non-digraph <e> for *ḡ: Eckhoff's digitisation suggests ⲉⲛⲁⲙⲓⲕ, ⲃⲱⲙⲉⲗⲏⲥ, ⲟⲩⲧⲉⲣ, ⲓⲛⲉ, ⲫⲉ, and ⲛⲉⲛⲁⲙⲓⲕⲟⲩⲙⲉ, but the last one ⲛⲉⲛⲁⲙⲓⲕⲟⲩⲙⲉ (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 Tarnanidēs (1988: 260) clearly shows ⲛⲉⲛⲁⲙⲓⲕⲟⲩⲙⲉ:



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

common also to the Germanic and sometimes Baltic cognates. The examples given by Meillet (1965: 136) include: (for $\sigma\tau\rho\omicron\gamma\tau\alpha$) Lith. dial. *srauja* next to Latvian *strauja*, then Germanic **straum-* (> Eng. *stream*, Old Norse *straumr* etc.); (for $\sigma\tau\rho\tau\epsilon$) Lith. *aštrus*, Gk. *ἄκρος* (here the **s* is from PIE **k*). As Meillet says, “*ce n'est pas un développement germano-balto-slave; d'une part, le développement d'un -t- dans le groupe sr est chose naturelle et se retrouve ailleurs (fr. pop. castrole de casserole) et, d'autre part, le développement de t en ces conditions n'est pas général en baltique: str est régulier 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. срѣдѣ <**sorm-omъ* spelling cited by Diels (1963: 122); otherwise new /sr/ from metathesised **sErC* groups is tolerated unchanged.

New occurrences 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-*, **jъz-* (e.g. Supr. 3sg. aor. взвѣдоу ‘roared’, from **vъz-ruti*), but also because of the clitic prepositions **jъz* and **bez*, which form one phonological word with whatever follows them and thus cause OCS spellings like Mar. Luke 1 срѣдѣ <**jъ.z ър.кѣ*. Meillet (p.136) also cites the Old Polish adverb *zdręki* <**jъz ър.кѣ*, 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 < зрѣкѣ > and so seems to be tolerated, even though Diels (p.122) cites prepositional forms like Supr. вездѣ <**bez ъorzuma*, вездѣ ¹⁶ <**bez ъordla*, 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 срѣдѣ (Diels 1963: 122). In Old Russian, the Uspenskij Sbornik is pretty consistent in keeping prefixed verb-forms like разрѣшитѣ <**orzrúšiti*, but by the time of the Laurentian Codex we get forms like взрѣшитѣ and нѣзрѣшитѣ . 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 ъdrъky/ can be identified with separate /bez/ and /rъky/.

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. срѣдѣ <**jъs-seče*, Mar. срѣдѣ <**jъs-šьdъ*, Assemanianus срѣдѣ <**ors-širĕjqtъ*, because that same drive towards restoration of the underlying shapes of the prefixes **jъs-/ors-* etc. can be seen in modern Russian *иссякнуть*, *расширять*, and Laurentian Codex расширѣши .¹⁷ This treatment is

16 For some baffling reason this form is missing from Eckhoff's text, so I link instead to the relevant folio of David Birnbaum's online edition.

17 Conversely, sequences of **sk*, **zg* at prefix-boundaries which show PV1 reflexes, like Mar., Zogr. срѣдѣ <**orš-čьtetъ* (ECS **skit-* > **ščit-*), Psal. срѣдѣ <**orž-žigajetъ* (ECS **zgig-* > **žžig-*) are kept as **šč*, **žž*. 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/ (> /žž/ and /šč/) that remained active until much more recently, especially given prepositional-phrase forms like Psal. срѣдѣ <**jъs-červa*, so this is arguably inconsistent with my treatment of **ss*, **sš* etc. My justification is firstly that **sk*, **zg* > **šč*, **žž* 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 <жж> spellings in compounds like *исчезнуть* and *разжечь*, the pronunciations are still arguably direct reflexes of LCS **šč* and **žž*, viz. [s:] and [z:] (or, in the conservative Moscow-dialect, the palatalised [z:] found also in *дожечь* <**dъžhъ*).



also more consistent with my handling of verbs like *jъs-kělitī (> OCS *ицѣлити/исцѣлити*) 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 *ѡцѣлѣ-* while Mar. and Psal. keep *ѡцѣлѣ-* (more discussion of the wider Slavic reflexes of this group, including the OCS <ѣ> spellings, can be found in Meillet 1965: 133).

1.6 Morphological innovations that scupper LCS reconstruction

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-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 possible 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/¹⁸. 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 *tatъ is a masc. i-stem noun with genitive *tati, as it still appears in the Suprasliensis translation of John Chrysostom's Homily for Holy Thursday (...*то кажетъ владѣицы ѿловѣколюбѣѣ ꙗко прѣданныка 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. **ТАТЪ** /tat'/) of this noun now ends on the same class of "soft" consonants as original jo-stem nouns like *koň > /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 *bornъ > /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)¹⁹. Consequently, a word like **ТАТЪ** has begun to take jo-stem endings, including the Old Russian /a/ reflex of LCS *Ā in the genitive/accusative singular.

LCS /Ā/, though, by definition can only occur after LCS palatal consonants (see above), so a reconstruction *tatĀ is just nonsensical. In the case of the dat. sg. /u/-desinence (which isn't attested in our Treaty but exists in modern Russian *матю*), 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 *zvěрь, then the structural change would not be so catastrophic, because our LCS system *does* contain a palatal *ř which any allophonically-softened LCS hard *r could easily be subsumed into. Indeed, interestingly

18 In some dialects (notably East Slavic) the PV3 reflexes *ś and *ž became palatalised 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 *ъхѣ, for example, turns up exclusively <ѣ> spellings in Marianus, just one <ѣ> in Zogr., and exclusively <ѣ> in Suprasliensis, with only Assem. and Psal. containing a significant minority of <ѣ> spellings.

19 Russian feminine i-stems like *весь* (<*vьсь, 'village') do not fall together with ja-stems in the way the masculines like *zvěрь 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. *вѣсям*, which contain the same LCS ja-stem *-Ā- vocalism which can only occur after LCS palatals, meaning they too end up totally unreconstructable due to an illegal **sĀ sequence.

Suprasliensis does in fact contain 3X gen. sg. **3r̥s̥p̥t̥s̥**, with what looks like a jo-stem reflex of *r̥Ā (spelt with jat' as an overhang of the Glagolitic tradition, cf. 2X **морѣ** vs 1X **морѣа** spellings), suggesting that Russian-style secondary palatalisation of *r > /r'/ may have occurred in the Bulgarian dialect underlying it²⁰. You don't, though, get anything like **ТАТА**²¹, so the system-wide development of secondary-palatalisation does not seem to have advanced enough to have caused the sort of fundamental structural reorganising which shifted *tatъ into the jo-stems in Old Russian.

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 ဖုခိဖုအယတဲဒဲး is given below:

Figure 1: TOROT annotation for Psal. Sin. Psalm 70 ဘုရားရှင်အရှင်အေးမြတ်စွာ၏အမည်ကို

A detailed explanation of each feature can be found in Section 6 of Eckhoff et al. (2018: 41), but here it suffices to say that in this example the tag “1spia---i” is telling us that ལ་མེད་ལ་འཁྱུག་པ་ is 1st

21 Except the numerous gen. sg. *господа* and dat. sg. *господѣ* for the i-stem **gospodъ*, but this word seems to be an isolated special case, because it bafflingly turns up even in early Glagolitic OCS with endings like *ꙗꙋꙋꙋ* (ju-stem dat. sg. *-*evi*, cf. Supr. *господаѣви*), *ꙋꙋ* (jo-stem dat. sg. *-*u*), and *ꙋ* (jo-stem gen. sg. *-*Ā*). See Van Wijk (1929).

23 The spreadsheet containing my reconstructions and inflection-class annotations for the TOROT OCS lemmas can be downloaded from https://github.com/12401453/torot_2023/blob/main/lemma_lists/chu_lemmas_master.xlsx

person, singular, present-tense, indicative-mood, active-voice, has no gender, case, degree, or strength features, and is inflextable rather than non-inflecting.

Of importance here is the *present* tense tagging, even though вззвѣтити (even in OCS) can be taken as a perfective verb, opposed to its imperfective counterpart вззвѣтити , and thus has future-tense meaning in its non-past indicative forms (and the Greek Septuagint here has $\text{ἀναγγεῖλῶ τὰ θαυμάσιά σου}$, with a morphologically future-formed ἀναγγεῖλῶ ‘I will proclaim’ from ἀναγγέλλω). The tagging thus follows the *inflectional*-morphology of вззвѣтити , rather than the future-meaning which is carried by the *derivational*-morphology. This is important because the Autoreconstructor works by *inflecting* LCS lemmas according to those morphology-tags; the annotation gives no information about its derivational-morphology or derivational relationship to its imperfective (and thus 1sg. *present* tense) counterpart вззвѣтити , since that is annotated with a separate lemma.

An even clearer example where this *form over function* annotation helps us is with the accusatives of animate nouns: it’s well known that a type of so-called differential-object-marking is beginning to take hold in Early Slavic, whereby syntactic accusatives use genitive endings to varying extents as a way of encoding the semantic **ensouledness** (and possibly definiteness) of the noun (the details aren’t important; for a recent thorough diachronic treatment of the topic see Eckhoff 2022), e.g. Supr. разбойника въ парадѣ ‘he brought the robber into paradise’. If these were all tagged as accusatives, the Autoreconstructor would produce the wrong ancestor to the text-form, i.e. *orzbojnik-ъ , because the semantic information needed to decide whether this differential-object-marking is needed is not available. Happily though all such animate-accusatives are marked as genitive in TOROT: the morphology-tag for разбойника above is “-s---mg--i”, so the Autoreconstructor produces *orzbojnik-a .

2.2 Computerised LCS inflectional-morphology and the Autoreconstructor

The Autoreconstructor reads the morphology-tag character-by-character and numerifies each field, and from that it computes a number which corresponds to the row of the inflection-table where the endings for that particular word’s inflection-class are stored. For verbs this will be a number between 1 and 44 (9 person/number combinations for each of present, aorist, imperfect, and imperative, plus 8 non-finite forms, $9 \times 4 + 8$), and for nominals between 1 and 63 (7 cases \times 3 numbers \times 3 genders). A version of the function which actually implements this tag-reading process can be seen in the `OcsServer::numerifyMorphTag()` function here:

https://github.com/12401453/ocs_server/blob/main/OcsServer.cpp#L2714.²⁴

Currently each inflection-class has up to three tables associated with it, the first of which is full (i.e. contains 44 or 63 entries) and holds the ‘basic’ or ‘correct’ (from the LCS perspective) endings. The second and third tables are ‘sparse’, in that they only contain entries for those parts of the paradigm where we expect to encounter alternative forms, with those I consider ‘deviant’ or ‘innovated’ held in table 2, and ‘alternative’ but still ‘correct’ endings (i.e. LCS allomorphs) in table 3. An example of some of the endings in the three tables for basic class 1 verbs like *rehi is given below:

²⁴ Extra handling is required for forms of *byti , since that has a separate future-paradigm (and future-participle) which TOROT does actually specify separately with an ‘f’ value for the *tense* feature, as well as two variant imperfect sets (3sg. *bě vs. *běaše , which aren’t tagged, but which I detect), and a ‘conditional’ *bimь , *bi , *bq etc. Participles require an extra step to read their nominal-features and add their adjective-like endings.

Full paradigms for the 4,500-ish OCS lemmas I have so far reconstructed can be dynamically generated here <https://ocstexts.co.uk/words>; these are constructed from LCS lemmas in the same way the Autoreconstructor reconstructs individual forms, except it produces every form in the paradigm, rather than just the one specified by a text-word's morphology-tag. The LCS forms are then converted into 'normalised' OCS by the browser using the `convertToOCS()` function here: https://github.com/12401453/ocs_server/blob/main/HTML_DOCS/LCS_to_OCS.js#L197.

☐ Corpus forms

	Present	Aorist	Imperfect	Imperative	Participles	
1st sg.	сѣтърѣ	сѣтърѣхъ	сѣтърѣахъ	сѣтърѣми	PRAP ¹	сѣтърѣи
2nd sg.	сѣтърѣши	сѣтърѣ сѣтърѣ	сѣтърѣаше	сѣтъри	PRAP ²	сѣтърѣци
3rd sg.	сѣтърѣтъ	сѣтърѣ сѣтърѣ	сѣтърѣаше	сѣтъри	PAP	сѣтърѣ
1st du.	сѣтърѣвѣ	сѣтърѣховѣ	сѣтърѣаховѣ	сѣтърѣвѣ	L-Part.	сѣтърѣа
2nd du.	сѣтърѣта	сѣтърѣта	сѣтърѣашета	сѣтърѣта	PPP	сѣтърѣтъ сѣтърѣнѣ
3rd du.	сѣтърѣте сѣтърѣта	сѣтърѣте сѣтърѣта	сѣтърѣашете сѣтърѣашета	сѣтърѣте	PrPP	сѣтърѣомѣ
1st pl.	сѣтърѣмѣ	сѣтърѣхомѣ	сѣтърѣахомѣ	сѣтърѣмѣ	Infinitive	сѣтърѣти
2nd pl.	сѣтърѣте	сѣтърѣте	сѣтърѣашете	сѣтърѣте	Supine	сѣтърѣтъ
3rd pl.	сѣтърѣтъ	сѣтърѣа сѣтърѣша	сѣтърѣахъ	сѣтърѣа		

Figure 4: Dynamically-generated paradigm for the OCS verb *сѣтъри*, based on the Autoreconstructor's computerised LCS inflectional-morphology

The accuracy of the generated-forms can then be gauged by comparing them to tables populated only by forms which actually occur in Eckhoff's corpus-texts, using the 'Corpus-forms' switch:

now I've left all such adjectives wholly uncontracted (though the Autoreconstructor does actually mark long-adjectivals that contain such problematic concatenations and this information could be used to exclude them from searches).

☒ Corpus forms

	Present	Aorist	Imperfect	Imperative	Participles	
1st sg.					PRAP ¹	
2nd sg.					PRAP ²	
3rd sg.	сѣтърѣтъ	сѣтърь			PAP	сѣтърьъ
	сѣтърѣтъ	сѣтърьѣ				сѣтърьъши
1st du.					L-Part.	сѣтърьъаъ
2nd du.					PPP	сѣтърьени
3rd du.					PrPP	
1st pl.				сѣтърьѣмъ	Infinitive	
2nd pl.					Supine	сѣтърьтъ
3rd pl.		сѣтърьша				

Figure 5: Forms of the same сѣтърѣти verb as they actually occur in Eckhoff's (Cyrillicised) TOROT corpus of Church Slavic texts

An advantage of basing computer-generated paradigms on converted LCS (rather than directly on normalised OCS) is that it allows for the possibility of dialect or manuscript-specific conversions: e.g. Kiev Folia-flavoured paradigms that convert **h*, **h̃* to <ц, з> and use <ѣ> for all **Ē*, or a Marianus-specific conversion that frequently shows vocalised strong-jers, or even an artificial pre-Jer Shift form of Old Russian that shows the denasalisation of **ę* via mixing up of <ѣ> and <а>:

	Sing.	Dual	Plural
Nom.	КЪНІАЗЬ КЪНАІГЪ	КЪНАЗІА	КЪНАЗІИ
Acc.	КЪНАЗЬ КЪНАІГЪ	КЪНАЗА	КЪНІАЗѢ КЪНАІГЪ
Gen.	КЪНАЗІА	КЪНАЗІЮ	КЪНІАЗЬ КЪНАІГЪ
Dat.	КЪНІАЗІЮ	КЪНІАЗЕМА КЪНАІГОМА	КЪНАЗЕМЪ КЪНАІГОМЪ
Loc.	КЪНІАЗІИ КЪНІАЗѢ	КЪНАЗІЮ	КЪНІАЗІИХЪ КЪНАЗѢХЪ
Instr.	КЪНАЗѢМЬ КЪНАІГѢМЬ	КЪНАЗЕМА КЪНАІГОМА	КЪНІАЗІИ КЪНАІГЪ
Voc.	КЪНІАЖЕ	КЪНІАЗА	КЪНІАЗІИ

Figure 6: Dynamically-generated paradigm of **кѣнегъ* but with an Old Russian rather than OCS conversion

	Sing.	Dual	Plural
Nom.	КЪНАСЬ КЪНАІГЪ	КЪНАСА	КЪНАСИ
Acc.	КЪНАСЬ КЪНАІГЪ	КЪНАСА	КЪНАСА КЪНАІГЪ
Gen.	КЪНАСА	КЪНАСОУ	КЪНАСЬ КЪНАІГЪ
Dat.	КЪНАСОУ	КЪНАСЕМА КЪНАІГОМА	КЪНАСЕМЪ КЪНАІГОМЪ
Loc.	КЪНАСИ КЪНАСѢ	КЪНАСОУ	КЪНАСИХЪ КЪНАСѢХЪ
Instr.	КЪНАСЕМЬ КЪНАІГОМЬ	КЪНАСЕМА КЪНАІГОМА	КЪНАСИ КЪНАІГЪ
Voc.	КЪНАЖЕ	КЪНАСА	КЪНАСИ

Figure 7: **кѣнегъ* with the canonical OCS conversion for comparison

2.3 Detecting and dealing with morphological innovations

217066	приведоша	3paia----i	privedošę	privedo
217101	въздѣхнѣвъ	-supamn-si	vъzdѣxnѣvъ	vъzdѣxъ
217112	разврѣзосте	3daia----i	orzvrѣzoste	orzvrѣzete
217261	бѣгивѣвъ	-supamn-si	bolgoslovivъ	bolgoslovilъ
217266	ѣша	3paia----i	jѣšę	jѣšę
217272	възаша	3paia----i	vъzęšę	vъzęšę
217302	начаша	3paia----i	načęšę	načęšę
217316	въздѣхнѣвъ	-supamn-si	vъzdѣxnѣvъ	vъzdѣxъ
217455	приведоша	3paia----i	privedošę	privedo
217600	начать	3saia----i	načęť	načę
217605	сѣноу	-s---md--i	synu	synovi
217635	начать	3saia----i	načęť	načę
217787	поѣтъ	3saia----i	pojęť	poję
217832	моѣомѣ	-s---mi--i	mosijomъ	mosijemъ
217856	моѣови	-s---md--i	mosijovi	mosiju
217869	быс	3saia----i	bystъ	by
217960	сѣнѣ	-s---ml--i	synę	synu
218067	невѣрьнѣ	-s---mvpsi	nevęrъnъ	nevęrъne
218108	быс	3saia----i	bystъ	by
218173	нѣмѣ	-s---mvpwi	nęmъjъ	nęmęjъ
218175	глоухѣ	-s---mvpwi	gluxъjъ	glušejъ
218198	быстъ	3saia----i	bystъ	by
218206	оумрѣтъ	3saia----i	umerť	umer
218388	ѣмени	-s---nl--i	jъmeni	jъmene
218419	ѣмени	-s---nl--i	jъmeni	jъmene
218561	окомѣ	-s---ni--i	okomъ	očesъmъ
218648	моѣю	-s---md--i	mōžu	mōžęvi
218688	моѣжа	-s---mg--i	mōžę	mōžu
218755	прѣлѣубѣ	-s---fa--i	perluby	perlubъvъ
218762	поустивѣши	-supafn-si	pustivъši	pušћъši


Figure 6: 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 of class 1 verbs: 3rd pl. **приведоша** vs. **приведѣ**, 3rd dual **разврѣзосте** vs. ***разврѣзете**²⁷
- unetymological extension of the RUKI-rule-produced *š in 3rd pl. primary sigmatic aorists: **ѣша** vs. **ѣса** <*jĕd-s-ę, **възаша** vs. **възаса** <*vъzym-s-ę, **начаша** vs. **начаса** <*načъn-s-ę (neither *d, *m, nor *n have ever been RUKI sounds)
- extension of the *-nq- suffix to the past. act. part. of class 2 verbs like **въздѣхнѣти**: **въздѣхнѣвъ** (cf. Mar. **въздѣхнѣ** from **оуздѣхнѣти**)

27 Koch (1990: 293) lists only sigmatic aorists as possibilities for the *-verz-/*-vřz- stem verbs, and it seems that outside of the 3rd sg. (e.g. Psal. **бѣгоуѣѣѣѣ**, Zogr. **ѣгоуѣѣѣѣѣ**) no root-aorists are attested in any Slavic text, so maybe I am wrong to set up asigmatic root-aorists like 3rd dual *-vřzete as a possibility alongside primary sigmatic *-verste (e.g. Mar. Mark 7 **бѣгоуѣѣѣѣѣѣѣ**). My justification is that the *-verg-/*-vřg- stem verbs *do* attest such root-aorists, e.g. 3rd pl. **ѣгоуѣѣѣѣѣѣѣѣѣ** 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.

- addition of the *-tQ suffix from the 3rd sg. pres. (see fn. 5 above) to 3rd sg. aorist forms: 
 НАУАТЪ, ПОАТЪ, ОУМРЪТЪ
- original u/ju-stem nouns taking o/jo-stem endings: dat. sg. ѿноу, мѣжо; loc. sg. ѿнѣ, gen. sg. мѣжа
- past act. part. of class 4 verbs using the suffix *-ivъ rather than *-jъ: бѣгивѣ, поустивѣши (cf. Mar. Mark 10 [ⲓⲁⲩⲱⲥⲱⲩⲩⲥ](#) <*pust-jъsi)

Deciding upon the “correct” morphological endings for an unattested language inevitably entails some uncertainty and controversy: for instance, *-ox- aorists occur in both OCS and Old Russian, so why do I consider them LCS deviances? Basically because they **never**²⁸ occur in Marianus, which would be quite improbable if they were a discarded archaism (especially given their ubiquity in the closely related Zographensis).

In other cases we are dealing with hodge-podge paradigms which are only ever attested with endings from multiple older classes, and sometimes dialectal or orthographic features of the manuscripts can make it difficult to distinguish potential LCS ancestors of certain endings. For instance, I have a *masc_tel* class used for agent-nouns like OCS [ДѢЛАТЕЛѢ](#) (i.e. Diels 1963: 166), which in the sg. and dual. behave exactly like masc jo-stems, but which in the gen. and instr. plural are attested also with basically consonant-stem endings on a hard *-tel- stem, e.g. Zogr. [ⲉⲧⲉⲗⲉⲧⲉⲗⲉ](#) <*tęžĚtelъ, Supr. [СВѢТИТЕЛѢ](#) <*svĕtitylъ. The nom. pl. appears also to take a consonant-stem *-e ending, but as Diels (op. cit.) points out, the spellings in Zogr. and Supr. (where use of the palatalisation-diacritic <^> to denote LCS *ĭ is consistent enough to suggest a real phonemic /ĭ/ in the underlying dialects) like [МѢЧИТЕЛѢ](#)²⁹ mean we can’t follow Meillet (1965: 426) in setting up *-tele as the ‘correct’ ending, because spellings like [ⲉⲧⲉⲗⲉⲧⲉⲗⲉ](#) in Mar. could just as easily descend from *žetelē as from *žetele. Absent a manuscript which both consistently marks *ĭ and doesn’t use such a mark in this nom. pl. desinence, there’s no hard evidence of this *-tele ending ever actually existing. I thus use *-telē in the ‘correct’ table, and the jo-stem *-telī in the ‘deviances’ table³⁰.

2.4 Overcoming poor lemmatisation practice

Sometimes Eckhoff’s lemmatisation is too coarse-grained, in that forms which clearly descend from distinct doublets are subsumed under one lemma. To demonstrate just one example of how the Autoreconstructor deals with this sort of problem, take the numeral [ЈЕДИНЪ](#) <*jedinъ, which in the earliest OCS has straightforward hard pronominal endings and which therefore goes in my *pron_hard* class alongside demonstratives like *онъ. There is, however, what must be viewed as an LCS doublet *jedъnъ³¹, which gives e.g. Serbian *jedan* and the modern Russian fem. nt. and oblique-case forms *одна*, *одного* etc., and which is used for the majority of non masc. nom./acc. sg. forms of the pronoun in Suprasliensis³², e.g. [ЈЕД’НОМОУ](#). Notwithstanding Eckhoff’s habit of

28 Having [Autoreconstructed](#) all of Marianus I can verify this (admittedly already well-established) fact far more quickly and easily than was possible just with Eckhoff’s morphology and part-of-speech annotation-information: using TOROT the smallest net you could cast would be one that caught all aorist-tense verbs, whereas I can just search for reflexes of *охъ, *охов, *охом, *оѣ, *osta\$, and *oste\$ (the latter two using ‘regular-expression’ mode and \$ to specify end-of-word), which for Mar. turns up nothing except the three occurrences of [ⲉⲩⲱⲩⲥ](#).

29 Searching my database for *telē\$ returns only a single [ВЛАСТЕЛѢ](#) in Supr. without the diacritic; everything in Zogr. has it.

30 Diels mentions only Psal. Psalm 26 [ⲉⲧⲉⲗⲉⲧⲉⲗⲉ](#), but the Autoreconstructor is intended for use with other early texts beyond canonical OCS which might also contain this type of assimilation to the jo-stems.

31 This despite Meillet’s (1965: 144) incoherent speculation about the /i/ in *jedinъ resulting from a phonetic development of strong *ъ, analogous to what happens in the vicinity of *j, because “or il s’agit d’un composé dont le second élément est *jīnŭ”. While this interpretation of *jedinъ’s derivation could be true (and according to Derksen’s (2008: 212) etymology of the *jъnъ pronoun, its *j is prothetic, meaning it wouldn’t develop if already attached to a stem ending in *d), the idea that a synchronic *-дъnъ by any purely phonological means could become /-din/, let alone early enough to completely displace by analogy the oblique forms in the earliest OCS where [ⲉⲩⲱⲩⲥ](#)-spellings are overwhelming, is ludicrous and contradicted by all the philological evidence.

32 According to my database there are 145 reflexes of *jedъn- in this category vs 46 from *jedin-, though all 107 reflexes of the masc. nom./acc. sg. are from *jedinъ. Interestingly the Uspenskij Sbornik, in contrast to later

lemmatising these with *ѣДИНЪ* instead of *ѣДЫНЪ*, it's trivial for the Autoreconstructor to check the form (which has in a previous step already been aggressively normalised to get rid of morphologically-irrelevant variation) for a <ДИН> sequence, which would never occur in the inflexional-ending and so always point to a *jedin-descended stem, and then replace our autoreconstruction with *jedьн- if such isn't found:

```
// check for *jedьн-
if (lemma_ref.lemma_id == compileTimeHashString("Рхѣдинъ") || lemma_ref.lemma_id == compileTimeHashString("Маѣдинъ"))
{
    if (Sniff(cyr_id, "дін", 20) == false)
    {
        stem = "jedьн";|
    }
}
```

Figure 7: Part of the Autoreconstructor's code which checks for reflexes of *jedьн- that TOROT has mistakenly lemmatised under *ѣДИНЪ*

In the long-term TOROT itself should fix such lemmatisation problems, but in the meantime checks like the above are computationally extremely cheap and prevent great swathes of the texts from being wrongly autoreconstructed.

3. Conclusion and prospects for future work

The foregoing article has laid out a method for leveraging existing morphosyntactic annotation-data to produce rigorous phonological annotations of Old Church Slavonic texts, and in so doing to make the results of large-scale corpus-building efforts like PROIEL and TOROT useful to historical phonologists, morphologists, and textologists, rather than just to "linguists" in the narrow cognitive-science sense of the word. Both the resulting annotations and the computational apparatus used to produce them have value not only for researchers, but also in their potential for creating innovative pedagogical tools for learners.

In subsequent articles I intend to show the usefulness of such annotations by using them to holistically and comprehensively test the predictions implied by previous researchers into the sound-systems of early Slavic, because for certain questions in this field there are simply too many data-points for exhaustive manual investigations to be feasible. One thinks for example of the so-called Jer Umlaut proposed by Jagić (REFERENCE ???), which holds that OCS texts show a tendency for weak back-jers to be fronted when followed by a front-vowel, or of the theory advanced by Totomanova (2014: 14-28) about a purported merger of the jers in the earliest Bulgarian after palatal and palatalised consonants, parallel to the better-known Middle Bulgarian merger of the nasals in the same environment. An LCS phoneme-index that allows the immediate retrieval of all the spellings of the relevant sound-groups is exactly what's needed to examine the correctness of such theories.

Though important OCS texts are still lacking the sort of annotation-data needed for the method of autoreconstruction outlined here, it should be noted that recent advances in neural-network-based tagging and lemmatisation (e.g. Rabus & Besters-Dilger 2021) mean that modern taggers trained on the existing manual-annotated corpus will perform extremely well on the remaining canonical OCS texts, which massively cuts down the time and effort required to do for, say, the Savinna Kniga what has already been done for the Marianus³³. As the volume and quality of training-data increases, the performance of such automatic taggers can only improve, which will open up more and more of the early Slavic written record to the sort of automatic 'phonological annotation' described here.

Russian, appears to completely lack *jedьн-, though all except the single *о.а.нн.о.н* are spelt with the OCS <ѣ/ѣ> reflex of initial *je-.

33 The Codex Assemanianus (Cyrillicised from Jouko Lindstedt's ASCII-encoded version) is included on octestexts.co.uk with a wholly automatic lemmatisation, tagging, and autoreconstruction that used a (slightly improved version of) the outdated method detailed in Berdičevskis, Eckhoff & Gavrilova (2016), and as shoddy as the autoreconstructions often are one can still glean useful insights from searching around in it.

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.
- Bethin, Christina. 1998. *Slavic prosody: Language change and phonological theory*. New York.
- Derksen, Rick. 2003. Slavic *j_b-. In Schaeken, Jos & Houtzagers, Peter & Kalsbeek, Janneke (eds.), *Dutch contributions to the Thirteenth International Congress of Slavists, Ljubljana: Linguistics*, 97-105. Amsterdam.
- Derksen, Rick. 2008. *Etymological Dictionary of the Slavic Inherited Lexicon*. Leiden.
- Diels, Paul. 1963. *Altkirchenslavische Grammatik*. 2. Aufl. Heidelberg.
- Durnovo, Nikolaj N. 1929. Mysli i predpoloženiia o proisxoždenii staroslavjanskogo jazyka i slavjanskix alfavitov. *Byzantoslavica* 1. 48-85.
- Eckhoff, Hanne Martine & Berdičevskis, Aleksandrs. 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta* 14-15.
- Eckhoff, Hanne & Bech, Kristin & Bouma, Gerlof & Eide, Kristine & Haug, Dag & Haugen, Odd Einar & Jøhndal, Marius. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation* 52. 29-65.
- Eckhoff, H.M. 2015. Animacy and differential object marking in Old Church Slavonic. *Russian Linguistics* 39(2), 233-254.
- Eckhoff, Hanne Martine. 2022. A Long-Haul Change: Differential Object Marking in Early Slavonic. *Journal of Historical Syntax, Volume 6, Article 8*, 1-40.
- Haug, Dag T. T. & Jøhndal, Marius L. 2008. 'Creating a Parallel Treebank of the Old Indo-European Bible Translations'. In Caroline Sporleder and Kiril Ribarov (eds.). *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, 27-34.
- Koch, Christoph. 1990. *Das Morphologische System des Altkirchenslavischen Verbums. I: Text*. München.
- Kortlandt, Frederik. 1979. On the history of the Slavic nasal vowels. *Indogermanische Forschungen* 84. 259-272.
- Le Feuvre, Claire. 1993. The Sound Change e > o in the Birchbark Letters of Novgorod and T. Fenne's "Manual" and the N.sg m. Ending -e. *Harvard Ukrainian Studies* 17(3/4). 219-250.
- Mareš, F.V (ed.). 1997. *Psalterii Sinaitici pars nova: monasterii s. Catharinae codex slav. 2/N*. Wien.

- Mathiesen, Robert. 2014. A new reconstruction of the original Glagolitic alphabet. In Flier, Michael S. & Birnbaum, David J. & Vakareliyska, Cynthia M. (eds.), *Philology broad and deep: In memoriam Horace G. Lunt*, 187–213. Bloomington.
- Meillet, Antoine. 1965. *Le slave commun. Seconde édition revue et augmentée avec le concours A. Vaillant*. Paris
- Nakonečnyj, Mykola F. 1962. Do vyvčennja procesu stanovlennja j rozvytku fonetyčnoï systemy ukraïns'koï movy. In Bilodid, Ivan K. (ed.), *Pytannja istoryčnoho rozvytku ukraïns'koï movy*, 125–165. Kharkiv.
- Olander, Thomas. 2015. *Proto-Slavic Inflectional Morphology*. Leiden.
- Schenker, Alexander M. 1995. *The dawn of Slavic*. New Haven.
- Severjanov, Sergey. 1922. *Синайская псалтырь. Глаголитический памятникъ XI вѣка*. Petrograd.
- Stojkov, Stojko. 1954. *Bălgarska dialektologija*. Sofija.
- Tarnanidēs, Iōannēs Chr. 1988. *The Slavonic Manuscripts Discovered in 1975 at St. Catherine's Monastery on Mount Sinai*. Thessaloniki:
- Teneva, Evelina. 2012. Das Personalpronomen 1. p. sg. nom. im Slavischen und das abweichende aksl. *azъ*: Eine interdisziplinäre Betrachtung und Alternativlösung im Licht der Soziolinguistik und Balkanistik. *Linguistique Balkanique LI(1)*. 61-104.
- Totomanova, Anna-Marija. 2014. *Iz bălgarskata istoričeska fonetika*. Sofija.
- 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).
- Trubačev, O.N. (ed.). 1974-. *Etimologičeskij slovar' slavjanskix jazykov*. Moskva.
- Trubetzkoy, Nikolaus S. 1954. *Altkirchenslavische Grammatik: Schrift-, Laut- und Formensystem*. Wien.
- Vaillant André. 1942. L'article en vieux slave. *Revue des études slaves, tome 20, fascicule 1-4*. 5-12.
- Van Wijk, N. 1929. Die aksl. Formen господѣ, господю und die Aussprache der Buchstaben ѣ, ю. *Zeitschrift für Slavische Philologie, Vol. 6, No. 3/4*. 363-368.
- Vermeer, William. 2014. Early Slavic dialect differences involving the consonant system. In Fortuin, Egbert & Houtzagers, Peter & Kalsbeek, Janneke & Dekker, Simeon (eds.), *Dutch contributions to the Fifteenth International Congress of Slavists, Minsk, 181-227*. Amsterdam.
- Wandl, Florian & Kavitskaja, Darya. 2023. On the reconstruction of contrastive secondary palatalization in Common Slavic. *Journal of Historical Linguistics 13(2)*. 220-254
- Winslow, Joseph J. 2022. Old Church Slavonic phonemes: The problem of /j/ and /ě, a/ after palatals. *Die Welt der Slaven 67(2)*. 296-323.

Zaliznjak, Andrej A. 2004. *Drevnenovgorodskij dialekt. 2-e izdanie, pererabotannoe s učetom materiala nachodok 1995–2003 gg.* Moskva.

Велчева, Боряна. 1981. Проблеми на глаголическата писменост: Асеманиево Евангелие. In *Константин-Кирил Философ: Материали от научните конференции по случай 1150-годишнината от рождението му, 167-171.* София.

Галинская, Елена. 2014. Прогрессивная палатализация и древненовгородское местоимение *въхѣ*. *Slavistica Vilnensis* 59. 7-16.

Поливаново, А.К. 2013. Старославянский язык - Грамматика · Словари. Москва.