

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 annotation¹. 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 1.2.

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. *dušich* vs Suprasliensis. *доуѣхъ* <*duxěxъ; Polish *wszak* vs Supr. *вѣсакъ*, Ru. *всяк[уй]* <*vьx-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* <*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 *свинѣ* <*svinьkъ and *полотѣнѣца* <*polъnъkа), and those of *g are not unknown: Zaliznjak (2004: 47) admits that palatalised forms of the Germanic loan *кѣнѣз-* <*kъnēg- 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 *оуѣрѣзѣ* (cf. Russian *серьга*, commonly assumed to be an Oghur, i.e. Bulgar, Turkic loan, cognate with e.g. Kazakh *сырға*) '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 *vьxъ 'all' which supposedly show a lack of PV3 by retaining both /x/ and back/hard desinences (e.g. fem. gen. sg. *вѣхѣ* <*vьxoĭĕ from letter №850), and which come from letters which otherwise correctly convey the jers (by writing <ѣ, ѣ> for *ъ and <о, ѣ> for *ь), the weak-jer is always written with <ѣ, ѣ>, unambiguously suggesting a /ь/ pronunciation. These forms therefore more likely point to a LCS doublet-form *vьxъ 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 /ь/ 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
	ĭ ѣ ĭ̇		ŷ ȳ ѣ ĭ̇	
Mid	e ɛ		ɔ o	
	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 ħ̣	k g
	s z	š ž	x
		č	
	l	ĺ	
	r	ř	
v		j	

Foreign sounds

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: /ḳ ġ ǰ f ü/, e.g. in respectively *κῆτις* <*ḳiṭṣ, *λεμονή* <*iǰemoṇḥ, *χιτών* <*x̣itoṇḥ, *ιοσιφ* <*ijosif̣³, and *μυρο* <*ṃyro. 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' (*eǰŭṛp̣ṭḥ): *ⲉⲓⲣⲣⲟⲩ* vs *ⲉⲓⲣⲣⲟⲩ* vs *ⲉⲓⲣⲣⲟⲩ* vs *ⲉⲓⲣⲣⲟⲩ* vs *ⲉⲓⲣⲣⲟⲩ*⁴, etc.. One might also ask whether a separate <ⲥ̣> letter for /ǰ/ (and the writing of <ⲥ̣>/<ⲥ̣> with the palatalisation-diacritic) could be linked to the inadmissibility in the native systems of soft [ḳ, g̣] sounds, and whether their replacement with regular <ⲥ̣> or <ⲥ̣> was more likely in systems with some level of native [ḳ, g̣] (for instance, in Rus' after the so-called Fourth Velar Palatalisation *ky, *gy, *xy > [ḳi, g̣i, x̣i], 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. *ⲣⲟⲃⲩ* <*orb-ɔ vs Uspenskij Sbornik nt. acc. sg. *ⲣⲁⲗⲟ* <*ordl-o, because for too large a proportion of the vocabulary this information is not sufficiently securely and

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, which is a work-in-progress web-interface for viewing and searching the annotated texts. 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 <ⲥ̣> with <ⲥ̣>, but Eckhoff then replaces Severjanov's roundy diacritic with a tilde, leading to nonsense transcriptions like Psalm 8 <ⲁⲛⲏⲣⲁⲗⲟ> for ms. <ⲁⲛⲏⲣⲁⲗⲟ>). Cleaned up digitisations, using the Glagolitic originals as the base-text, and a mechanism for displaying manuscript-images, are planned for the near future.

5 Other troublesome pre-LCS morphological isoglosses reflected in the texts include the masc./nt. instr. sg. *o- and *jo-stem endings *-ѣмь/*-ѣмь (N.Sl., e.g. KF ѡѣмѡ ѡѣмѡ, Uspensk. Sbor. КНАЗѢМЬ) and *-омь/*-емь (S.Sl., e.g. Supr. ѡѣрѡмь, КНАЗѢМЬ), 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 *-тъ (N.Sl.) vs *-тъ (S.Sl.) verbal endings of 3rd sg. and pl. present (plus its extension to 2nd and 3rd sg. aorists like OCS НАУАТЪ, OR (Uspensk. Sbor.) БЫТЪ, НАУАТЪ). 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.

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



The syllabic liquids /r̥ r̩ l̥ l̩/ are included as unitary vocalic phonemes, following Schenker (1995: 94), rather than as combinations of /ɤ ʁ/ + /r̥ r̩ l̥ l̩/, 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 /r̥ r̩/ even in dialects with secondarily-palatalsed consonants before fallen weak LCS /ɤ/). The need for both front and back *r̥ *r̩ is unambiguously shown by the East Slavic reflexes /er/ and /or/ (Ru. *смерть*, *морковь*), but *l̥ vs *l̩ is more complicated: PIE *pl̥nos, *wł̥k^wos > Lithuanian *pilnas* 'full', *wilkas* 'wolf' (LCS *pr̥l̥n̥s, *v̥l̥k̥s) vs Lith. *stulpas* (LCS *stl̥p̥s '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 мѣлѣко), and thus only has /ol/ reflexes here: Ru. *волк*, *столб*, *полный*. It's true that Polish has *wilk* and *milczeń* (<*m̥l̥čĕti), but the Polish reflexes are complicated and likely have more to do with the surrounding consonants: *pr̥l̥n̥s by contrast gives *pełny* with hardened /l/ and the Polish non-palatalsing-/e/ reflex of *ɤ, and the differing reflexes in *wierzch* <*v̥r̥x̥s, *śmierć* <*sm̥r̥t̥s and *marty* <*mr̥tv̥j̥s 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 <рз> and <лз>, the Kiev Folio, 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 *r̥ *r̩ and *l̥ as one would expect: *р̆р̆л̆н̆с*- *м̆л̆ч̆с*- *в̆л̆к̆с*- *п̆л̆п̆с*- <*r̥ *r̩ and *l̥ (Winslow 2022: 313), and even Zographensis spells all 5 occurrences of *v̥l̥k- 'wolf' with *ўл̆к̆с*/ўл̆к̆с- and all 15 instances of its *-m̥l̥č- root with *м̆л̆ч̆с*- (e.g. *м̆л̆ч̆с*-*м̆л̆ч̆с*). 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 symbols⁷ rather than as the sequences /ɤ ʁ ɤ ʁ/⁸.

- 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 <*bъčĒla, 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 ѿвѣщанъ of the Kipchak word /jovšan/ 'wormwood' in the Hypatian Codex, whose modern cognates (Turkmen ýowşan /jowšan/, 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 **Ē*.
- 7 In the database I will have to use the single Unicode characters <ř ř ĩ ḷ>, rather than what's shown in my table, since the latter cannot actually be rendered without using the letters for /r ř l ḷ/ 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 /ǵ ɣ/, which I will have to replace with <ğ ŷ>.
- 8 To my mind the only evidence in support of a genuine jer + liquid stage comes from the paradigms of verbs like OCS сътърѣти <*sътŕĭti, where the syllabic /ŕ/ in the stem alternates with /ьr/ depending on the vocalicity of the following morpheme: the e.g. 3sg. pres. *сътъреть (Zogr., Supr. сѣтьреть) or (one possibility of the) 3rd sg. aorist *сътѣре (Supr. сѣтърьє) must have /ьre/, while the 3rd pl. aorist *сътŕĭšę (Supr. сѣтъръша) and the other possibility for the 3rd sg. aorist *сътŕĭ (Psal. рѣшѣ, or with a different prefix Mar. рѣшѣ) <*otŕĭ, being word-final or pre-consonantal, must be syllabic /ŕ/. The same alternation occurs in the zero-grade forms of verbs like *umerti, as is clear from the Polish reflexes umarł <*umŕĭł vs umře <*umŕĭro. The argument could be that at some stage, before

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 /ń́ í́ / 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. /t̚ d̚/ (as used in e.g. Olander 2015) or /k̚ g̚/ (as used by me in Winslow 2022), or variations thereof, are visually too close to symbols used elsewhere in the system. /k̚, g̚/ are anyway already used in my system for foreign /k, g/ before front-vowels, and /t̚ d̚/ 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 ѡ> for both sets (i.e. alongside attested  < ἡγεμών would have been **ѡѡѡѡѡѡ < *osqĥeni, and alongside attested  < *dĥerĥ would have been **ѡѡѡѡѡѡ < κῆνσοϛ⁹), 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 *šĥ and *žĥ 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: očīšĥenĥje vs. jĥskati: jĥščĥ, jĥzđiti: jĥžžĥ vs jĥzgĥnati: jĥžžĥenĥ.

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 ń́ í́ / 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, ɟ] 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 ń́ í́ / in my system strictly speaking would really just be cover-symbols for the pre-jotation sequences, but such notation is preferable

the LCS tendency towards Open Syllables became dominant, the stems in these paradigms were surely unitary /tĥr/, /mĥr/, i.e. 3sg. aor. /sŭ.tĥ.re/ vs 3rd. pl. aor. /sŭ.tĥr.šĥe/, and that the latter's closed /tĥr/ syllable was only forced to open itself up by changing to /tĥ/ 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 /tĥr/ to /tĥ/ 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* < *[xĥ]ĥe[tĥ]; *ѓон* < Tk. *gön* ‘leather’, *меѓу* < *meĥu.

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.

Word-initial *jĀ-/ *a-

Like Derksen, I assume that roots going back to PIE jot-less long *ē or diphthongal *oi-, e.g. the root for ‘to eat’, PIE *h₁ēd-, all took prothetic *j and merged with *jĀ- 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 ѣти etc.). Isolated nominal forms like Ru. *язва* (which Derksen (2008: 155) 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. *есть*, Pol. *jeść* are later generalisations from prefixed forms like OR *ѣтънѣтънѣ*, where no jot-prothesis could take place (cf. Schenker 1995: 88, Winslow 2022: 302 fn.14).

As explored more fully in Winslow (2022: 313-315), OCS spellings seem to suggest that free-variation between <ѣ, ѥ> and <и, ѡ> was a feature of the pre-Jer Shift Urkirchenslavisch orthographic system for conveying the reflexes of the sound-groups *ъj and *ыj (so-called ‘tense jers’), regardless of whether they were in strong or weak position. The examples given were: Zogr. Ѡрѧѡѡѧѧѧ vs Ѡрѧѡѡѧѧѧ <*znamenĭjĕ, ѡѡѧѧѧѧѧ <*udařŋ-jь vs ѡѡѧѧѧѧѧ <*omočŋ-jь; Mar. ѡѡѧѧѧѧѧ ꙗ vs ѡѡѧѧѧѧѧ Ѯ <*osodęť ѱjь; KF ѡѡѧѧѧѧѧѧѧ vs ѡѡѧѧѧѧѧѧѧ <*milostъjо, ѡѡѧѧѧѧѧѧѧ vs ѡѡѧѧѧѧѧѧѧ <*vъxomogy-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 ѡѡѧѧѧѧѧ, ѡѡѧѧѧѧѧ, ѡѡѧѧѧѧѧѧѧ for what in the live dialect underlying Zogr. must surely have been /udařij/, /boľij/, /veštij/, are not infrequent¹⁰. The fact that the same alternation occurs in the pre-Jer Shift KF (i-stem gen. plurals Ѡрѧѡѡѧѧѧѧѧѧѧ <*zapověďjь vs ѡѡѧѧѧѧѧѧѧ <*ludъjь) suggests that it is a common inheritance from the Urkirchenslavisch spelling system, and thus that in pre-Jer Shift Slavic the difference between /ъ ѣ/ and /и y/ was neutralised before /j/, and we should perhaps posit archiphonemes (which I call /Ĥ/ 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-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 **дѡѡволѡ** (< διάβολος), however, do show up in OCS with jer-spellings: Supr. **дѡѡвола**, Zogr. Luke 8 and Psal. Psalm 108 **дѡѡвола**, 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. **вѡснѡѡ** (jo-stem **вѡснѡѡ** < 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: *diĭjĀvolъ, *vasilijъ, *marijĀ etc.

Word-initial *jъ-/*ji-/*i-

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

I make an exception for certain forms of the personal-pronoun *jъ, however, and write *jimъ, *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 **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 *jisusъ, 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 /Ŷ/

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. **дрѡѡѡѡ** <*drugi-ji, 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 *jъjĀkovъ/ > /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 (шунца) 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 *ję vs *ę, where in Zogr. and Mar. and partially in Assem. (Велчева 1981: 168) the full front-nasal digraph <ѣе> is reserved for *ję, while just the second 'nasalising component' <ѣ> is used for post-consonantal *ę, e.g. Mar. 3rd pl. aorist **ѡѣѡѥ** <*jęse, as opposed to KF **рѣтѣѡѥ** <*prijeti vs **вѣѡѥѣѡѥ** <*vъželi¹⁴. 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 <ѣ̑> 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*, *evanġelъje*, *eġŭрътъ*, *elisavetъ*, *elinъ*, *evanġelistъ*, *eġŭрътъskъ*, *elinъskъ*, *episkupъstvo*, *evrejskъ*, *elisejъ*, *етѣтауsъ*, *etiјоръskъ*), the only spellings with <ѣ̑> are **ѣ̑лиси**, **ѣ̑ппъ**, **ѣ̑лини**, and **ѣ̑лина**, 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 <ѣ̑>, vs 3070 with <ѣ̑>¹⁵. Thus 97.2% of native word-initial *je- in Suprasliensis is spelt with <ѣ̑>, while 97.5% of the occurrences of the clearly post-LCS Greek-mediated foreign borrowings listed above instead use plain <ѣ̑>, 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 **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 *гъѣвѣнѣти* <**gyb-nqti* > Ukr. *гу́нути*, vs OCS *оуѣзѣнѣти* <**usъr-nqti* (cf. 3sg. aor. *оуѣзѣ*), Ru. *тонуть* <**top-nqti*, though see Meillet (1965: 142)). Geminate consonants were banned and either simplified (*иѣвшѣти* <**jъs-sekti*) or dissimilated (*процѣвѣти* <**prokviť-ti*).

The ban on **sr/*zr* is dealt with by insertion of **t* and **d* respectively, but the commonly-cited examples of **str* <**sr* (*сѣстра*, *строуѣа*, *остръ*) all concern root-internal **sr* where insertion of **t* is

५१०६९३४८३०७२३५३३

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 $\sigma\tau\rho\zeta$) 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 <зрѣк> 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-šъdъ, 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-čъtetъ (ECS **skūt-* > **ščīt-*), Psal. срѣдѣ <*orž-žigajetъ (ECS **zgīg-* > **žžīg-*) 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 <ѡ> spellings, can be found in Meillet 1965: 133).

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 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 владѣкъ ѡловѣколюбѣѣ 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. **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 *ъхѡ, for example, turns up exclusively <ѡѣ> spellings in Marianus, just one <ѡѣ> in Zogr., and exclusively <ѡѣ> in Suprasliensis, with only Assem. and Psal. containing a significant minority of <ѡѣ> spellings.

19 Russian feminine i-stems like *вѣсь* (<*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. **385¹ѣѣ**, with what looks like a jo-stem reflex of *rĭĕ (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.

1.2 LCS Morphology and the Autoreconstructor

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

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 below 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 lemma²² and thus to its LCS reconstruction and inflexion-class information²³. 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.

20 Numerous spellings in Supr. like *бурѣ* <*burĕ, *оукарѣтъ* <*ukarĕjetŭ, *моуѣ* <*moŭ etc., however, point to a hardening of LCS palatal *r̥ to plain /r/, so it's difficult to know whether *зѣръ* spellings stem from a genuine /zvĕra/ form in the history of the language, or if they instead represent synchronic /zvĕra/, i.e. with a hard o-stem ending, but with confusion by the scribe between <ra> and <ra>/<rb> spellings for what in his/her dialect would've all been /ra/.

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 вззвѣтиати , 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 вззвѣтиаше , 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. разбоѣника въ породаѣ въведе ‘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 разбоѣника above is “-s---mg--i”, so the Autoreconstructor produces *orzbojnik-a .

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](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).

сѣти

Random

Corpus forms

	Present	Aorist	Imperfect	Imperative
1st sg.				
2nd sg.				
3rd sg.	сѣтърѣтъ	сѣтърь		
	сѣтърьѣтъ	сѣтъре		
1st du.				
2nd du.				
3rd du.				
1st pl.				сѣтърьѣмъ
2nd pl.				
3rd pl.		сѣтърьша		

Participles	
PRAP ¹	
PRAP ²	
PAP	сѣтърьъ сѣтърьши
L-Part.	сѣтърьлъ
PPP	сѣтърьени
PrPP	
Infinitive	
Supine	сѣтърътъ

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

Deviance detection

217066	приведоша	3paia----i	privedošę	privedo
217101	въздѣхнѣвъ	-supamn-si	vъzdxnqovъ	vъzdxъ
217112	разврѣзосте	3daia----i	orzvrъzoste	orzvrъzete
217261	бѣгивѣ	-supamn-si	bolgoslovivъ	bolgoslovъ
217266	ѣша	3paia----i	jĕšę	jĕšę
217272	възаша	3paia----i	vъzęšę	vъzęšę
217302	начаша	3paia----i	načęšę	načęšę
217316	въздѣхнѣвъ	-supamn-si	vъzdxnqovъ	vъzdxъ
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	моѣомъ	-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	окомъ	-s---ni--i	okomъ	očesъmъ
218648	моѣю	-s---md--i	mōžu	mōžęvi
218688	моѣжа	-s---mg--i	mōžĕ	mōžu
218755	прѣлѣубѣ	-s---fa--i	perľuby	perľubъvъ
218762	поустивѣши	-supafn-si	pustivъši	pušĕbъši

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 **оуцѣхнѣти**)

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. **ѣгоуѣѣѣѣ**) 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 **бѣгоуѣѣѣѣѣѣ**). My justification is that the *-verg-/*-vřg- stem verbs *do* attest such root-aorists, e.g. 3rd pl. **ѣгоуѣѣѣѣѣѣ** 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, мѣжоу; loc. sg. ѿнѣ, gen. sg. мѣжа
- past act. part. of class 4 verbs using the suffix *-ivъ rather than *-jъ: бѣгивѣ, поуѣтивѣши (cf. Mar. Mark 10 [ѣхашѣ](#) <*pust-jъši)

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. [ѣхѣ+ѣ+ѣ](#) <*tęžĀtelĕ, Supr. [свѣтитѣлѣ](#) <*svĕtitely. 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 [мѣхѣлѣ](#)²⁹ mean we can’t follow Meillet (1965: 426) in setting up *-tele as the ‘correct’ ending, because spellings like [ѣхѣлѣ](#) 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 *-telĭ in the ‘deviances’ table³⁰.

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 [ѣдинъ](#) <*jedīnъ, 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. [ѣд’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, *оста\$, and *осте\$ (the latter two using ‘regular-expression’ mode and \$ to specify end-of-word), which for Mar. turns up nothing except the three occurrences of [ѣстоу](#).

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

30 Diels mentions only Psal. Psalm 26 [ѣхѣлѣ+ѣ+ѣ](#), 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 *jedīnъ 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 *jedīnъ’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 [ѣдѣ](#)-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 *jedīn-, though all 107 reflexes of the masc. nom./acc. sg. are from *jedīnъ. 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("Рѣдинъ") || 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.

References

- Bethin, Christina. 1998. *Slavic prosody: Language change and phonological theory*. New York.
- Derksen, Rick. 2003. Slavic *jь-. In Schaeken, Jos & Houtzagers, Peter & Kalsbeek, Janneke (eds.), *Dutch contributions to the Thirteenth International Congress of Slavists, Ljubana: 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. Mysli i predpoloženija 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.

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

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.

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 *въхъ*. *Slavistica Vilnensis* 59. 7-16.