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

0. Introduction

Much progress has been made in the last twenty years in early Slavonic corpus linguistics as a result of the Old Church Slavonic part of the PROIEL project (Haug & Jøhndal 2008) and its subsequent expansion as the TOROT treebank (Eckhoff & Berdičevskis 2015), such that currently just over 240,000 words of canonical OCS have been manually lemmatised, part-of-speech and morphologically-tagged, and syntactically parsed. The focus of these projects, however, has been exclusively on the higher-level linguistic domains of syntax, semantics, and pragmatics: surface-morphology has been of only incidental concern, for example in investigations into differential-object marking (Eckhoff 2015, 2022). No inflection-class data is included in these corpora, and phonology has been totally ignored to the extent that some of the texts (esp. Kiev Folia, Codex Suprasliensis, and partially Codex Zographensis) contain quite severe typographical inconsistencies and errors that make them dangerous to use without reference to the manuscripts.

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

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

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

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

---

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

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

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

The premise of my chosen form of "phonological annotation" is that the earliest Slavic texts reflect languages which are **structurally** close enough to the broadly-agreed-upon system of Late Common Slavonic that the forms underlying the manuscript-spellings are more or less trivially derivable (by the application of sound-change rules) from their theoretical LCS ancestors.
By 'structurally' I am referring to the structure at the phonological level; structural changes at higher levels of analysis (i.e. inflectional morphology, derivational morphology) are of no concern unless they are **made possible only by intervening phonological changes**.
My contention is that before about 1100 not enough of these structural changes are in evidence in any Slavic text, and thus they can be relatively straightforwardly indexed using a well-chosen LCS system. Before giving examples of structural changes that are problematic for such an indexing-system, it's necessary to first lay out my LCS system in full:

1.1 Late Common Slavonic as a "phonological index"

In order to account for as much of the subsequently attested Slavic as possible, a point after the monophthongisation of diphthongs, but before the Second and Third Velar Palatalisations (PV2 and PV3) is chosen as the point of departure, because of the difference between the West Slavic /š/ and South/East /ś/ reflex of these two palatalisations of *x (Cz. loc. pl. *duších* vs Supraslienses. доусѣхъ <*duxěxъ; Polish *wszak* vs Supr. вьсакъ , Ru. *всяк[ий]* <*vьx-akъ), as well as the probable complete absence of PV2[2] in northern East Slavic (Old Novgorodian, see Zaliznjak 2004: 42-45 for the evidence), 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 свинеце <*svinьkь and полотенеца <*poltьnьka), and those of *g are not unknown: Zaliznjak (2004: 47) admits that palatalised forms of the Germanic loan кънѧз- <*kъnęg- are the rule, but considers this to be a "supradialectal" word originating outside of the Novgorodian dialect-area; Galinskaja (2014: 10) is less convinced and adduces the form оусьрѧзи 'earrings' from letter №429 as a word of "вполне бытового характера" which thus supposedly shows a native Novgorodian reflex of PV3 of *g. (This is commonly assumed to be a Turkic loan, cognate with e.g. Kazakh *сырға*, but the fact that it appears in Slavic with front-vowels (Ru. *серьга*), unlike its back-voweled Common Turkic cognates, and the fact that it was borrowed early enough to undergo PV3 at all, suggests that Vasmer's derivation of it from "Old Chuvash" (i.e. some form of Oghur or Bulgar Turkic) is correct, and it thus belongs to an earlier layer of Turkic loans than those borrowed from the Kipchak dialects of the Polovtsians (e.g. Ru. *камыш* < *qamış (> Kaz. *қамыс*).)
More importantly, as Galinskaja (op. cit.) points out, in all of the well-known Novgorodian forms of the pronoun *vьxъ 'all' which supposedly show a lack of PV3 by retaining both /x/ and back/hard desinences (e.g. fem. gen. sg. въхоѣ <*vьxoję̂ from letter №850), and which come from letters which otherwise correctly convey the jers (by writing <ь,е> for *ь and <о,ъ> for *ъ), the weak-jer is always written with <ъ,о>, unambiguously suggesting a /ъ/ pronunciation. These forms therefore more likely point to a LCS doublet-form *vъxъ which would never contain the conditioning environment for PV3 anyway, and thus you can't use them as evidence of a lack of PV3 in Novgorodian (on the plausibility of such a doublet see Galinskaja (2014: 14), though cf. Zaliznjak's (2004: 54) less convincing explanation of the /ъ/ in these words as an assimilation of original /ь/ to the back-vowels of the following syllable).

Table 2: LCS Vowels after the monophthongisation of diphthongs

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

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

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

In addition, the following symbols are used to represent phonemes of wholly foreign origin in order to represent badly-integrated foreign borrowings, whose level of integration into the native system we deliberately do not take a position on: /ḱ ǵ x́ f ü/, e.g. in respectively ќитъ <*ḱitъ, іѓемонъ <*iǵemonъ, χитонъ <*x́itonъ, иосифъ <*ijosifъ[3], and мѵро <*müro. Almost none of the words containing these symbols would actually have existed in the language during Common Slavonic times, but they need to be included in the indexing-system because they often contain native Slavic elements (f.ex. inflectional endings). Normally they represent specific sounds in the source-language (usually Greek), so including them is useful for investigating the process of these sounds' integration into the native systems. For instance, the extent to which Greek /ü/ is integrated into either native /i/ or /u/ can be seen in variations in the OCS spellings of the word for 'Egypt': ⰵⰼⱛⱂⱅⱏ vs ⰵⰳⰺⱂⱅⱏ vs ⰵⰳⰹⱂⱅⱐ vs єгоѵптъ vs ⰵⰼⰹⱂⱅⱏ[4], etc.. One might also ask whether a separate <ⰼ> letter for /ǵ/ (and the writing of <ⰽ> with the palatalisation-diacritic) could be linked to the inadmissibility in the native systems of soft [kʲ, gʲ] sounds, and whether their replacement with regular <ⰳ,ⰽ> or <г, к> was more likely in systems with some level of native [kʲ, gʲ] (for instance, in Rus' after the so-called Fourth Velar Palatalisation, or in Novgorod due to the retention of native velars before front-vowels because of the non-action of PV2, etc.); 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.

I have deliberately not included accentual information in my reconstruction of vowels, even though such information is in fact required to explain certain differing manuscript-reflexes, e.g. Russkaja Pravda fem. acc. sg. робꙋ < *orb-ǫ vs Uspenskij Sbornik nt. acc. sg. рало < *ordl-o, because for too large a proportion of the vocabulary this information is not sufficiently securely and uncontroversially reconstructed to justify its inclusion, and anyway the (often post-LCS) derivational processes which are responsible for most of the actual words in the attested texts (and the inevitable accentual levelling processes likely to have occurred in the course of these derivations) complicate things even further.

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

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

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

Slavic remains high and backed, e.g. Supr. ȥовꙑ <*zovy̨, Psalterium Sinaiticum ⱄⱅⱃⱑⰳⱔⰻ <*stergy̨-jь, Codex Marianus ⱑⰴⱔⰻ <*jǢdy̨-jь (as is clear from these forms, some dialects of early OCS retained some kind of nasal character in this vowel and may even have developed the special "hooked" nasal letter <ⱔ> for it), but in North Slavic lowered to /a/: Old Polish (Kazania Świętokrzyskie) *reca* /r'eka/ <*reky̨, Ru.Ch.Sl. (Vita Methodii) вьсємогаи <*vъxemogy̨-jь. /ę̌/ is responsible for the NSl. /ě/ vs SSl. /ę/ shapes of jo-stem masc. acc. pl. and the ja-stem nom./acc. pl. and gen. sg. endings (Kortlandt 1979), which are reflected in respectively the post- and pre-revolutionary spellings of the Russian nom./acc. pl. long-adjective endings *-ые* < *yjě < *yję̌ vs *-ыя* < *yja < *yję < *yję̌.

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

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

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

3.) the evidence of certain modern Bulgarian dialects, which have reflexes of LCS *ě in words like *ж'еба* <*žǢba 'toad',

all together point very strongly towards there having occurred a split on the Southeastern periphery of Slavic between LCS dialects which have /ě/ < *Ǣ and the majority of the rest which got /a/, and that original OCS ('Urkirchenslavisch') was an *Ǣ > /ě/ dialect. I posit that *Ǣ remained until the opposition /a/:/ě/ after palatal consonants was reintroduced when PV2 and PV3 brought new soft-consonants /c ś dź/ into the system, which could be followed by both /a/ and /ě/: /stьdźa/ < *stьga (PV3) vs /nodźě/ <*nogě (PV2) (Winslow 2022: 304-305). Thus since my LCS system is based on a point just *before* PV2 and PV3, I must also retain the *Ǣ archiphoneme.[5]

The syllabic liquids /ŕ̥ ĺ̥ r̥ l̥/ are included as unitary vocalic phonemes, following Schenker (1995: 94), rather than as combinations of /ь ъ/ + /ŕ ĺ r l/, because these groups descend from PIE syllabic liquids and many descendant South Slavic dialects which retain syllabic liquids in this position (including most of those underlying canonical OCS) do not show any evidence of an intervening oral-vowel + liquid stage (such a view is shared by Bethin 1998: 71-72; cf. also Bulgarian dialectal

5 It's possible to argue that the short *Æ counterpart to *Ǣ persisted in ESl. until after the Fall of the Jers, and that the 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, e.g. Ukr. *бджола* <*pъčÆla, or was newly phonemicised, e.g. Ru. *вёсла* <*v'Æsla <*vesla), and that there was never a stage when these words had /e/ (based among other things on <o> spellings regardless of stress after palatal-letters in very early texts, and even after the letters for secondarily-soft LCS plain consonants in the Birchbark documents (Le Feuvre 1993, Nakonečnyj 1962), but there isn't space to elaborate on the issue here (see Winslow 2022: 304 fn.16). Unlike the situation with long *Ǣ, OCS shows no sign of anything but an /e/ reflex of short *Æ (and indeed the fact that the 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 ѥвшанъ of the Kipchak word /jovšan/ 'wormwood' in the Hypatian Codex, whose modern cognates (Turkmen *ýowşan* /jowşan/, Kazakh *жусан* /žuwsan/, Azeri *yovşan*) unambiguously point to a Kipchak /o/), and the history of the East Slavic /o/ reflexes remains the subject of much disagreement, so it's simpler for everyone if I continue the traditional practice of writing LCS *e after palatals, even if that strictly speaking is inconsistent with my use of *Ǣ.

evidence in Stojkov 1954: 130-131, where hard consonants precede reflexes of the LCS /ĺ̥ ŕ̥/ even in dialects with secondarily-palatalised consonants before fallen weak LCS /ь/).
The need for both front and back *ŕ̥ *r̥ is unambiguously shown by the East Slavic reflexes /er/ and /or/ (Ru. *смерть*, *морковь*), but *ĺ̥ vs *l̥ is more complicated: PIE *pl̥nos, *wl̥kʷos > Lithuanian *pilnas* 'full', *wilkas* 'wolf' (LCS *pĺ̥nъ, *vĺ̥kъ) vs Lith. *stulpas* (LCS *stl̥pъ 'pillar') suggests that Balto-Slavic had differentiated front/back variants of the PIE syllabic *l̥ (Bethin 1998: 69), but the ancestor to East Slavic backed all vowels preceding tautosyllabic /l/ (Ru. молоко < Proto-ESl. *molko < LCS *melko > OCS млѣко), and thus only has /ol/ reflexes here: Ru. *волк*, *столб*, *полный*. It's true that Polish has *wilk* and *milczeć* (<*mĺ̥čǞti), but the Polish reflexes are complicated and likely have more to do with the surrounding consonants: *pĺ̥nъ by contrast gives *pełny* with hardened /l/ and the Polish non-palatalising-/e/ reflex of *ъ, and the differing reflexes in *wierzch* <*vŕ̥xъ, *śmierć* <*sъmŕ̥ťь and *martwy* <*mŕ̥tvъjь rule out any explanation based on the nature of the LCS syllabic-liquid alone (for more discussion see Bethin op. cit.: 73-75).
While most OCS shows no sign at all of a front-back distinction in the syllabic-liquids and writes the reflexes of these groups overwhelmingly with <ρъ> and <λъ>, the Kiev Folia, which is the only OCS text that reflects a pre-Jer Shift stage and is very nearly flawless in its etymologically correct rendering of the jers, also spells *ŕ̥ *r̥ and *ĺ̥ as one would expect: ⱄⰲⱐⱃⱎ- ⱅⰲⱐⱃⰴ- ⱄⱐⱃⰴⱃⰲ- ⰴⱐⱃⰶ- <*ŕ̥, ⱄⱐⱃⱎⱃⱅ <*r̥, and ⱂⰾⱏⱀⱁⱅⱁⱂ <*ĺ̥ (Winslow 2022: 313), and even Zographensis spells all 5 occurences of *vĺ̥k- 'wolf' with ⰲⰾⱐⰽ-/ⰲⰾⱐⱍ- and all 15 instances of its *-mĺ̥č- root with -ⰿⰾⱐⱍ- (e.g. ⱁⰿⰾⱐⱍⰰⱎⰵ). Therefore, taken as a whole the Slavic evidence pretty securely points to front and back variants of both syllabic liquids, and for searching purposes it's far preferable to denote them with separate symbols[6] rather than as the sequences /ьr ъr ьl ъl/ (which there is zero evidence for anyway).

Dejotation
The dejotated reflexes of *tj, *kt+front-vowel and *dj are denoted using the modern Serbian Cyrillic letters /ћ/ and /ђ/ respectively, because the commonly used alternatives, i.e. /ť ď/ or /ḱ ǵ/, or variations thereof, are visually too close to symbols used elsewhere in the system. /ḱ, ǵ/ (as used for /ћ ђ/ at the behest of the editor in Winslow 2022) are anyway already used in my system for foreign /k, g/ before front-vowels.
The compelling hypothesis, first proposed by Durnovo (1929: 55-58) but most recently elaborated by Vermeer (2014: 209-214), and accepted by Mathiesen (2014: 197 fn. 22) and Winslow (2022: 310 fn.25), according to which the Urkirchenslavisch reflexes of *ђ,ћ were close enough to foreign /g k/ before front-vowels that the original Glagolitic system used <ⰼ ⰽ> for both sets (i.e. alongside attested ⰺⰼⰵⰿⰵⱀⰵ < ἡγεμών would have been **ⱁⱄⱘⰼⰵⱀⰺ <*osǫђeni, and alongside attested ⰴⱏⰽⰵⱃⱐ <*dъћerь would have been **ⰽⰺⱀⱁⱄⱏ < κῆνσος[7]), does not prevent us from keeping the foreign sounds separate for our LCS stage, since clearly they differed enough in all the dialects underlying actually attested OCS to be written separately.
Pre-dejotation *stj and *zdj are differentiated from the PV1 reflexes of *sk and *zg by writing the former as *šћ and *žђ and the latter as *šč and *žǯ, even though their modern reflexes do not differ from each other anywhere so they must've fallen together in the CS period, because they often alternate with their respective un-palatalised counterparts morphologically and derivationally, e.g. očistiti:očišћenьje vs. jьskati:jьščǫ, jǞzditi:jǞžђǫ vs jьzgъnati:jьžǯenǫ.
Therefore the only dejotation-reflexes which are not fully "undoable" under my notation are those of *sj and *zj and *kt, viz. /š ž ћ/.

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

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

Autoreconstructed forms are actually built out of a pre-dejotation stage, with dejotation applied as a post-processing step, because this greatly simplifies the inflectional morphology in places like the 1sg. pres and past-active-participle of class IV (-iti) verbs: we can just use the desinences *-jǫ and *-jь regardless of stem-consonant, and then apply dejotation later in a post-processing step that every word undergoes, rather than needing a whole set of consonant-mutation rules for these endings. Therefore it would be possible to allow searching based on pre-dejotation forms, but in the case of *sj, *zj wider Indo-European evidence is needed to distinguish their LCS /š ž/ reflexes from the identical outputs of PV1 (e.g. Gothic *siujan* confirms an ECS form *sjū-tei for the verb *šiti), which it is outside the scope of this project to consider, as the goal here is to enable investigations of actual texts, for which such ECS differences are irrelevant. I cannot therefore consistently offer pre-dejotation reconstructions, because stems containing reflexes of *sj and *zj are only ever reconstructed with š ž //unfinished
some brief words about dejotated labials and epenthetic Í, which in my system is taken as the regular LCS outcome, removed only by later dialect-specific developments

There are convincing arguments for PV2/3 having preceded dejotation, at least in more central areas, most recently presented in e.g. Vermeer (2014: 197) and Wandl & Kavitskaya (2023 244-247), and therefore it could be objected that my system, which contains the dejotation reflexes /ћђńĺŕ/ but not the PV2/3 reflexes /c s’ dz’/, is ahistorical. However it should be reemphasised that the primary goal of my LCS reconstructions is to act as an index allowing 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 cannot 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 Serbian-like palatal /c’ dj’/ reflexes of *tj and *dj, and also no sequences of [tj, dj], could not convicingly be argued to have undergone dejotation at the phonemic level, as these new units would just be phonetic realisations of /tj, dj/. Analysed like that, the symbols /ћђńĺŕ/ in my system strictly speaking would really just be cover-symbols for the pre-jotation sequences, but such notation is preferable since it prevents searches for groups containing /j/ alone from returning results polluted by all the dejotation-groups. As I explored in my previous article (Winslow 2022), the status of /j/ as a phoneme in the earliest OCS texts is an intricate problem, so the ability to investigate the reflexes of *j in isolation from the dejotation-reflexes is important.

Word-initial *j-

The tendency for ECS *ā- to have taken prothetic /j/ by LCS times (in accordance with the drive towards open syllables) can make it difficult to distinguish these groups from *jĀ- in the absence of wider Indo-European evidence. Normally I’ve followed Derksen (2009), or the ESSJA, but for certain lexemes, e.g. *ama ‘pit’, which in OCS is spelt overwhelmingly with ѧм- or ꙗм-, the single Greek cognate ἄμη adduced by ESSJa I p.70 in favour of jot-less *am- is not enough to categorically exclude the alternative *jĀma. In particular the 1sg. nom. pronoun *azъ/jĀzъ is particularly problematic: I follow ESSJA I p.100 which ultimately plumps for *azъ, but Derksen doesn't discuss it at all. A lengthy discussion of the evidence can be found in Teneva's (2012) article on the subject

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

With word-initial *ji-/*jь-, I follow Derksen's (2009: 16) practice of writing *jь-, even though Derksen himself (2003) has argued for a split between *ji- and *jь- conditioned 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. (before vowels) *йдуть* appear to have treated the weak-jer in *jьdǫtь just like any other and retained the /j/, and Ukr. *ськати* <*jьskati (with the restricted meaning ‘look for nits/fleas in someone's hair’ after the base-meaning ‘seek’ was transferred to the Polonism шукати) shows the expected Ukr. softening of the /s/ after fallen weak-jer in *ьsk groups (cf. польський).

//I've just realised that the entire argument below is wrong because I forgot how Havlik's Law works: the middle-jer in these forms would be strong so the thing I'm talking about never showing up as <ⰵ> would be weak and thus not expected to vocalise anyway.
I make an exception for certain forms of the personal-pronoun *jь, however, and write *jimь, *jima, *jixъ *jimъ and *jimi for the masc/nt. instr. sg. and dat./instr. dual/pl., because Czech here has *jim jich jimi*, and even those OCS manuscripts which often show an <ⰵ> reflex of strong *ь before *j never show such in the gen pl. forms of soft-stem long-adjectives and participles (which we'd expect if these forms reflected LCS *ь-jьxъ; one exception is Zogr. John 5 (check ms. because TOROT Zogr. is unreliable and this is from an autotagged bit) ⱍⱑⱙⱋⰵⰻⱈⱏ <*čǢjǫťь-jixъ, but Zogr. otherwise never(?) writes strong tense front-jer as <ⰵ>; the mss. which often do, like Assem. ⰲⱔⱋⰵⰻ <*vęťьjь, ⰲⰵⰾⰵⰻ <*velьjь, ⱃⱁⰶⰴⰵⰻ <*roђь-jь, Mar. ⰱⱁⰾⰵⰻ <*boĺьjь, ⰾⱓⰴⰵⰻ <*ĺudьjь, ⱆⰴⰰⱃⰵⰻ <*udaŕь-jь, Psal. ⰴⰵⱀⰵⰻ <*dьnьjь, etc., never have <ⰵ> for the first jer in Gpl. long-adjectival forms.

Prefixed forms like *do-jьti 'to come, arrive' for morphological reasons have to be distinguished from the class 4 verb *dojiti/dojiši/dojimъ etc. 'to breastfeed' (and its derived noun *dojilika), a difference which is reflected in the modern Ukrainian *дійти* (<*dojьti with compensatorily-lengthened /o/ > /i/) vs *доїти*. //this is not principled or well-grounded, I should probably reference Shevelov's book here

Foreign *e- are all reconstructed identically to native words as *je-, (jeǵüpьtъ etc.), because no distinction between them is detectable in the Glagolitic mss. and the total ban on word-initial *e- in the native LCS sysytem would make adpation of foreign /e-/ to /je-/ automatic, but the innovation of the <ѥ> letter in Cyrillic and the detectable distinctions in Supr. between foreign <e-> and native <ѥ-> spellings suggests that this decision should perhaps be reversed.

*u-, *ju- : ju- only for inherited PIE *y-

treatment of clusters like jьs-šьdъ jьs-kě, which in OCS are simplified but kept separate by me, vs ot-xod, ob-xod, which I reconstruct as simplified to ox-, ox-, issues of reformation with jer-containing prefixes like otъ- in many words

An example of morphological change contingent upon structural phonological change, leading to manuscript forms which preclude any valid reconstruction of their direct LCS-stage ancestors, is the replacement of i-stem endings with those of the corresponding jo- or jā-stems, in nouns whose stems end on labials or the subset of LCS dental consonants which lack palatal counterparts, viz. /d t s z/. Evidence for such a change is furnished by the Old Russian masc gen./acc. form **тата** from the 1229 Treaty between Smolensk, Riga and Gotland (Version A). LCS *tatь is a masc. i-stem noun with genitive *tati, as it still appears in the Codex Suprasliensis translation of John Chrysostom's Homily for Holy Thursday (…то кажетъ владъıкъı чловѣколюбьıѥ· ꙗ́ко прѣданника разъ́йника 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 *pastyŕь > /pastyr'/, where the original LCS palatal *ŕ has fallen together with secondarily-palatalised /r'/ from plain LCS *r before LCS front-vowels, in e.g. the original i-stem *zvěrь > /zvěr'/. This system thus no longer distinguishes between descendants of the original LCS palatals and the newly secondarily-palatalised consonants like /t'/, both are now together in the set of 'soft' consonants, opposed to their 'plain' or 'hard' counterparts, and so tend towards taking the same set of inflectional endings (in this case those of the original jo-stems). Consequently, a word like **татъ** has begun to take jo-stem endings, including the Old Russian /a/ reflex of LCS *Ǣ in the genitive/accusative singular. LCS /Ǣ/, though, by definition can only occur after LCS palatal consonants (see above), so a reconstruction *tatǢ is just nonsensical. In the case of the dat. sg. /u/-desinence (which isn't attested in our Treaty but it exists in modern Russian *matю*), 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.

//this is just rough unstructured ideas, some of which may already have been incorporated into the text above

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

For example, whether or not there actually existed at the LCS stage a mechanism for deriving secondary-imperfective verbs like OCS разарꙗти <*orzaŕǢti from the prefixed разорити *orzoriti is irrelevant, because LCS /orzaŕǢti/ does not violate the rule of LCS phonotactics: palatal /ŕ/ can be followed by /Ǣ/ because such a combination exists in the paradigms of wholly securely reconstructable jo-stem nouns, e.g. nt. gen. sg. *moŕǢ (> Pol. morza, OCS морꙗ, Ru. моря, etc.)

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

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