Handling non-Roman characters / alphabets / writing systems in repositories

Milica Ševkušić
EIFL (Lithiania) &
Institute of Technical Sciences of SASA (Serbia)

A little bit of history: examples

- Capturing vernacular languages in the Latin alphabet in early medieval Europe
- Missionaries and the translations of the Bible to the indigenous languages

Legacy character encoding systems

- missing characters
- conversion from one encoding system to another, not without errors

- Examples:
  - ISO-8859-1, Windows-1252 - Latin, missing characters with diacritics
  - Windows-1250 - Central European, Latin
  - Windows-1251 - Cyrillic, Slavic languages, also adjusted to other languages
  - Legacy encoding systems for Chinese (e.g. Guobiao, Big5)
  - ...

Transliteration: rendering characters as they are written in a source language in a target language by corresponding characters that sound similar.

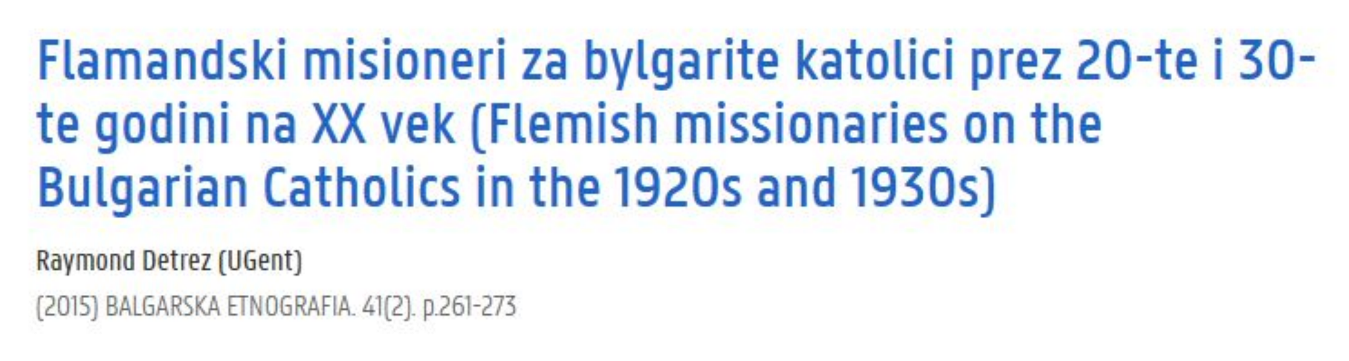
Transcription: conversion of a word or phrase based on the pronunciation.

Romanization: transliteration of a foreign language into Latin letters.

Greek word	Ευαγγέλιο
Transliteration	Euaggelio
Transcription	Evangelio

Example

Bulgarian transliterated, in a repository:
http://hdl.handle.net/1854/LU-7011210

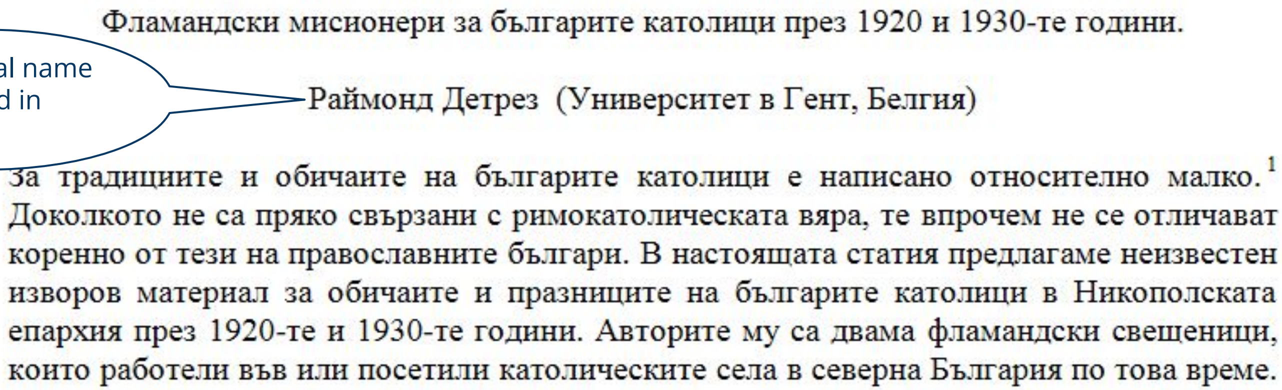


Flamandski misioneri za bylgarite katolici prez 20-te i 30-te godini na XX vek (Flemish missionaries on the Bulgarian Catholics in the 1920s and 1930s)

Raymond Detrez (UGent)

(2015) BALGARSKA ETNOGRAFIA. 41(2). p.261-273

Metadata in the repository



Фламандски мисионери за българите католици през 1920 и 1930-те години.

Раймонд Детрез (Университет в Гент, Белгия)

За традициите и обичаите на българите католици е написано относително малко.[1] Доколкото не са пряко свързани с римокатолическата вяра, те впрочем не се отличават коренно от тези на православните българи. В настоящата статия предлагаме неизвестен изворов материал за обичаите и празниците на българите католици в Никополската епархия през 1920-те и 1930-те години. Авторите му са двама фламандски свещеници, които работели във или посетили католическите села в северна България по това време.

The original name transcribed in Bulgarian

The paper, full text

Transliteration standards: examples

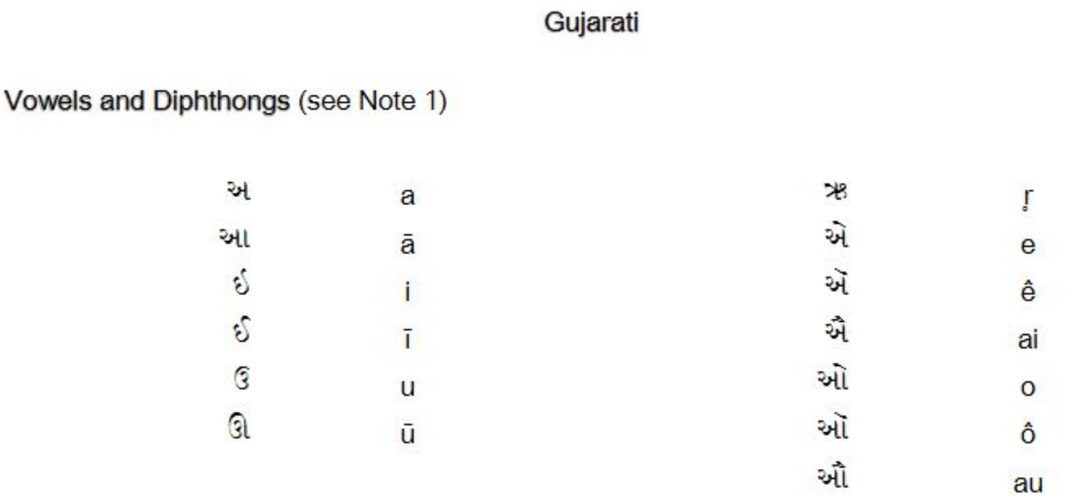
Original:	**Градският билигвизъм и училищното дело**
ISO 9:1995:	**Gradskiât biligviz”m i učiliŝnoto delo**
Scientific ISO 9 (1968):	**Gradskijat biligvizăm i učilištnoto delo**
ALA/LC:	**Gradskii͡at biligvizŭm i učilištnoto delo**
British Standard (1958):	**Gradskiyat biligvizŭm i uchilishtnoto delo**
Official Bulgarian (2006):	**Gradskiyat biligvizam i uchilishtnoto delo** (alsoUN (2012); BGN/PCGN (2013))
Danchev:	**Gradskiyat biligvizum i uchilishtnoto delo**

ii͡at

Characters used in transliteration are not always easy to render and users tend to omit or simplify them, which makes it difficult to reconstruct the original text.

- Gradskiiat = Градскииат
- biligvizum = билигвизум

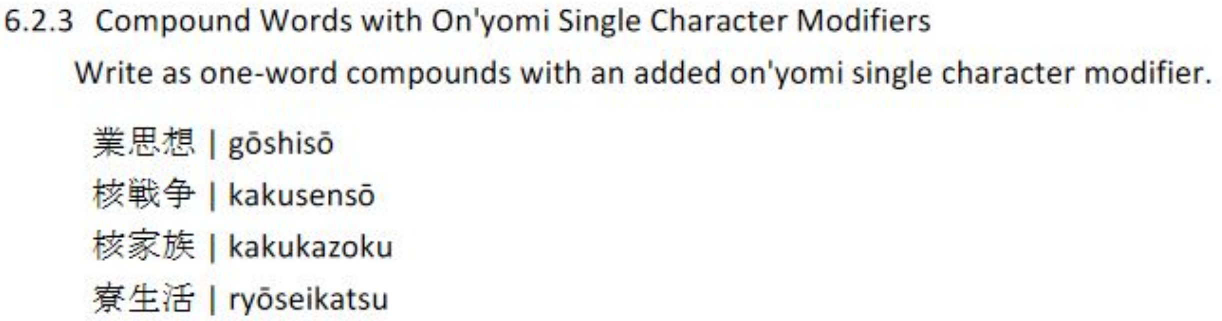
ALA-LC Romanization Tables



Gujarati

Vowels and Diphthongs (see Note 1)

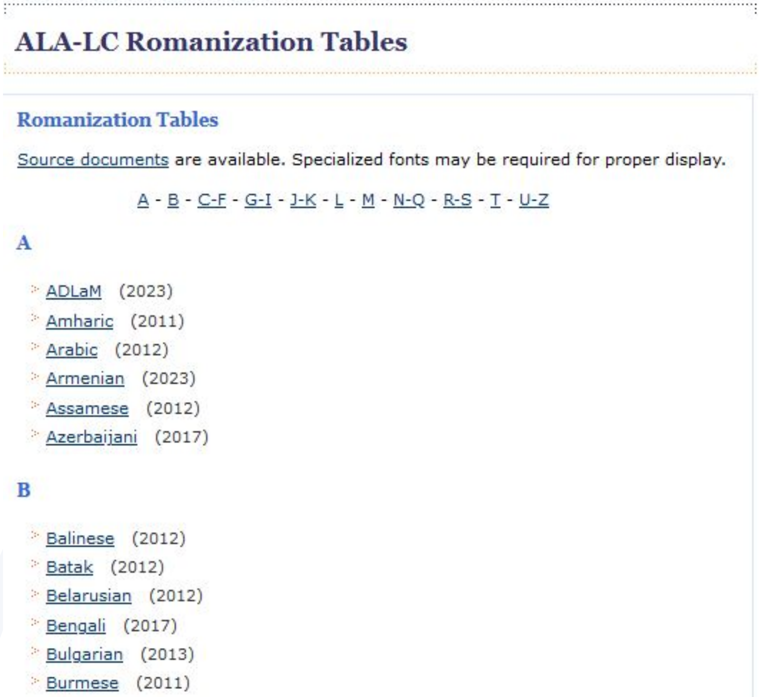
અ	a	ઋ	ṛ
આ	ā	એ	e
ઇ	i	ઍ	ê
ઈ	ī	ઐ	ai
ઉ	u	ઓ	o
ઊ	ū	ઑ	ô
		ઔ	au



6.2.3 Compound Words with On'yomi Single Character Modifiers

Write as one-word compounds with an added on'yomi single character modifier.

業思想	gōshisō
核戦争	kakusensō
核家族	kakukazoku
寮生活	ryōseikatsu



ALA-LC Romanization Tables

Romanization Tables

Source documents are available. Specialized fonts may be required for proper display.

A - B - C-F - G-I - J-K - L - M - N-Q - R-S - T - U-Z

A

- ADLaM (2023)
- Amharic (2011)
- Arabic (2012)
- Armenian (2023)
- Assamese (2012)
- Azerbaijani (2017)

B

- Balinese (2012)
- Batak (2012)
- Belarusian (2012)
- Bengali (2017)
- Bulgarian (2013)
- Burmese (2011)

https://www.loc.gov/catdir/cpso/roman.html

More challenges

- Languages and writing systems have their histories
  e.g. differences between the ways that the Hebrew alphabet is used in Jewish languages such as Ladino and Yiddish, and the way it is used in modern Hebrew:
  https://sites.lsa.umich.edu/translatemidwest/2021/04/27/the-multivalence-of-hebrew-alphabet-across-jewish-languages-in-the-hathitrust-repository/
- Transliteration is not limited to romanization
  e.g. Arabic to Tamil: http://ir.lib.seu.ac.lk/handle/123456789/5117

Problems

- transliterated data imported from other systems
- inconsistent approaches:
  - some records transliterated, some not
  - different standards used in the same repository (impossible to automate conversion)
- transliteration errors make it difficult to reconstruct the original form of the metadata
- errors may occur during conversion (resulting in data corruption)

Recommendations

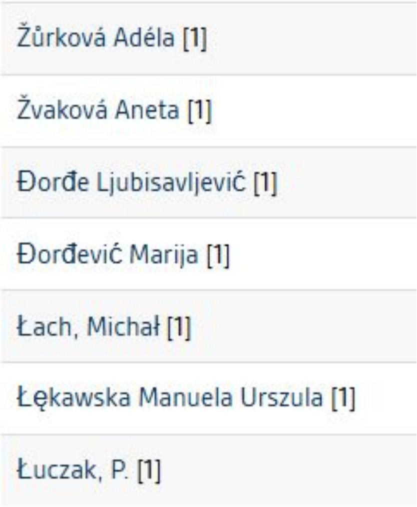
- Enable UTF-8 support in the repository
- use the original alphabet / writing system whenever possible
- avoid transcription whenever possible
- if transliteration is inevitable, choose one standard and declare it in the repository's FAQ / user manual / about pages.
- If this is not possible, declare all used standards in the FAQ / user manual / about pages.

Recommendations

- To ensure that readers can reconstruct the original spelling, provide links to relevant transliteration guidelines (e.g. Library of Congress) and/or tools (e.g. https://alittlehebrew.com/transliterate/) in FAQ / user manual / about pages.
- If author names are transliterated, identifiers such as ORCID should be used to connect different name variants.

More problems

Sorting



Žůrková Adéla [1]

Žvaková Aneta [1]

Đorđe Ljubisavljević [1]

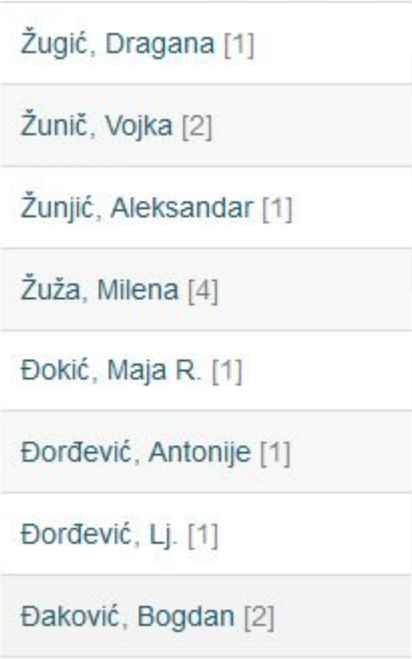
Đorđević Marija [1]

Łach, Michał [1]

Łękawska Manuela Urszula [1]

Łuczak, P. [1]

A
B
C
Č
Ć
D
Dž
Đ
E
…
U
V
Z
Ž



Žugić, Dragana [1]

Žunič, Vojka [2]

Žunjić, Aleksandar [1]

Žuža, Milena [4]

Đokić, Maja R. [1]

Đorđević, Antonije [1]

Đorđević, Lj. [1]

Đaković, Bogdan [2]

Questions?

scmilica@gmail.com

Twitter: @lessormore4

Template Concentric Blue, designed by Jimena Catalina; Slides Carnival. CC BY.