Peer-reviewed article

Beyond Havlik's Law: a dataset for quantitative investigation

Aleksandrs Berdicevskis[1]

Språkbanken, Department of Swedish, multilingualism, language technology, University of Gothenburg

aleksandrs.berdicevskis@gu.se

Hanne Eckhoff

Faculty of Medieval and Modern Languages, Faculty of Linguistics, Philology and Phonetics, University of Oxford

hanne.eckhoff@mod-langs.ox.ac.uk

Joseph Winslow[2]

Faculty of Linguistics, Philology and Phonetics, University of Oxford

joseph.winslow@trinity.ox.ac.uk

Abstract

This article explores a methodology for large-scale quantitative study of the jer change as it is reflected in Contemporary Standard Russian (CSR) as represented by Russian Wiktionary. The aim is to find out what proportion of the modern forms can be accounted for by a pure unmodified application of Havlik's Law, so that exceptions can be more systematically identified and investigated. We

[1] Berdicevskis has been supported by Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161).

[2] The authors are listed in alphabetical order. This article would not be possible without Winslow's pre-existing project on using LCS reconstructions as phonological 'annotation', and he is responsible for all coding and conversion scripts, inflection tables as well as the vast bulk of the LCS reconstructions (with some assistance from Eckhoff). Eckhoff and Berdicevskis did all manual post-correction and matching of TOROT and CSR lemmas and all manual annotation. Eckhoff did the main bulk of the writing and most of the pilot study. The key scripts and datasets, as well as a full account of the reconstruction principles, can be found in the TROLLing Dataverse at https://doi.org/doi:10.18710/YACE1Z. The complete run of scripts and documentation can be found at https://github.com/hanneme/jer_project

CC BY 4.0

employ a computerised inflectional morphology of Late Common Slavic (LCS) to generate the ancestors of CSR forms, as well as a forward-reconstruction algorithm that attempts to convert these into CSR, with a strict implementation of Havlik's Law. Where the forward-reconstructed forms do not match the Wiktionary forms, we manually annotate them for the potential causes of the missed match, including the ones most frequently invoked in the literature: "difficult" consonant clusters, analogy and Church Slavonic influence. We find that especially consonant clusters and Church Slavonic influence are prominent potential causes. Overall we find that Havlik's law holds in a large majority of cases. We also provide a small pilot study where we test some of the morphophonological rules proposed by Isačenko (1970) to demonstrate the methodology's potential.

In 2015 Tore Nesset and Aleksandrs Berdicevskis co-authored an abstract entitled "YERs are Yearning for Empirical Research: A corpus approach to diachronic phonology", where they proposed to do a large-scale corpus investigation of deviations from the jer change, i.e. cases where Havlik's Law does not come through and the deviation cannot be explained by occurrence in TЪRT or TRЪT groups. These deviations are usually ascribed to "difficult" clusters, analogical levelling and Church Slavonic influence. Nesset and Berdicevskis's ambition was, in particular, to formulate criteria of "difficulty", and to find ways of predicting whether analogical levelling is expected to take place, and if yes, which form will serve as its basis. Their methodological idea was to use Old East Slavonic lemmas from the TOROT treebank (Eckhoff and Berdicevskis 2015) as the point of departure due to TOROT's archaising lemmatisation policy where weak and strong jers were retained in all positions. The abstract never resulted in a published piece of work, but in this article, we pick up on this research idea and propose a related, but more robust methodology for historical corpus phonology that will allow researchers to identify, and thus allow systematic study of, all deviations from the regular jer-change rules across morphological paradigms. Our method relies on reconstructing Late Common Slavonic (LCS) lemmas, using them as a basis to generate reconstructed LCS morphological forms, and then forward-reconstructing Contemporary Standard Russian (CSR) forms. The predicted forms are then compared with actual CSR forms. The main contribution of this article is thus methodological, but it also includes an annotated dataset and a small pilot study on cluster avoidance and analogical change.

1. Introduction

It is well known that many modern forms do not match the expected outcome of Havlik's Law (Havlik 1889) and also cannot be explained by the presence of etymological TЪRT or TRЪT groups, and various generalisations have been proposed (see e.g. Galinskaja 2009: 92–94, Žuravlev 1977: 40–41). Some mismatches are believed to be caused by "difficult" consonant clusters that lead to vocalisation or cluster simplification, such as *дъска* 'board' yielding *доска* in CSR rather than expected *цка* (which is in fact found both in early texts and Russian dialects). Isačenko (1970) goes a long way towards formulating detailed morphophonological rules for these deviances. Other mismatches are put down to analogical levelling, which is clearly exemplified in OES *шьвьць*, *шьвьца* 'tailor, cobbler' yielding CSR *швец*, *швеца* rather than *швец*, *шевца*. Another considerable group of deviations is ascribed to Church Slavonic influence, most frequently invoked for vocalisation of weak jers in prefixes and adjective suffixes, such as *съгрѣшити* 'to sin' yielding CSR *согрешить* and *грьчьскыи* 'Greek' yielding *греческий* alongside the expected *гречский* (cf. *грецкий орех* 'walnut'). Isačenko (1970) tends to (over-)use this explanation when his morphophonological rules are violated. Overall, there is no clear consensus on whether these generalisations can be formulated as clear rules, and if so, how, and we believe a more systematic examination of all documented deviations is warranted.

In this paper we therefore propose a strict corpus-based methodology to study the jer change as it played out in the history of Russian, comparing forward-reconstructed expected CSR forms to actual CSR forms and discussing the main groups of discrepancies. Importantly, our forward reconstruction is formalized and performed automatically. A similar methodology (but with a broader focus) was used by Marr and Mortensen (2020) to analyze phonological changes from Latin to French. We take a goal-oriented approach, asking how a set of CSR lemmas and their inflected forms came to have the forms they have, given their known (and most closely matching) jer-containing etymon.

We recognise that our goal-oriented approach differs from a more traditional source-oriented approach, which is more common in historical phonology, namely to look for the best-fitting outcome of each etymon in all available sources in a particular linguistic community, including e.g. dialect documentation. Choosing CSR lemmas as found in modern dictionaries (and deemed by Berdicevskis as a native speaker not to be overly archaic) as our endpoint certainly has its drawbacks, such as the obviously normative and to some extent written-language-biased way this inventory came into being. However, forms from a modern dictionary is the closest we can get to a full description of the system. A large-scale study, where a large range of forms from different East Slavonic standard languages, dialects and earlier language stages were taken into account, would run into delimitation problems as well as severe lack of dialect and historical data. In any case tracking down the best outcomes in any relevant source is

well beyond the scope of this largely methodological experiment. Focusing on CSR allows us to extract full morphological paradigms for any lemma found in Russian Wiktionary. Thus we can avoid the potentially vicious circle of finding CSR matches by forward reconstruction alone, which would inevitably limit the number of deviant outcomes detected.

2. Method

We extracted lemmas from the TOROT Old East Slavonic and Middle Russian Treebank that *either* have a lookup form that contains a non-final jer, *or* has at least one form in the attested form inventory that contains a non-final jer,[3] with the following limitations:

> -We exclude all proper nouns, since they often have unclear etymology and/or are loans
> -We exclude all lemma forms that contain a whitespace (such as *не быти*, complex numerals)
> -We only check for jer-containing forms of verbs and common nouns, and limit them to indicative forms of verbs and genitive forms of common nouns. Words from other parts of speech were included on the basis of their lemma form only.

This produced a list of 5455 lemmas that fulfilled the criteria. Some of these lemmas are still not relevant for our purposes, for example, the noun *сестра* 'sister' is attested with the acc sg form *сьтру* (<**sestrǫ*), which is a Novgorodian spelling found in BBL 531. These lemmas were excluded later. We also excluded all complex numerals (beyond those with lemma forms with whitespace) since most have had highly idiosyncratic developments and/or were only lemmatised separately as a consequence of early Cyrillic numeral notation.

The next step was to establish modern Russian counterparts for each of the relevant lemmas, wherever such a counterpart exists. This was done semi-automatically, using the integrated Old East Slavonic and Middle Russian dictionary found at https://syntacticus.org/dictionary/syntacticus:20180920:orv. The dictionary was created by providing ca. 8000 of the lemmas in the TOROT Old East Slavonic and Middle Russian treebank with Russian and English glosses (Eckhoff and Berdicevskis 2016). The dictionary naturally focussed on finding the best meaning match, not the best etymological match. For our list of counterparts, we used Levenshtein distances to compare the OES lemmas with the Russian glosses, automatically selecting the gloss with the lowest Levenshtein distance from the OES lemma form. The list was then

[3] An advantage of this approach is that we are able to include a number of forms with non-etymological jers, such as gen sg *льда* 'ice' from *leda.

manually checked, corrections were made and missing Russian counterparts were supplied, if possible, for the 1800 most frequent OES lemmas from our original list. The matching was done according to the following principles: i) only one etymon per CSR lemma, ii) the matching is based on form, not semantics, iii) reflexive markers on verbs are ignored, since they are not included in the TOROT OES lemmatisaton.

This produced a list of 1513 modern CSR matches for OES etymon lemmas. The frequency criterion was chosen simply to limit the manual workload. It has some disadvantages, in particular the fact that many stems and derivational affixes are repeated on the list in various combinations, to the exclusion of less frequent stems and affixes that have potentially interesting jer features. It also includes a number of derivations from proper nouns (adjectives such as *твѣрьскыи*, nouns such as *хоросаньць*), which were excluded from the list in their underived form. Since our goal is to account for the modern language, it might have been better to choose lemmas based on CSR frequency, but we would still be limited to ones with OES ancestors attested in TOROT.

Every OES lemma was then matched to its reconstructed and inflection-class-marked LCS ancestor lemma, in a dataset originally created by Winslow for the purposes of annotating PROIEL/TOROT OCS texts with LCS reconstructed forms (see https://ocstexts.co.uk). The idea behind such reconstructions is to encode all the phonological information necessary to produce descendant forms, so a point after the Common Slavonic monophthongisation of diphthongs but just before the second and third velar palatalisations was chosen. Existing reconstructions had been created for TOROT OCS lemmas, so these were automatically matched to their OES counterparts by forward-reconstructing and converting to TOROT's OES lemma-orthography, though several hundred still had to be specifically reconstructed to cover all our target lemmas. A full description of the reconstruction conventions is beyond the scope of this article, but is included in the data repository. Of importance for this article is our decision to reconstruct LCS syllabic liquids as unitary phonemes /ŕ̥ ĺ̥ r̥ l̥/, rather than as biphonemic sequences of liquid + jer, as is often done, so OES *vъlkъ* 'wolf' is reconstructed as *vĺ̥kъ rather than *vьlkъ. These sequences develop distinctly from other jer groups, and the extent to which their reflexes fall together with those of the jers is coincidental and varies dialectally across Slavic (cf. e.g. Bethin 1998: 69–75).

In most cases the matching of lemmas with LCS reconstructed lemmas was unproblematic. However, some lemmas had to be excluded because no plausible LCS reconstruction was possible.[4] These issues largely stem from the fact that the TOROT OES lemma inventory also contains lemmas from Middle Russian texts, some of which must clearly have been derived (and/or borrowed) later than LCS. An example is the

[4] The lemmas in question were *хрьстиянинъ*, *хрьстияньскыи*, *батюшька*, *тюрьма*, *миленькыи*, *приказьщикъ*, *хорошенько*, *матушька*, *овьсяныи*, *бѣлозерьць*, *отъсюда*, *хрьстити*, *бѣлозерьскыи*, *обезияна*, *хрьстияньство*, *сибирьскыи*, *кърмьщикъ*, *дъмъра*, *лямъка*, *помаленьку*, *багъряныи*, *тряпезьница*, *псалътыря*, *полъть*, *дрьгъвичи*, *метельникъ*.

OES lemma *батюшька*, which only occurs in the Life of Avvakum and 18th-century Muscovite correspondence in the TOROT treebank. It is derived from LCS *batę with the suffixes *их* + *ька*. However, it can only have been derived after the rise of phonemically palatalised /t'/, which is a post-LCS development. In such cases the lemma was removed from the list. Lemmas that were unlikely to have existed in LCS (e.g. derivations from stems that are likely to be post-LCS loans), but are phonotactically feasible for LCS, were not excluded, for example the Kipchak Turkic loans *деньга* 'coin', cf. Kazakh *теңге*, and *тамъга* 'seal, mark', cf. Kazakh *таңба*. Clearly these did not exist in LCS, but the derived adjectives *денежный* (<**деньžьпъјь*) and *таможный* (<**тамъžьпъјь*) suggest that their closed syllables were adapted to pre-jer-shift OES phonotactics by jer-insertion. This left us with 1464 OES lemmas with LCS reconstructed counterparts.

Full paradigms of each LCS lemma are produced using tables containing inflectional endings associated with each lemma's inflection class.[5] These tables were originally created to produce the reconstructed LCS forms of words in the TOROT OCS texts found at https://ocstexts.co.uk, and they cope with post-LCS morphological restructuring by including alternative tables on top of the base idealised LCS one. The morphological innovations occurring in canonical OCS texts are different and far less extensive than those responsible for the modern Russian paradigms, so the alternative tables were specially adapted for this project.

Table 1 gives an example of the reconstructed LCS paradigm for the lemma *дъчи/дъщи* 'daughter'.[6] This noun is etymologically a consonant stem (specifically we use the classification-label *fem_R*, used also for *мати* 'mother'), but even in the very earliest sources it is attested with ĭ-stem endings. The alternative tables therefore contain regular LCS ĭ-stem forms (e.g. gen/loc sg, gen/loc du, nom, gen, voc pl), as well as the short nom/acc sg LCS form *dъ̀ħь to be able to produce the later shortened form *дочь*, which is not observed in the TOROT data until the 16th century. For both this type and for regular ĭ-stems we also produced alternative ja-stem forms in the dat, ins and loc pl, to account for the modern forms in *-ям*, *-ями*, *-ях*.

[5] The inflection routine is more sophisticated than simply adding endings to stems; the guts of how it works can be seen in LcsFlecterJerProj.cpp

[6] The Old East Slavonic lemmas in TOROT often come as doublets of one Eastern and one Southern form. This is because Church Slavonic forms frequently occur in early East Slavonic texts: in the Primary Chronicle (Laur.), for example, we find forms such as *дъщери* and *тъчери* occurring within a few lines of each other, and they are lemmatised accordingly. Associating both lemmas with the same LCS reconstructed lemma is an efficient and informative way of grouping such doublets.

	sg	du	pl
nom	dъĥi: дчи, дщи dъĥь: дочь, дощь, доч, дощ	dъĥeri: дчери, дщери, дчори, дщори	dъĥere: дчере, дщере, дчоре, дщоре dъĥeri: дчери, дщери, дчори, дщори
acc	dъĥerь: дчерь, дщерь, дчорь, дщорь dъĥь: дочь, дощь, доч, дощ	dъĥeri: дчери, дщери, дчори, дщори	dъĥeri: дчери, дщери, дчори, дщори
gen	dъĥere: дчере, дщере, дчоре, дщоре dъĥeri: дчери, дщери, дчори, дщори	dъĥeru: дчеру, дщеру, дчору, дщору dъĥerьju: дчерью, дщерью, дчорью, дщорью	dъĥerъ: дчер, дщер, дчор, дщор dъĥerьjь: дчерей, дщерей, дчорей, дщорей
dat	dъĥeri: дчери, дщери, дчори, дщори	dъĥerьma: дчерма, дщерма, дчорма, дщорма	dъĥerьmъ: дчерем, дщерем, дчорем, дщорем dъĥerǢmъ: дчерям, дщерям, дчорям, дщорям
loc	dъĥere: дчере, дщере, дчоре, дщоре dъĥeri: дчери, дщери, дчори, дщори	dъĥeru: дчеру, дщеру, дчору, дщору dъĥerьju: дчерью, дщерью, дчорью, дщорью	dъĥerьхъ: дчерех, дщерех, дчорех, дщорех dъĥerǢхъ: дчерях, дщерях, дчорях, дщорях

ins	dъħerьjǫ: дчерью, дщерью, дчорью, дщорью	dъħerьma: дчерма, дщерма, дчорма, дщорма	dъħerьmi: дчерми, дщерми, дчорми, дщорми dъħerǢmi: дчерями, дщерями, дчорями, дщорями
voc	dъħi: дчи, дщи dъħь: дочь, дощь, доч, дощ	dъħeri: дчери, дщери, дчори, дщори	dъħere: дчере, дщере, дчоре, дщоре dъħeri: дчери, дщери, дчори, дщори

Table 1. Generated LCS forms and their forward-reconstructed modern Russian forms for the LCS reconstructed lemma *dъħi 'daughter'

To these reconstructed LCS paradigms we applied an algorithm[7] to forward-reconstruct the expected CSR forms. Since our aim here is to study specifically the jer changes, the forward reconstruction had to try to account for all possible variation in the application of sound laws from LCS onwards (and their final orthographic representation in CSR), but keep constant just the outcome of Havlik's Law. The latter was applied in a strict Northeastern form, where all jer-groups except those resulting from LCS syllabic liquids and in word-initial *jь- were treated equally, meaning no special treatment was given either to the position before *j (so-called 'tense jers'; see Popov 2020: 374)[8] or to *TRЪT-groups (Galinskaja 2009: 89–91). Such forward reconstruction was no trivial task, as right from the very beginning of East Slavic there is lexically specific variation in the reflexes of LCS sound-groups.

Most obviously this applies to borrowed Church Slavonic forms containing Bulgarian reflexes of phonemes like *ħ and *ǧ (<щ, жд> instead of native Russian <ч, ж>), but we also have things like variation in the orthographic representation of voicing assimilation: *везде* /v'ezd'e/ < *vьśьde vs *сделать* /zd'elat'/ < *sъdělati. The so-called /e/ > /o/ shift is especially problematic, because its final outcome in Russian depends on stress placement, for which we had no data, *and* it seems to get re-applied[9] *after* the hardening of what originally would have been following soft consonants (*дёшево*,

[7] The full forwards-reconstruction script is included in the data repository as LCS_to_CSR.js

[8] An exception was made only for the nom. sg. masc. long-form adjectival endings *-ъjь and *-ьjь of participles (but not adjectives), where we permit SSl. /yj/ and /ij/ reflexes as variants, since these are often their only listed form in Wiktionary.

[9] There are various views on when the /e/>/o/ shift happened, only some of which would entail re-application, see e.g. Galinskaja (2009: 135).

молодёжь). This, combined with the possibility of <o> spellings in words like *горячо́* <*gorę̆he (but not *ещё* <*ješče), meant we felt compelled to generate variants with <o> reflexes of *e after all palatal letters, regardless of the following consonant, hence the rather ridiculous variant forms like *дчорь* <*dъ̆herъ in Table 1.

At each step in the forward reconstruction where variant outcomes like the above were possible, all possible reflexes are generated and passed onto the next step.[10] This systematic overgeneration maximises the chance of matching our generated forms with their actual CSR equivalents, and ensures that mismatches are caused *only* by deviations from the pure Havlikian outcome. In the *dъ̆i example in Table 1 the overgeneration occurs due to three factors: a) both Russian and Bulgarian reflexes of *ĥ (*дчи* vs. *дщи*); b) the <o> reflexes of *e after palatals mentioned above (*дчери* vs. *дчори*); and c) word-final *-ĥь yields forms with and without the soft sign (*дочь*, *доч*, *дощь*, *дощ*).

Our next step was to extract CSR inflected forms from Russian Wiktionary for our list of 1464 CSR lemmas that had both an OES lemma match and an LCS reconstructed lemma. We used a script (ru_wiktionary_scraper.js) to scrape inflectional forms directly from the HTML of the webpages, which resulted in deduplicated unstructured lists of inflectional forms linked to each lemma in JSON format.[11] Given that the shape of paradigms has changed significantly between LCS and CSR (e.g. loss of dual, emergence of gerunds), trying to align our LCS inflection-labels with their descendant CSR Wiktonary-labels would in any case have been unfeasible. This JSON data was then filtered against our list of matched CSR lemmas.

In several cases this produced multiple paradigms for the same lemma. In most cases these paradigms were identical, such as for the two second-declension homonyms *белка* I ‘squirrel’ and *белка* II ‘delirium tremens’. In such cases the duplicate was removed from the dataset. Duplicate paradigms that only differed in stress, with no other orthographical consequences, were also removed, such as *ну́жник/нужни́к* ‘toilet’. When paradigms of different inflection classes were listed as variant forms for the same lemma, the variant paradigms were retained, as in the case of *изымать* ‘take out’ which is listed with two alternate present-tense paradigms: *изъемлю, изъемлешь* and *изымаю*, *изымаешь*. The same was done when paradigms with two variant prefixation patterns were listed. For example, *изъять* ‘take out’ is listed with two present-tense paradigms with different reflexes of the prefix: *изыму* vs. *изойму*. In cases where one Old East Slavonic lemma was the etymon of several CSR lemmas with different paradigms, all

[10] The largest number of variants generated from a single LCS form in our dataset was 24, from the nom. sg. masc. def. present-participle *sъ̆dorvьstvujǫ̆ĥjь > *сдоровствующей*, *сдоровствующей*, *сдравствующей*, *сдоровствующий*, *сдоровствующий*, *сдравствующий*, *сдравствующий*, *здоровствующей*, *здоровствующей*, *здравствующей*, *здравствующей*, *здоровствующий*, *здоровствующий*, *здравствующий*, *здравствующий*, *сдоровствующой*, *сдоровствующей*, *сдравствующей*, *сдравствующей*, *здоровствующей*, *здоровствующей*, *здравствующей*, *здравствующей*. Full data can be found in lcs_converted.json

[11] Morphologically labelled inflection-table data *has* been extracted from Russian Wiktionary by Tatu Ylonen using his Wiktextract tool (Ylonen 2022), but Ylonen’s paper focused on English Wiktionary, and datasets from other language-versions are provided with the caveat that they are a “work in progress and may still contain many errors or omissions” (https://kaikki.org/ruwiktionary/Русский/index.html).

those paradigms were also retained. The extreme case of this is the OES lemma *вихъръ* ‘whirlwind, which was matched with CSR *вихрь*, *вихор* and *вихорь*, the latter of which has two Wiktionary paradigms, one with fleeting vowel and one without. All these four paradigms were retained.

This resulted in a dataset of 1445 CSR paradigms and a total of 17905 individual CSR word forms. The next step was to see how much of these paradigms could be accounted for from their corresponding forward-reconstructed LCS paradigms. Forms retained in our LCS paradigms which do not survive into CSR (e.g. lost verb forms such as the aorist and imperfect) are unproblematic for the matching algorithm, since it only has to account for the Wiktionary forms; forward-reconstructed forms like aorist 1sg *взях* ‘took’ (<*vъzęхъ) are simply surplus to requirements. Conversely, CSR inflectional-categories which did not exist in LCS, such as gerund forms, often straightforwardly match the forward-reconstructed output of masculine nominative singular short form present and past participle forms, which is, of course, where they originate. For example, the gerund forms of *праздновать* ‘celebrate’, *празднуя* and *праздновав*, match straightforwardly with the forward-reconstructed outputs of LCS *porzdьnujǫ (present active participle, masc nom sg short form) and *porzdьnovavъ (past active participle, masc nom sg short form). Similarly, retention of the LCS dual-forms is judicious given that modern Russian plurals occasionally descend from such forms (e.g. *плечи* ‘shoulders’, *уши* ‘ears’).

3. Dataset and annotation

In the resulting dataset, 11731 CSR Wiktionary forms (65.5% of the extracted set) found straightforward forward-reconstructed matches. 6174 forms did not. Eckhoff and Berdicevskis manually classified each missed match according to the assumed cause, into the categories listed in Table 2. A missed match could be assigned to several categories at once.

Label	Explanation	Jer realisation issue?	Count
cluster	Weak jer vocalisation breaks up a consonant cluster that is difficult to pronounce. We relied on our intuitions about which clusters are "difficult to pronounce". Our hope is that our data could potentially be used to develop a more formal definition. Example: SF nom. sg. masc. *тесен* ‘tight’ predicted form: **тесн*	yes	1990
analogy	Deviance from Havlik’s Law is in line with at least	yes	637

	one other form in the paradigm (interparadigmatic analogy is ignored) Example: gen. pl. *служб* ‘service’ (cf. nom.sg. *служба*, nom.pl. *службы* etc.) predicted form: **служеб*		
Church Slavonic influence	Deviance from Havlik’s Law is likely to be due to Church Slavonic influence Example: *собор* ‘cathedral, clerical council’ predicted form: **сбор*	yes	1638
imperative	The conversion script does not produce the full set of innovative CSR imperative forms Example: *снимите* ‘take off' predicted form: **снимете*	no	559
tense jer	A jer yields *и* or *ы* before /j/ (tense position), while no special reflex is expected in the East Slavonic dialects underlying CSR (this includes all masc nom sg LF adjectives that are not end-stressed) Example: *веселий* gen. pl. ‘merriness’ predicted form: *веселей	yes	1559
soft short form	Masculine singular soft adjective short forms were forward-reconstructed with a word-final soft sign which is not retained in the CSR forms. Example: *древен* ‘ancient' (the form is infrequent in CSR) predicted form: **древень*	no	30
prefix issue	The prefix occurred in a different form to that in the forward-reconstructed form Example: past sg. masc. *собирал* ‘collect’ predicted form: **сбирал*	yes	1817
category change	The lemma had changed inflectional pattern Example*: азбука* ‘alphabet’ (changed from ū-declension to ā-declension)	no	495
irregular change	The mismatch was due to irregular (morpho)phonological change *Example: пскович* ‘Pskovian’ predicted form: **пльскович* (cluster simplification)	some	1184
orthography	The mismatch was due to CSR orthography (in a wide sense) actual form: *влезть* ‘to climb in’ predicted form: **влесть*	no	378
other	Various other causes	no	756

Table 2. Classification of missed matches with counts and indication of whether they are due to differences in jer realisation

Some of the annotation categories mark non-matches that can be ignored, either because of their systematic nature (tense-jer spellings of nouns in *-ие* and masc nom sg LF adjectives, the systematic absence of soft stem-final consonants in masc nom sg SF adjectives), because of their idiosyncrasy (irregular change, category change, orthography) or because of a limitation of the forward-reconstruction script (such as the missing generation of various innovative imperative forms). The three first annotation categories (cluster, analogy, Church Slavonic influence), however, address the three most common generalisations about deviations from the regular jer change, which we will focus on in our small pilot study.

An important limitation of the current annotation is that is has been done only for mismatches, which means that cases when one or more of the factors listed in Table 2 (for instance, a consonant cluster, or potential for analogical levelling, or Church Slavonic influence) *are* present, but *no* deviation from Havlik's Law is observed, cannot be automatically identified from the dataset. A possibility to discover such cases is crucial for quantitative analysis of the predictive power of the explanations, and we plan to add this annotation later.

4. Pilot study

With this dataset we are, however, able to test some hypotheses about actually occurring CSR deviations from Havlik's Law. We used a cleaned-up dataset that excluded occurrences tagged "tense jer", "imperative", "orthography", "category change", "soft SF", and "irregular change". This produced a dataset of 14127 forms, of which 11701 (82.9%) were matched and 2426 (17.1%) were not.

As mentioned in the introduction, Isačenko (1970) goes far in stating firm morphophonological rules for the long-term fallout of the jer change. We will use the subset of our data annotated "cluster" (see Table 2) to assess whether the rules cited below are confirmed by our data. Rule B is a jer insertion rule, the other rules affect vocalisation of weak jers.

- Rule B: A jer is inserted between stem-final obstruent+sonorant (Isačenko 1970: 79)[12]: *огнь* 'fire' > *огънь* > *огонь*
- Rule F: Verbal roots are syllabic, i.e. they may not have the shape C or CC(C) (Isačenko 1970:88): *чьстити* 'honour' > *честить*[13]
- Rule H: A fleeting vowel becomes a fixed vowel before stem-final consonant clusters CC(C) before a grammatical morpheme (Isačenko 1970:92): *тъска* 'longing' > *тоска*

[12] Where the inventory of sonorants is /r/, /l/, /n/, /m/ and /v/.

[13] Our example, note that Isačenko does not provide any examples of this, nor does he refer back to the rule when discussing CSR verbs such as *мстить* 'avenge' (Isačenko 1970: 93), which clearly violate this rule.

- Rules I and J: Multiple fleeting vowels in a single lexeme are only allowed when two adjacent morphemes containing a fleeting vowel are separated by a morpheme boundary. In stems not divided by a prefix/preposition boundary and containing more than one vowel/zero alternation, only the last of these alternations is preserved (Isačenko 1970: 94)

1513 non-matching forms in our cleaned dataset were labelled "cluster" (10.7%), i.e. we thought it likely that a weak jer had been vocalised to avoid a consonant cluster. This included examples where the jer was not etymological. In a separate dataset (cluster.csv) these were manually tagged for

- cluster form (the exact cluster that had been broken up)
- cluster resolution type: what type of cluster was resolved by vocalising the weak jer, and was the jer etymological?
- position (prefix, suffix, verb root, stem-inital, stem-final)
- whether there were multiple potential fleeting vowels in the word and if so whether a morpheme boundary was involved and what position the cluster was in

We look exclusively at cluster resolution in roots, i.e. not in derivational prefixes and suffixes. We therefore exclude the 646 occurrences of resolved clusters in prefixes and 260 occurrences of resolved clusters in suffixes. 267 occurrences had vocalisation in a TRЪT stem cluster. Since this seems to be consistent in the dataset, as expected from the literature, these were also excluded. We also exclude occurrences in verb roots, as Isačenko's Rule F is not clear and verbs overwhelmingly have stem suffixes. This leaves us with 224 forms belonging to 40 lemmas.

Looking first at Isačenko's Rules I and J, we found that only 12 of the forms[14] in our nominal root subset had the potential for two fleeting vowels, ten of them across a suffix boundary. In all of these ten,[15] the vocalised weak jer was found in the root-final syllable, and was vocalised in addition to the strong jer in the following derivational suffix, and in all of them the explanation can also be analogy. For example, CSR gen pl *ложек* (predicted **лжек*) 'spoons' from LCS *lъžьkъ vocalises both the weak jer in the root, in keeping with the rest of the paradigm, and the strong jer as a fleeting vowel in the diminutive suffix.

There were also two examples with potential for two fleeting vowels in the root, masc nom sg short adjective form *лёгок* (predicted **льгок*) 'light' from LCS *lьgъkъ and gen

[14] See Isačenko 1970:96–97 for discussion of some of these forms.

[15] *бочек, душеполезен, ложен, ложек, любовен, молебен, отечеств, солнечен, точен, честен* vs. predicted **бчек*, **душепользен*, **лжен*, **лжек*, **любвен*, **мольбен, *отчеств*, **солнчен*, **тчен*, **чстен.*

pl *мечт* (predicted **мчет*) ‘dream’ from LCS *mьčьtъ.[16] Again, both examples are also tagged for potential analogy, since the vocalisation of the weak jer in the first syllable is in keeping with the rest of the paradigm. While the first example does preserve the last fleeting vowel, but turns the first one into a constant vowel, in accordance with Isačenko’s Rule J, the second example does not, but rather eliminates the last fleeting vowel and turns the first vowel into a constant vowel. This is then a violation of Rule J, but in keeping with an analogy-based explanation.

Turning to Rule H, there were 212 nominal forms with potential for only one fleeting vowel, belonging to 19 noun lemmas and nine adjective lemmas. Since the deviant adjective forms[17] are mostly either due to a corresponding noun with the same behaviour or part of the root being reinterpreted as a suffix, we will ignore them here and discuss only the 132 noun forms. The 19 CSR lemmas in question are *бисер* ‘pearl’ < LCS *bisьrъ, *вопль* ‘howl’ < LCS *vъpĺь, *весь* ‘village’ < LCS *vьsь, *гот* ‘Goth’ < LCS *gъtъ, *дверь* ‘door’ < LCS *dvьrь, *дождь* ‘rain’ < LCS *dъžđь, *доска* ‘board’ < LCS *dъska, *дочь* ‘daughter’ < LCS *dъđь , *дебри* ‘wilderness’ < LCS *dьbrь, *жезл* ‘staff’ < LCS *žьzlъ, *земля* ‘land’ < LCS *zemĺĀ, *кознь* ‘craftiness’ < LCS *kъznь, *лесть* ‘flattery’ < LCS *lьstь, *месть* ‘revenge’ < LCS *mьstь, *прелесть* ‘delight’ < LCS *perlьstь, *сноха* ‘daughter-in-law’ < LCS *snъxa, *тоска* ‘longing’ < LCS *tъska, *тесть* ‘father-in-law’ < LCS *tьstь, *честь* ‘honour’ < LCS *čьstь.

59 of these forms follow Rule H, i.e. a fleeting vowel becomes a fixed one before a consonant cluster directly before a grammatical morpheme. The nine lemmas involved are *дождь, доска, лесть, месть, прелесть, тоска, тесть* and *честь*.[18] By definition, all of those can also be accounted for by analogy, since at least one form in each paradigm has the jer in question in strong position. We may also note that if the weak jer is not vocalised, it will in all cases produce a CCC stem before the grammatical morpheme (preceded by a prefix in the case of *прелесть*). This is in fact something they have in common with several of the lemmas that do not follow Rule H, namely *дверь, сноха; вопль, дебри, жезл* and *кознь* accounting for a further 43 forms. Of course, the latter four *do* have a stem-final CC group, but in this instance the CC group consists of obstruent+sonorant, and we would have expected jer-insertion by Rule B instead.[19] It is therefore tempting to suggest that a generalisation stating that weak jers may be vocalised to avoid CCC stems before grammatical morphemes may account for more of the data than Isačenko’s Rule H does.

[16] While gen pl of *мечта* is often considered to be a paradigmatic gap by dictionaries, the form *мечт* does occur, and, more importantly, there is no rival potential form with a different outcome of the jer change.

[17] From CSR lemmas *благочестивый*, *жёсткий*, *земельный*, *напрасный*, *нечестивый*, *пресный*, *полный*, *тощий*.

[18] All these lemmas are discussed by Isačenko (1970: 92–93 and 102–103 for the deviant gen pl *досок* of *доска* ‘board’).

[19] We include *вопль* here since gen pl *земель* of *земля* demonstrates that j-palatalisation products of labials are not immune to jer insertion.

Many more observations may be made even in this very small subset of our data, and it is clear that the full dataset has great potential for further hypothesis testing and generalisations.

5. Conclusion

The main contribution of this article is an exploration of a methodology for large-scale quantitative study of the jer change as it is reflected in CSR as represented by Russian Wiktionary. The idea was to find out exactly what proportion of the modern forms can be accounted for by a pure unmodified application of Havlik's Law, so that exceptions could be more systematically identified and investigated. To this end we employed a computerised inflectional morphology of LCS to generate the ancestors of CSR forms, as well as a forward-reconstruction algorithm that tries to convert these into CSR, but with a strict implementation of Havlik's Law. Our final dataset included 17905 CSR forms from 1445 different paradigms, which were automatically matched against their computer-generated predicted forms. The non-matches in the dataset were manually annotated for the potential causes of the missed match, including the ones most frequently invoked in the literature: “difficult” consonant clusters, analogy and Church Slavonic influence. We found that especially consonant clusters and Church Slavonic influence were prominent potential causes, while there were also a number of other prominent factors that could prevent a match, such as irregular (morpho)phonological change. In a smaller version of the dataset that excluded mismatches that may not be directly due to jer realisation, we found that 82.9% of the forms were matched and 17.1% not. We thus see that Havlik’s law held in a large majority of the cases in this more focused dataset. To further demonstrate the value of the methodology and dataset we provided a small pilot study where we tested some of the morphophonological rules proposed by Isačenko (1970) against this smaller dataset. The pilot study is meant as nothing more than a proof-of-concept example, and it is worth noting that the two methodologies cannot be directly compared, as Isačenko provides no quantitative estimates. A more systematic approach to the role of consonant clusters would be a promising avenue for further research. In any case, the pilot immediately shows the potential of the method for the jer change in particular, as well as for systematic empirical study of *any* phonological change. The current dataset can easily be expanded by processing *all* the jer-containing TOROT lemmas without a frequency restriction. It would also make a good point of departure for a broader study of East Slavonic if dictionary data for Ukrainian and Belarusian were to be added.

References

Bethin, Christina. 1998. *Slavic prosody: Language change and phonological theory*. New York.

Eckhoff, Hanne and Aleksandrs Berdicevskis. 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta* 14–15. 9–25.

Eckhoff, Hanne and Aleksandrs Berdicevskis. 2016. From diachronic treebank to dictionary resource: the Varangian Rus' project. In *Proceedings of the EURALEX 2016 conference*, 335–340. Tbilisi: Ivane Javakhishvili Tbilisi State University

Galinskaja, E.A. 2009. *Istoričeskaja fonetika russkogo jazyka*. Second edition. Moscow: Izdatel'stvo Moskovskogo universiteta.

Isačenko, A.V. 1970. East Slavic morphophonemics and the treatment of the jers in Russian: a revision of Havlík's Law. *International Journal of Slavic Linguistics and Poetics* 13. 73–124.

Marr, Clayton and David R. Mortensen. 2020. Computerized Forward Reconstruction for Analysis in Diachronic Phonology, and Latin to French Reflex Prediction. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, 28–36, Marseille, France. European Language Resources Association (ELRA).

Popov, M.B. 2020. Reducirovannye. In V.B. Krys'ko (red.): *Istoričeskaja grammatika russkogo jazyka: enciklopedičeskij slovar'*, 369–383. Moscow.

Ylonen, Tatu. 2022. Wiktextract: Wiktionary as Machine-Readable Structured Data *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC)*, 20–25 June 2022, 1317–1325, Marseille.

Žuravlev, V.K. 1977. Pravilo Gavlika i mexanizm padenija slavjanskix reducirovannyx. *Voprosy jazykoznanija* 6. 30–43.