Expanding the Unicode Repertoire

Unencoded Scripts of Africa and Asia

Deborah Anderson, SEI, Department of Linguistics, UC Berkeley

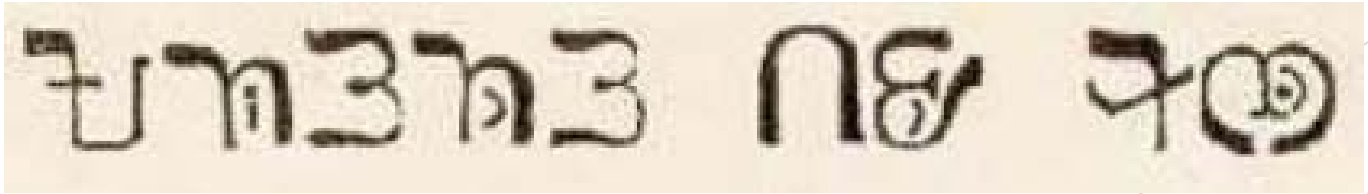
Anshuman Pandey, Department of History, University of Michigan

IUC 38 • November 5, 2014

Already Encoded Scripts (12)

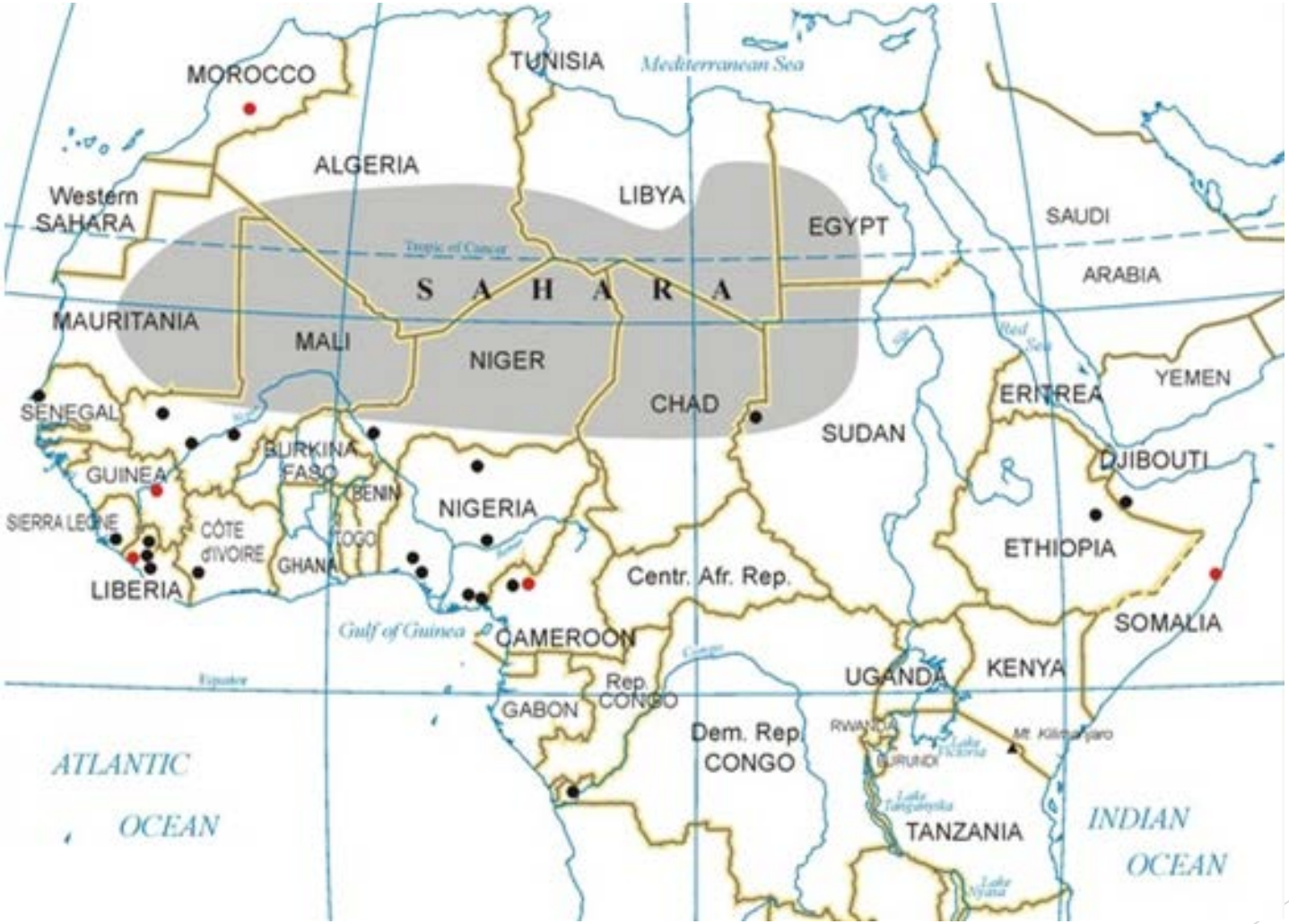
- "Modern" use (8)
  - ***Bamum/Bamum Supplement***
  - ***Bassa Vah***
  - Ethiopic/Ethiopic Supplement and Extensions
  - ***Mende Kikakui***
  - ***N'Ko***
  - Osmanya
  - Tifiangh
  - ***Vai***
- Historic use (3)
  - ***Egyptian Hieroglyphs***
  - ***Meroitic Cursive***
  - ***Meroitic Hieroglyphs***
- Liturgical use (1)
  - ***Coptic***

Note: Scripts in ***bold italic*** had assistance from SEI



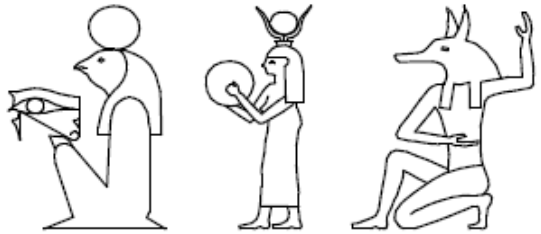
Bassa Vah (Unicode 7.0)

Scripts of Africa



Unencoded scripts (historical) – possible candidates for encoding

- Additions to Egyptian Hieroglyphs (Ptolemaic) – over 7K characters



- Hieratic?

- Demotic?

Source: Chicago Demotic Dictionary

- Numidian?

Unencoded scripts (modern or near-modern) – good candidates (13)

- Adlam * (1978)
- Bagam (1910)
- Beria (1980s)
- Bete (1956)
- Borama (Gadabuursi) (1933)
- Garay (Wolof) (1961)
- Hausa Raina Kama (1990s)
- Kaddare (1952)
- Kpelle (1930s)
- Loma (1930s)
- Mandombe (1978)
- Mwangwego (1979)
- Nwagu Aneke Igbo (1960s)
- Oberi Okaime (1927)

* Approved by UTC

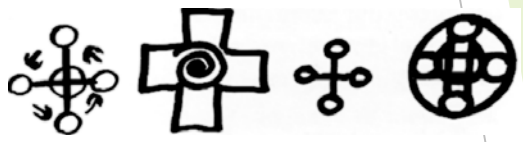
Unencoded scripts – not currently good candidates for encoding (21)

- Aka Umuagbara Igbo (1993)
- Aladura Holy alphabet (1927)
- Bassa (1836)
- Esan oracle rainbow (1996)
- Fula (2 scripts) (1958/1963)
- Hausa (2 scripts) (1970/1998)
- Kii (2006)
- Kru alphabet (1972)
- Luo (2 scripts)
- Masaba (1930)
- Ndebe Igbo (2009)
- New Nubian (2005)
- Nubian Kenzi (1993)
- Oromo (1956)
- Soni (2001)
- Wolof Saalliw wi (2002)
- Yoruba FaYe (2007)
- Yoruba holy script (undeciphered) (20c)

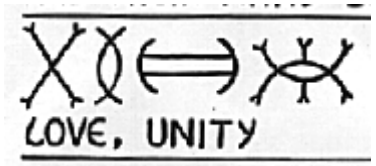
Unencoded scripts – non-phonetic graphic symbols (10)

- Adinkra
- Akan
- Bogolanfini
- Cenda
- Dogon cosmograms
- Gicandi
- Hu-ronko
- Kongo cosmograms
- Nsibidi
- Poro symbols

Adinkra

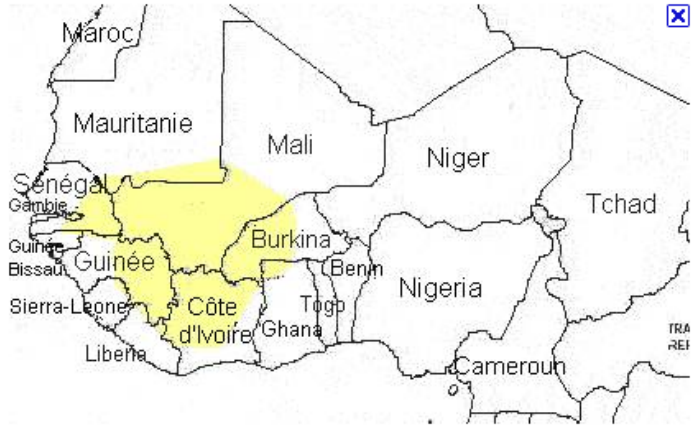


Kongo cosmograms



Nsibidi

Poster child for modern script: N’Ko



- Created in 1949 by Solomane Kante
- Used for Mande languages (18-20m speakers)
- Used in religious materials, newspapers, books, Internet

Poster child for modern script: N’Ko

Key traits:

- Many active users (used in 10 countries)
- Significant written text materials
- Taught in schools

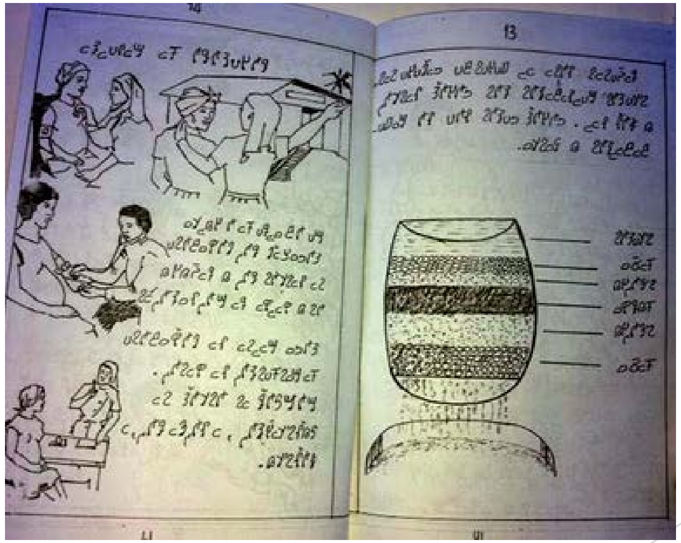
  (e.g., Guinea and Mali)
- Funding support
- Tireless proponent: M. Doumbouya
- Has iPhone app, but still some issues in browsers and other software

PUBLISHED

2006

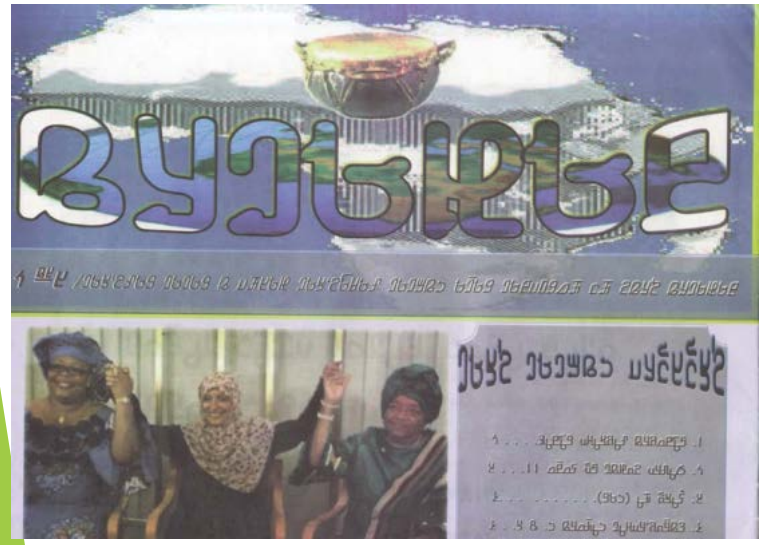
Case study: Adlam

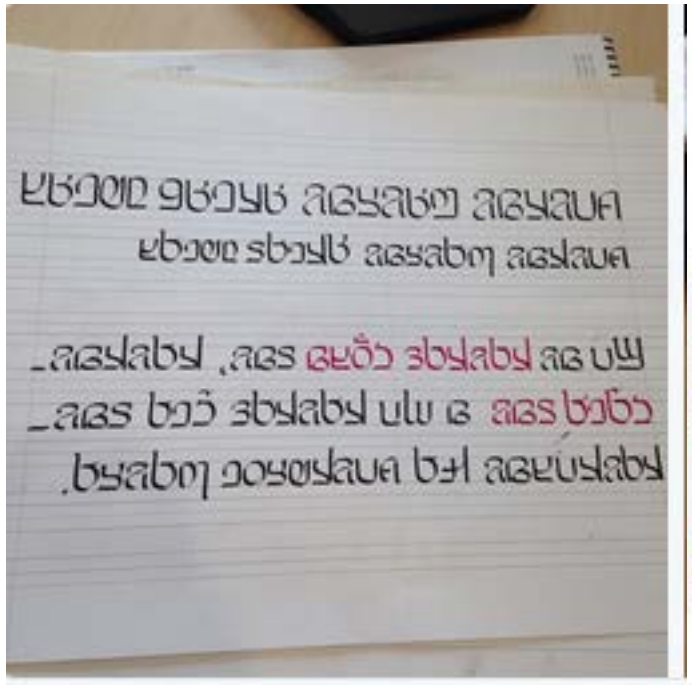
- Created in 1980s by A. and I. Barry
- Alphabetic script used for Fulani language (Pular / Fulfulde) spoken by 40m people across Africa



Case study: Adlam

- Used in 9 countries across West Africa
- Learning materials and monthly periodical are published in the script





Case study: Adlam

- Unicode Technical Committee, Sunnyvale, CA

  October 27 2014

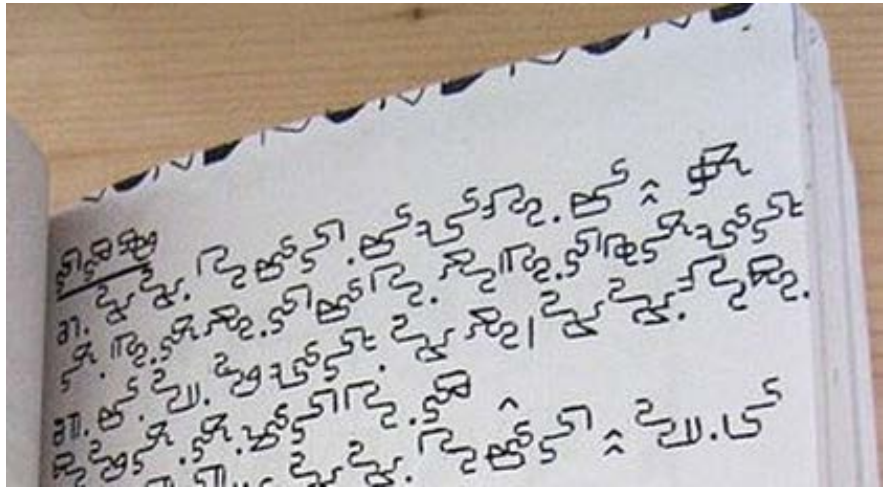
Case study: Adlam

- Unicode Technical Committee, Sunnyvale, CA

  October 27 2014

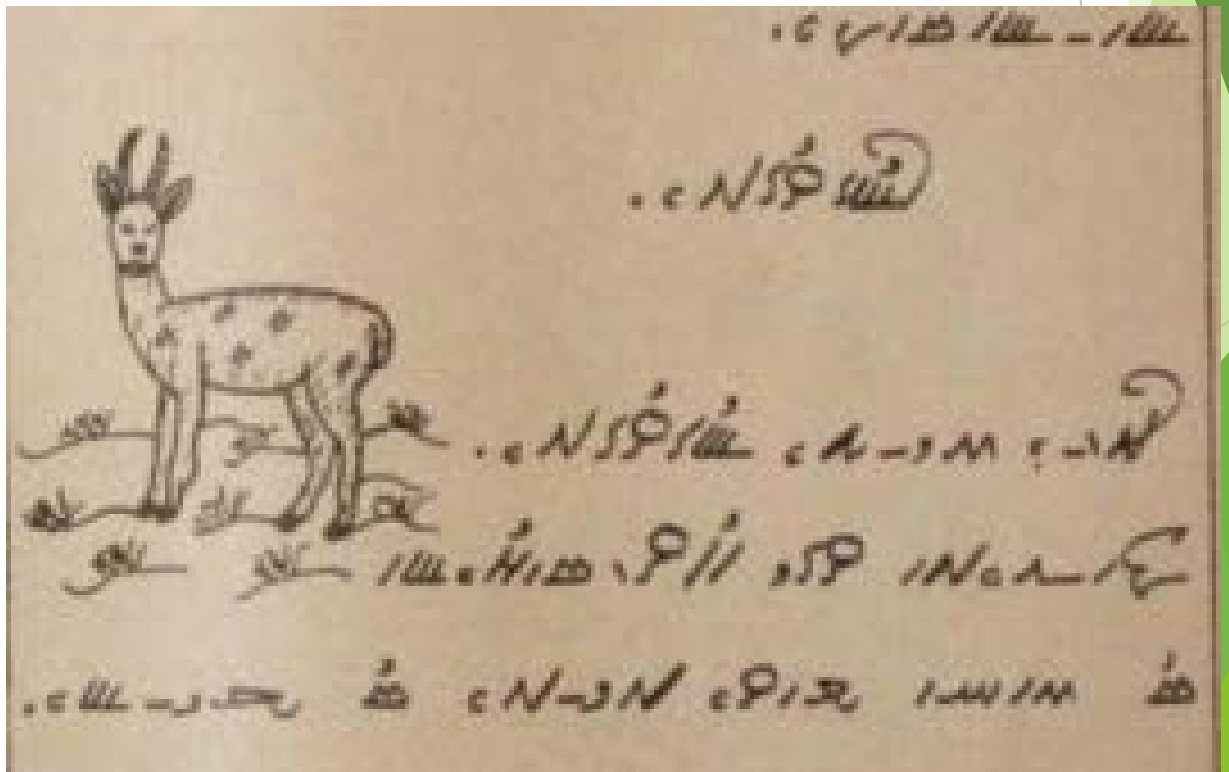
Case study: Mandombe

- Created in 1978
- Used in Democratic Republic of Congo and surrounding countries for Bantu languages of the Congo
- Connected to Kimbanguist Church
- Copyright issue affecting its encoding



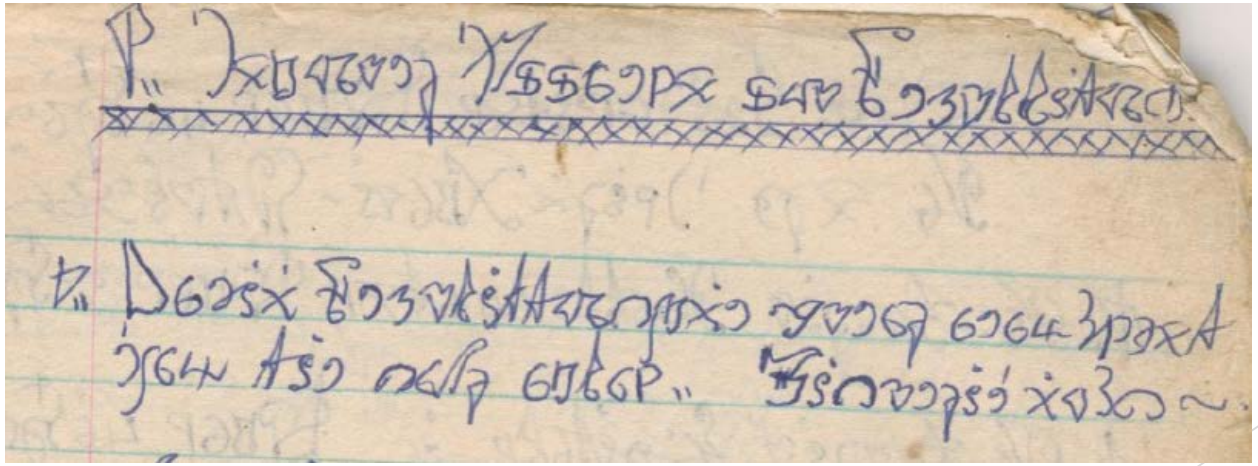
Case study: Garay (Wolof)

- Developed in 1961
- Creator (Assane Faye) still alive
- Used for Wolof (4 million speakers in West Africa)
- Taught in classes



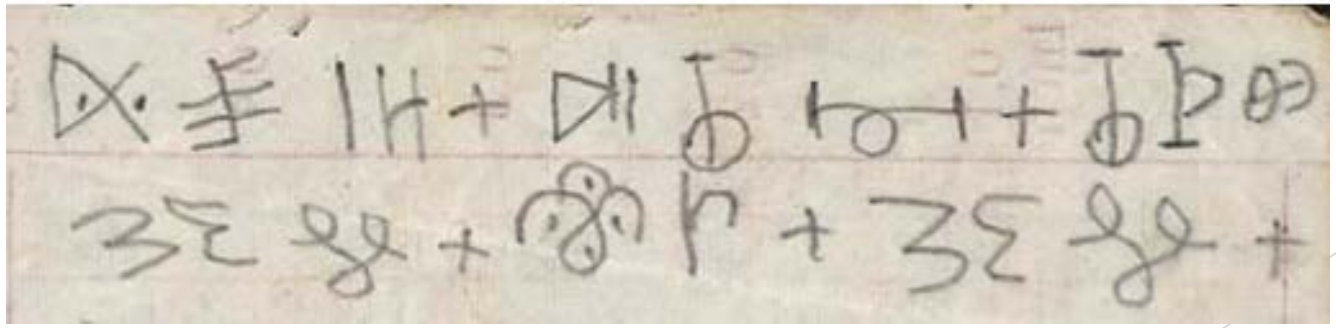
Case study: Oberi Okaime (Church "freely given")

- Created ca. 1927, fl.1930-1980
- Used for Medefaidrin language, a "spirit language" spoken by a Christian group in SE Nigeria
- Limited use today but linguists and community are interested in documenting and preserving it



Case study: Loma

- Created in 1930s
- Used in 1930s and 1940s for Loma language, spoken in Guinea and Liberia by 195,000
- Scarce primary material, primarily personal correspondence or record-keeping
- Small group of interested users



Problems

- Difficult to get information on the scripts and their use
  - Fieldwork may be required

  - Some scripts have scarce source material, so need to rely on secondary material

Problems

From standards committees' perspective:

- Need to provide rationale for encoding the script:
  - Is there an interested group of scholars or users?
  - Are there ongoing digitization projects?

- Need to show (newer) scripts will take hold, not be ephemeral or limited to very few people

Other challenges

- Many of the unencoded scripts are in remote areas in West Africa; may be difficult to get a timely response to questions

- Most of the scripts have no official government support

Approaches to gather information

- Rely on users in diaspora for information

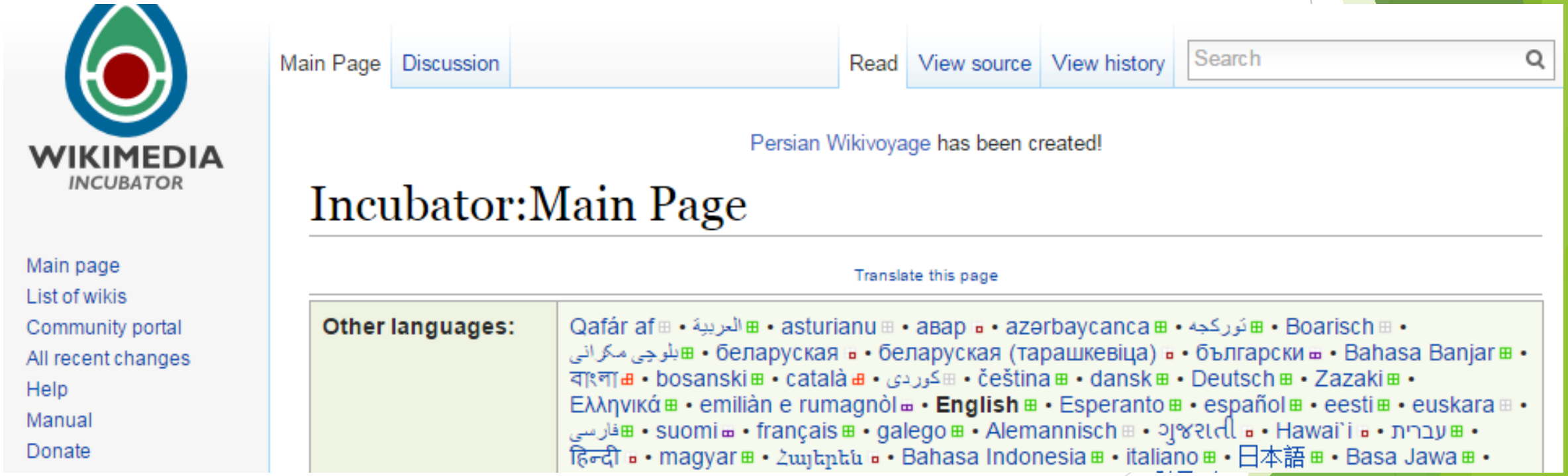
- Use social media to locate members of the community and gauge interest

New possibilities for encoded scripts

- Growth of mobile phones may encourage use of local scripts (once encoded)

New possibilities for encoded scripts

- Wikimedia Incubators as a way to spawn interest in local scripts



Summary

- Egyptian hieroglyphs (Ptolemaic): need research
- Various modern African scripts still need:
  - adequate text materials

  - information on use of characters

  - verification script is used today (and stable)
  - rationale for encoding the script

Acknowledgements

- Andrij Rovenchak, author of *African Writing Systems of the Modern Age* (with J. Glavy)
- Chuck Riley, Catalog Librarian for African Languages, Yale University Library
- Prof. Konrad Tuchscherer, St. John's University
- Don Osborn, Bisharat

Bamum

Scripts of Asia

Scripts of (Non-Ideographic) Asia

South Asia: already encoded (30)

- Bengali
- ***Brahmi***
- Gujarati
- Grantha
- Gurmukhi
- ***Kaithi***
- ***Kharoshthi***
- Kannada
- ***Khojki***
- ***Khudawadi***
- ***Lepcha***
- Limbu
- ***Mahajani***
- Malayalam
- ***Meetei Mayek***
- ***Modi***
- Mro
- ***Ol Chiki***
- Oriya
- ***Saurashtra***
- ***Sharada***
- ***Siddham***
- Sinhala
- ***Sora Sompeng***
- Syloti Nagri
- ***Takri***
- Telugu
- Thaana
- ***Tirhuta***
- ***Warang Citi***

Note: Scripts in ***bold italic*** had assistance from SEI

South Asia: unencoded (23)

- Ahom *
- Bhaiksuki *
- Balti ‘A’
- Balti ‘B’
- Bhujinmol
- Chalukya
- Chola
- Dhives Akuru
- Dogra
- Gondi
- Gunjala Gondi
- India Valley script
- Kadamba
- Landa
- Multani *
- Nandinagari
- Newa *
- Pallava
- Ranjana (Landzya)
- Satavahana
- ‘Shankha lipi’ (shell script)
- Sindhi scripts
- Tulu (Tigalari)

* Approved by UTC

South Asia: unencoded - new scripts (15)

- Bagada
- Coorgi Cox
- Dhimal
- Jenticha
- Khambu Rai
- Gurung (Khema & Phri)
- Kirat Rai
- Magar Akkha
- Tangsa (2 scripts)
- Tani Lipi
- Tikamuli
- Tolong Siki
- Zou

Southeast Asia: already encoded (22)

- ***Balinese***
- ***Batak***
- ***Buginese***
- Buhid
- ***Cham***
- Hanunoo
- ***Javanese***
- ***Kayah Li***
- Khmer
- Lao
- Myanmar
- New Tai Lue
- ***Pahawh Hmong***
- ***Pau Cin Hau***
- ***Rejang***
- ***Sundanese***
- Tagalog
- Tagbanwa
- Tai Le
- ***Tai Tham***
- ***Tai Viet***
- Thai

Note: Scripts in ***bold italic*** had assistance from SEI

Southeast Asia: unencoded (9)

- Eskaya
- Gangga Malayu (cipher?)
- Kawi
- Leke
- Makassrese Bird Script
- Pau Cin Hau Syllabary
- Pyu
- Rakhawunna
- Rohingya

Central Asia: already encoded (5)

- ***Manichaean***
- Mongolian
- ***Old Turkic***
- Phags-pa
- Tibetan

Note: Scripts in ***bold italic*** had assistance from SEI

Central Asia: unencoded (8)

- Khatt-i Baburi (cipher?)
- Khotanese (Turkestani)
- Marchen *
- Old Uyghur
- Sogdian
- Soyombo
- Tocharian
- Zanabazar Square *

* Approved by UTC

Number Systems: unencoded

- North Indian ‘Letter Numbers’
- South Indian ‘Letter Numbers’
- Siyaq Numbers
  - Arabic (Diwani)
  - Ottoman
  - Persian
  - North Indian
  - South Indian (Dakkhani)

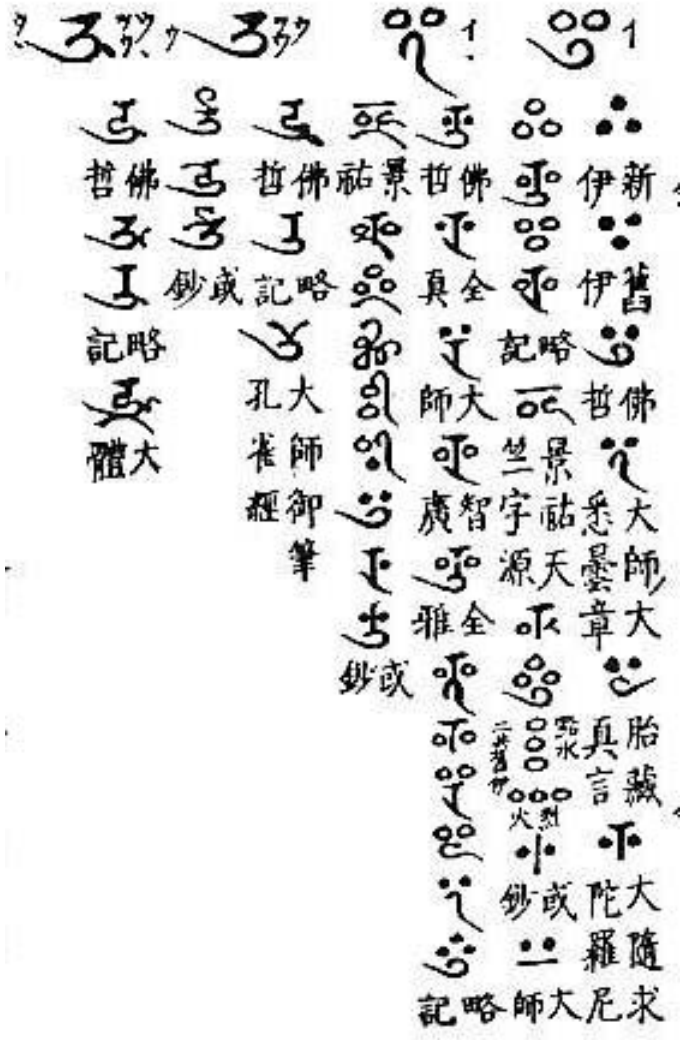
Recent Success: Siddham

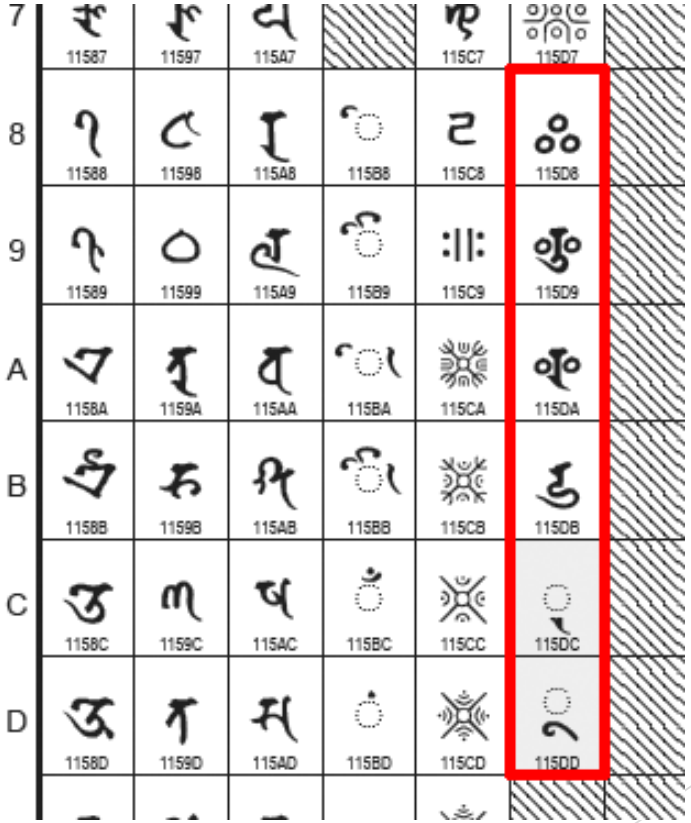
Recent Success: Siddham

- East Asia, since 9th c. CE, predominantly in Japan
- Brahmi-based, left to right
- Liturgical: Buddhist texts in Sanskrit
- Challenges for encoding:
  - Alphasyllabic script, but is analyzed from an ideographic perspective
  - Features have different semantics in Japanese context
  - Meeting in Tokyo, November 2013 with experts

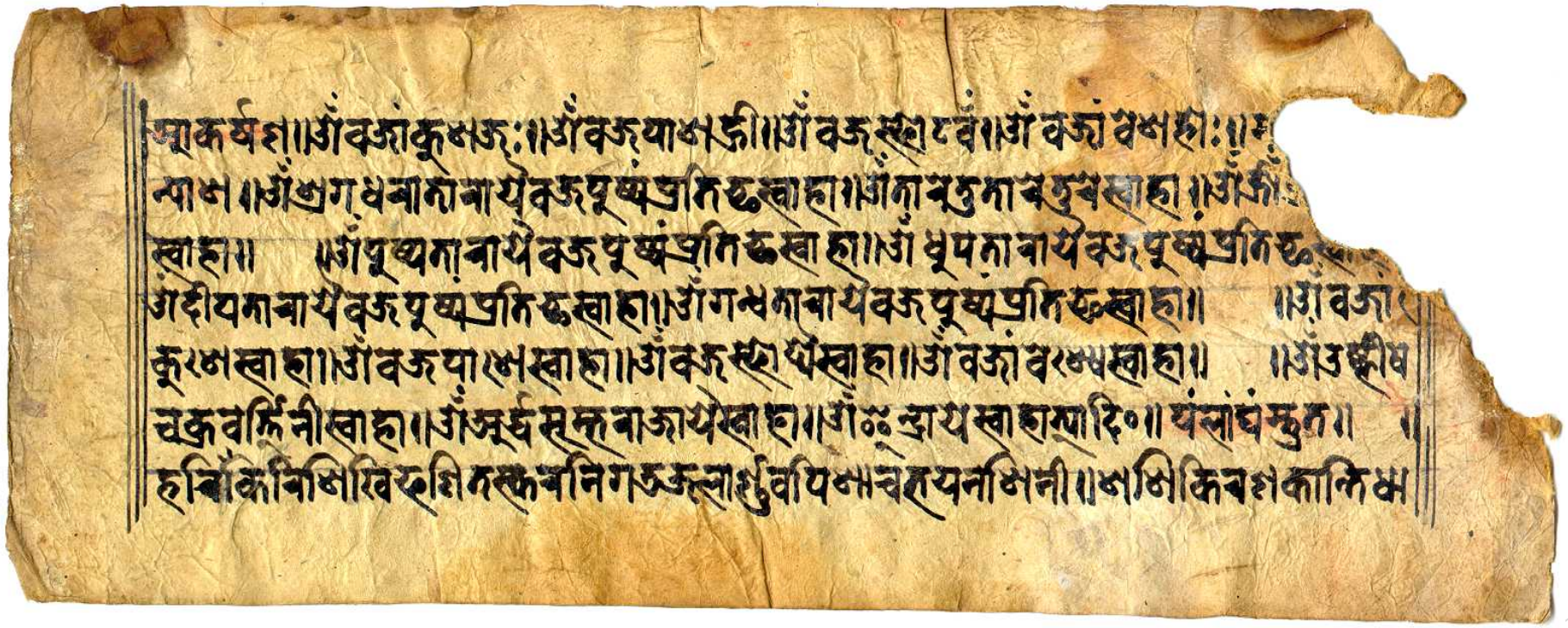
Recent Success: Siddham

APPROVED





Recent Success: Newa



Recent Success: Newa

- Nepal, 10th century to 20th century
- Brahmi-based
- Used for writing Sanskrit, Maithili, Nepalese, Nepal Bhasa (Newar)
- +100,000 records (manuscripts, inscriptions, books)
- Challenges for encoding:
  - Historical script being revived and reformed
  - Ethno-political issues
  - Adaption of Brahmi-based script for writing Tibeto-Burman

Recent Success: Newa

- First proposed in 2012
- Wikimedia funded trip to Kathmandu to meet with user community
- Consensus developed during meeting and remotely after
- Approved for encoding at UTC October 2014

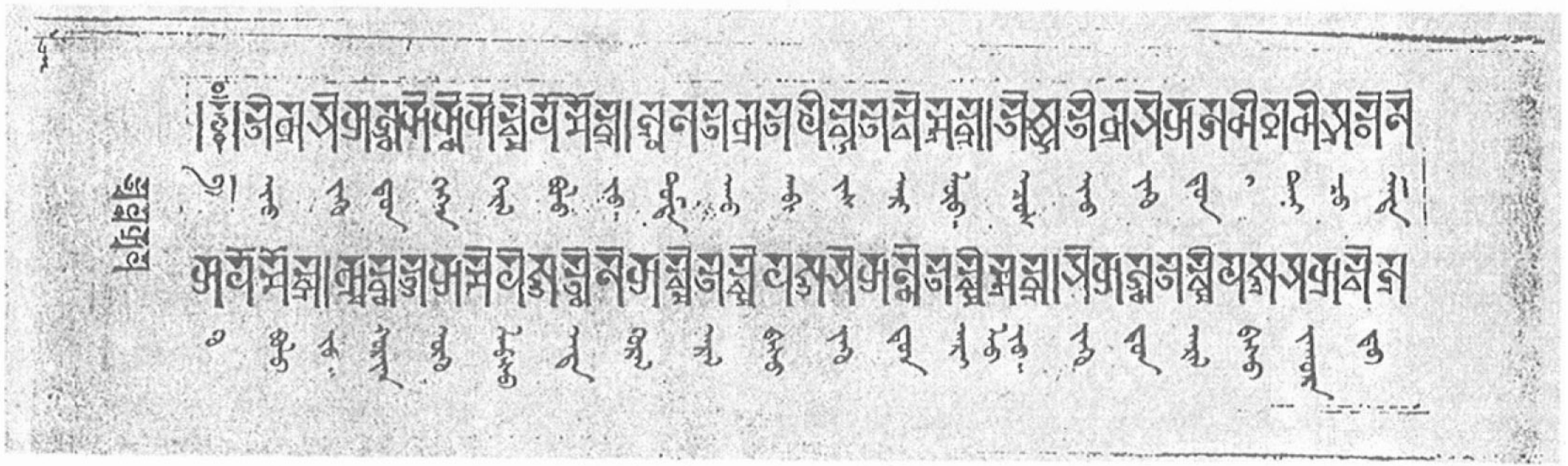
APPROVED

Challenges: Bhujinmol

Challenges: Bhujinmol

- Nepal, parts of northern India, 12-17th centuries CE
- Brahmi-based: structure identical to Newa script
- Glyph repertoire nearly identical to Newa
- Distinguished by head-stroke (*bhujinmol* = “fly-headed”)
- Challenges for encoding:
  - Unify as style of Newa or encode as independent script for plain text?

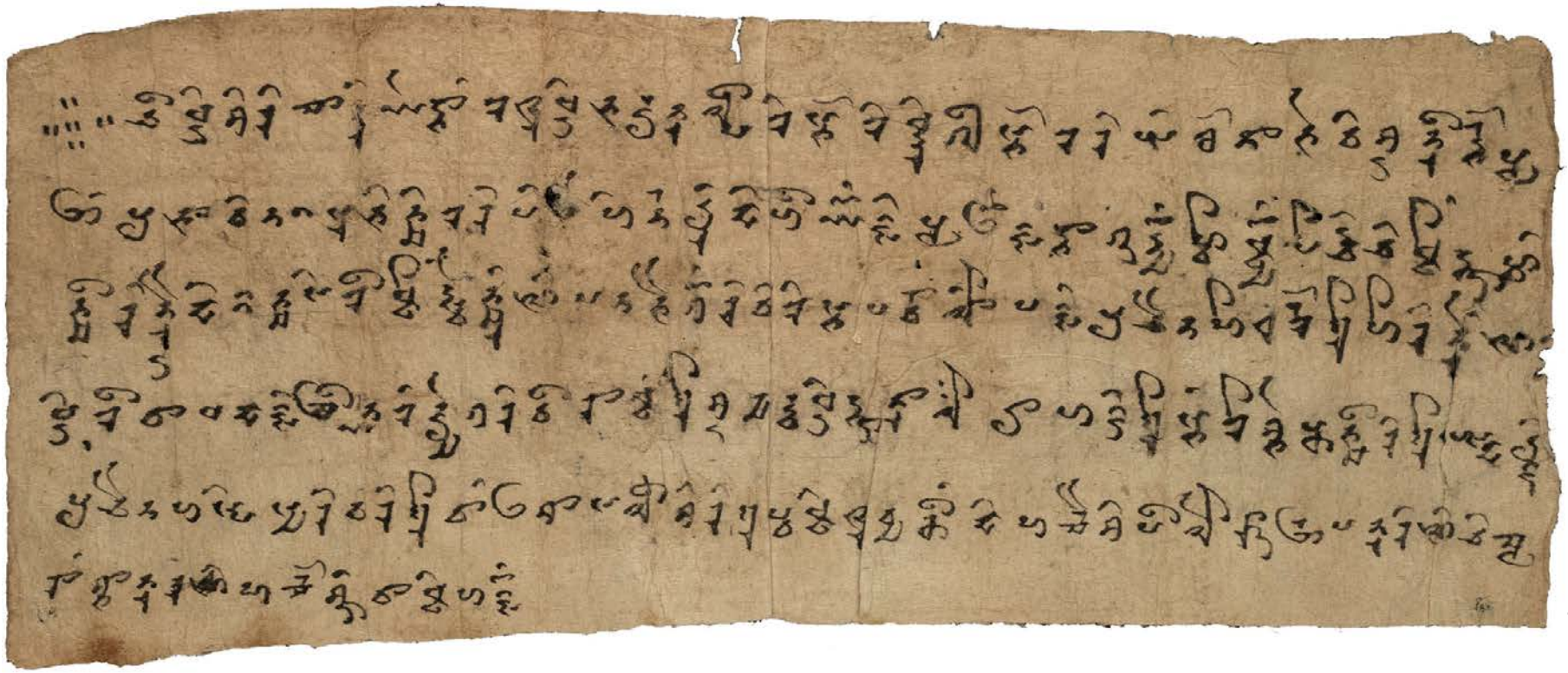
Unencoded: Soyombo



Unencoded: Soyombo

- Liturgical script developed by Zanabazar, 17th c. CE
- Brahmi-based, modeled upon Ranjana and Tibetan
- Used for writing Sanskrit, Tibetan, Mongolian
- Writing system has language-specific features
- Challenges for encoding:
  - Access to user community
  - Access to sources

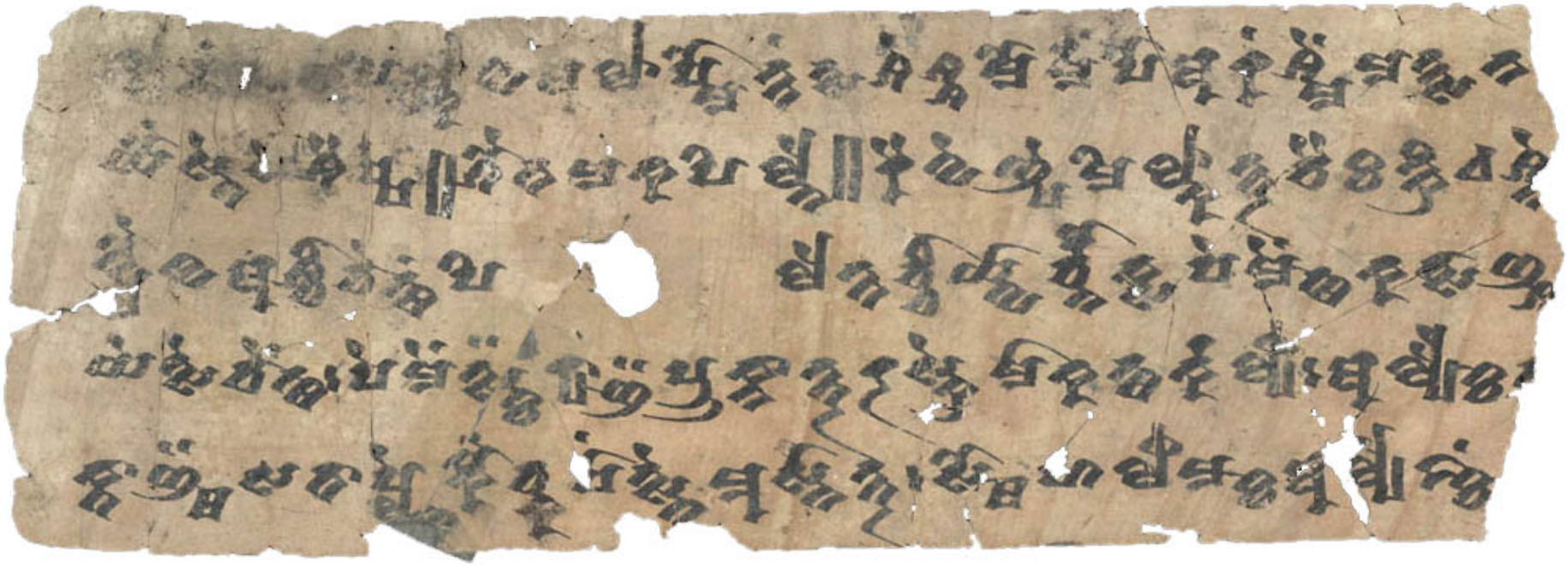
Unencoded: Khotanese



Unencoded: Khotanese

- Western China, 4th-11th c. CE
- Brahmi-based script, left to right
- Used for Gandhari, Khotan
- Challenges for encoding:
  - Unify with Brahmi?
  - Access to sources

Unencoded: Tocharian



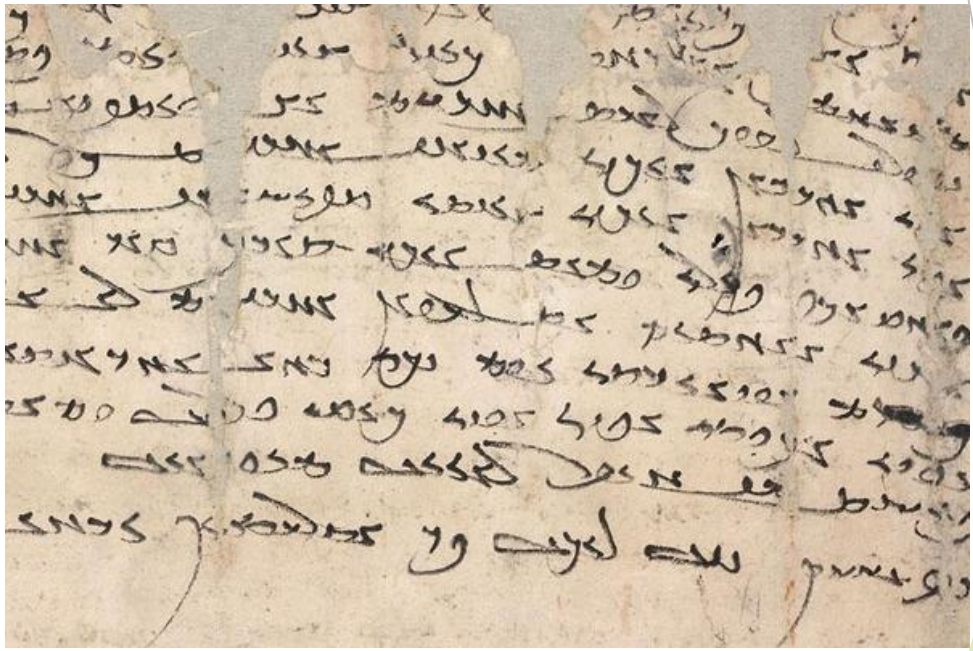
Unencoded: Tocharian

- Western China, 9th century
- Brahmi-based script, left to right
- Used for writing Sanskrit, Tocharian
- Buddhist and Manichaean texts, administrative documents,
- Challenges for encoding:
  - Unification with Brahmi?
  - Further analysis of sources

Unencoded: Sogdian

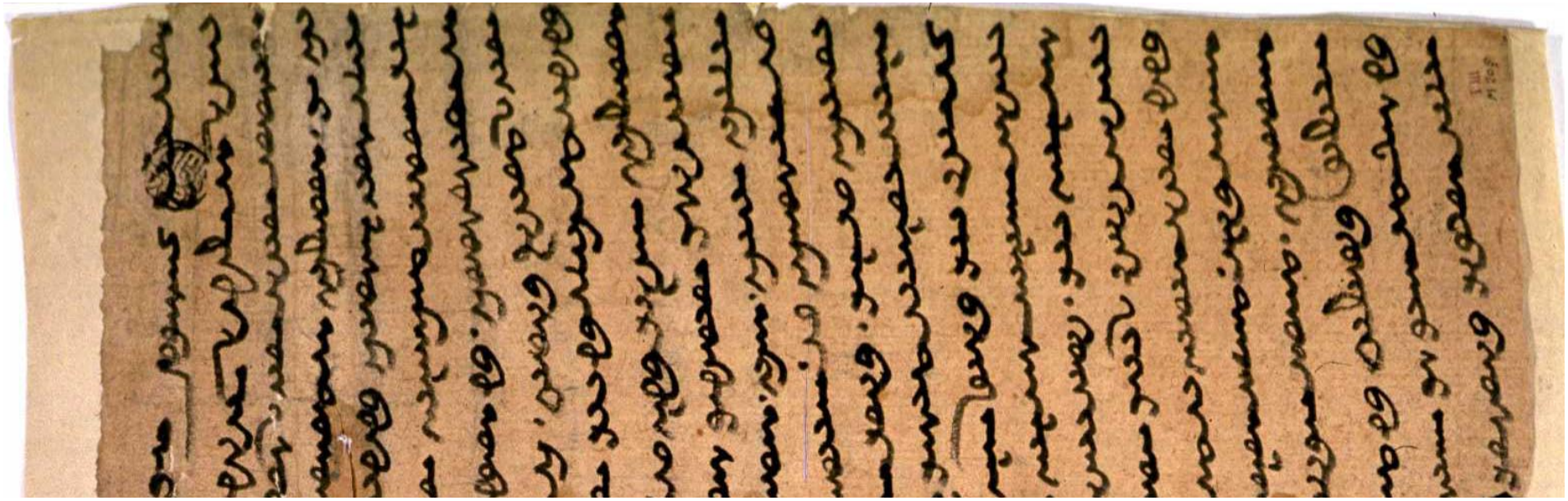
Unencoded: Sogdian

- Iran to China, 2nd-13th c. CE
- Abjad, alphabet; right to left, derived from Syriac
- Used for writing Sogdian
- Religious texts of Buddhism, Manichaeanism, Christianity
- Challenges for encoding:
  - Unification with Syriac?
  - Analysis of logograms
  - Further analysis of sources



"I'd rather be a dog's or a pig's wife than yours" – Sogdian lady writing to her husband, 314 CE
(source: International Dunhuang Project, British Library)

Unencoded: Old Uyghur



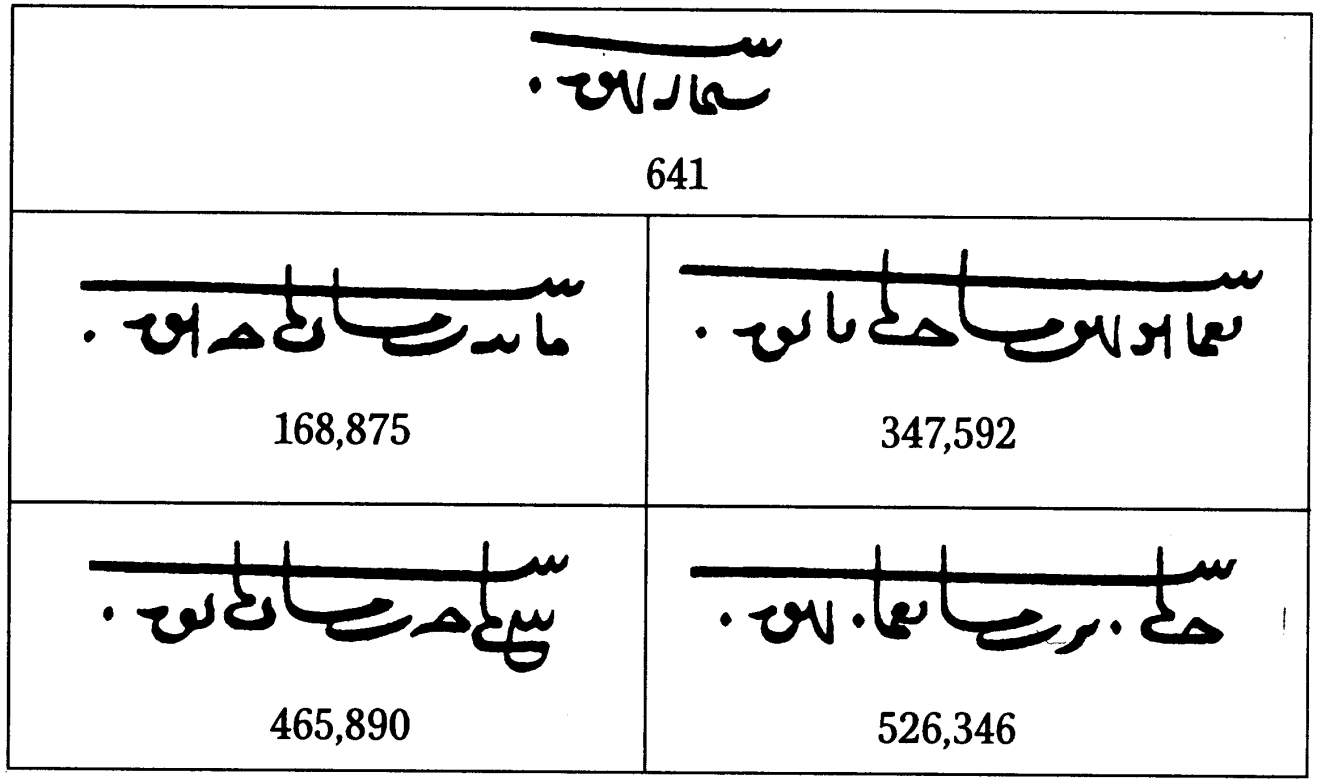
Unencoded: Old Uyghur

- Used in western China, predominantly in Xinjiang region, 7th-19th c. CE
- Abjad, alphabet; vertical orientation
- Derived from Sogdian, basis for Mongolian
- Challenges for encoding:
  - Accommodating sub-regional styles and orthographies
  - Access to sources and user community
  - Political sensitivities

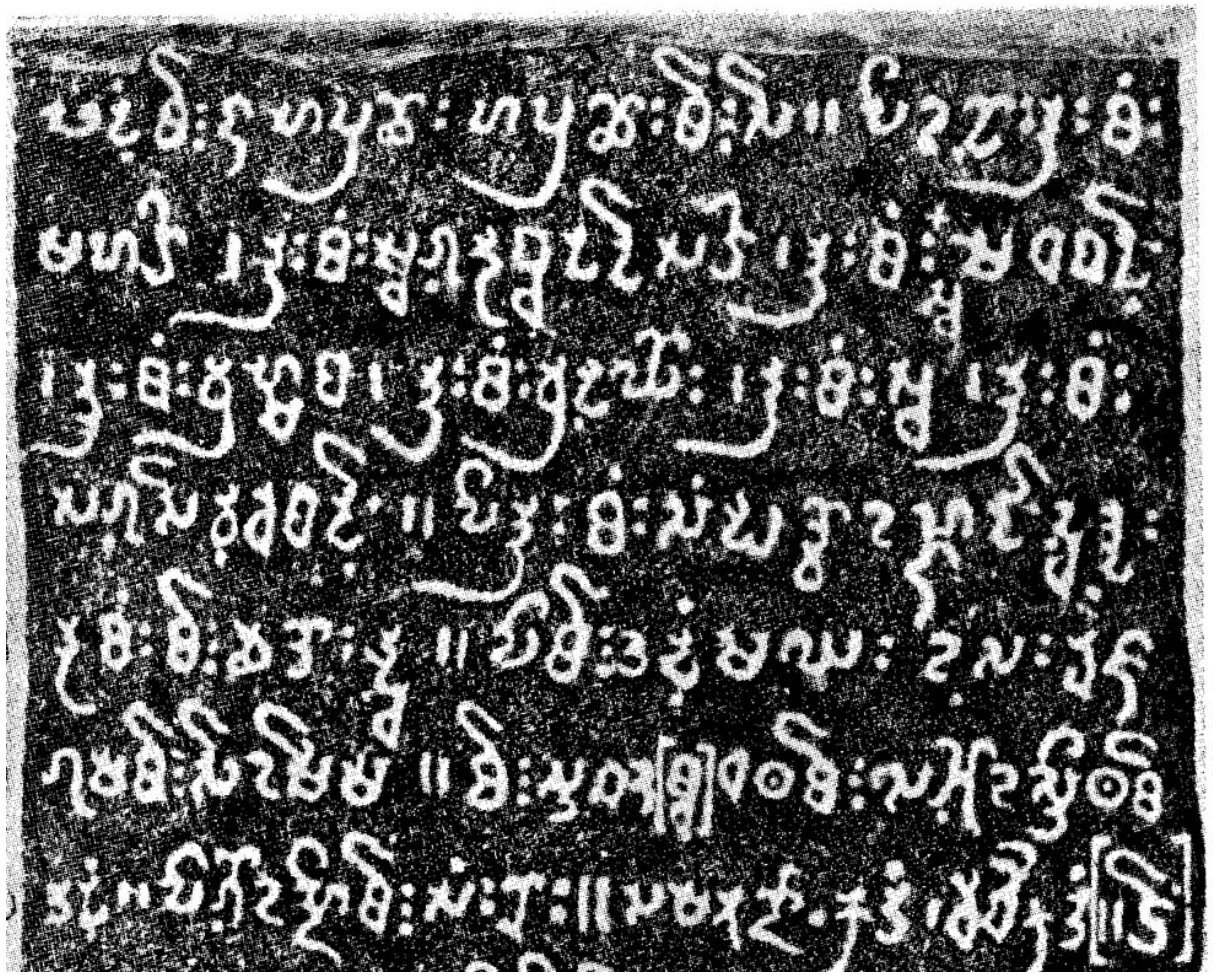
Unencoded: Siyaq Numbers

Unencoded: Siyaq Numbers

- Specialized subset of Arabic used for numerical notation
- Highly stylized abbreviations for Arabic names of numbers
- Middle East to South Asia
- Different styles, same underlying principle
- Challenges for encoding:
  - Model for encoding
  - Fractions, unit marks
  - How much to unify?



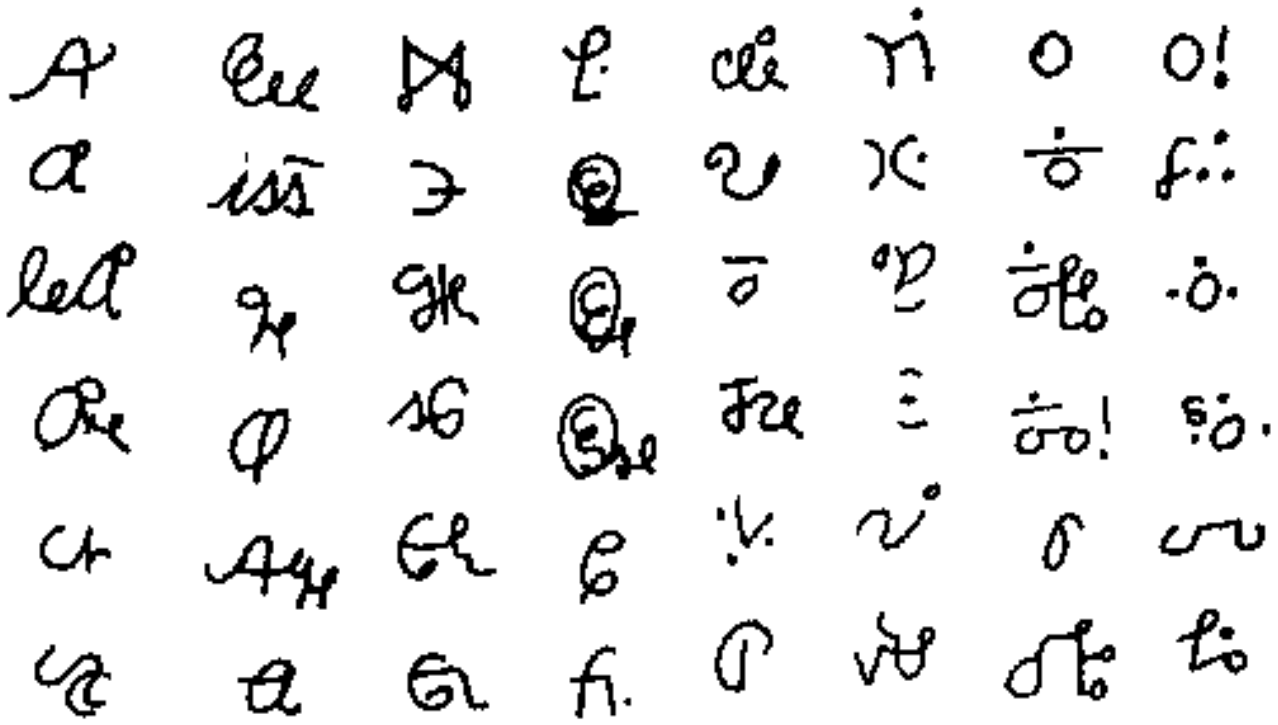
Unencoded: Pyu



Unencoded: Pyu

- Myanmar, 5th c. CE
- Brahmi-based, left to right
- Used primarily for inscriptions: gold leaf, terracotta, stone
- Two styles: Pyu Pali & Pyu Tircul
- Challenges for encoding:
  - Could be unified with the Pallava script
    - Requires encoding the Pallava script
  - Access to and analysis of sources

Unencoded: Eskaya



Unencoded: Eskaya

- Created by Mariano Datahan, early 20th c.
- Syllabary, 1,065 letters
- Used for writing Eskayan, an artificial language used on Bohol
- Challenges for encoding:
- Determining suitability for encoding
  - Investigation of sources
  - Extent of usage
  - Current status

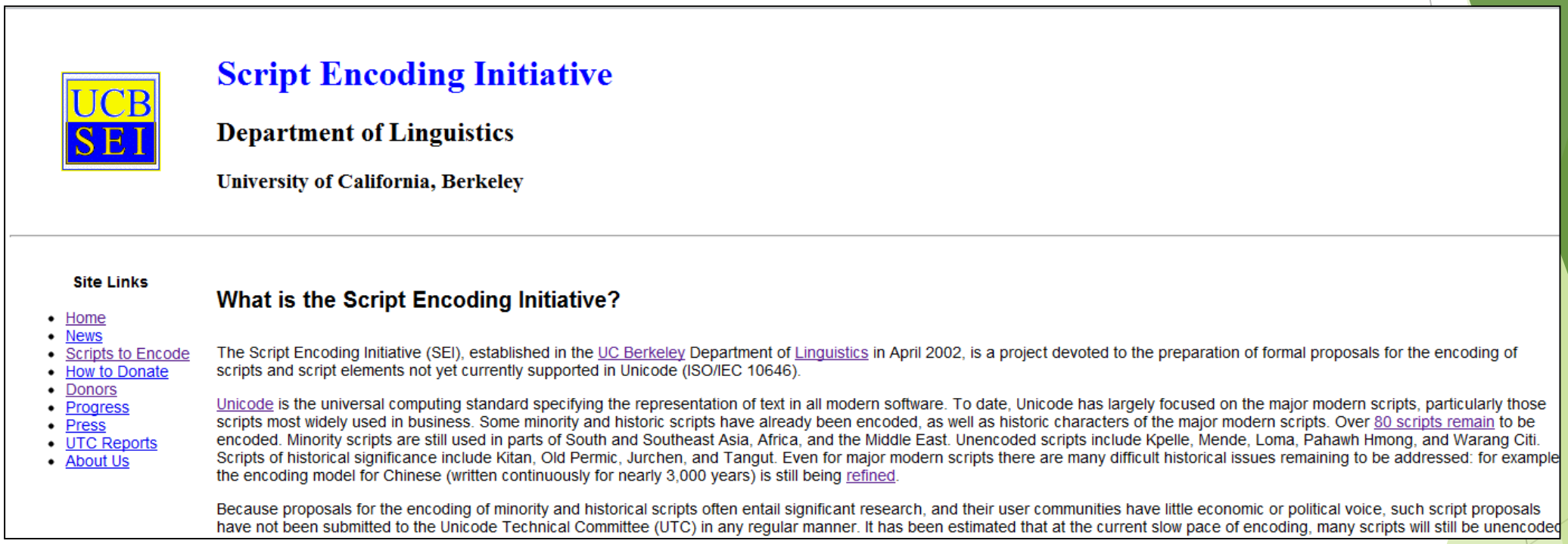
Filling in the Gaps

- Bengali: weights and measures
- Buginese: Ende, Bimanese extensions
- Devanagari: invocation signs, vowel signs, Vedic extensions
- Gujarati: Arabic transliteration marks
- Khojki: additional letters, Arabic transliteration marks
- Malayalam: weights and measures
- Mongolian: head marks
- Oriya: invocation signs, fraction signs, ‘letter-numbers’
- Rejang: Kerinci, Minangkabau, Lampung, Angka Bejagung numeral extensions
- Sharada: various signs, Vedic tone marks
- Takri: disunification of some regional scripts
- Tirhuta: fractions, currency, weights, measures marks

Expanding the Repertoire

- Unencoded scripts: +102
  - Africa: 47
  - Asia: +55
- Challenges
  - +8 years: from preliminary research for proposal to publication in Unicode
  - New universal shaping engine will speed up implementation
  - Access to user community, sources, and funding affect encoding projects

Script Encoding Initiative at UC Berkeley



Script Encoding Initiative

Department of Linguistics

University of California, Berkeley

Site Links

- Home
- News
- Scripts to Encode
- How to Donate
- Donors
- Progress
- Press
- UTC Reports
- About Us

What is the Script Encoding Initiative?

The Script Encoding Initiative (SEI), established in the UC Berkeley Department of Linguistics in April 2002, is a project devoted to the preparation of formal proposals for the encoding of scripts and script elements not yet currently supported in Unicode (ISO/IEC 10646).

Unicode is the universal computing standard specifying the representation of text in all modern software. To date, Unicode has largely focused on the major modern scripts, particularly those scripts most widely used in business. Some minority and historic scripts have already been encoded, as well as historic characters of the major modern scripts. Over 80 scripts remain to be encoded. Minority scripts are still used in parts of South and Southeast Asia, Africa, and the Middle East. Unencoded scripts include Kpelle, Mende, Loma, Pahawh Hmong, and Warang Citi. Scripts of historical significance include Kitan, Old Permic, Jurchen, and Tangut. Even for major modern scripts there are many difficult historical issues remaining to be addressed: for example the encoding model for Chinese (written continuously for nearly 3,000 years) is still being refined.

Because proposals for the encoding of minority and historical scripts often entail significant research, and their user communities have little economic or political voice, such script proposals have not been submitted to the Unicode Technical Committee (UTC) in any regular manner. It has been estimated that at the current slow pace of encoding, many scripts will still be unencoded

http://linguistics.berkeley.edu/sei

Email: dwanders@berkeley.edu

“One standard to rule them all”