ALCTS
Association for Library Collections & Technical Services
a Division of the American Library Association

Managing Transliteration of Bibliographic Data

Deborah W. Anderson, University of California, Berkeley
Steven R. Loomis, IBM Global Foundations Technology Team
Moderator: Margaret Hughes, Stanford University Libraries

June 27, 2015

Sponsor: Committee on Cataloging Asian and African Materials

Co-Sponsors:

Africana Librarians Council

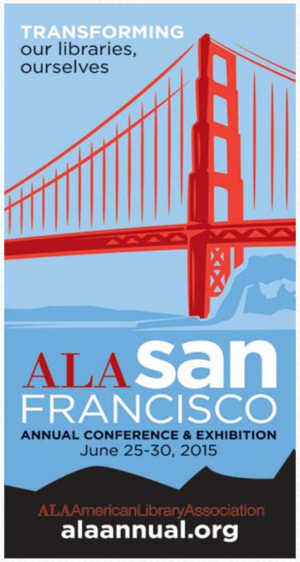
Asian, African, Middle Eastern Section - ACRL

Committee on Research Materials on Southeast Asia

Committee on South Asian Libraries & Documentation

Middle East Librarians Association

#alctsAC15



Character encoding in Unicode, transliteration, and the future of multilingual search

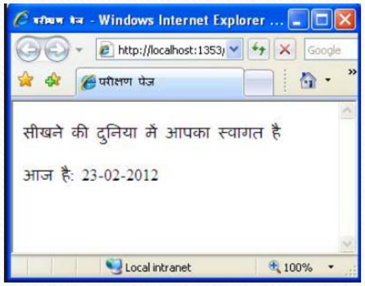
Deborah Anderson

Researcher, Dept. of Linguistics, UC Berkeley

ALA June 2015

The Unicode Consortium

- The Unicode Standard (and related specs)
- Unicode website: http://unicode.org
- Other projects, including CLDR (locale data)
  - Includes some text transliteration data



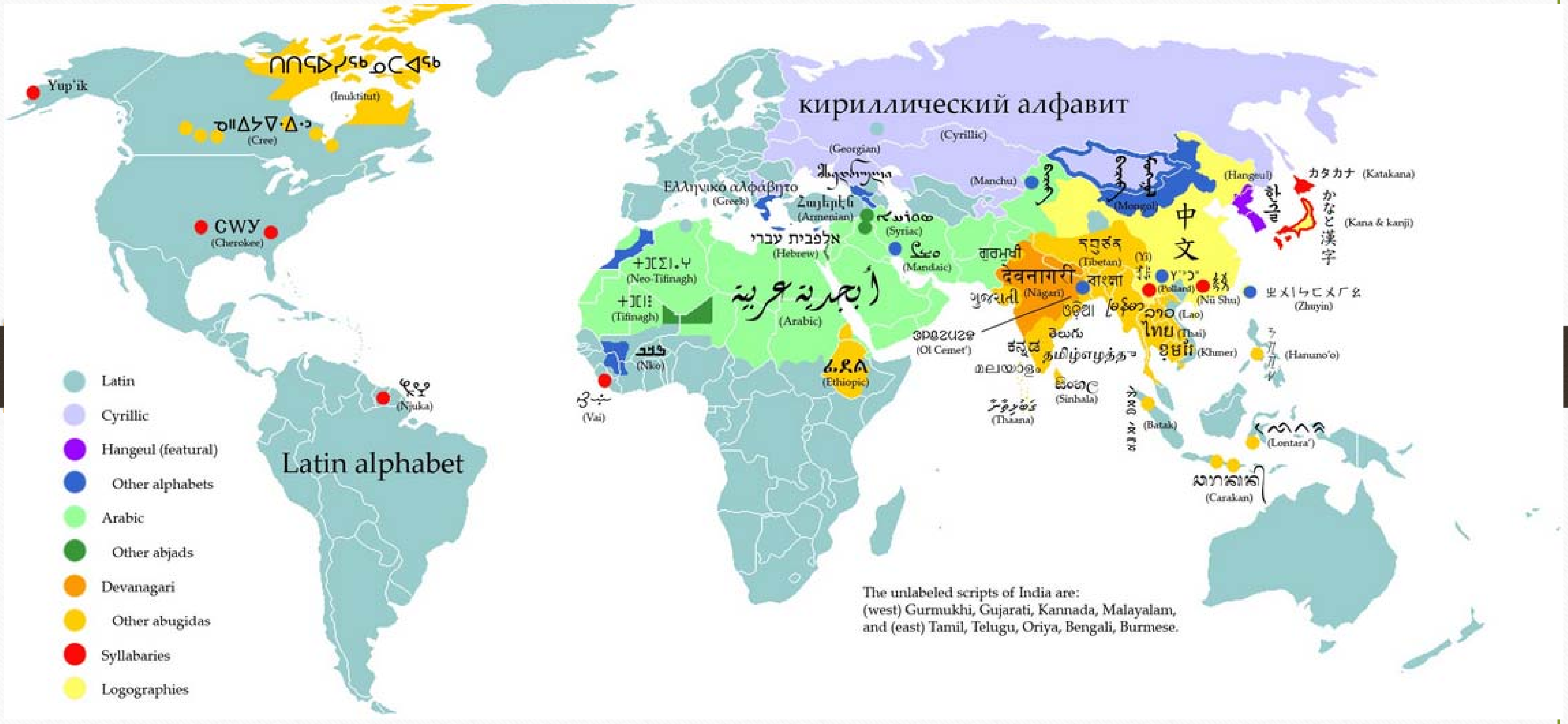
सीखने की दुनिया में आपका स्वागत है

आज है: 23-02-2012

Unicode Standard

- International standard, synchronized with ISO/IEC 10646
- Supported on modern browsers, mobile devices, and computers

- Backbone of multilingual text representation on the Internet, in email, text messages, word-processing docs, etc.
- Basis of Unicode-enabled fonts, keyboards, and OCR



The unlabeled scripts of India are:
(west) Gurmukhi, Gujarati, Kannada, Malayalam,
and (east) Tamil, Telugu, Oriya, Bengali, Burmese.

Unicode basics -1

- Unicode Standard assigns to letters and symbols of the world's writing systems a unique number (**code point**)

Latin letter **b** is "0062"

Devanagari भ is "092D"

- Numbers (code points) stay the same on any modern device, whether an iPhone, on Android device, tablets, computers, etc.

Unicode basics -2

- **New script/characters must be approved by two standards committees**

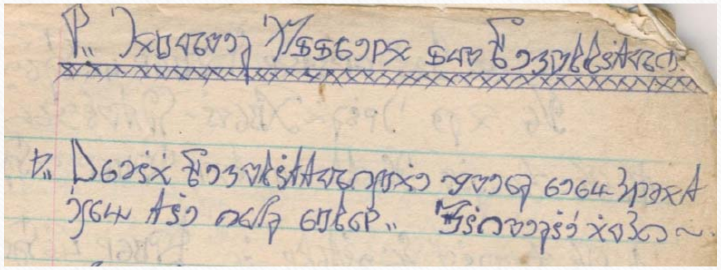
- **Proposals provide information on**
  - characters, glyphs and names
  - sort order (i.e., a, A, b, B, c, C, etc.)
  - directionality of the script
  - other information needed to implement the script on computers

Languages and Scripts

- Number of languages: over 6,000 (*Ethnologue*)
- Number of scripts: ca. 223 (modern and historical)
  - Number in Unicode: 123
  - Not yet in Unicode: over **100** (approximately 35 modern)

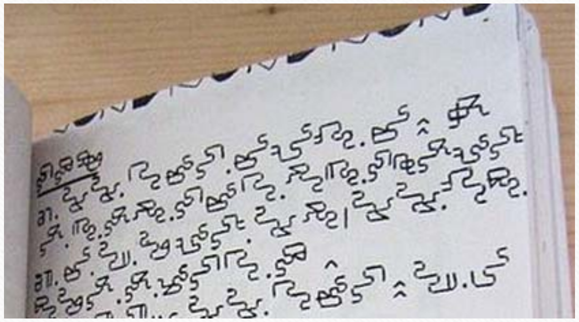
UC Berkeley Script Encoding Initiative

- Works with users to get eligible characters and scripts into Unicode
- Remaining modern unencoded scripts are primarily in Africa, S/SE Asia

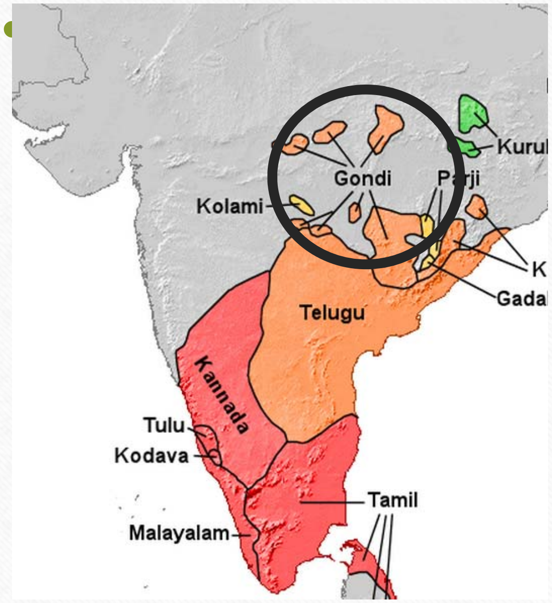


Medefaidrin
(Nigeria)

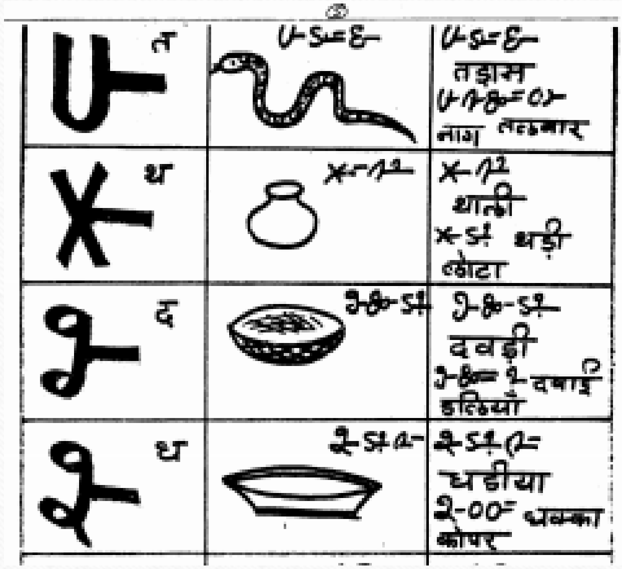
Mandombe
(Congo)



UC Berkeley Script Encoding Initiative



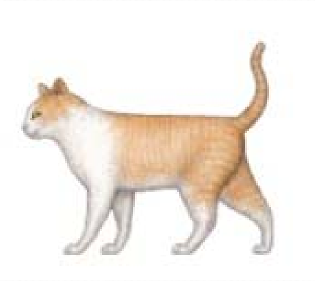
Masaram Gondi


त
तड़ास
नाग
थ
थाली
थड़ी
लोटा
द
दवड़ी
दवाई
हड़ियाँ
ध
धड़ीया
धमका
कोपर

Components

- Language
- Script
- Orthography (for non-Latin script=transliteration scheme)
- Text representation (fonts, keyboard/IME, rendering, software)
  - Example of rendering: क + ◌् + ष → क्ष
- Unicode code points (<0915, 094D, 0937>)

Example 1:
Language: English



- Script: Latin

  Orthography 1: Standard English Spelling

  Text representation: **cat**

  Unicode: <0063, 0061, 0074>

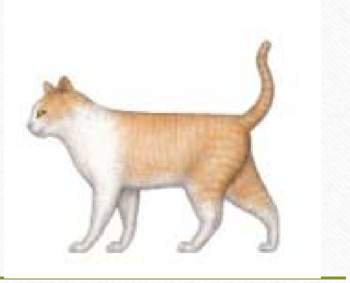
- Script: Latin

  Orthography 2: IPA (phonetic)

  Text representation: **kʰæt̚** (with Unicode-compliant font, etc.)

  Unicode: <006B, 02B0, 00E6, 0074, 031A>

Example 2:
Language: Modern Greek



- Script: Greek

  Orthography 1: Standard Modern Greek Spelling

  Text representation: **γάτα** (with Unicode-compliant font, etc.)

  Unicode: <03B3, 03AC, 03C4, 03B1>

- Script: Latin

  Orthography 2: ALA-LC Greek Romanization table

  Text representation: **gata**

  Unicode: <0061, 0041, 0074, 0061>

Example 3: Language: Japanese

Script: Han

Orthography 1: Standard Japanese (as kanji)

Text representation: 猫 (with Unicode-compliant font, etc.)

Unicode: <732B>

Script: Hiragana

Orthography 2: Standard Japanese (spelled out in hiragana)

Text representation: ねこ (with Unicode-compliant font, etc.)

Unicode: <306D, 3053>

Script: Latin

Orthography 3: Standard Romanization of Japanese

Text representation: **neko**

Unicode: <006E, 0065, 006B, 006F>

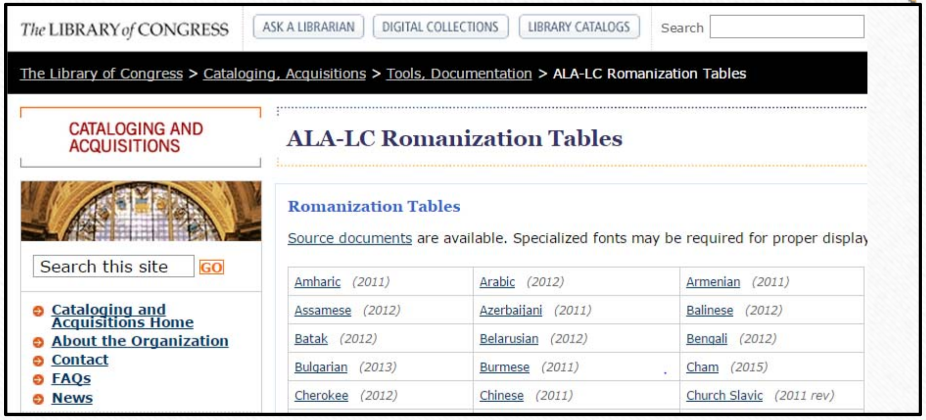
Transliteration Tables for non-Latin scripts (Romanization tables)

• ALA-LC:	ca. 129 tables for languages	40 different scripts
• BGN/PCGN:	45 tables	17 scripts
• UNGEGN:	45 tables	26 scripts
• ISO standards:		21 scripts
[Total number of scripts		220+ scripts]

Background on Romanization tables -1

ALA-LC Romanization tables* page:

- Tamil (2011)
- Romanian (in Cyrillic) (2014)
- Mande languages (in N'ko script) (2015)


The LIBRARY of CONGRESS
ASK A LIBRARIAN
DIGITAL COLLECTIONS
LIBRARY CATALOGS
Search
The Library of Congress > Cataloging, Acquisitions > Tools, Documentation > ALA-LC Romanization Tables
CATALOGING AND ACQUISITIONS
Search this site
GO
Cataloging and Acquisitions Home
About the Organization
Contact
FAQs
News

ALA-LC Romanization Tables

Romanization Tables

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

Amharic (2011)	Arabic (2012)	Armenian (2011)
Assamese (2012)	Azerbaijani (2011)	Balinese (2012)
Batak (2012)	Belarusian (2012)	Bengali (2012)
Bulgarian (2013)	Burmese (2011)	Cham (2015)
Cherokee (2012)	Chinese (2011)	Church Slavic (2011 rev)

*http://www.loc.gov/catdir/cpso/roman.html

Background on Romanization tables -2

LC Guidelines*:

- "should enable machine-transliteration as much as possible and preferably reversible transliteration"
- take equivalent Latin letter used from MARC Basic Latin, avoid rarer letters
- diacritics can be used to accommodate pronunciation; when using diacritics, avoid those not widely supported or whose position may interfere with printing/display of Latin letter (i.e., those diacritics occurring below).

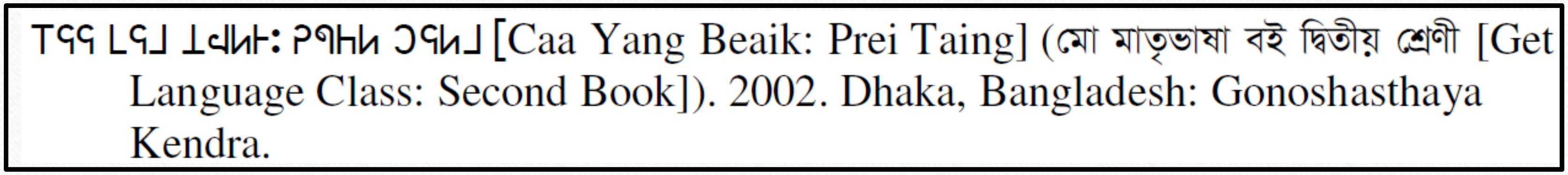
* http://www.loc.gov/catdir/cpso/romguid_2010.html

ALA-LC Romanization Tables: Adding New Tables

- 6 months - 1 year (typically)
- If controversial, can take 2-4 years (or longer)

Transliteration: Advantages

- Consistent set of rules to follow
  - Can find book title if script is not in Unicode or if no Unicode-enabled font is available

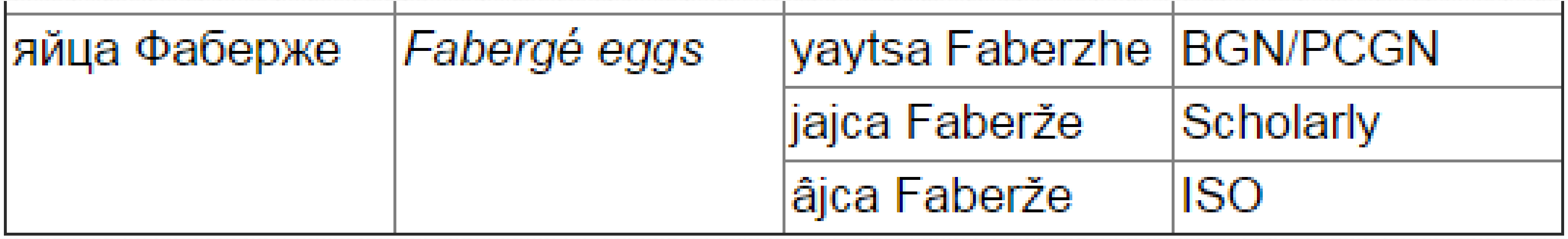


[illegible] [Caa Yang Beaik: Prei Taing] (ম্রো মাতৃভাষা বই দ্বিতীয় শ্রেণী [Get Language Class: Second Book]). 2002. Dhaka, Bangladesh: Gonoshasthaya Kendra.

  - Can find book if there is an error in a record in the original script (in Unicode), example for Arabic

Transliteration: Problems

- Different transliteration schemes (and legacy data) not conformant with ALA-LC Romanization may make it hard to find a title



яйца Фаберже	*Fabergé eggs*	yaytsa Faberzhe	BGN/PCGN
		jajca Faberže	Scholarly
		âjca Faberže	ISO

- Many scripts missing from ALA-LC Romanization tables
- Takes time to propose transliteration table and get approved

ALA-LC Romanized Tables: Exs. of Missing Scripts with Printed Materials

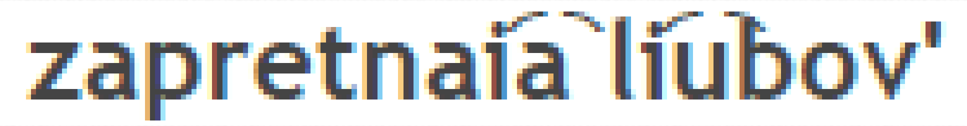
- **Africa (4)**: Bamum, Bassa Vah, Mende Kikakui, Osmanya
- **South Asian (15)**: Chakma, Grantha, Kaithi, Khojki, Khudawadi Mahajani, Meetei Mayek, Modi, Mro, Saurashtra, Siddham, Syloti Nagri, Takri, Tirhuta, Warang Citi
- **SE Asian (7)**: Kayah Li, New Tai Lue, Pahawh Hmong, Pau Cin Hau, Tai Le, Tai Tham, Tai Viet
- **Indonesia and Oceania (3)**: Buginese, Rejang, Sundanese
- **E Asia (3)** : Lisu, Miao, Yi

Components

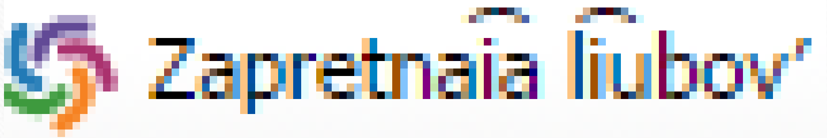
- Language
- Script
- Orthography (for non-Latin script=transliteration scheme)
- Text representation (fonts, keyboard/IME, rendering, software)
- Unicode code points (<XXXX, XXXX>)

Issues with fonts, keyboards, and software

- Font issue



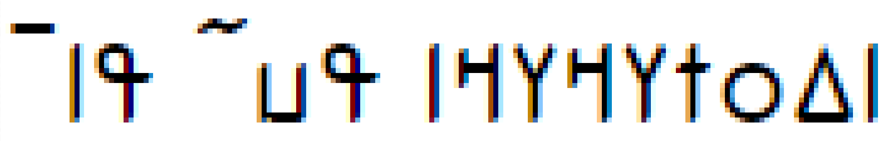
zapretnai︠a︡ li︠u︡bov'



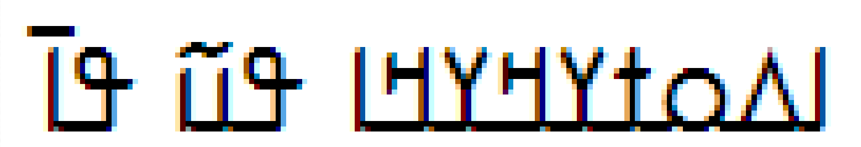
Zapretnai︠a︡ li︠u︡bov'

Issues with fonts, keyboards, and software

N’Ko: Using older rendering engine software/OS:



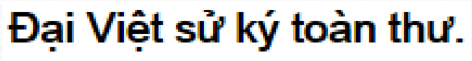
On Windows 8:



Issues with fonts, keyboards, and software (or messy data?)

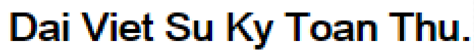
- Vietnamese

Correct:

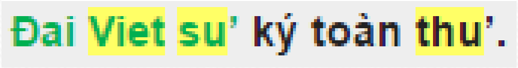


Đại Việt sử ký toàn thư.

on OCLC FirstSearch:



Dai Viet Su Ky Toan Thu.



Đai Viet su’ kỷ toản thu’.

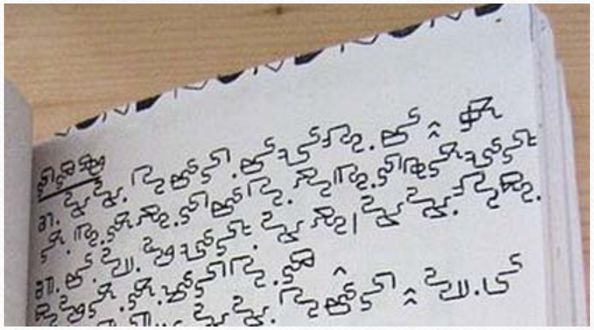
Components

- Language
- Script
- Orthography (for non-Latin script=transliteration scheme)
- Text representation (fonts, keyboard/IME, rendering, software)
- Unicode code points (<XXXX, XXXX>)

Issues with Unicode

- Missing scripts or characters
  - About 100 scripts are known to be missing

Jurchen



Mandombe

Transliteration tools -1

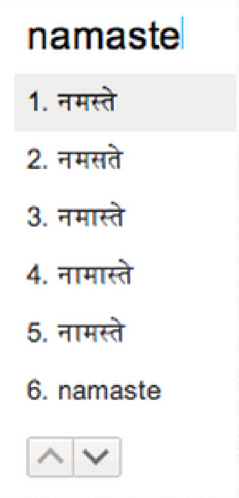
- CLDR has 16 script-script transliteration tables*, possible to have more added
- Process of adding more tables requires submitting rules in a special syntax which needs to catch the edge cases, like casing (UTR #35)

*See http://www.unicode.org/cldr/charts/latest/transforms/index.html

Transliteration tools -2

- Google transliteration input tool* has 25 languages, but is not rule-based
- Type the word in phonetically in Latin, pick from list:

Google Input Tools



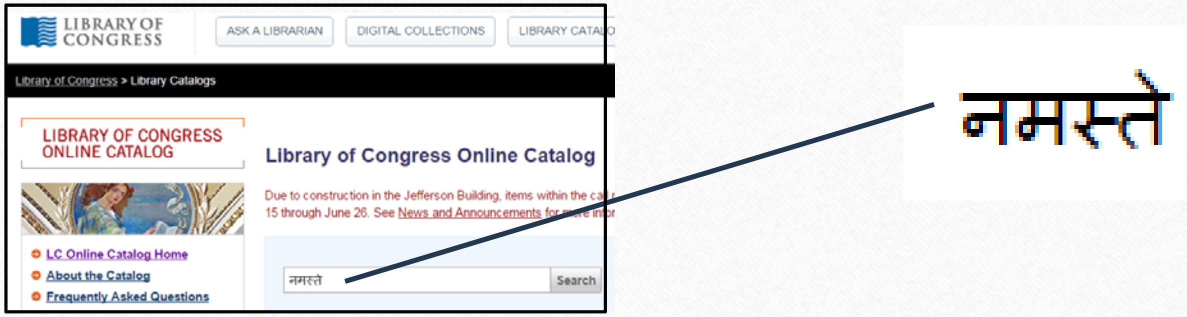
namaste

1. नमस्ते
2. नमसते
3. नमास्ते
4. नामास्ते
5. नामस्ते
6. namaste

*http://www.google.com/inputtools/services/features/transliteration.html

The Future….

- Will fonts/software support the world's scripts?
- Be able to search in more of the original scripts?

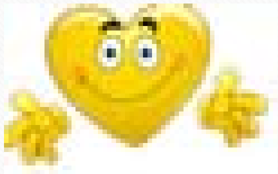


Library of Congress Online Catalog

Due to construction in the Jefferson Building, items within the ca
15 through June 26. See News and Announcements for more info

- Add ALA-LC transliteration schemes to CLDR?

Thank you



U+1F917

- Thanks to Bruce Johnson (LC); UC Berkeley librarians Shayee Khanaka, Virginia Shih, Adnan Malik, Haiqing Lin, Noriko Nishizawa, and Jaeyong Change; Google Input Tools members Xiangye Xiao, Yuanbo Zhang, Yingbing; Unicode Technical Director Ken Whistler
- Support for SEI project comes from a Google Research Award and NEH grant PR-50205-15

Questions?

U+1F914

Debbie Anderson

dwanders@berkeley.edu

Script Encoding Initiative project: http://linguistics.berkeley.edu/sei