

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 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. (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 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* <*duxěx̥; Polish *wszak* vs Supr. *вѣакъ*, Ru. *всѣакъ* [уѣ] <*vьx-akъ), as well as the probable complete absence of PV2² 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* <*vьxoјѣ from letter №850), and which come from letters which otherwise correctly convey the jers (by writing <ѣ,e> for *ь and <о,ъ> for *ъ), the weak-jeŕ 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
	ɨ ʏ ɪ		ʏ ɨ ʏ ɪ
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	

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: /k̑ ġ x̑ f ü/, e.g. in respectively кѢТЪ <*kit̑, ѿѬМОНЪ <*ġemont̑, хѢТОНЪ <*x̑iton̑, ѱѢФЪ <*ijosif̑³, and мѹро <*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': ѧѼѲѳѸ vs ѧѼѲѳѸ vs ѧѼѲѳѸ vs егѹѲѳѸ vs ѧѼѲѳѸ⁴, etc.. One might also ask whether a separate <Ѽ> letter for /ğ/ (and the writing of <ѣ> under the palatalisation-diacritic) could be linked to the inadmissibility in the native systems of soft [k^j, g^j] sounds, and whether their replacement with regular <ѡ, ѣ> or <г, к> was more likely in systems with some level of native [k^j, g^j] (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. **ꙋꙋбѣ** < *orb-q vs Uspenskij Sbornik nt. acc. sg. **ꙋꙋꙋꙋ** < *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 /ɣ/ and /ě/ are required to account for the split between North (East and West) and South Slavic forms of certain inflectional-endings: *y for the nom. sg. masc./nt. pres. act.

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 justified in any context.





6 It's possible to argue that the short **Ē* 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* <**ръčĒla*, or was newly phonemicised, in e.g. Ru. *вёсна* <**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 **Ē* (and indeed the fact that

Reflexes of the so-called jot-palatalisation are all written either as unitary palatal phonemes, or in the case of jot-palatalised labials as /vǐ mǐ bǐ pǐ/, rather than as sequences of consonant + /j/, hence /ń ǐ ǐ/ for *nj *lj *rj. The ‘dejotated’ reflexes of *tj (and *kt+front-vowel) and *dj are denoted using the modern Serbian Cyrillic letters /h/ and /ĥ/ respectively, because the commonly used alternatives, i.e. /t̥ d̥/ (as used in e.g. Olander 2015) or /k̥ ġ/ (as used by me in Winslow 2022), or variations thereof, are visually too close to symbols used elsewhere in the system. /k̥, ġ/ are anyway already used in my system for foreign /k, g/ before front-vowels, and /t̥ d̥/ look too similar to the common denotations of secondarily-palatalised post-Jer Shift /t' d'/, as used in discussions of systems like Russian or Eastern Bulgarian where they arise.

The compelling hypothesis, first proposed by Durnovo (1929: 55-58) but most recently elaborated by Vermeer (2014: 209-214), and accepted by Mathiesen (2014: 197 fn. 22) and Winslow (2022: 310 fn.25), according to which the Urkirchenslavisch reflexes of *ĥ, ĥ were close enough to foreign /g k/ before front-vowels that the original Glagolitic system used <ꙗ ꙗ> for both sets (i.e. alongside attested ꙗꙗꙗꙗꙗꙗ < ἡγεμών would have been **ꙗꙗꙗꙗꙗꙗ < *osq̃heni, and alongside attested ꙗꙗꙗꙗꙗꙗ < *dъherъ would have been **ꙗꙗꙗꙗꙗꙗ < 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:očišhenje vs. jъskati:jъščq, jĀzditi:jĀžĥq vs jъzgъnati:jъžženq.

There are convincing arguments for PV2/3 having preceded dejotation, at least in more central areas, most recently presented in e.g. Vermeer (2014: 197) and Wandl & Kavitskaya (2023 244-247), and therefore it could be objected that my system, which contains the dejotation reflexes /ĥĥńǐ/ but not the PV2/3 reflexes /c ś dź/, is ahistorical. However it should be reemphasised that the primary goal of my LCS reconstructions is to act as an index which allows reflexes in texts to be found, not to be a historically realistic description of some actually-existing LCS dialect. The absence of PV2 in Novgorodian shows that it 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 /ĥĥńǐ/ in my system strictly speaking would really just be cover-symbols for the pre-jotation sequences, but such notation is preferable since it prevents searches for groups containing /j/ alone from returning results polluted by all the dejotation-groups. As I explored in my previous article (Winslow 2022), the status of /j/ as a phoneme in the earliest OCS texts is an intricate problem, so the ability to investigate the reflexes of *j in isolation from the dejotation-reflexes is important.

/j/

Word-initial *jĀ-/a-

The tendency for ECS *ā- to have taken prothetic /j/ by LCS times (in accordance with the drive towards open syllables) can make it difficult to distinguish these groups from *jĀ- in the absence of

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

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 $\Delta\text{ѡ}$ - or $\text{ѡ}\Delta\text{М}$ -, the single Greek cognate $\alpha\mu\eta$ adduced by ESSJa I p.70 in favour of jot-less *am- is not enough to categorically exclude the alternative *j $\bar{\text{A}}$ ma. In particular the 1sg. nom. pronoun *azъ/j $\bar{\text{A}}$ zъ 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.)

Jers before *j

10 Marianus and Psalterium Sinaiticum, on the other hand, frequently show a Russian-style /ej/ reflex of strong tense *bj: Psal. $\text{vbrb}^{\text{b}}\text{ej}^{\text{b}}$ <*vorbyjь, $\text{rpb}^{\text{b}}\text{ej}^{\text{b}}$ <*rplybyjь, $\text{m}^{\text{b}}\text{nj}^{\text{b}}\text{ej}^{\text{b}}$ <*mnybyjь; Mar. $\text{z}^{\text{b}}\text{apov}^{\text{b}}\text{ej}^{\text{b}}$ <*zapovědyjь, $\text{adaf}^{\text{b}}\text{ej}^{\text{b}}$ <*udařbyjь, $\text{g}^{\text{b}}\text{voz}^{\text{b}}\text{ej}^{\text{b}}$ <*gvozdybyjny-jě.

or as nativised *marъjĒ, *stadъjъ, but there are no occurrences of jer-spellings in these words in the OCS texts in TOROT. Other similarly-Greek words like дѣволъ (< διάβολος), however, do show up in OCS with jer-spellings: Supr. дѣволъ, Zogr. Luke 8 and Psal. Psalm 108 ѡѡдѡѡѡ, which (alongside the modern Macedonian *ѓавол* with the reflex of *ђ produced by the Macedonian so-called ‘new jotation’ of /d/ after the fallen jer brought it into contact with /j/) clearly suggest an early adaption of this foreign /ij/-group to native /ĭj/. Old Russian texts even show spellings of мариѡ suggestive of full nativisation: Laurentian Primary Chronicle *мѣрьѡ*, *мѣрьѡ*, Zadonschina *мѣрьѡ*, *мѣрьѡ*, as well as First Novgorod Chronicle gen. sg. *вѣснѣѡ* (jo-stem *вѣснѣнн* < 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 *ъj occurred: *diĭĒvolъ, *vasiliĭъ, *mariĭĒ etc.

Word-initial *jъ-/*ji-/i-

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) *йдуть* appear to have treated the weak-jer in *jъdъtъ 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*).

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 *иѡковъ* (borrowed via Gk. *Ἰακώβ*), or *ѡемонъ* (< 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 *иѡсѡвъ* < Gk. *Ἰησοῦς*, because of the greater likelihood that Slavs will have heard of Jesus even before the first biblical translations, and because spellings like Zogr. *ѡѡ ѡсѡ* suggest the same /Ŷ/ archiphoneme reflex of *ъ before *j as you get in native Mar. *вѡ иѡтинѡ* < *vъ ѡ jъstinъ (see above).

Prefixed forms like *do-jъti ‘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 *доїму* (<*dojъti 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 (шюица) and *vojinъ ‘warrior’ (воинъ, as opposed to *vojъnъ, the gen. pl. of *vojъna), or the loc. sg/pl. desinences of any jo-stem noun whose stem ends on /j/, e.g. Psal. *ѡѡѡѡѡѡ* <*žerъji.

Word-initial *je-/e-

No Glagolitic text makes any effort to distinguish /je/ (after vowels or word-initially) from post-consonantal /e/, writing both with <ѡ>, unlike the situation with the reflexes of *jĕ vs *ĕ, where in Zogr. and Mar. and partially in Assem. (Velcheva 1981: p.168) the full front-nasal digraph <ѡѡ> is

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

reserved for *je, while just the second 'nasalising component' <ɛ> is used for post-consonantal *ɛ, e.g. Mar. 3rd pl. aorist **ʒɛɛ** <*jese, as opposed to KF **ɲɪɹɛ** <*prijeti vs **ɲɪɹɛ** <*vɲɛli¹⁴.

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, but Suprasliensis, with its jotated <ѣ> letter, does 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*, *eġŭrъtъ*, *elisavetъ*, *elinъ*, *evanġelistъ*, *eġŭrъtъskъ*, *elinъskъ*, *episkupъstvo*, *evrejъskъ*, *elisejъ*, *etъmausъ*, *etijorъskъ*), the only spellings with <ѣ> are ѣлисеи, ѣѡпъзъ, ѣлини, and ѣлина, 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 <ѣ>, vs 3070 with <ѣ>¹⁵. Thus 97.2% of native word-initial *je- in Suprasliensis is spelt with <ѣ>, while 97.5% of the occurrences of the clearly post-LCS Greek-mediated foreign borrowings listed above instead use plain <ѣ>, 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-jotated *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.

Prefixes

The last particularity of my LCS indexing-system worth mentioning relates to the handling of consonant-clusters in prefixes: as exhaustively exemplified by Diels (1963: 121-125), Common Slavic permitted only a restricted set of consonant-combinations in the syllable onset, generally either combinations of the continuants */*z plus obstruent or sonorant (except */r, see below), or of obstruents plus sonorant (with some curiosities such as the seeming tolerance of */bn but not */pn: OCS *ръзѣнѣти* <*gyb-*noti* but *оуѣзѣнѣти* <*usъr-*noti*, cf. 3sg. aor. *оуѣзѣ*).

Geminate consonants were banned and either simplified (и́сѣшти <*j̥s-sekti) or dissimilated (про́цвисти <*prokvit-ti).

The ban on *sr/*zr is dealt with by insertion of *t and *d respectively, but the commonly-cited examples of *str <*sr (цестра, ерговѣа, ѡстръ) all concern root-internal *sr where insertion of *t is common also to the Germanic and sometimes Baltic cognates: the examples give by Meillet (1965: 136) include: (for ерговѣа) Lith. dial. *srauja* next to Latvian *strauja*, then Germanic *straum- (> Eng. *stream*, Old Norse *straumur* etc.); (for ѡстръ) Lith. *aštrus*, Gk. ἄστρος (here the *s is from PIE *ḱ). As Meillet says, “*ce n’ est pas un developpement germano-balto-slave ; d’une part, le developpement d’un -t- dans le groupe sr est chose naturelle et se retrouve ailleurs (fr. pop. castrole de casserole) et, d’autre part, le developpement de t en ces conditions n’est pas general en baltique: str est regulier en lette, mais sr subsiste couramment en lituanien.*”, so we can’t really be sure when the Slavic change took place or whether it was still active during our LCS stage. The only indication of its activity in OCS is the single Psal. 𐌱𐌰𐌹𐌳𐌵𐌿𐌽𐌾𐌰𐌺𐌴 <*sorm-omъ spelling cited by Diels (p. 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. **вѣздъ** ‘roared’, from *vъz-ruti), but 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 **ѡбѣзѣ** <*jъ.z ѱr.q.kъ. Meillet (p. 136) also cites the Old Polish adverb *zdręki* <*jъz ѱr.q.ky, which proves that the phenomenon is not limited to SSL or OCS. Curiously, though, despite this overwhelming evidence of a synchronic *zr > /zdr/ rule in OCS, /zr/ from the

14 Psalterium Sinaiticum contains just six occurrences of non-digraph <ε>: **ἔ**κλυσε, **ἔ**κλυσε, **ἔ**κλυσε, **ἔ**κλυσε, **ἔ**κλυσε, and **ἔ**κλυσε, according to Eckhoff's digitisation, five of which I've confirmed with Altbauer's (1971) facsimile of the manuscript. **ἔ**κλυσε is from Psalm 151 in the newly-discovered part and thus not included in Altbauer's facsimile.

15 The leftover 14 are things like 1st. pres. dual. *ймавѣ* which Eckhoff's corpus wrongly lemmatises as **jъmāti* instead of **jъmĕti*, and which thus get reconstructed as **jĕmĕvē* instead of **jъmavĕ*. 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.

metathesised *zork- root is never spelt <ЗДРАК> and so seems to be tolerated, even though Diels cites prepositional forms like Supr. БЕЗДРАЗУМА, БЕЗДРАЛА which come from metathesised *orT-groups <*bez ◡*orzuma, <*bez ◡*ordla 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 **ЗѢБѢТЪ ѿ ꙗзѹѡ** (Diels 1963: 122). The Old Rus. Uspenskij Sbornik is pretty consistent in keeping prefixed verb-forms like **РАЗДРУШИТЬ** <*orzrušitъ, but by the time of the Laurentian Codex we get forms like **ВЪЗРАДУЕМ** and **НЕЗРЕЧЕНОЕ**.

Because things like the *zr > *zdr, or the *ss > *s occur most frequently at transparent prefix-boundaries, and because of the clear tendency, even in the earliest texts, to undo them, I prefer to reconstruct them **will** strictly speaking illegal *zr and *ss groups, because that way an investigator can see for themselves the extent of the adherence to the expected phonological development vs restoration

јъskĕliti (Meillet 1965: 133)
ОБНОВИТИ/ОХОДИТИ ОШЬЛЪ

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 *tati 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 ВЛАДЫКЪ ѿЛОВЪКОЛЮБЬІЕ ІАКО ПРѢДАННИКА РАЗВОЙНИКА ТАТИ**...), but in the dialect underlying the 1229 Treaty the rise of phonemically palatalised /t'/ after the Jer Shift means that the stem (and the nom. sg. **ТАТЬ** /tat'/) of this noun now ends on the same class of "soft" consonants as original jo-stem nouns like *pastyrъ > /pastyr'/, where the original LCS palatal *ř has fallen together with secondarily-palatalised /r'/ from plain LCS *r before LCS front-vowels, in e.g. the original i-stem *zvĕrъ > /zvĕr'/. 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 **ТАТЬ** 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 it 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.

Forms like **ТАТА**, then, though they frustrate our goal of reconstructing entire texts, do provide us some objective measure of 'linguistic distance' between stages of a language, because

//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 /řĚ/ (palatal /ř/ < *rj + the archiphoneme /Ě/) to occur, then a morphological change which replaces the sequence /ri/ with /řĚ/ 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 *разрѣти* < *orzařĚti from the prefixed *разорити* *orzoriti is irrelevant, because LCS /orzařĚti/ does not violate the rule of LCS phonotactics: palatal /ř/ can be followed by /Ě/ because such a combination exists in the paradigms of wholly securely reconstructable jo-stem nouns, e.g. nt. gen. sg. *mořĚ (> Pol. *morza*, OCS *морѣ*, Ru. *морѣ*, etc.)

In the case of Supr. Gsg. masc. *звѣрѣ*, for an original i-stem (*звѣри* < *zvěri), a direct LCS ancestor for the attested form can still be given (*zvěřĚ), because palatal /ř/ 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 /ř/, so that the nom. sg. *zvěř became /zvěř'/, and its stem now ended on the same consonant /r'/ as original ja- and jo-stems ending on LCS *ř like *морѣ* and *воуриѣ*, 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,

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

-Could use the *овсяяныи* OR adjective as another example of derivational-morphology made possible only by the rise of soft /s’/ (if it can be confirmed that the *ěнь adjectival suffix in e.g. OCS *оловънь* ‘leadene’ is LCS)

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



LCS Morphology and the Autoreconstructor

- Consonant-stems – with the *tel- suffix agent-nouns, I mostly follow people like Meillet (1965: 426) in taking consonant-stem endings in most of the plural, but the nom. pl. it’s difficult to agree with his positing of a plain /le/, as opposed to palatalised /Ě/ desinence (i.e. with the consonant-stem vowel on the jo-stem stem), because Zographensis and Suprasliensis are consistent in marking such forms with their palatalisation-diacritic.

- related is derivation-morphology difficulties such as whether adjective *volъnъ should have a palatal /l̥/, or whether the *volĀ is specifically differentiated from the root *vol- by a *-jĀ noun-forming suffix. Spellings are similarly suggestive of *vol-

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

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

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 past-active-participle and 1sg pres. indic. of class IV (-iti) verbs: we can just use the desinences *-jъ and *jq 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.