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 Ѣѣ, Ѳѳ, Ii, Vv 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, ѳ→ф, i→и, v→и
  - motivation: “simplify the spelling”

Old Russian orthography: example

Nabokov, *Storm* (1930)

ГРОЗА

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. ALL 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 somwhat 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: *somwhat*, *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 ≈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 Ѣѣ, Ѳѳ, Ii, Vv
  - Unicode support was not widespread
- No keyboard layouts, no screen fonts, few vector fonts
  - LaTeX font support: nonstandard font encoding by AMS
- No spelling checker support, no OCR, no hyphenation

Back to 2015: State of the art

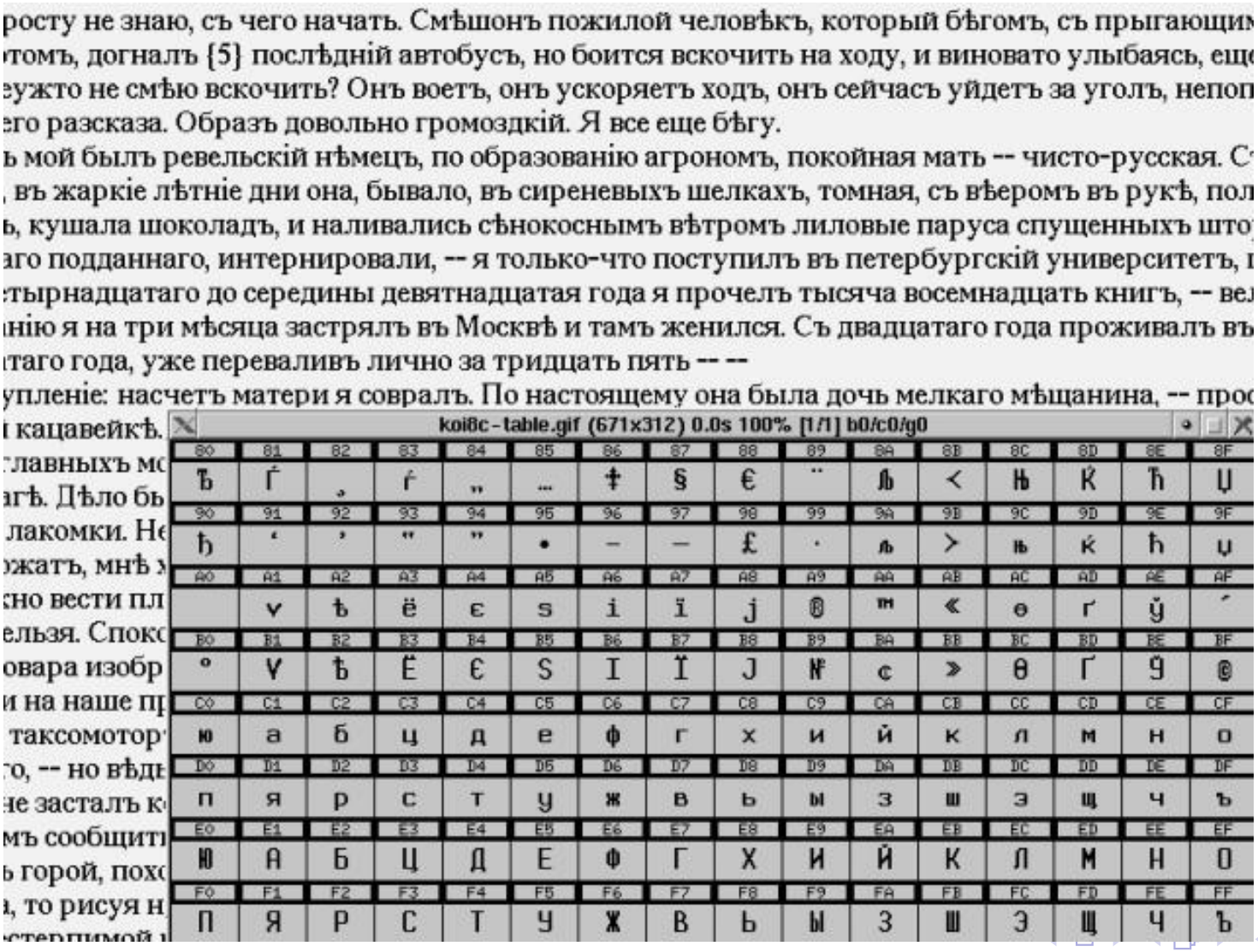
- The letters Ѣѣ, Ѳѳ, Ii, Vv are present in most Cyrillic fonts
  - typesetting, printing supported through Unicode
  - LaTeX font support: OT2 font encoding only!
- Some spelling checker support (ispell, AbiWord, Open Office)
- Some language support for OCR
- No hyphenation support, no standard keyboard 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 $\rightleftarrows$ old orthography (Perl)

The KOI8-C encoding and screen fonts

Viewing Nabokov, *Despair* (1934) in Netscape using KO8-C



Working with the oldrus-ispell dictionary

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

File Edit Search Preferences Shell Macro Windows

Shell Command in Progress -- Press Ctrl+. to Cancel

"Еженедѣльникъ Ч. К." *(6) никогда ничего больше {39} не было опубликовано. А между тѣм мы знаем, что людей в эти дни в Москвѣ по общим свѣдѣніям 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".
присутство
автомобили
каждом авт
"с вещами"
связывать
сидѣл в эт
кошмары. В
"В пам

ispell

без File: spellTmp

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

0: бей
1: безе
2: безо
3: безъ

Main linguistic challenges

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

- disambiguate word pairs that became homographs...
  - самого - самаго, синее - синѣе, оселъ - осѣлъ
- ...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), мvро (myrrhe)
  - рассвирепел → разсвирѣпѣлъ, нежнейшего → нѣжнѣйшаго

Main linguistic challenges

Disambiguation requires:

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

Main challenges