

# Reviving the Traditional Russian Orthography for the 21st Century

Sergei Winitzki

Text By The Bay 2015

April 24, 2015

# Old Russian Orthography: what and why

- The Russian orthography was reformed in 1918
  - ▶ 4 letters Ъъ, Ѡѡ, Ѣѣ, Ѥѥ were physically removed from printing presses
    - ★ Often, Ъъ was also removed by mistake
- Classic Russian literature used the old orthography:
  - ▶ Pushkin, Turgenev, Dostoyevsky, Tolstoy, Chekhov, Nabokov, ...
- Replace Ѣ→е, Ѡ→ѡ, ѣ→ѣ, ѥ→ѥ
  - ▶ motivation: “simplify the spelling”

## Old Russian orthography: example

Nabokov, *Storm* (1930)

### Г Р О З А

На углу, подъ шатромъ цвѣтущей липы, обдало меня буйнымъ благоуханіемъ. Туманныя громады поднимались по ночному небу, и когда поглощенъ былъ послѣдній звѣздный просвѣтъ, слѣпой вѣтеръ, закрывъ лицо рукавами, низко пронесся вдоль опустѣвшей улицы. Въ тусклой темнотѣ, надъ желѣзнымъ ставнемъ парикмахерской, маятникомъ заходилъ висячій щитъ, золотое блюдо.

Вернувшись домой, я засталъ вѣтеръ уже въ комнатѣ: — онъ хлопнулъ оконной рамой и поспѣшно от-

# Natural language orthography changes naturally...

Hobbes, *De cive* (1651)

I. **A**LL Authors agree not concerning the definition of *the Naturall Law*, who notwithstanding doe very often make use of this terme in their Writings. The Method therefore, wherein we begin from definitions, and exclusion of all equivocation, is only proper to them who leave no place for contrary Disputes; for the rest, if any man say, that somewhat is done against the Law of Nature, one proves it hence, because it was done against the generall Agreement of all the most wise, and learned Nations: But this declares not who shall be the judg of the wisdome and learning of all Nations:

Some letters are now added: *somewhat*, *judg*

Some letters are now removed: *doe*, *generall*, *wisdome*

## ...or when guided by a committee of experts

- A proposal for English, similar to the Russian spelling reform
  - ▶ Omit the mute “e” in endings, replace vowels to “simplify”
    - ★ some → sum, made → maid, gate → gaitdd, breathe → breath
  - ▶ Replace the letters c→k/s, j→g, x→ks, y→i
    - ★ place →plais, played → plaid, hence → hens, john → gon
  - ▶ Physically remove the letters C, J, X, Y from all computer keyboards
    - ★ because they are “a heritage of the horrible old times”

# What happens if you ban a few letters...

- One generation later, nobody remembers anything was ever different
- You lose effective access to your own culture
  - ▶ Cannot print adequate critical editions of classic literature
  - ▶ Nobody wants to read even 100-year old books
  - ▶ Reprinting each old book is hard work for editors
- In Russia:
  - ▶ People born after  $\approx 1950$  are loath to read texts in the old orthography
  - ▶ Such texts are perceived to be irrevocably obsolete
    - ★ (regardless of what the texts say)

# Main challenges in processing old Russian orthography

- As of 2000, no 8-bit encodings contained the letters Ъъ, Ѡѡ, Ѣѣ, Ѥѥ
  - ▶ Unicode support was not widespread
- No keyboard layouts, no screen fonts, few vector fonts
  - ▶ L<sup>A</sup>T<sub>E</sub>X font support: nonstandard font encoding by AMS
- No spelling checker support, no OCR, no hyphenation

# Back to 2015: State of the art

- The letters Ъ ѡ, Ѡ ѡ, І і, V v are present in most Cyrillic fonts
  - ▶ typesetting, printing supported through Unicode
  - ▶ L<sup>A</sup>T<sub>E</sub>X font support: OT2 font encoding only!
- Some spelling checker support (ispell, AbiWord, Open Office)
- Some language support for OCR
- No hyphenation support, no standard keyboard layouts



# The oldrus-ispell project (1999-2003)

- <http://oldrus-ispell.sourceforge.net/koi8-extended.html>
- KOI8-C encoding ([IETF draft](#)), localization for the [links](#) browser
- A keyboard layout for X-Window and MS Windows 95
- Bitmapped X-Window fonts (BDF), the `xcyr` package
  - ▶ Became part of [ucs-fonts](#) and [xfonts-cyrillic](#)
- `ispell` dictionary for old Russian orthography
- A PostScript typesetter using BDF as Type 3 fonts (Perl)
- A primitive converter, new  $\Leftrightarrow$  old orthography (Perl)

# The KOI8-C encoding and screen fonts

## Viewing Nabokov, *Despair* (1934) in Netscape using KOI8-C

росту не знаю, съ чего начать. Смѣшонъ пожилой человекъ, который бѣгомъ, съ прыгающими  
томъ, догналъ {5} послѣдній автобусъ, но боится вскочить на ходу, и виновато улыбаясь, еще  
ужто не смѣю вскочить? Онъ воетъ, онъ ускоряетъ ходъ, онъ сейчасъ уйдетъ за уголъ, непо  
его рассказа. Образъ довольно громоздкій. Я все еще бѣгу.

В мой быль ревельскій нѣмецъ, по образованію агрономъ, покойная мать -- чисто-русская. С  
въ жаркіе лѣтніе дни она, бывало, въ сиреневыхъ шелкахъ, томная, съ вѣромъ въ рукѣ, пол  
ь, кушала шоколадъ, и наливались сѣнокошнымъ вѣтромъ лиловые паруса спущенныхъ што  
аго подданного, интернировали, -- я только-что поступилъ въ петербургскій университетъ, и  
тырнадцатаго до середины девятнадцатая года я прочелъ тысяча восемнадцать книгъ, -- ве  
нію я на три мѣсяца застрялъ въ Москвѣ и тамъ женился. Съ двадцатаго года проживалъ въ  
такого года, уже переваливъ лично за тридцать пять -- --

упление: насчетъ матери я совралъ. По настоящему она была дочь мелкаго мѣщанина, -- про  
кацавейкѣ.

koi8c-table.gif (671x312) 0.8s 100% [1/1] b0/c0/g0																											
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F												
Љ	Г	„	Г	„	„	„	„	„	„	„	„	„	„	„	„												
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F												
ћ	‘	’	”	”	•	—	—	£	·	љ	>	њ	ќ	ћ	џ												
AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP												
ѡ	Ѣ	Ѥ	ѥ	Ѧ	ѧ	Ѩ	ѩ	Ѫ	ѫ	Ѭ	ѭ	Ѯ	ѯ	Ѱ	ѱ												
BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP												
°	Ѱ	ѱ	Ѳ	ѳ	Ѵ	ѵ	Ѷ	ѷ	Ѹ	ѹ	Ѻ	ѻ	Ѽ	ѽ	Ѿ												
CA	CB	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL	CM	CN	CO	CP												
ю	а	б	ц	д	е	ф	г	х	и	й	к	л	м	н	о												
DO	DA	DB	DC	DD	DE	DF	DG	DH	DI	DJ	DK	DL	DM	DN	DO												
п	я	р	с	т	у	ж	в	ь	ы	з	ш	э	щ	ч	ъ												
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF												
Ю	А	Б	Ц	Д	Е	Ф	Г	Х	И	Й	К	Л	М	Н	О												
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF												
П	Я	Р	С	Т	У	Ж	В	Ь	Ы	З	Ш	Э	Щ	Ч	Ъ												

# Working with the oldrus-ispell dictionary

## Proofreading Melgunov's "Red Terror in Russia" (1924)

The screenshot shows a text editor window with a menu bar (File, Edit, Search, Preferences, Shell, Macro, Windows) and a status bar (Shell Command in Progress -- Press Ctrl+. to Cancel). The main text area contains a passage from Melgunov's "Red Terror in Russia" (1924) in Old Russian orthography. The text is: "Еженедѣльникъ Ч. К." \*(6) никогда ничего больше {39} не было опубликовано. А между тѣм мы знаем, что людей в эти дни в Москвѣ по общимъ свѣдѣніямъ было разстрѣлено больше 300 \*(7). Тѣ, которые сидѣли в эти поистинѣ мучительные дни в Бутырской тюрьмѣ, когда были арестованы тысячи людей из самых разнообразных общественных слоев, никогда не забудут своих душевных переживаній. было время террора".

A vertical sidebar on the left shows a list of words from the dictionary: без, присутствие, автомобили, каждом авт, "с вещами", связывать, сидѣл в эт, кошмары. В, "В пам".

A small window titled "ispell" is open, showing the word "без" and its suggestions: 0: бей, 1: безе, 2: безо, 3: безъ. The word "без" is highlighted in the main text.

The text continues: Невольно вновь вспоминаешь слова В. Г. Короленко, мимолетно б, , которая может ее молчаливо терпѣть без протеста. И пожалѣешь, как Г

# Main linguistic challenges

To convert texts to the old Russian orthography, we...

- disambiguate word pairs that became homographs...
  - ▶ самого - самага, синее - синѣе, осель - осьль
- ...and also some homophones that became homonyms
  - ▶ есть - ѣсть, нежить - нѣжить, миръ - міръ, некогда - нѣкогда
- restore old grammatical endings
  - ▶ они, туманные звезды → онѣ, туманныя звѣзды
- restore old letters in word stems, prefixes, and suffixes
  - ▶ Ѳедоръ (Theodor), орѳографія (orthography), мѣро (myrrhe)
  - ▶ рассвирепел → разсвирѣпѣль, нежнейшего → нѣжнѣйшаго

# Main linguistic challenges

Disambiguation requires:

- identifying part of speech
- identifying inflection (number, gender, case)
- for some words, even identifying semantic content
  - ▶ pronunciation details (stress, e/ë) do not always suffice

# Main challenges