

Encoding Diversity for Asian and African Languages

The Script Encoding Initiative

Michael Everson, Everttype
Westport, Co. Mayo, Ireland

Geneva, Switzerland • 9 May 2006

Current State of the Unicode Standard

- Unicode 4.1 defines over 97,000 characters
- Unicode covers over 50 scripts (often used for languages with over 5 million speakers)
- Unicode enables millions of users to view web pages, send e-mails, converse in chat-rooms, and share text documents in their native script
- Unicode is widely supported by current fonts and operating systems, but...

Over 80 scripts are missing!

Missing Modern Minority Scripts

India, Nepal, Bangladesh:

- Chakma
- Methei/
Manipuri
- Newari
- Sorang
Sompeng
- Varang Kshiti

Southeast Asia (excluding China):

- Batak
- Cham
- Javanese
- Pahawh
Hmong
- Viet Thai

China:

- Lanna
- Naxi Geba
- Naxi Tomba
- Pollard

Africa:

- Bamum
- Bassa
- Mende

Over 80 scripts are missing!

Missing Modern Minority Scripts

- Ahom
- Alpine
- Aramaic
- Avestan
- Aztec Pictograms
- Balti
- Brahmi
- Büthakukye
- Byblos
- Chalukya
- Chola
- Cypro-Minoan
- Egyptian Hieroglyphs
- Elbasan
- Elymaic
- Grantha
- Hatran
- Iberian
- Indus Valley
- Jurchin
- Kaithi
- Kawi
- Khotanese
- Kitan Large Script
- Kitan Small Script
- Landa
- Linear A
- Luwian
- Mandaic
- Manichaean
- Mayan Hieroglyphs
- Meroitic
- Modi
- Nabataean
- North Arabic
- Numidian
- Old Hungarian
- Old Permic
- Orkhon
- Pahlavi
- Palmyrene
- Proto-Elamite
- Pyu
- Rongorongo
- Samaritan
- Satavahana
- Sharada
- Siddham
- South Arabian
- Soyombo
- Takri
- Tangut Ideograms
- Uighur
- Vedic accents

Current State of the Unicode Standard: New Script Additions

For Unicode 5.0 (2006):

N’Ko (*West Africa*)

Balinese (*Indonesia*)

Phags-pa (*historical*)

Phoenician (*historical*)

Cuneiform (*historical*)

For Unicode 5.1 (2008):

Lepcha (*India*)

Ol Chiki (*India*)

Vai (*Liberia*)

Saurashtra (*India*)

Myanmar minorities (*Myanmar*)

Kayah Li (*Myanmar*)

Rejang (*Indonesia*)

Sundanese (*Indonesia*)

Carian, Lycian, Lydian (*historical*)

Three Case Studies (Modern Scripts)

ကုမ္ပဏီအဖွဲ့အစည်း၏ ပစ္စည်းအသုံးပြုမှုကို

အသုံးပြုမှုကို အသုံးပြုမှုကို

အသုံးပြုမှုကို အသုံးပြုမှုကို

Case 1: Balinese

- Used for the Balinese language, an Austronesian language with 3.8 million speakers
- Used in many traditional literary and cultural works (ritual choruses, dramatic recitations)
- Considered by some Balinese to be “endangered”
- Taught in primary and secondary schools as a mandatory subject, about 2 hours a week
- Encoded as of Unicode 5.0 – thanks to support from UNESCO

Street signs in Bali



Building signs in Bali



Case 2: The Vai script of Liberia



- Some 105,000 Vai people live mainly in Liberia
- The Vai script was invented ca. 1833. In 1962 a “Standard Vai Syllabary” was published in Monrovia
 - Schoolbooks are available in Vai script
 - Work to encode Vai began in April 2005; it is under ballot and will appear in Unicode 5.1 in 2008

Case 3: N'Ko

- Used to write a number of Manden languages, comprising 18 to 20 million speakers

- Used in Côte d'Ivoire, the Gambia, Guinea, Liberia, Mali, Senegal, and Sierra Leone



𞤎𞤵𞤲 𞤵𞤴𞤲 𞤵𞤴𞤲 𞤵𞤴𞤲

N'Ko

- Devised in the late 1940s by Solomana Kante of Guinea to be used for the Manden languages of West Africa

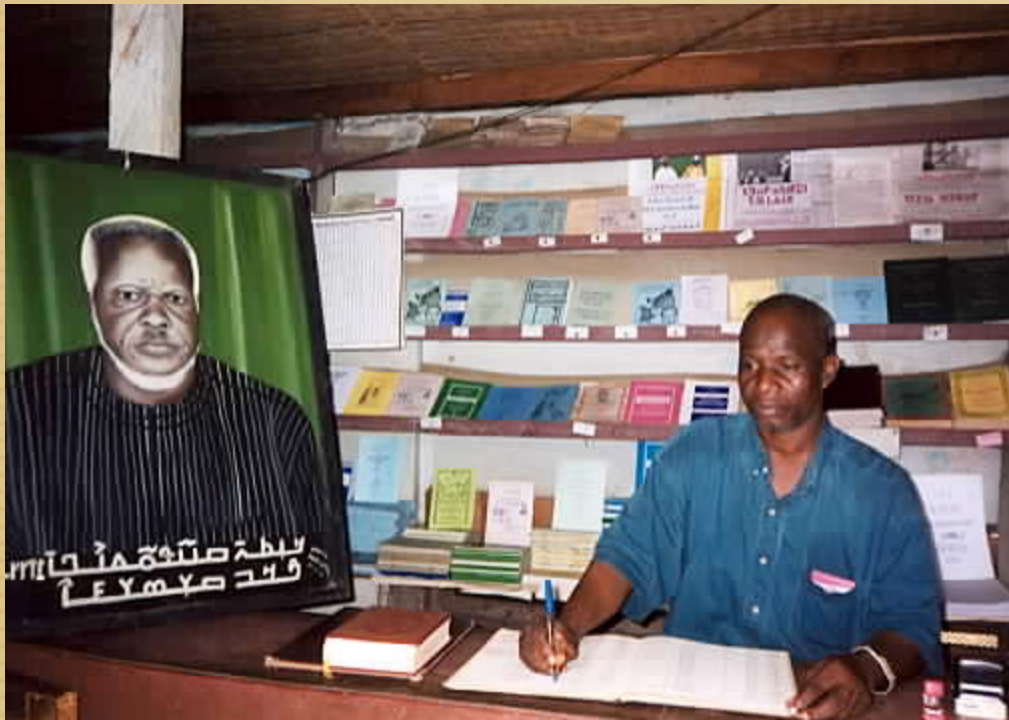
- Has a vigorous and active user community

- Encoded as of Unicode 5.0 – thanks to support from UNESCO



Solomana Kante

Bookstore in Guinea



N'Ko school in Kankan, Guinea



Stage II: N'Ko font design

- Requires complex shaping behaviour
- Requires precise diacritic placement
- Development supported by UNESCO



The image displays three examples of N'Ko script characters, each featuring a diacritic mark. The first character has a tilde (~) above a circle, with a vertical stem and a dot below. The second character has a tilde (~) above a circle, with a vertical stem and a loop to the right. The third character has a tilde (~) above a circle, with a vertical stem and a loop to the right, and a horizontal line extending to the right.

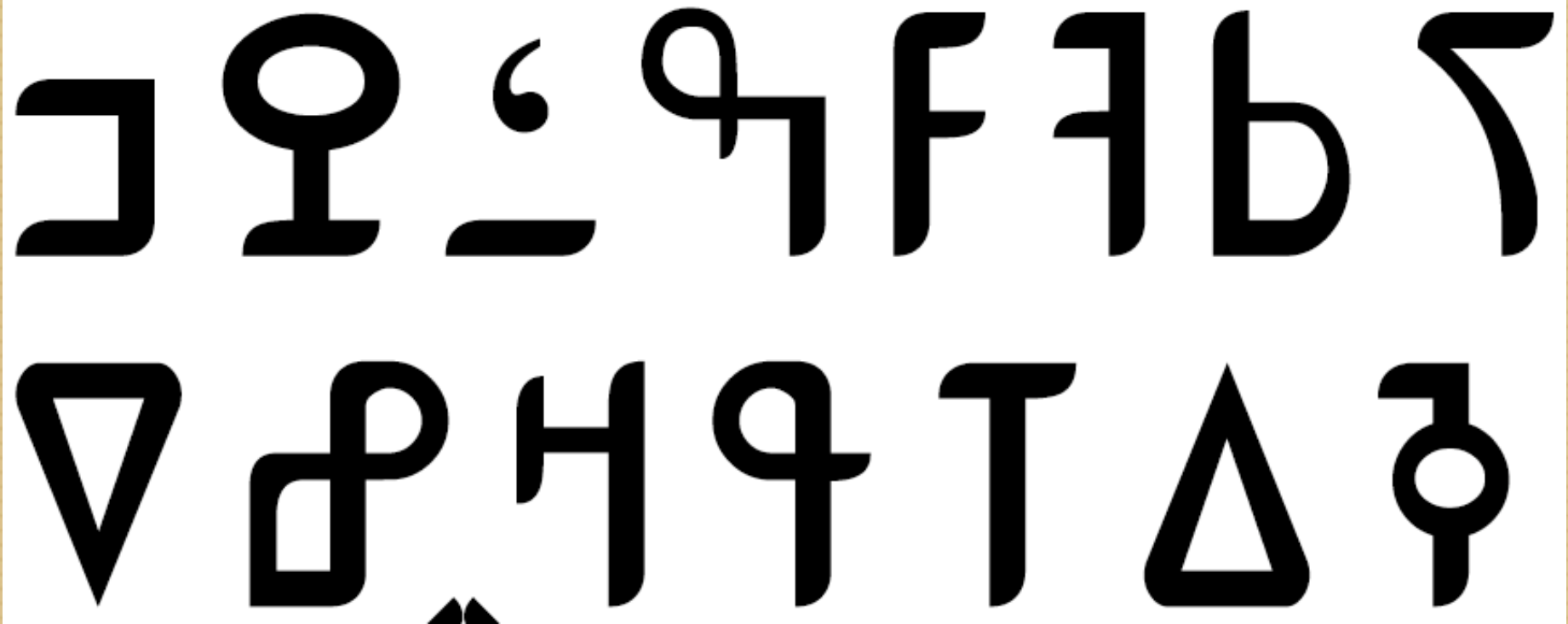
Stage II: N'Ko font design

- Unrectified glyph shapes



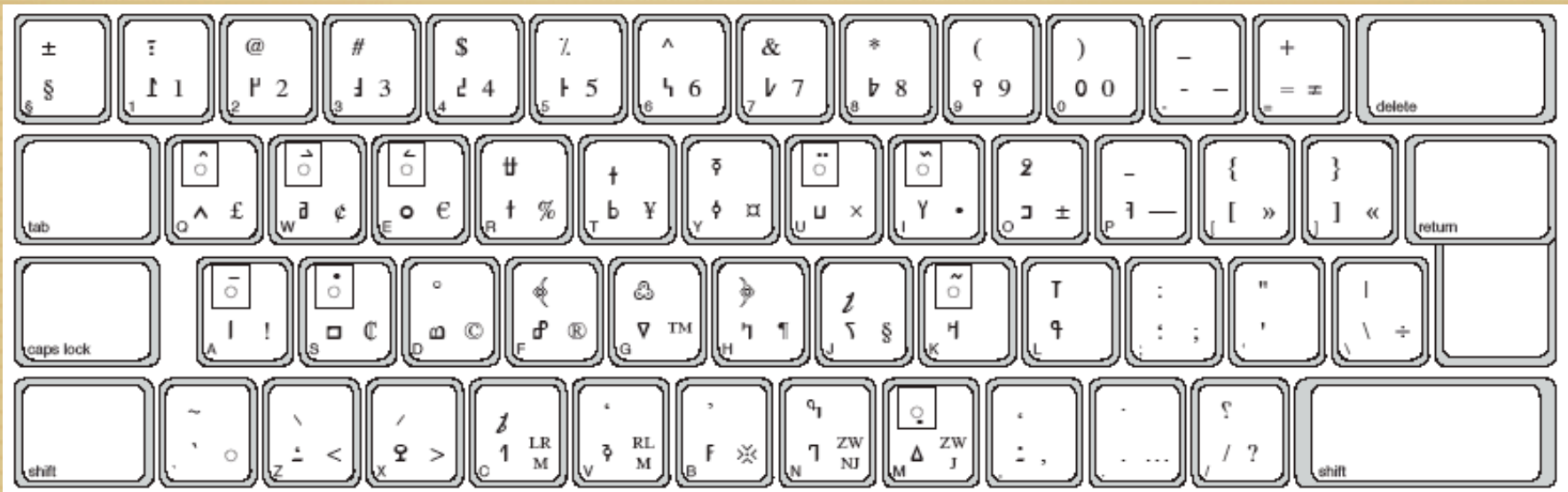
Stage II: N'Ko font design

- Rectified glyph shapes
- Development supported by UNESCO



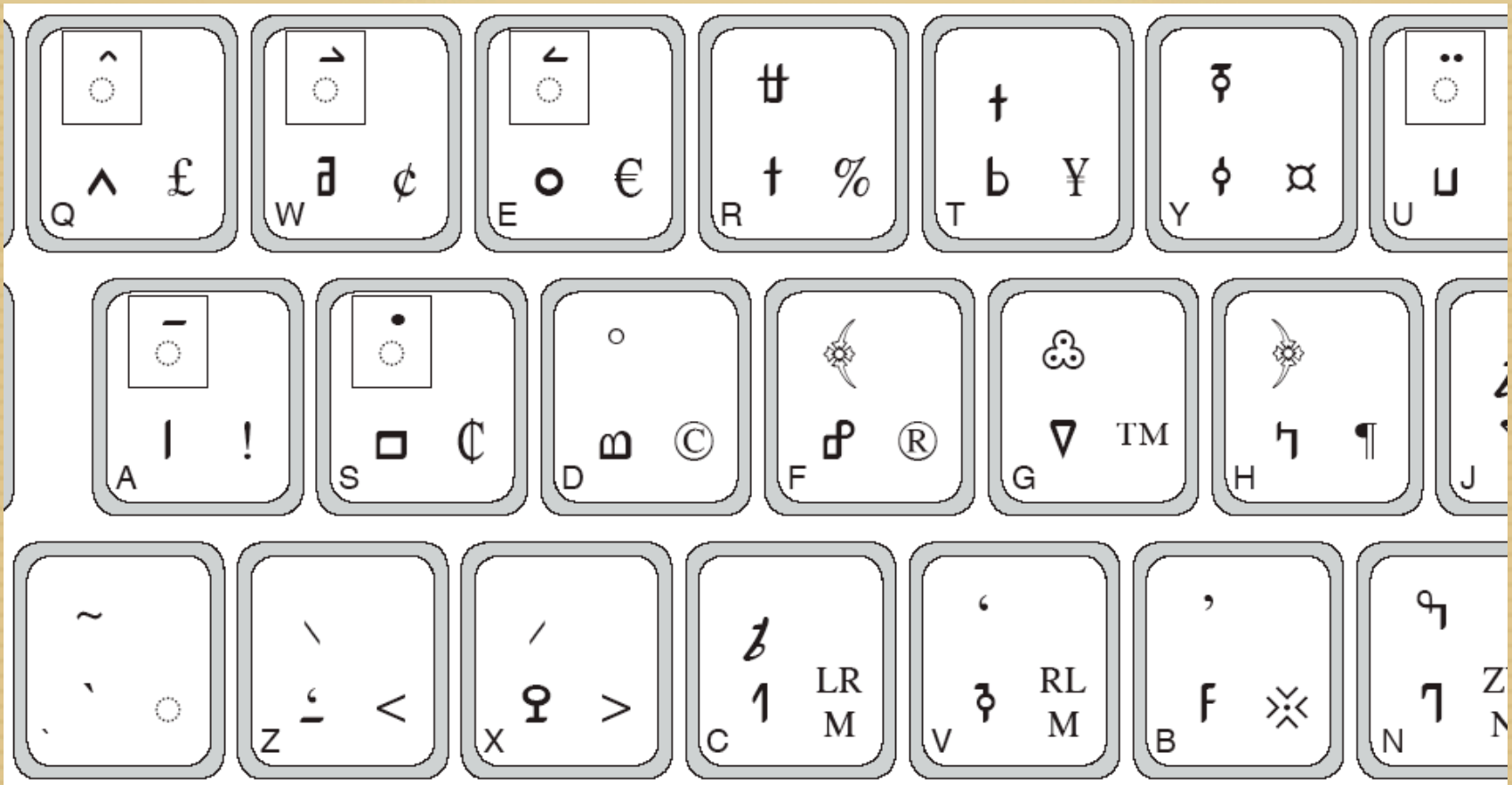
N'Ko keyboard design

- Three keyboard layouts designed
- Ergonomic, QWERTY, AZERTY



N'Ko keyboard design

- Primary support: N'Ko letters and bidi characters
- Secondary support: Generic symbols and ASCII



Work is by no means complete!

Missing Modern Minority Scripts

India, Nepal, Bangladesh:

- Chakma
- Methei/
Manipuri
- Newari
- Sorang
- Sompeng
- Varang Kshiti

Southeast Asia (excluding China):

- Batak
- Cham
- Javanese
- Pahawh
Hmong
- Viet Thai

China:

- Lanna
- Naxi Geba
- Naxi Tomba
- Pollard

Africa:

- Bamum
- Bassa
- Mende

Why are scripts missing?

- Modern script users are minorities and tend to be less affluent than users of majority scripts
- It is difficult for users to attend international standardization meetings
- There is no large consumer base and so minority scripts are not of much interest to corporations
- Governments may be reluctant to help certain minority groups

Why are scripts missing?

- Historic scripts have no “large consumer base”
- It is difficult to get universities to support because they do not understand the problem fully
- Scripts (both historic and modern) are often less well-known
- Additional research is needed to encode such scripts properly

Why are scripts missing?

- In the past, most work has been done by volunteers; proposals have appeared sporadically
- As more scripts are encoded, the focus is increasingly on implementation, maintenance of the standard, and locale data collection, with less of a focus on encoding. This leaves behind those groups whose script is not yet in Unicode, because they don't represent an economically viable market for computer companies.
 - Unicode already covers all of the “economically interesting” scripts for computer companies. Some people have estimated that within 5 years all of the implementation for the major scripts and locale data will have been collected.

Why are scripts missing?

- While many of the computer companies will continue to be involved in maintaining the standard, they will not be active in encoding new scripts — and it will become more difficult to pass new encodings through committees.
 - The Unicode Consortium will still be active: it enjoys tremendous support in the industry and national bodies, and is expanding. Maintenance of the standard represents a long-term commitment by all of those involved.
- But in order to get the remaining scripts encoded, users of those scripts need to participate and help fund the project.
 - Computer companies are not going to do that.

What it means when a script is encoded

- Ethnic pride and identity is promoted
 - Literacy efforts can be encouraged
- The study of historic scripts is kept alive
 - Communication between and amongst members of the community is promoted

6. What it means when a script is encoded

- Encoding allows communication in times of emergency (disease, war, natural disaster) with people throughout the world

- Encoding a script can permit the creation of health materials in local languages

 back
 home

PREVENTING RICKETS IN BREASTFED BABIES (TIGRINYA VERSION)

August 1995


ንዑ-ላድክን ጡብ ተጥብዎ ኣዴታት ኣስተብህላ

ንዘጠብው ዑ-ላድክም ካብ ሪኬትስ ተኸላኸሉ

ሪኬትስ እንታይ ኢዩ? ሪኬትስ ድኹም ኣዕጽምቲ ዘበዕብ ሕጻን ኢዩ። መጥጥሊኡ ጠንቁ ድግ ሕጽሪት ናይ ቪታሚን ዲ ኣብ ኣካላት ኢዩ። ኣብ ሰጊናዊ ፖሶሬብ ኣሚሪካ እዚ ሕጻን መብዛሕትኡ ጊዜ ብሕጽሪት ናይ ጸሓይ መቐት ይመጽእ።

መን ኢዩ በዚ ሕጻን ዝጥቃዕ? ኩሎም ቆልዑ ናይ ኣዴኦም ጡብ ጥራይ ካብ ሹዱሽተ ወርሒ ንላዕሊ ዝጠብዉ፣ ኩሎም ጸላምቲ ቆልዑ ጡብ ኣዴኦም ጥርይ ዝጥቀሙ፣ ብዘይመጻልቲም ዝተወልዱ ህጻውንቲ ከምኡውን ኣታኸልቲ ዘይበልፁ ቆልዑ ዝኣመሰሉ ኢዮም።

ሪኬትስ ከመይ ዝኣመሰሉ ፖልክታት ኣብ ኣካላት የርኢ? ድኹም ኣካላት፣ ጠጠው ፖባል ወይ ፖሰንም ይኸልእ፣ ፖድንጓይ ዕብዮት ኣካላት፣ ሃንደበታዊ ላካል ዘይምቁጽጻር፣ ቀጨውጫው የብል።


ፖልክታት ሪኬትስ

What it means when a script is NOT encoded

- Communication for those whose script is outside Unicode will be difficult
- Implementations for unencoded scripts will be more costly to make interoperable with major platforms and software
- Knowledge of the various scripts of the world will be incomplete

Solution: The Script Encoding Initiative

Script Encoding Initiative

Department of Linguistics
U C Berkeley

Home Page

Welcome to the home of the Script Encoding Initiative

Contents of this Page:

[News](#)

[What is the Script Encoding Initiative?](#)

[List of Scripts Needing Encoding](#)

[How to Help](#)

[Who We Are](#)

[How To Contact Us](#)

<http://linguistics.berkeley.edu/sei>

The role of The Script Encoding Initiative

- To work with users on script proposals
- If needed — it is always needed — to raise money for script proposals to be written and free fonts to be created
- To work collaboratively with other groups (such as SIL) to ensure there is no duplication of effort
- To seek experts to review proposals

The role of The Script Encoding Initiative

- To participate at standards meetings on behalf of minority groups and scholars
- To explain the role and importance of Unicode to scholars, to users, and to the general public

The Script Encoding Initiative: Current Progress

- SEI has helped approximately 12 scripts through the standards process — so far
- Work continues actively on 8 scripts, and over 47 await review and expert input
- SEI has received funds from UNESCO's Initiative B@bel (for three projects) and the U.S. National Endowment for the Humanities — funding runs out at the end of 2006

The Script Encoding Initiative: Plans (and hopes) for the future

- To get adequate, stable funding
- To continue to work on proposals
- To promote the need to encode the missing scripts into Unicode

Conclusion

Finishing the job of encoding the world's minority scripts and historic scripts is a task that will build the infrastructure for world-wide computerization and literacy

We need assistance
from government,
from NGOs,
from UN organizations,
and from the private sector
in order to be able to accomplish this

Script Encoding Initiative:

<http://linguistics.berkeley.edu/sei>

Evertype website:

<http://www.evertype.com>

Unicode website:

<http://www.unicode.org>