ISMLA Multilingual Session 7: Building a Useful Russian Transliterator

Johannes Dellert

Tübingen University

December 20, 2017

Plan

1. Cyrillic: An Overview
2. Stress in Russian: Basic Facts
3. Exercise 06: A Useful Russian Transliterator

Cyrillic: Consonants

- trivial if you know the Latin alphabet:
  к k **т** t **м** m **б** b
- easy if you know the Greek (uppercase) alphabet:
  г g **д** d **л** l **п** p
  р r **с** s **ф** f **х** x/kh
- not very difficult to remember for most-learners:
  в v **з** z **й** j/y **н** n
- difficult for most learners: Russian-specific sibilants
  ж ž/zh **ш** š/sh **ч** č/ch
  ц c/ts **щ** šč/shch
- we will use a variant of **scholarly transcription** (the first variant)

Cyrillic: Vowels

- the five basic vowels exist in a palatalizing variant (second row):

  а a **э** e **ы** y **о** o **у** u

  я ja **е** je **и** i **ё** jo **ю** ju

- the palatalizing vowels except и are pronounced with a [j] at the beginning of words and after vowels
- otherwise, they have the effect of **palatalizing the preceding consonant** (we all know the word **нет**)
- the ы/и pair is special in not only differing in palatalization
- э is very infrequent (mostly loanwords)
- ё is a stressed allophone of е,
  these two are not necessarily distinguished in writing!

Cyrillic: More on Palatalization

Usage of the letter **ь**:

- consonants can be palatalized even if not followed by a vowel
- in this case, palatalization is written **ь** (the "soft sign")
- minimal example: брат "brother" vs. брать "to take"
- transcription of **ь**: ′

Usage of the letter **ъ**:

- sometimes, we want a non-palatalizing [j] (mostly in loanwords and with certain verbal prefixes)
- for this, palatalization is canceled by **ъ** (the "hard sign")
- example of loanword: объект
- minimal example: сесть "to sit down" vs. съесть "to eat (up)"
- transcription of **ъ**: ″

Stress in Russian: Introduction

- Russian has **phonemic variable stress**
- stress may shift within paradigms in very complex ways
- stress is lexical and must in principle be learned with every word:
  - свиса́ть “to dangle (ipf)”:
    я свиса́ю “I am dangling”, ты свиса́ешь “you are dangling”
  - списа́ть “to copy (prf)”:
    я спишу́ “I will copy”, ты спи́шешь “you will copy”
- crucially: stress is only written in educational materials for L2 learners!
- if you learn Russian mostly by reading (like me), your intuition for stress will be underdeveloped; an enrichment tool will help

Stress in Russian: Effect on Vowels

Problem: knowing the correct stress is **absolutely crucial** to pronouncing Russian correctly or even comprehensibly!

- example 1: the name Колмогоров (Kolmogorov)
  - Ко́лмогоров would be pronounced *Kólməgərəv*
  - Колмо́горов would be pronounced *Kalmógərəv*
  - Колмого́ров would be pronounced *Kəlmagórəv*
  - Колмогоро́в would be pronounced *Kəlməgaróv*
  - the third one is correct (and there is no way you can infer this)
- example 2: the personal pronoun ее "her"
  - good to know it is actually её
  - pronunciation in our transcription format: *jijó*
- example 3: the ambiguous form города
  - го́рода *górədə* means "of the city"
  - городá *gəradá* means "(the) cities"

Vowel Quality in Russian: Simplified Rules

If the stress is known (marked either by ë or the acute), vowel quality becomes predictable by the following (simplified) set of rules:

- in stressed position, all vowels maintain their quality
- in unstressed position, some vowels change:
  - *ja* becomes *jə* at the end of a word, and *ji* otherwise
  - *je* becomes *ji* in any syllable before the stress
  - *o* becomes *a* immediately before the stress, and *ə* otherwise
  - *a* becomes *ə* at the end of a word

Exercise 06: Backend Requirements

Requirements for the backend servlet you are going to implement:

- maintain a map of non-accented Russian forms into accented ones
- ability to tokenize Russian text (taking care of punctuation this time)
- looks up each token to assign stress (normalization to lowercase might be necessary), and returns the reassembled text
- provides the option to transliterate according to the tables on the first slides (build the table yourself to learn some Cyrillic!)
- provides the option to adapt vowel qualities in the transliteration according to the simple rules on the previous slide

Exercise 06: Frontend Requirements

- basic interface: text area to paste Russian text, “Transcribe” button, output text displayed in addition to input (not replacing it)
- option panel consisting of three checkboxes:
  - Romanization: scholarly Transliteration if checked, Cyrillic if unchecked
  - Stress Marks: stress marks are removed after the operation if unchecked
  - Vowel Adaptation: activates rules (Romanization only)
- optional: react to option changes by regenerating and updating the transliteration without having to click on “Transcribe” again

Exercise 06: The Data

- in `russian-forms.txt`, you find 1.5M Russian wordforms with accent marks (covering entire paradigms, not many names)
- these were collected from various sources (especially helpful: Wiktionary, full paradigms by Андрей Усачёв (2004))
- to build the necessary map from these, you need to remove the stress marks (U+0301 COMBINING ACUTE ACCENT), and map these forms to the originals (ignore the few ambiguous forms for now)
- special treatment is needed for the ё, which does tend to appear in texts (but you cannot expect it to!)

Exercise 06: Implementation Hints

- for implementing the vowel quality rules, it might help to introduce temporary symbols; you will need several back-and-forth passes
- optional removal of stress marks can be done in a trivial last step
- for testing: the Wiktionary provides IPA for most forms you will find
- reference implementation available here:
  [to be announced in class]

Exercise 06: Alternative for Native Speakers

For native speakers, the task might be boring or too easy.
Alternative suggestion: how about a transliterator for Sakha?

Олоҥхо — саха уус-уран айымньыта. Олус былыргы кэмнэргэ үөскээбит. Олоҥхоһуттар араас айымньылары холбоон уһун олоҥхолору айаллара. Үгүс олоҥхо 10—15 тыһ. хоһоон устуруоκаларыттан турар эбит. Бөдөҥ олоҥхолор 20 тыһ. тахса буолаллар үһү. Хас биирдии кэпсэнэр герой туспа куолаһынан ылланар. Сорох олоҥхолор хас да күн ылланаллар үһү. Ордук биллиилээх олоҥхо буолар Дьурулуйар Ньургун Боотур.

Exercise 06: Questions

Questions?