

# 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<sup>1</sup>. 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.

---

<sup>1</sup> 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. *dušich* vs Suprasliensis. *доуѣхъ* <\*duxěx; Polish *wszak* vs Supr. *вѣакъ*, Ru. *всѣакъ* [уѣ] <\*vьx-akъ), as well as the probable complete absence of PV2<sup>2</sup> 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 *свинѣцѣ* <\*svinьkъ and *полотѣнѣца* <\*polьtьnъka), and those of \*g are not unknown: Zaliznjak (2004: 47) admits that palatalised forms of the Germanic loan *кѣнѣзъ* <\*kьneg- 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 *оуѣрѣзѣ* '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. *вѣхѣ* <\*vьxoĭĕ from letter №850), and which come from letters which otherwise correctly convey the jers (by writing <ѣ,e> for \*ь and <о,ъ> for \*ѡ), the weak-jer is always written with <ѣ,o>, 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 Galinskaja (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
	ǐ Ǘ ǐ̇		ǣ ǣ̇ ǐ̇	
Mid	e	ɛ	ɔ	o
	ě	ě̇		
Low	æ̃			
		a		

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

Labial		Dental		Palatal		Velar	
m		n		ń			
b	p	t	d	ħ	ḥ	k	g
		s z		š ž		x	
				č			
		l		ĺ			
		r		rí			
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: /ḳ ǵ ǰ f̣ ü/, e.g. in respectively *кѣтъ* <\*kítǵ, *ѣгемонъ* <\*ǵgemonǵ, *хитонъ* <\*xítǵ, *иосифъ* <\*ijosifǵ<sup>3</sup>, and *мүро* <\*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': *ѣгѣп̣тѣ* vs *ѣгѣп̣тѣ* vs *ѣгѣп̣тѣ* vs *ѣгѣп̣тѣ* vs *ѣгѣп̣тѣ*<sup>4</sup>, etc.. One might also ask whether a separate <ǵ> letter for /ǵ/ (and the writing of <ǵ> 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 <ǵ, ǵ> or <ǵ, ǵ> was more likely in systems with some level of native [ḳ, g̣] (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.

## 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. *рѣбѣ* <\*orb-ǵ vs *Uspenskij Sbornik* nt. acc. sg. *рѣбѣ* <\*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 /y/ and /ǵ/ are required to account for the split between North (East and West) and South Slavic forms of certain inflectional-endings: \*ǵ for the nom. sg. masc./nt. pres. act.

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

<sup>4</sup> 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.

participle of certain verb-classes whose present-stem ends on a hard-consonant, which in South Slavic remains high and backed, e.g. Supr. **ꙗꙋꙋꙗ** <\*zovy, Psalterium Sinaiticum **ꙗꙋꙋꙗ** <\*stergy-**ꙗꙋ**, Codex Marianus **ꙗꙋꙋꙗ** <\*jĀdy-**ꙗꙋ** (these forms lead Kortlandt (1979:260) to posit that some dialects of early OCS retained some kind of nasal character in this vowel and may even have developed the special "hooked" nasal letter <ѥ> for it), but in most of North Slavic lowered to /a/: Old Polish (Kazania Świętokrzyskie) has both *reca* and *recø* (with the special Old Polish letter for the merged reflex of \*ę and \*o) <\*reky, and in other texts also *biorø* <\*bery; Russkaja Pravda **ꙗꙋꙋꙗ**, Uspenskij Sbornik **ꙗꙋꙋꙗ** <\*dojdy, Ru.Ch.Sl. (Vita Methodii) **ꙗꙋꙋꙗ** <\*vъxemogy-**ꙗꙋ**. Kortlandt's positing of a CS \***ꙗ** (which he writes as \*a<sup>N</sup>) is far from universally accepted, and others consider these forms the result of various dialect-specific analogical process; see references and discussion in Olander (2015: 88-92). Whatever the truth of the matter, our \***ꙗ** is a convenient placeholder which allows all the relevant evidence to be retrieved.

The need for the retention of the /Ē/ 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 /Ė/) was explicitly posited by Kortlandt as far back as 1979 (p.266) as part of his ECS system. In short, a combination of:

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

5 Other annoying pre-LCS morphological isoglosses reflected in the texts include the masc./nt. instr. sg. \*o- and \*jo-stem endings \*-ѣмь (N.Sl.) and \*-омь (S.Sl.), 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 \*въчера ‘yesterday’, and the \*-тъ (N.Sl.) vs \*-тѣ (S.Sl.) verbal endings of 3<sup>rd</sup> sg. and pl. present (plus its extension to 2<sup>nd</sup> and 3<sup>rd</sup> sgl. aorists like OCS НАУАТЬ, OR (Uspenskij Sbornik) НАУАТЬ). Here I have no choice but to index them with dummy-symbols in the database: \*-Omь for the instr. sg. ending and \*-tQ for the verb-endings.

The syllabic liquids /ʃ̥ ʃ̣ r l/ are included as unitary vocalic phonemes, following Schenker (1995: 94), rather than as combinations of /ɤ ɐ/ + /ʃ̥ ʃ̣ r l/, because these groups descend from PIE syllabic liquids and many descendant South Slavic dialects which retain syllabic liquids in this position (including most of those underlying canonical OCS) do not show any evidence of an intervening oral-vowel + liquid stage (such a view is shared by Bethin 1998: 71-72; cf. also Bulgarian dialectal evidence in Stojkov 1954: 130-131, where hard consonants precede reflexes of the LCS /ǰ̯ ǰ̣/ even in dialects with secondarily-palatalised consonants before fallen weak LCS /ɤ/).

The need for both front and back  $*r_{\text{f}}^{\text{f}}$   $*r_{\text{b}}^{\text{b}}$  is unambiguously shown by the East Slavic reflexes /er/ and /or/ (Ru. *смерть*, *морковь*), but  $*r_{\text{f}}^{\text{f}}$  vs  $*r_{\text{b}}^{\text{b}}$  is more complicated: PIE  $*p_{\text{h}}\text{nos}$ ,  $*w_{\text{h}}k^{wos}$  > Lithuanian *pilnas* 'full', *wilkas* 'wolf' (LCS  $*p_{\text{h}}\text{ľn̥}$ ,  $*v_{\text{h}}\text{ľk̥}$ ) vs Lith. *stulpas* (LCS  $*st_{\text{h}}\text{ľp̥}$  'pillar') suggests that Balto-Slavic had differentiated front/back variants of the PIE syllabic  $*r_{\text{f}}^{\text{f}}$  (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* (< $*m_{\text{h}}\text{ľč}^{\text{f}}\text{Ēti}$ ), but the Polish reflexes are complicated and likely have more to do with the surrounding consonants:  $*p_{\text{h}}\text{ľn̥}$  by contrast gives *pełny* with hardened /l/ and the Polish non-palatalising-/e/ reflex of  $*r_{\text{b}}^{\text{b}}$ , and the differing reflexes in *wierzch* < $*v_{\text{h}}\text{ľx̥}$ , *śmierć* < $*s_{\text{h}}m_{\text{h}}\text{ľt̥}$  and *martwy* < $*m_{\text{h}}\text{ľtv̥j̥}$  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 <ꙗ> and <ꙋ>, 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 \**ĭ*, \**r̥* and \**l̥* as one would expect: *ѡѧѣѡѧѣ ѡѧѣѡѧѣ ѡѧѣѡѧѣ ѡѧѣѡѧѣ* <\**ĭ*, *ѡѧѣѡѧѣ* <\**r̥*, and *ѡѧѣѡѧѣ* <\**l̥* (Winslow 2022: 313), and even Zographensis spells all 5 occurrences of \**vľk-* ‘wolf’ with *ѡѧѣѡѧѣ*- and all 15 instances of its \*-*mľč-* root with *-ѡѧѣѡѧѣ* (e.g. *ѡѧѣѡѧѣѡѧѣ*). 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<sup>7</sup> rather than as the sequences /*ɤr ʔr ɤl ʔl*/ (which there is zero evidence for anyway).

## Consonants

The so-called ‘dejotated’ reflexes of \*tj (and \*kt+front-vowel) and \*dj are denoted using the modern Serbian Cyrillic letters /h/ and /h̑/ respectively, because the commonly used alternatives, i.e. /t̑ d̑/ (as used in e.g. Olander 2015) or /k̑ g̑/ (as used by me in Winslow 2022), or variations thereof, are visually too close to symbols used elsewhere in the system. /k̑, g̑/ are anyway already used in my system for foreign /k, g/ before front-vowels, and /t̑ d̑/ 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 \*h, ħ were close enough to foreign /g k/ before front-vowels that the original Glagolitic system used <ꙗ ѡ> for both sets (i.e. alongside attested  $\text{ꙗꙗꙗꙗꙗꙗ} < \eta\gamma\epsilon\mu\acute{o}\nu$  would have been  $^{**}\text{ꙗꙗꙗꙗꙗꙗ} < *osoheni$ , and alongside

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 *ѡвѡдъ* of the Kipchak word /*ovʂan*/ ‘wormwood’ in the Hypatian Codex, whose modern cognates (Turkmen *ýowşan* /*owşan*/, Kazakh *жуған* /*žuwan*/, 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 <ɾ ɽ ɺ>, rather than what's shown in my table, since the latter cannot actually be rendered without using the letters for /r ɾ l ɺ/ 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 /ɛ̥ y/, which I will have to replace with <ɛ̥̯ y̯>.

attested  $\text{ѡѡѡѡ} < *d\bar{h}er\bar{e}$  would have been  $**\text{ѡѡѡѡ} < \kappa\eta\nu\sigma\omicron\varsigma^8$ ), does not prevent us from keeping the foreign sounds separate for our LCS stage, since clearly they differed enough in all the dialects underlying actually attested OCS to be written separately.

Pre-dejotation  $*stj$  and  $*zdj$  are differentiated from the PV1 reflexes of  $*sk$  and  $*zg$  by writing the former as  $*šh$  and  $*žh$  and the latter as  $*šč$  and  $*žž$ , even though their modern reflexes do not differ from each other anywhere and so must've fallen together in the CS period, because they often alternate with their respective un-palatalised counterparts morphologically and derivationally, e.g.  $očistiti:očišen\bar{n}je$  vs.  $j\bar{b}skati:j\bar{b}ščq$ ,  $j\bar{A}zditi:j\bar{A}žžq$  vs.  $j\bar{b}zg\bar{n}ati:j\bar{b}žženq$ .

Therefore the only dejotation-reflexes which are not fully “undoable” under my notation are those of  $*sj$  and  $*zj$  and  $*kt$ , viz.  $/š ž h/$ .

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  $*-j\bar{q}$  and  $*-j\bar{b}$  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  $*sj\bar{u}-tei$  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  $š ž$ .

As regards the reflexes of labial consonant =  $*j$ , all such are reflected in my system as  $/m\bar{l} b\bar{l} p\bar{l}/$ . and subsequent loss of  $*\bar{l}$  is regarded as a dialect-specific development. //unfinished

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  $/h\bar{h}\bar{n}\bar{l}\bar{l}/$  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 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  $/h\bar{h}\bar{n}\bar{l}\bar{l}/$  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.

---

8 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* <  $*[x\bar{b}]he[t\bar{b}]$ ; *ѓон* < Tk. *gön* ‘leather’, *меѓу* <  $*me\bar{h}u$ .

### 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 **ѧѡ-** or **ѧм-**, 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 especially problematic: I follow ESSJA I p.100 which ultimately plumps for \*azъ, but Derksen doesn't discuss it at all. A lengthy discussion of the evidence can be found in Teneva's (2012) article on the subject

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 \*h<sub>1</sub>ēd, 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 **ѧсти** 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 \*ē reflexes in the modern forms of verbs like Ru. *есть*, Pol. *jeść* are later generalisations from prefixed forms like OR **ѣнѣти**, where no jot-prothesis could take place (cf. Schenker 1995: 88, Winslow 2022: 302 fn.14).

With word-initial \*ji-/\*jъ-, I follow Derksen's (2009: 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) *їдуть* appear to have treated 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 transferred to the Polonism *шукати*) shows the expected Ukr. softening of the /s/ after fallen weak-jer in \*ъsk groups (cf. польський).