From Local Script to Global Standard

The Lifecycle of a Script in Unicode

SF Globalization MeetUp
3 August 2015

Deborah Anderson
Director, Script Encoding Initiative
Department of Linguistics, UC-Berkeley

Anshuman Pandey
Post-Doctoral Researcher
Department of Linguistics, UC-Berkeley

The Lifecycle of a Unicode Script

- There are +100 scripts, symbol sets, and number blocks in Unicode
- Transparent implementation of most into operating systems
  - As to be expected from a modern i18n and i10n perspective…
- Nuts and bolts of script and language support also available
  - Code charts; fonts; and transliteration, collation, locale data
- How is a Unicode standard for a script developed?

Script Encoding Process: Overview

1. Users, linguists, others identify a script not yet encoded in Unicode/ISO 10646 standard
2. Research script and develop script proposal (often with revisions)
3. Two standards committees review proposals; may request changes; vote to approve or disapprove
4. Publication of script in Unicode/ISO 10646 standard
5. Create fonts, keyboards, update software

SEI: Overview

- Started 2002 in Department of Linguistics, UC Berkeley

- Assists users with encoding characters and scripts into Unicode

- Support: NEH (PR-50205), Google Research Award, other sources

Script Encoding Process: Role of SEI

1. Users, linguists, others identify a script not yet encoded in Unicode/ISO 10646 standard

SEI 

2. Research script and develop script proposal (often with revisions)

SEI 

3. Two standards committees review proposals; may request changes; vote to approve or disapprove

4. Publication of script in Unicode/ISO 10646 standard

5. Create fonts, keyboards, update software

SEI: Goals

- Assist in preservation of physical documents and in making them electronically accessible
- Contribute to the creation of a global digital repository
- Enable users to take advantage of electronic communication (text messaging, email, etc.)

SEI: Successes

Since 2002, SEI has helped encode over **70** scripts and individual characters in the Unicode standard

Warang Citi

Ho language, India

N’Ko

Manding languages, west Africa

SEI: Work remaining

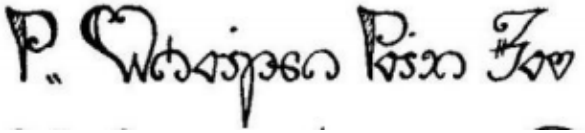
Over **100** scripts remain unencoded, including:

- *Historical scripts:*

Khitan Small Script, China

Mayan

- *Modern minority scripts:*



Medefaidrin script, Nigeria

SEI: Plan for 2015-16

- Soyombo* (Mongolia)
- Masaram Gondi* (India)
- Gunjala Gondi (India)
- Dogra (India)
- Nandinagari (India)
- Hanifi Rohingya (Myanmar)
- Pau Cin Hau Syllabary (Myanmar)
- Makasar (Indonesia)
- Siyaq (4 forms) (various)
- Old Sogdian (Central Asia)
- Old Italic additions*
- Ptolemaic additions (to Egyptian hieroglyphs)
- Late Latin additions
- Proto-Cuneiform
- Eebee Hmong (Thailand/U.S.)
- Medefaidrin (Nigeria)
- Cham additions (Vietnam and Cambodia)

* indicates scripts/characters approved by Unicode Technical Committee

SEI: Challenges

- Time
  - 2 years minimum to develop encodings, to seek approval and publication in Unicode
  - Several additional years for enabling full support for end users
- Access:
  - Experts and user communities
  - Source materials
- Funding
  - Short funding cycles (1-2 years) impair shepherding of scripts through the encoding process

Script Encoding Process: Overview

1. Users, linguists, others identify a script not yet encoded in Unicode/ISO 10646 standard
2. Research script and develop script proposal (often with revisions)
3. Two standards committees review proposals; may request changes; vote to approve or disapprove
4. Publication of script in Unicode/ISO 10646 standard
5. Create fonts, keyboards, update software

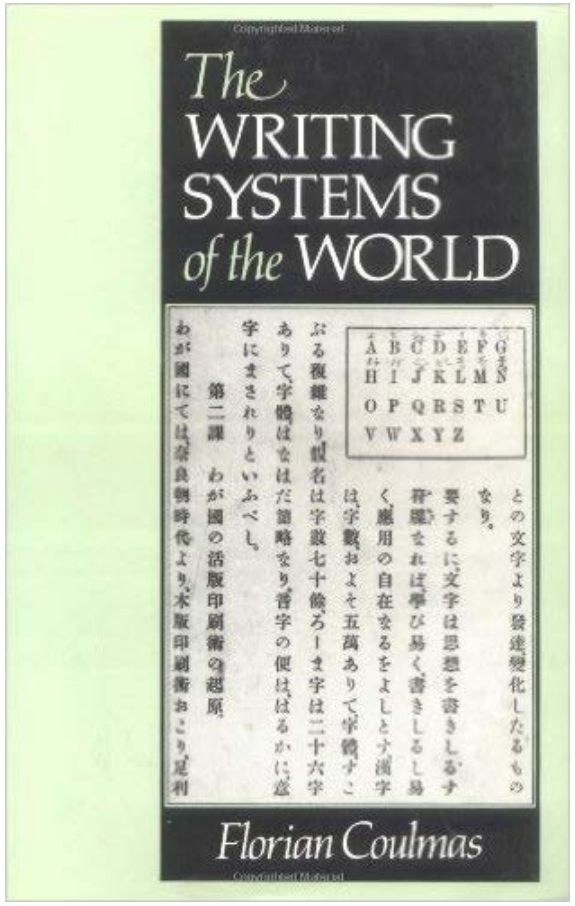
Script Encoding Lifecycle

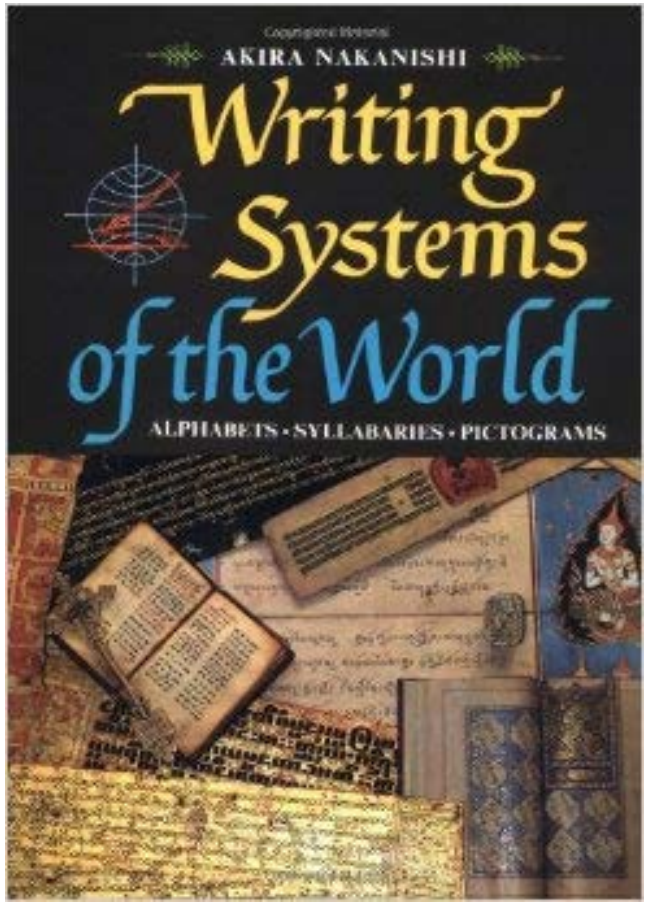
- Identification
- Research and proposal development
- Proposal review and approval
- Publication
- Implementation
- Repeat

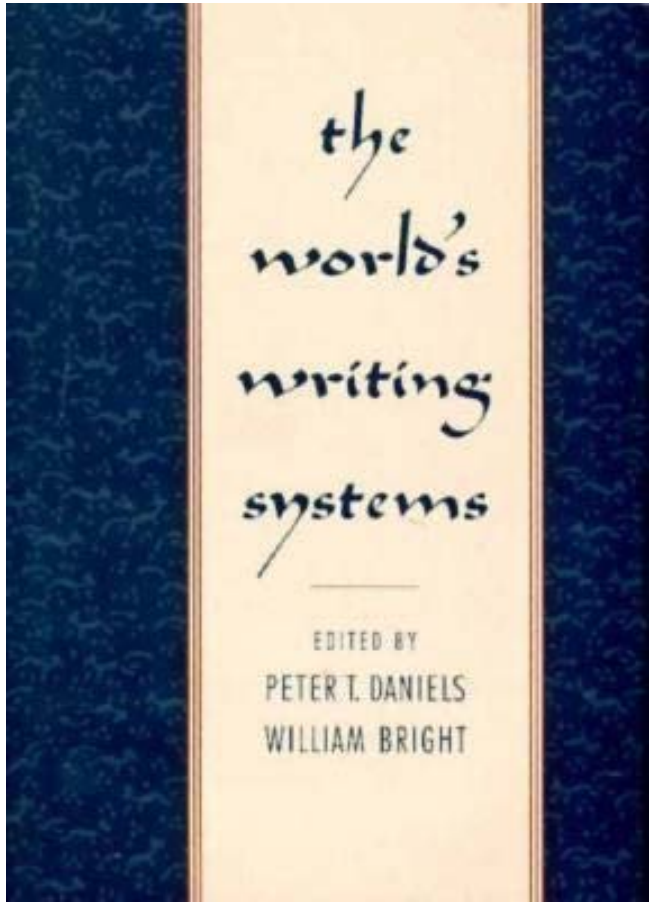
Lifecycle: Identification

- Incorporation of existing character-encoding standards
- Request from a native user community
- Proposal from scholarly user community
- Submission from writing-systems researchers and enthusiasts
- Fieldwork and archival research

Lifecycle: Identification



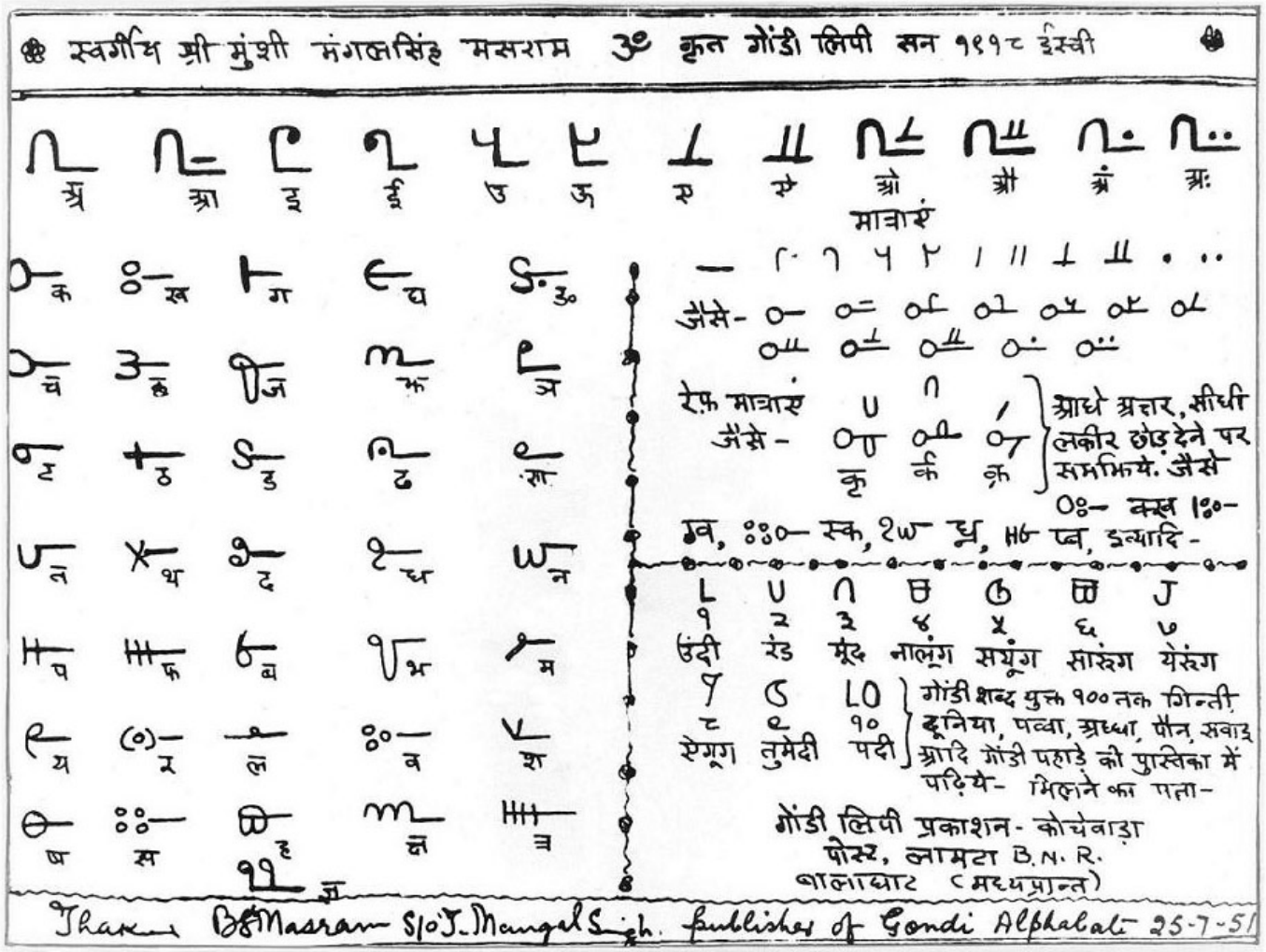




Lifecycle: Identification - Challenges

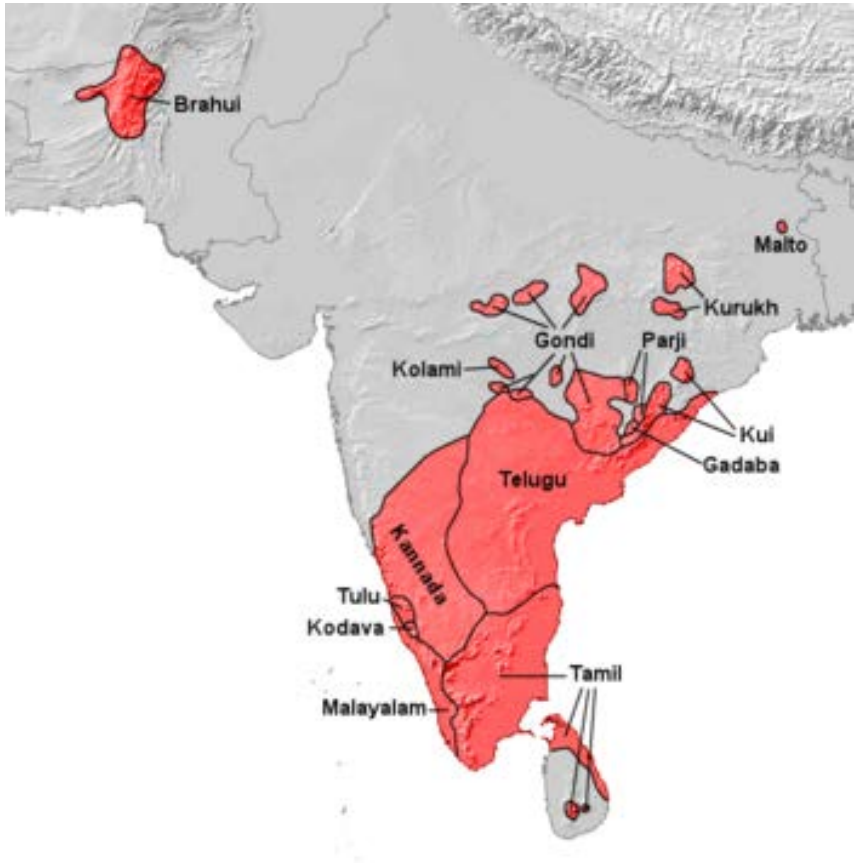
- Many historical and local scripts not in published compendia
- User communities not aware of Unicode
- Unicode specialists not aware of indigenous scripts

Lifecycle: Identification



Sent to Pandey by a scholar from the Central Institute of Indian Languages (CIIL), Mysore in 2010

Lifecycle: Case studies - Gondi



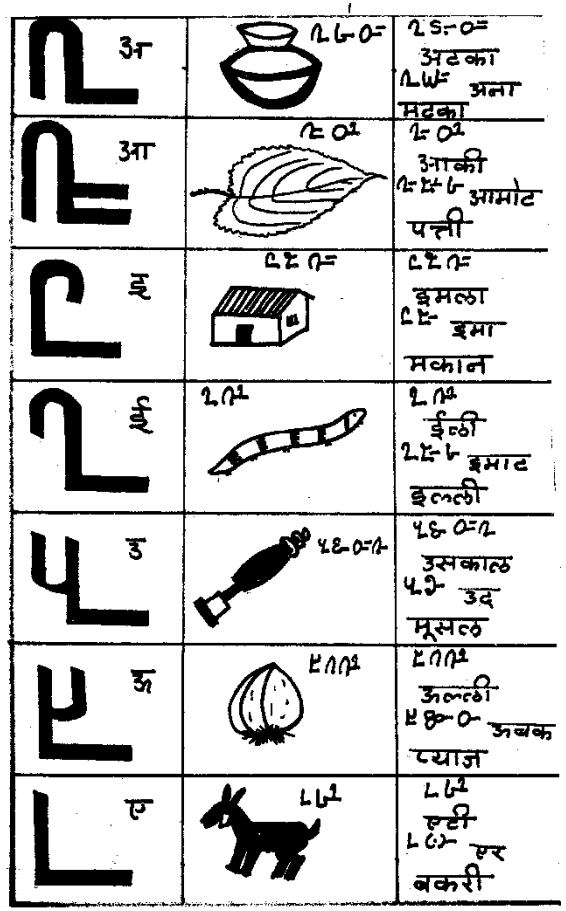
- Gondi (ISO 639-3: gon); northern & southern
- 2 million speakers
- India: Madhya Pradesh, Andhra Pradesh, Telangana
- Devanagari, Telugu

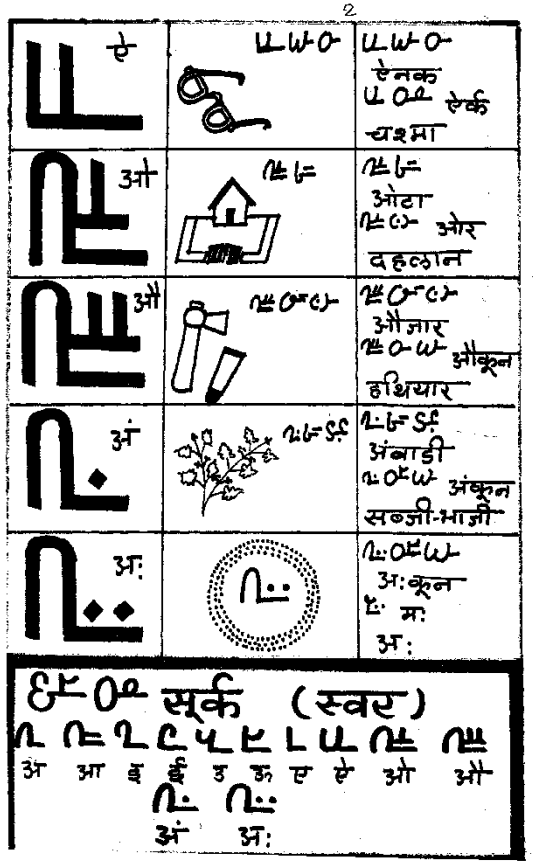
- Two indigenous scripts:
  - ‘Masaram’ Gondi
  - ‘Gunjala’ Gondi

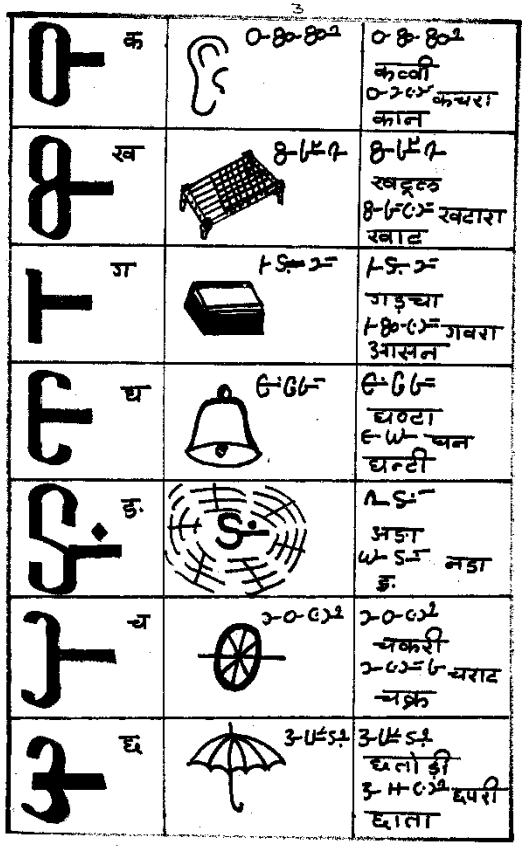
Lifecycle: Research

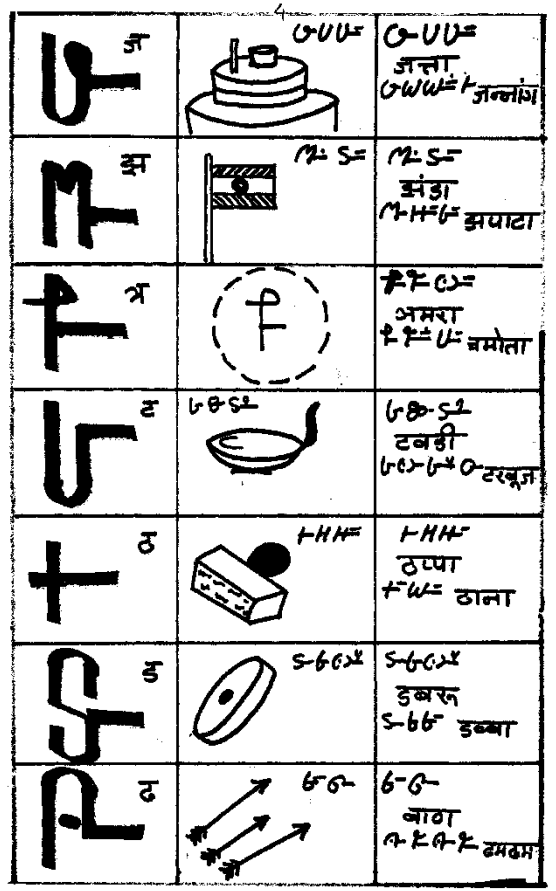
- Analysis of script and its grammar
- Grammar and orthography of associated languages
- Analysis of existing literature
- Outreach to user community to understand current usage

‘Masaram’ Gondi: Attestations

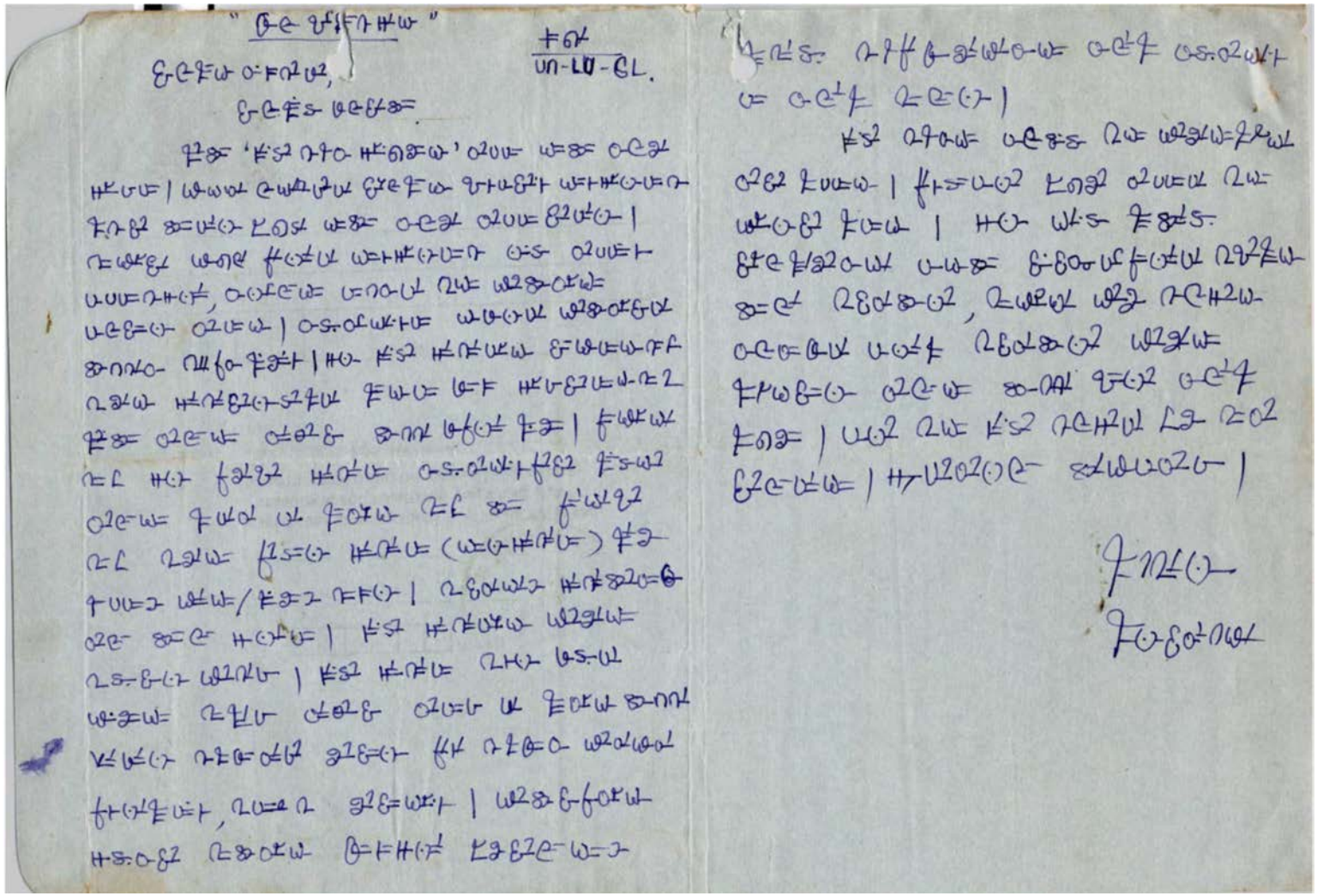




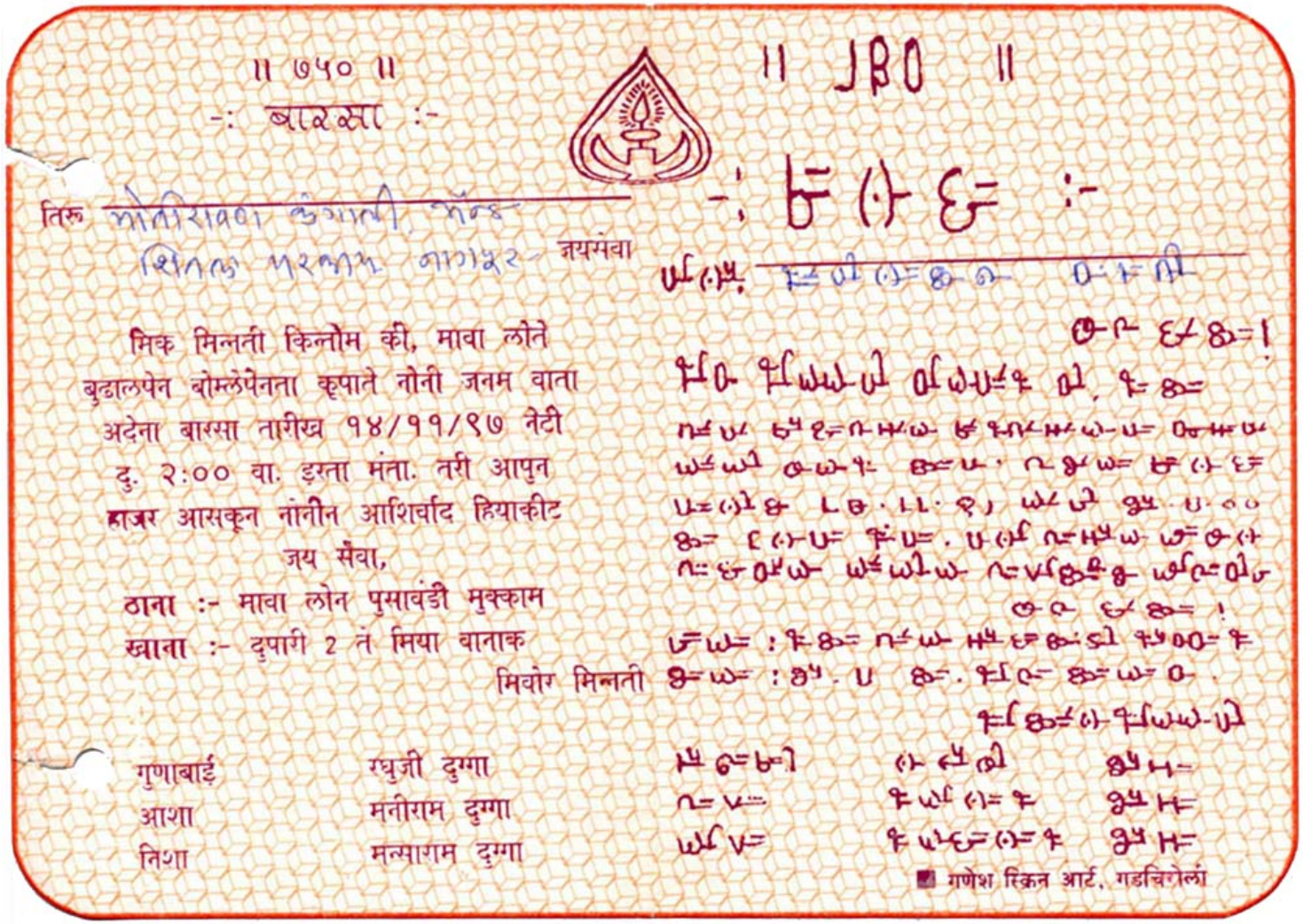




‘Masaram’ Gondi: Attestations



‘Masaram’ Gondi: Attestations



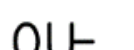
‘Masaram’ Gondi: Attestations

‘Masaram’ Gondi: Developing an encoding

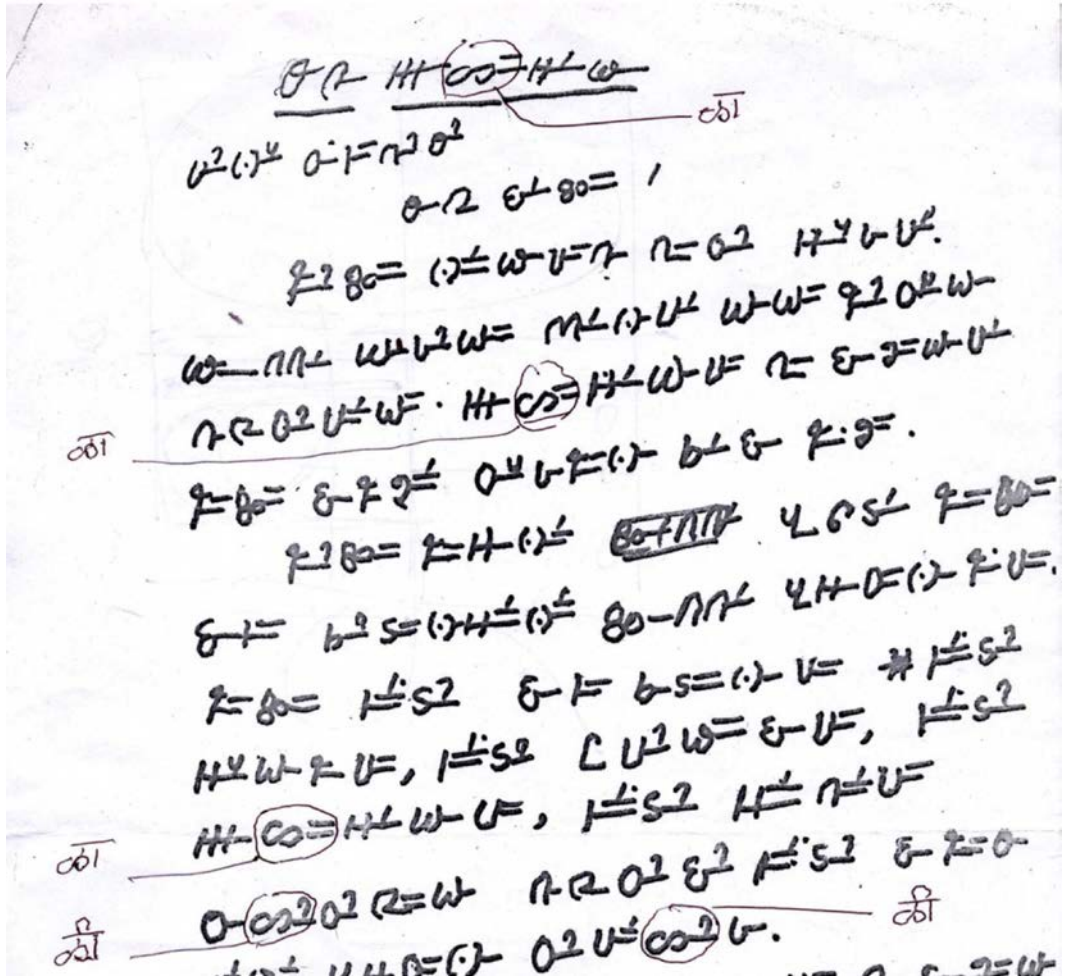
- Making use of source materials

‘Masaram’ Gondi: Analysis

- Understanding script logic, eg. representation of consonant clusters

ka		<KA>
kta		<KA, VIRAMA, TA>
ktva		<KA, VIRAMA, TA, VIRAMA, VA>
ktvya		<KA, VIRAMA, TA, VIRAMA, VA, VIRAMA, YA>

‘Masaram’ Gondi: Analysis

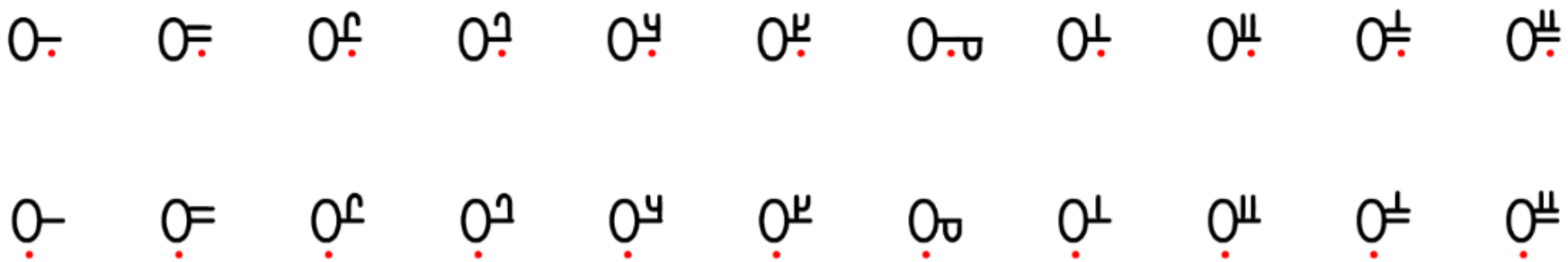


- Identification of innovations

‘Masaram’ Gondi: Analysis

- Identification of user preferences

Nukta positioning

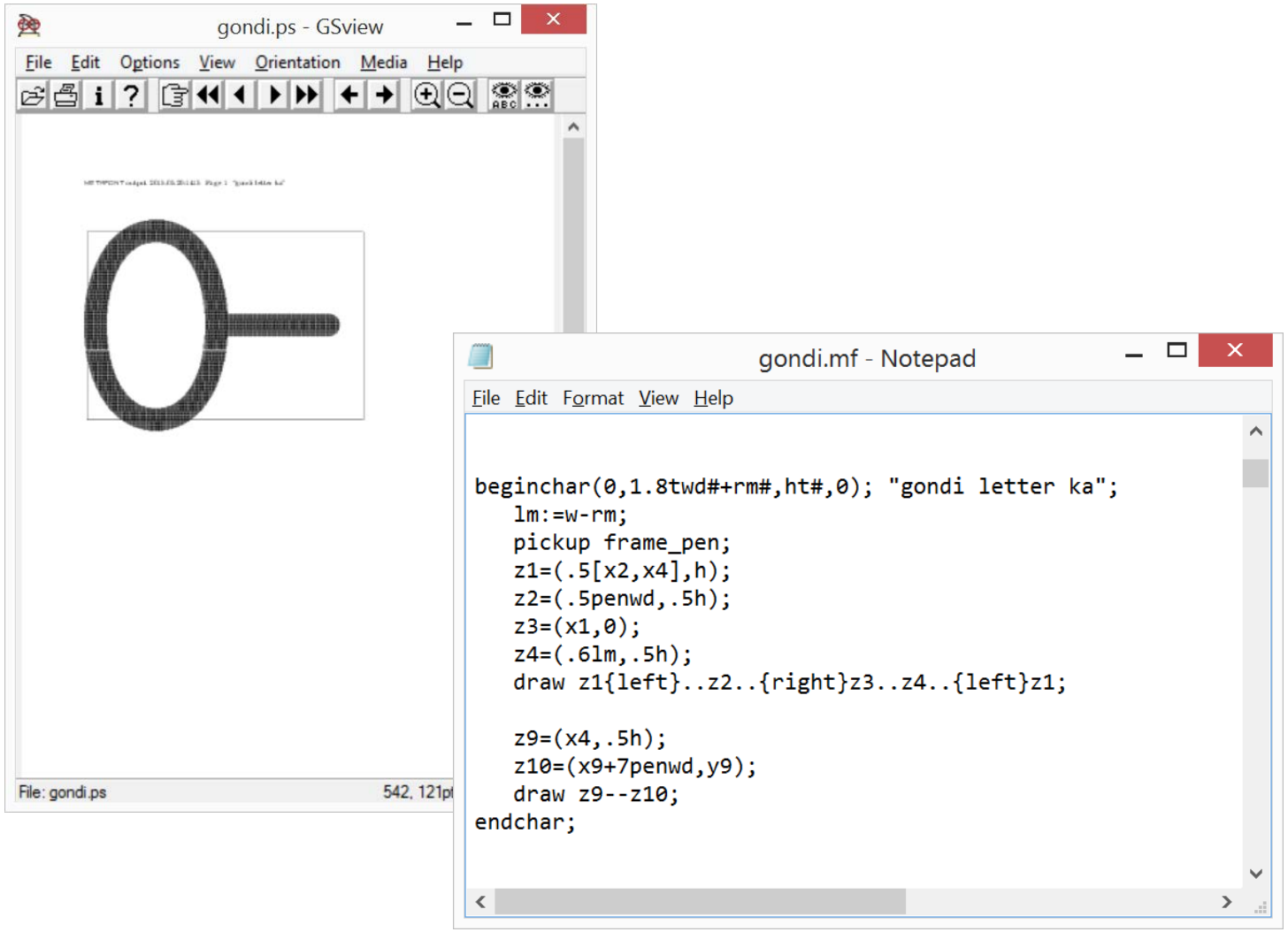


‘Masaram’ Gondi: Proposal development

- Representative font for Unicode code charts

- Translation of qualitative analysis into technical data
  - Display: linebreaking
  - Fonts: mark positioning
  - Security: confusable detection

‘Masaram’ Gondi: Font prototyping



‘Masaram’ Gondi: Linebreaking

```
11D00..11D06; AL    # MASARAM GONDI LETTER A .. MASARAM GONDI LETTER E
11D08..11D09; AL    # MASARAM GONDI LETTER AI .. MASARAM GONDI LETTER O
11D0B..11D30; AL    # MASARAM GONDI LETTER AU .. MASARAM GONDI LETTER TRA
11D31..11D36; CM    # MASARAM GONDI SIGN AA .. MASARAM GONDI SIGN VOCALIC R
11D3A..; CM         # MASARAM GONDI VOWEL SIGN E
11D3C..11D3D; CM    # MASARAM GONDI SIGN AI .. MASARAM GONDI VOWEL SIGN O
11D3F; CM           # MASARAM GONDI VOWEL SIGN AU
11D40..11D44; CM    # MASARAM GONDI SIGN ANUSVARA .. MASARAM GONDI SIGN HALANTA
11D45; CM           # MASARAM GONDI VIRAMA
11D36; AL           # MASARAM GONDI REPHA
11D47; CM           # MASARAM GONDI RA-KARA
11D50..11D59; NU    # MASARAM GONDI DIGIT ZERO .. MASARAM GONDI DIGIT NINE
```

‘Masaram’ Gondi: Syllabic Categories

```
# Indic_Syllabic_Category=Bindu
11D40 ; Bindu # Mn MASARAM GONDI SIGN ANUSVARA

# Indic_Syllabic_Category=Visarga
11D41 ; Visarga # Mc MASARAM GONDI SIGN VISARGA

# Indic_Syllabic_Category=Nukta
11D42 ; Nukta # Mn MASARAM GONDI SIGN NUKTA

# Indic_Syllabic_Category=Virama
11D45 ; Virama # Mn MASARAM GONDI VIRAMA

# Indic_Syllabic_Category=Pure_Killer
11D44 ; Pure_Killer # Mn MASARAM GONDI SIGN HALANTA

# Indic_Syllabic_Category=Vowel_Independent
11D00..11D0B ; Vowel_Independent # Lo [10] MASARAM GONDI LETTER A .. AU

# Indic_Syllabic_Category=Vowel_Dependent
[...]
```

‘Masaram’ Gondi: Positional Categories

```
# Indic_Positional_Category=Top
11D31..11D35 ; Top # Mn [5] MASARAM GONDI VOWEL SIGN AA .. UU
11D3A ; Top        # Mn MASARAM GONDI VOWEL SIGN E
11D3C..11D3D ; Top # Mn [2] MASARAM GONDI VOWEL SIGN AI .. O
11D3F ; Top        # Mn MASARAM GONDI VOWEL SIGN AU
11D40 ; Top        # Mn MASARAM GONDI SIGN ANUSVARA
11D41 ; Top        # Mn MASARAM GONDI SIGN VISARGA
11D43 ; Top        # Mn MASARAM GONDI SIGN CANDRA


# Indic_Positional_Category=Bottom
11D36 ; Bottom     # Mn MASARAM GONDI VOWEL SIGN VOCALIC R
11D42 ; Bottom     # Mn MASARAM GONDI SIGN NUKTA
11D44 ; Bottom     # Mn MASARAM GONDI SIGN HALANTA
11D47 ; Bottom     # Mn MASARAM GONDI RA-KARA
```

‘Masaram’ Gondi: Confusables

```
11D31 MASARAM GONDI VOWEL SIGN AA ;  0304 COMBINING MACRON
11D21 MASARAM GONDI LETTER PHA    ;  1109D KAITHI LETTER NNA
11D2A MASARAM GONDI LETTER SSA    ;  0398 GREEK CAPITAL LETTER THETA
11D52 MASARAM GONDI DIGIT TWO     ;  0055 LATIN CAPITAL LETTER U
```

Lifecycle: Formal submission

L2/15-090R
2015-06-02

Proposal to Encode the Masaram Gondi Script in Unicode

Anshuman Pandey
Department of Linguistics
University of California, Berkeley
Berkeley, California, U.S.A.
anshuman.pandey@berkeley.edu

June 2, 2015

1 Introduction

This is a proposal to encode the Gondi script created by Mangal Singh Masaram in 1918 in Unicode. It replaces and supersedes the following documents:

- L2/10-207 “Preliminary Proposal to Encode the Gondi Script in the UCS”
- L2/12-235 “Revised Preliminary Proposal to Encode the Gondi Script”
- L2/15-005 “Proposal to Encode the Gondi Script”

This document provides a description of the writing system, a code chart and names list, character properties,

‘Masaram’ Gondi: Approval and publication

L2/15-090R
2015-06-02

Proposal to Encode the Masaram Gondi Script in Unicode

Anshuman Pandey
Department of Linguistics
Univers
Berk
anshur

1 Introduction

This is a proposal to encode the Gondi sc
replaces and supersedes the following docu

- L2/10-207 “Preliminary Proposal to E
- L2/12-235 “Revised Preliminary Prop
- L2/15-005 “Proposal to Encode the G

This document provides a description of the

File Edit View History Bookmarks Tools Help

Proposed New Characters ...

www.unicode.org/alloc/Pipeline.html

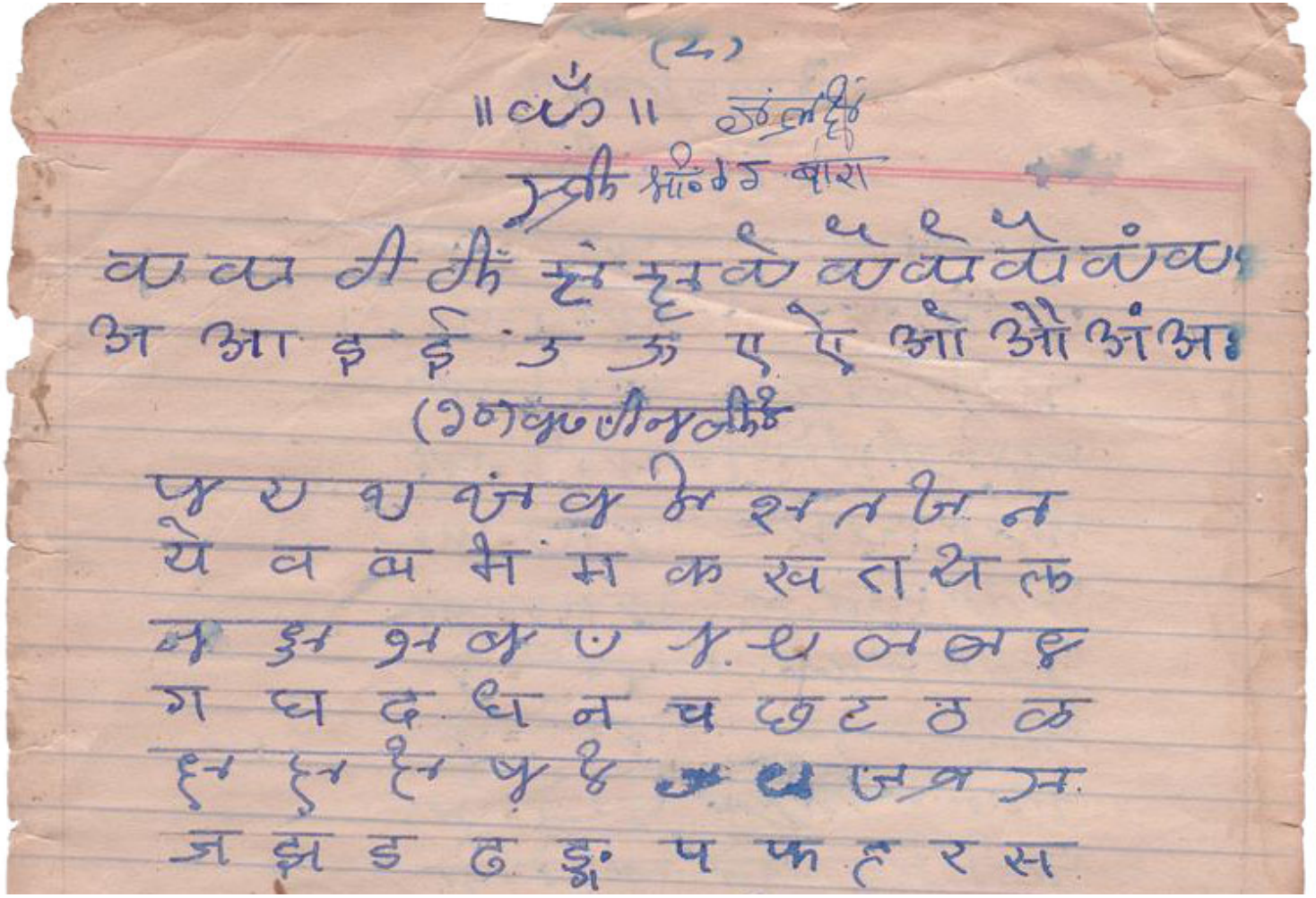
Search

11C70..11CB6	68	Marchen (Marchen block 11C70..11CBF)	2014-Feb-03 Accepted; 2015-Feb-05 One character removed	2015-Jul-06 Stage 6
11D00..11D59	75	Masaram Gondi (Masaram Gondi block: 11D00..11D5F)	2015-May-06 Accepted	2015-Jun-01 Stage 5
11FC0..11FD0	17	Tamil historic fractions (Tamil Supplement block: 11FC0..11FFF)	2013-May-09 Accepted; 2015-Feb-05 Spelling of various names changed	
11FD1..11FEF	31	Tamil historic signs	2013-May-09 Accepted; 2015-Feb-05 Spelling of various names changed	
		TAMIL CURRENT SIGN		

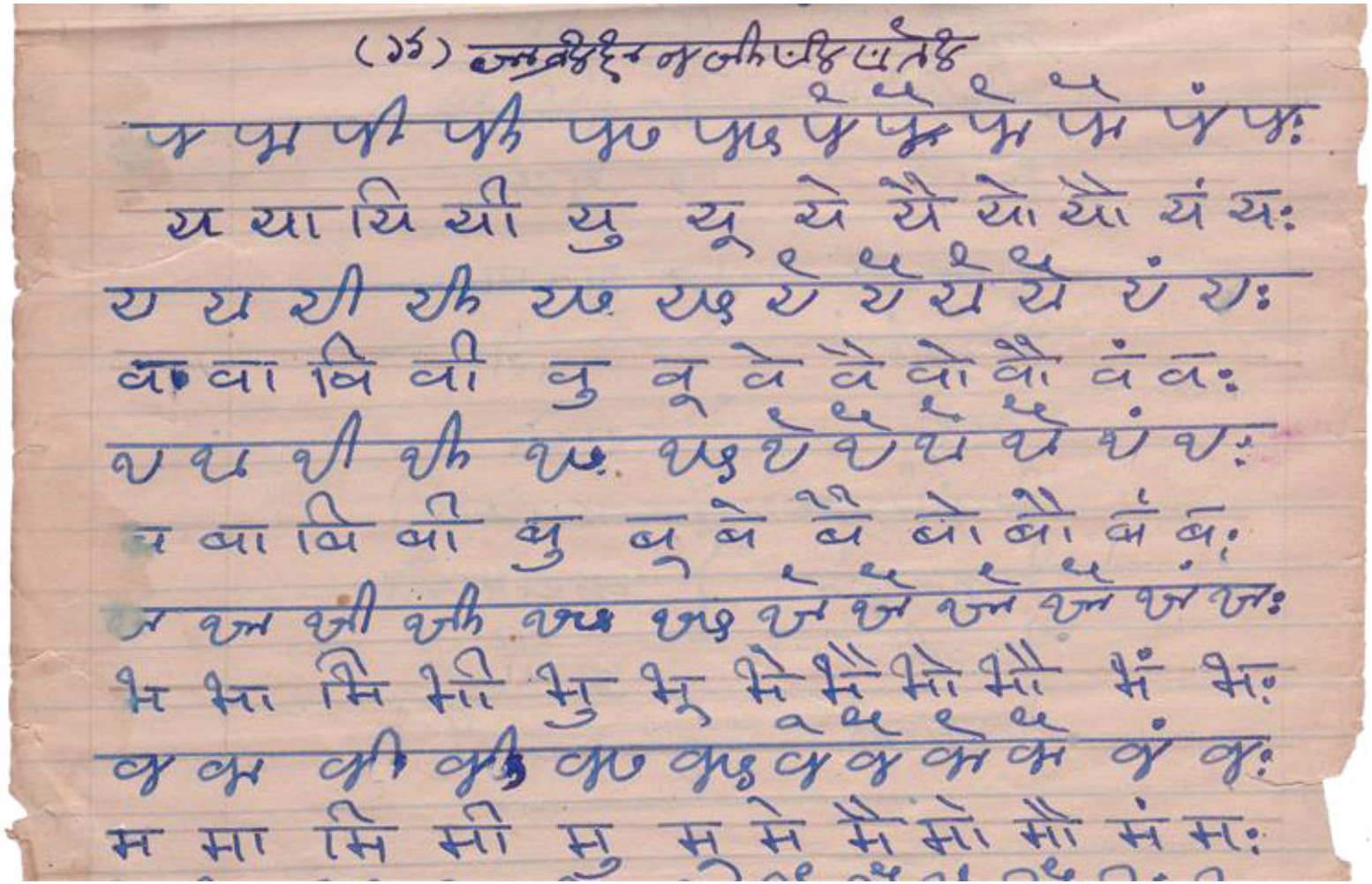
11D00 **Masaram Gondi** **11D5F**

	11D0	11D1	11D2	11D3	11D4	11D5
0	11D00	11D10	11D20	11D30	11D40	11D50
1	11D01	11D11	11D21	11D31	11D41	11D51
2	11D02	11D12	11D22	11D32	11D42	11D52

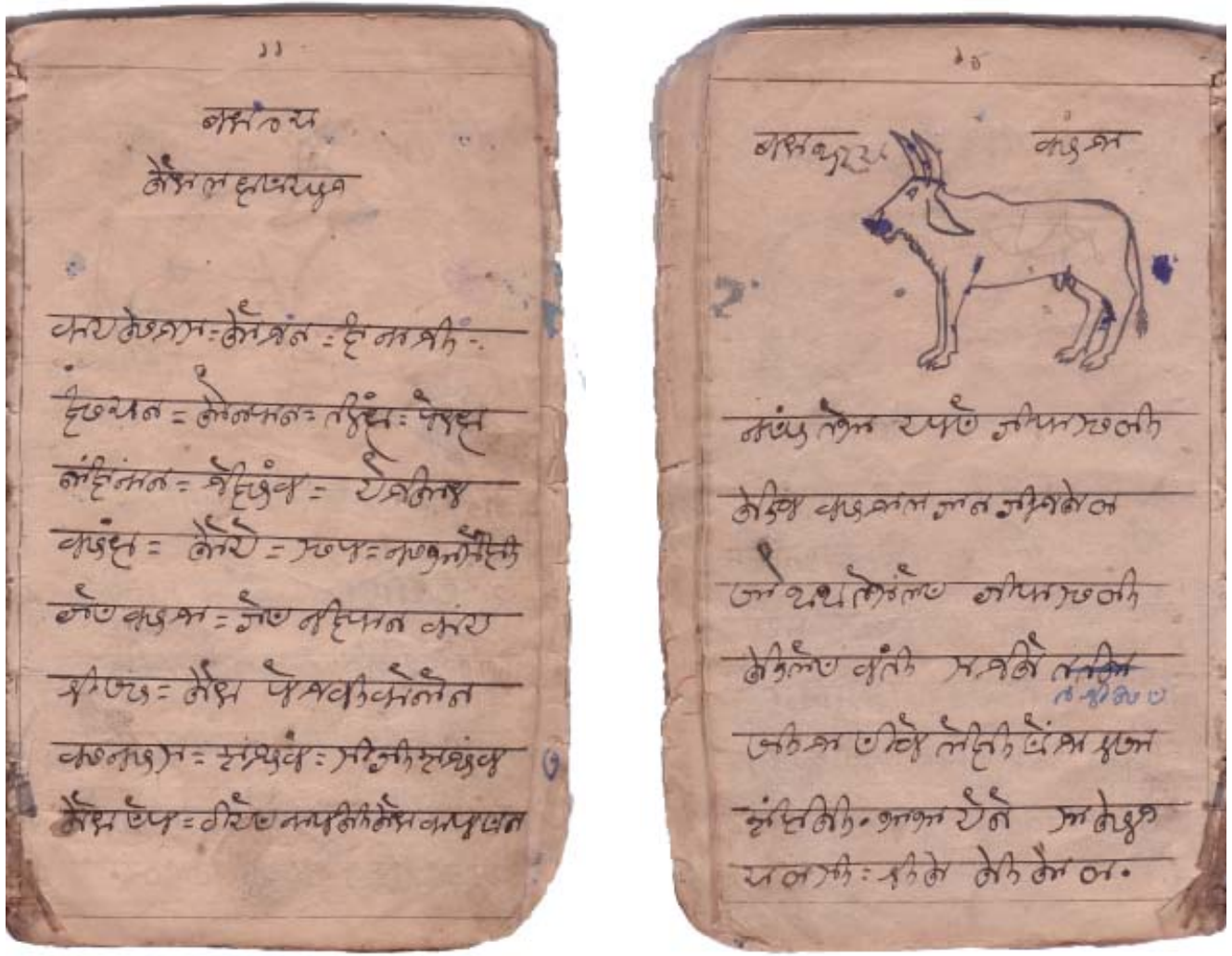
‘Gunjala’ Gondi: Overview



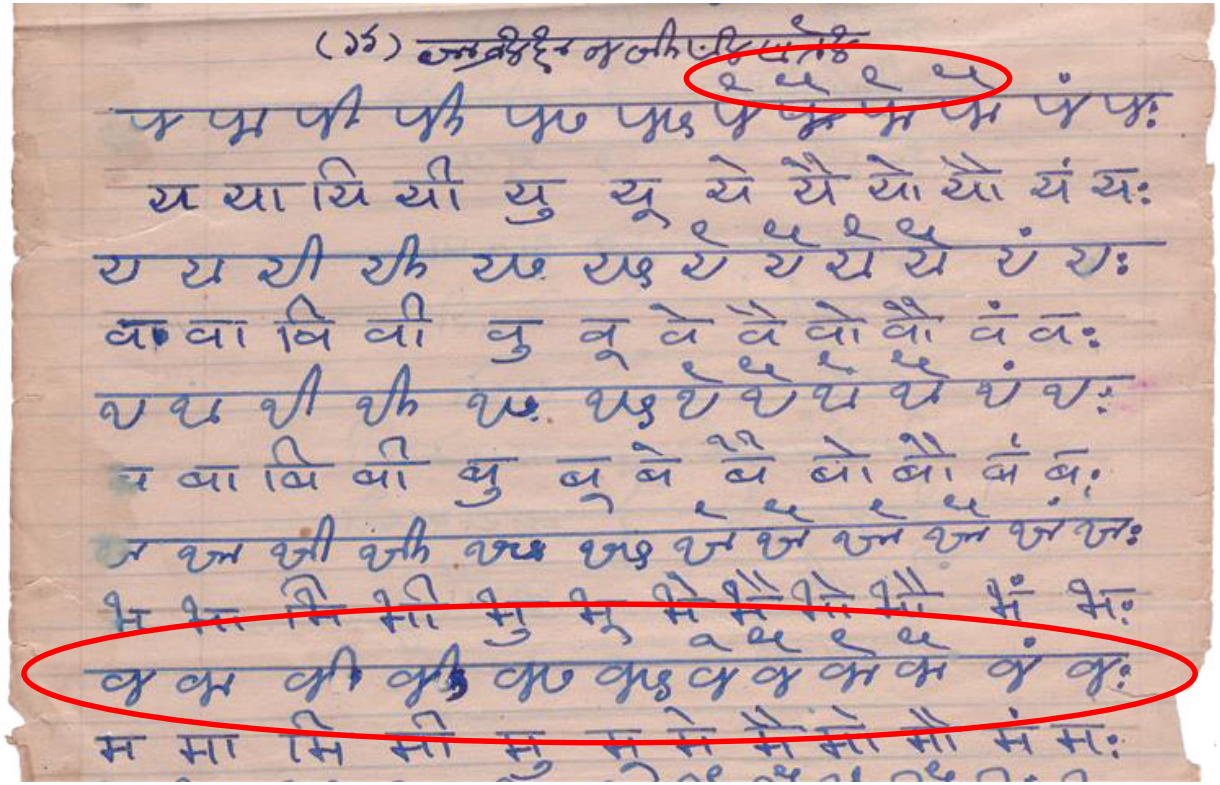
‘Gunjala’ Gondi: Overview

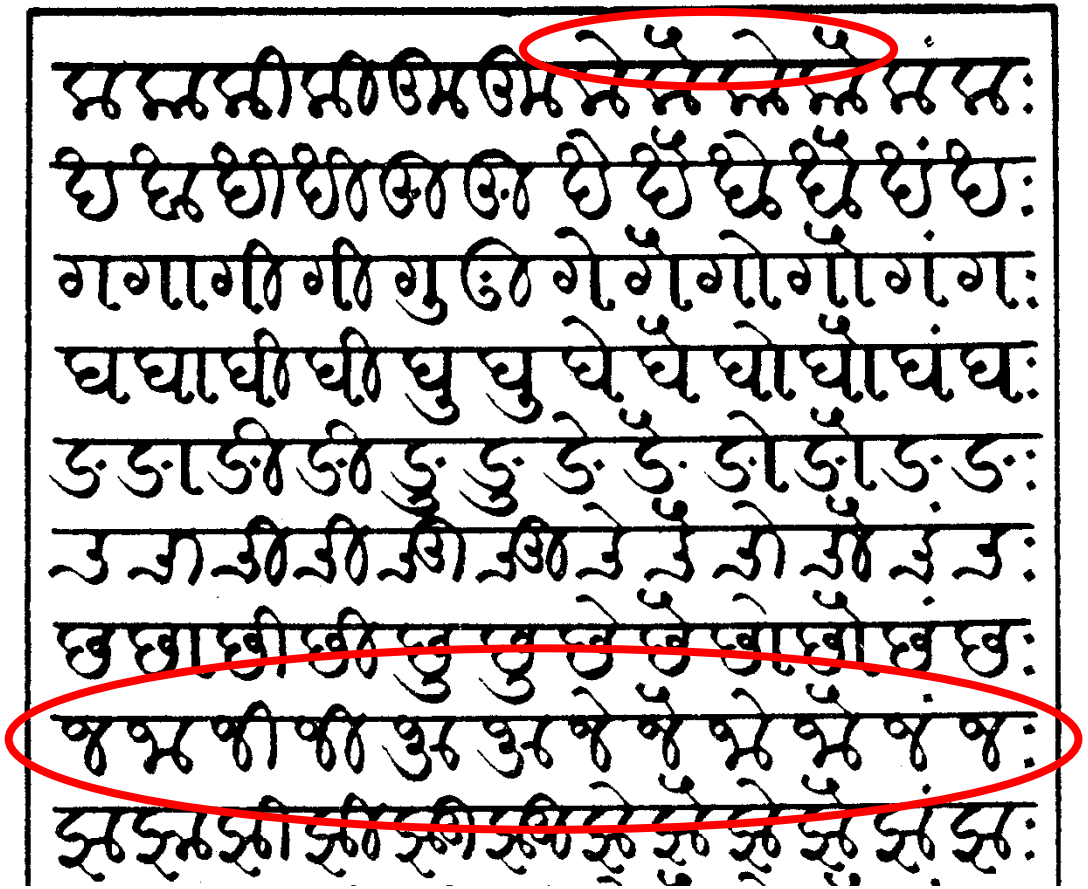


‘Gunjala’ Gondi: Attestations



‘Gunjala Gondi’: Mistaken identity?





‘Gunjala’ Gondi: In the news

“Is the Gunjala Gondi script actually the extinct script as is being made out to be or is the obsolete and archaic Modhi script being passed off as Gondi? A controversy has erupted in the tribal heartland of Adilabad with Adivasi elders and those involved in development of the Gondi language disputing the claim of discovery of the Gondi script.”

‘Gunjala’ Gondi: In the news

Gondi Script Gets a Fresh Leas...

www.thehansindia.com/posts/index/2015-03-07/Gondi-script-gets-a-fresh-lease-of-life-136020

Search

THE HANS INDIA

Prof K Nageshwar

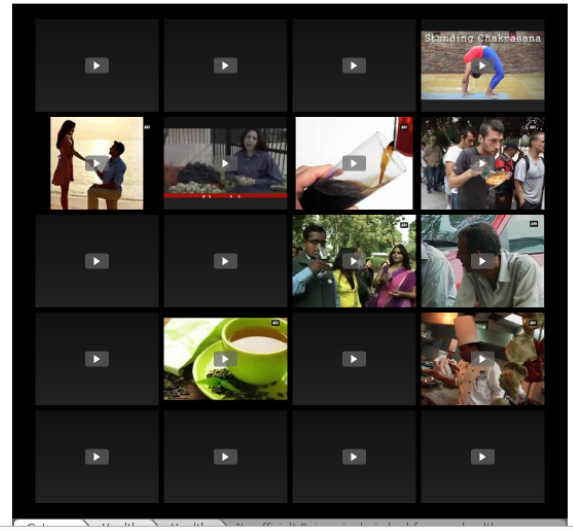
HOME | AP | TELANGANA | INDIA | WORLD | NRI | BUSINESS | SPORTS | CRIME | LIFE STYLE | HANS | VIDEOS

CAREERS

Gondi script gets a fresh lease of life

March 07,2015, 11.22 PM IST | | THE HANS INDIA

Like 2 Tweet 7 Share



As part of the efforts to revive a dying language, it is being taught in schools and a magazine is being launched in the month of April

The Gondi script that was brought to light in 2006 when a joint survey conducted by the National Mission for Manuscripts, Delhi and AP State Oriental Manuscripts Library and Research Institution also threw up interesting facts that there were just four members of Gunjala village of Narnoor mandal who knew to read and write the script. Thanks to the efforts of the Centre for Dalit and Adivasi Studies and Translation (CDAST), University of Hyderabad, the ancient script would soon be spoken as well as written by the present generation.

Fifteen instructors were appointed by Integrated Tribal Development Agency (ITDA), Utnoor, Adilabad, to teach koyatur language in 11 schools under its jurisdiction and another four mandal schools. The script is being taught as an extra period but is not part of the curriculum. CDAST

Script Encoding Lifecycle

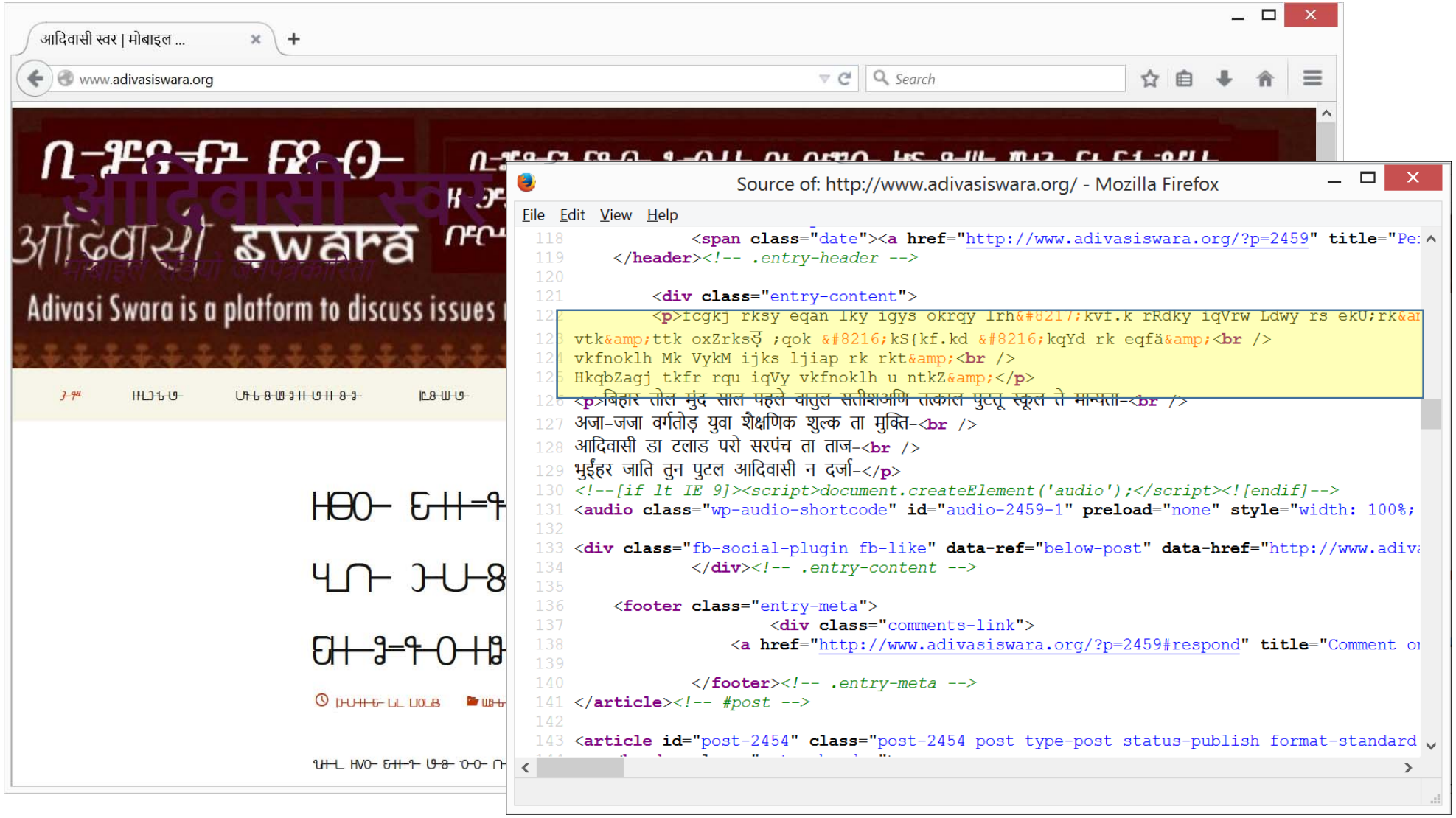
- Identification
- Research and Proposal development
- Proposal review and approval
- Publication
- Implementation
- Repeat

Lifecycle: Implementation

- After long proposal and approval process, a script is in Unicode
  - The actual Unicode encoding is only part of a larger process
- Encoding does not mean a script is ready for use
  - Universal Shaping Engine now makes out of the box support possible
- Development of support is the next step after encoding
  - Input methods need to be designed and developed
  - Proper OpenType fonts need to be designed

‘Masaram’ Gondi: Desired usage

‘Masaram’ Gondi: Desired usage



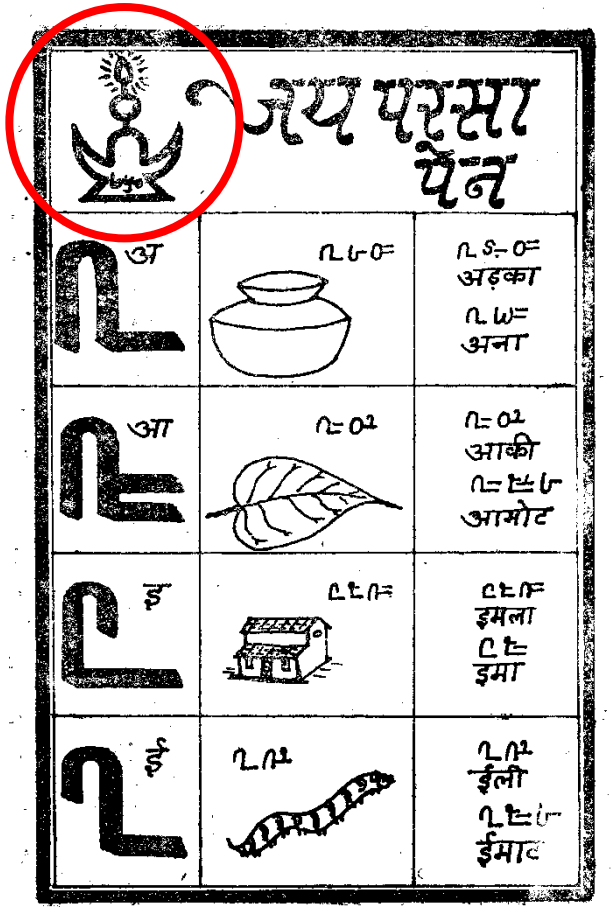
Lifecycle: Repeat

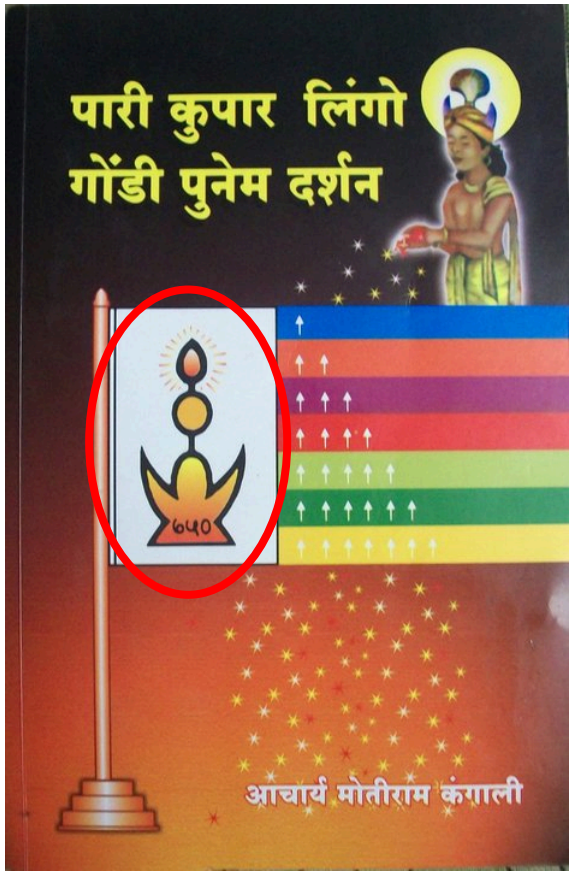
- The Unicode encoding is the first step for enabling language support

- Full support of writing system:
  - archaic characters
  - additional stylistic variants
  - cultural symbols

- Development of locale data:
  - transliteration systems
  - calendar systems
  - date and number formats

Lifecycle: Repeat





Lifecycle: Repeat

L2/15-111
2015-04-14

Proposal to Encode the 'Parsapen' Symbol in Unicode

Anshuman Pandey
Department of Linguistics
University of California, Berkeley
Berkeley, California, U.S.A.
anshuman.pandey@berkeley.edu

April 14, 2015

1 Introduction

This is a proposal to encode the 'Parsapen' symbol in Unicode. The character is proposed for inclusion in the block 'Miscellaneous Symbols and Pictographs' (U+1F300). Basic details of the character are as follows (the actual code point will be determined if the proposal is approved):

glyph	code point	character name
	U+1F9xx	PARSAPEN

2 Description

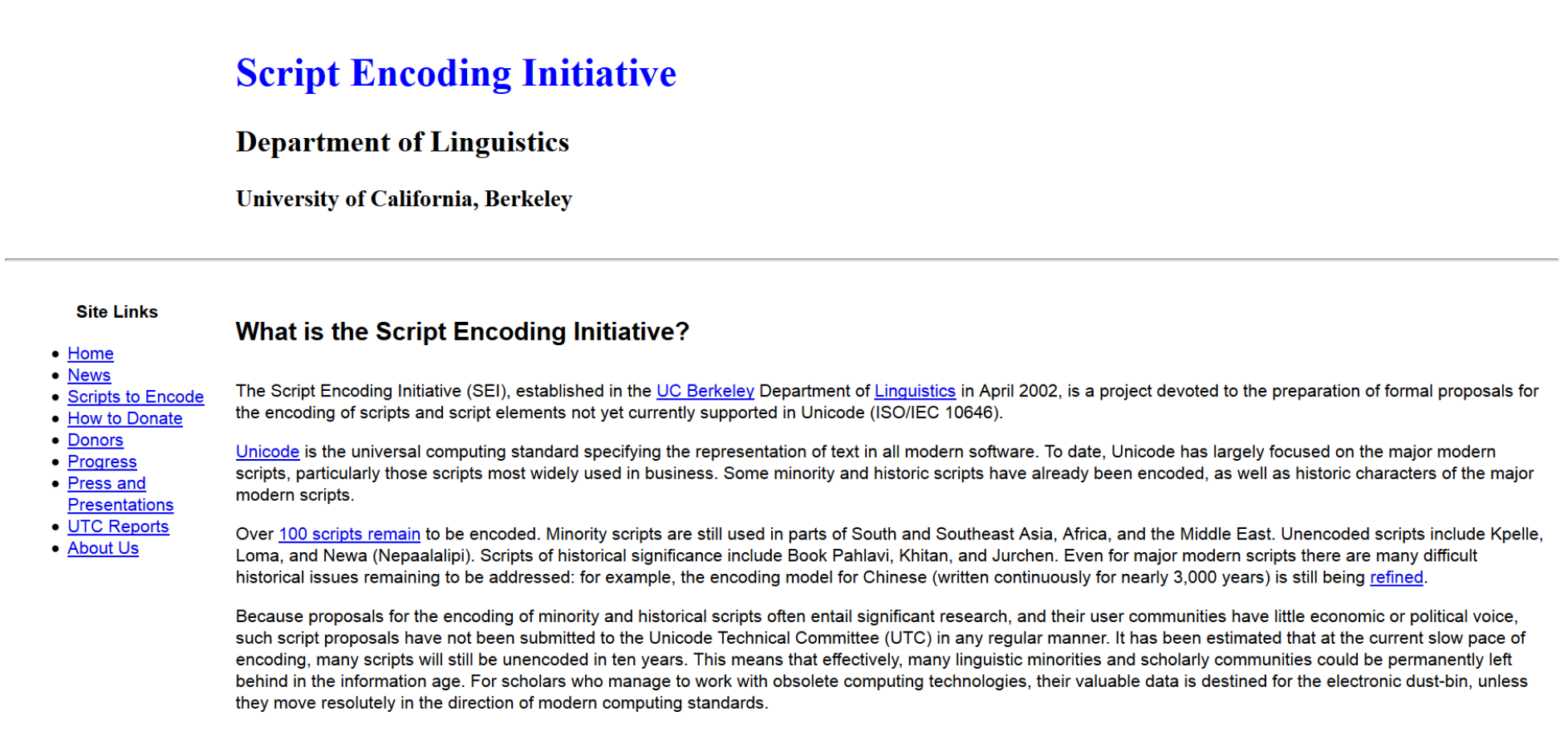
The PARSAPEN symbol encapsulates the fundamental principles of the religious culture of the *ādivāsī* Gond community of India. The indigenous philosophical system of the Gonds is called *koya punem*. The symbol is wrought in statuary form, drawn on religious altars, and printed in books, used upon flags, and on emblems of Gond associations (see the accompanying figures).

The symbol is comprised of three basic components positioned in vertical orientation. The first component is a *triśūla* 'trident' that consists of three elements: 1) a trapezoidal foundation; 2) an upward-facing crescent in profile, which rests upon the foundation; 3) a pillar with curved top, which rests within the crescent. The second and middle component is an orb. The third component is a pointed spheroid. The three are joined together with a vertical strut between the base and the orb, and between the orb and spheroid. The number '750' is often written on the face of the base trapezoid.

1

?

SEI: Webpage



Script Encoding Initiative

Department of Linguistics

University of California, Berkeley

Site Links

- Home
- News
- Scripts to Encode
- How to Donate
- Donors
- Progress
- Press and Presentations
- 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).

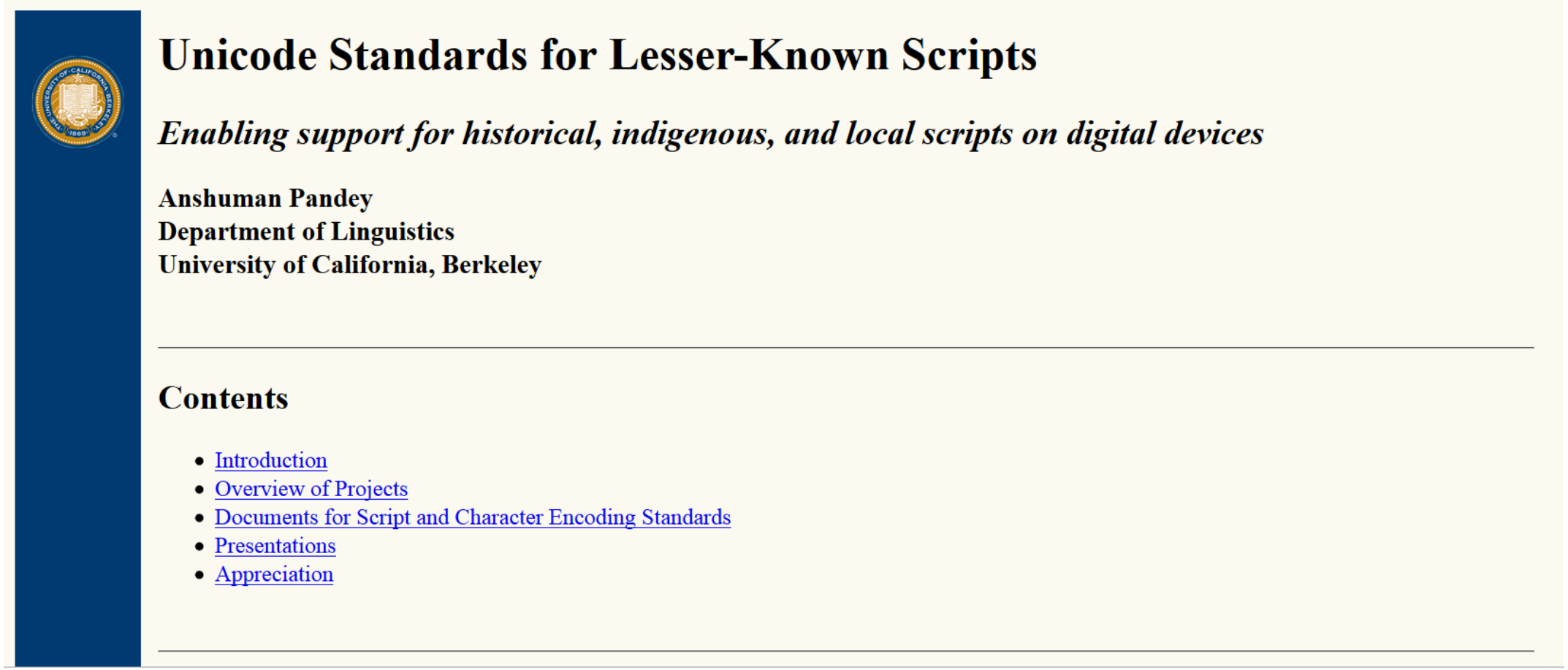
Unicode is the universal computing standard specifying the representation of text in all modern software. To date, Unicode has largely focused on the major modern scripts, particularly those scripts most widely used in business. Some minority and historic scripts have already been encoded, as well as historic characters of the major modern scripts.

Over 100 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, Loma, and Newa (Nepaalalipi). Scripts of historical significance include Book Pahlavi, Khitan, and Jurchen. 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 in ten years. This means that effectively, many linguistic minorities and scholarly communities could be permanently left behind in the information age. For scholars who manage to work with obsolete computing technologies, their valuable data is destined for the electronic dust-bin, unless they move resolutely in the direction of modern computing standards.

linguistics.berkeley.edu/sei

Pandey: Webpage



Unicode Standards for Lesser-Known Scripts

Enabling support for historical, indigenous, and local scripts on digital devices

Anshuman Pandey
Department of Linguistics
University of California, Berkeley

Contents

---

linguistics.berkeley.edu/~pandey

Contact us

Deborah Anderson

dwanders@berkeley.edu

@sei_berkeley

Anshuman Pandey

anshuman.pandey@berkeley.edu

@anshumanpandey_