Script Encoding Initiative: Past and Future

Unicode Technical Workshop
November 8, 2023

Debbie Anderson
Anushah Hossain

Debbie Anderson
Researcher, Department of Linguistics
University of California, Berkeley

Anushah Hossain
Postdoctoral Scholar, Digital Civil Society Lab
Stanford University

What is the Script Encoding Initiative?

A project hosted by UC Berkeley’s Linguistics department that helps prepare proposals of historic and minority scripts for inclusion in the Unicode Standard and ISO 10646

What is the Script Encoding Initiative?

1. Has held a unique position in the Unicode Consortium as an academic member for 19 years
2. Persists despite opposing incentive structures
3. Extremely successful! Over 100 scripts already encoded through SEI’s help and many more in the pipeline

SEI’s history and unique position

Unicode’s early priorities

Wide coverage

Consistency and efficiency in a limited codespace

Buy-in from companies and governments

Unicode’s early priorities

Wide coverage

Consistency and efficiency in a limited codespace

Buy-in from companies and governments

Round-trip backwards compatibility

Characters, not glyphs

Stability policy

Unicode's early priorities

What is needed is a new international/multilingual text encoding standard that is as workable and reliable as ASCII, but that covers all the scripts of the world.

For reference, the table below ranks the world's writing systems roughly in order of commercial importance, as measured by the total GNP of countries using each system:

Rank	Writing System	Languages	% of World GNP
1	Latin	English, German, French, Spanish, Italian, Portuguese, Indonesian/Malay, ...	68
2	CJK ideographs	Chinese, Japanese, (Korean)	14
3	Cyrillic	Russian, Ukrainian, ...	14
4	Arabic	Arabic, Persian, ...	3
5	Devanāgarī family	Hindi, Bengali, Punjabi, Marathi, ...	1
6	Korean (Hangul)	Korean	1
7	Dravidian family	Telugu, Tamil, ...	ε
8	Greek	Greek	ε
9	Khmer	Thai, Lao, Khmer	ε
10	Hebrew	Hebrew	ε

Excerpt from "Unicode 88"

Unicode's early priorities

Distinction of "modern-use" characters: Unicode gives higher priority to ensuring utility for the future than to preserving past antiquities. Unicode aims in the first instance at the characters published in modern text (e.g. in the union of all newspapers and magazines printed in the world in 1988), whose number is undoubtedly far below $2^{14} = 16{,}384$. Beyond those modern-use characters, all others may be defined to be obsolete or rare; these are better candidates for private-use registration than for congesting the public list of generally-useful Unicodes.

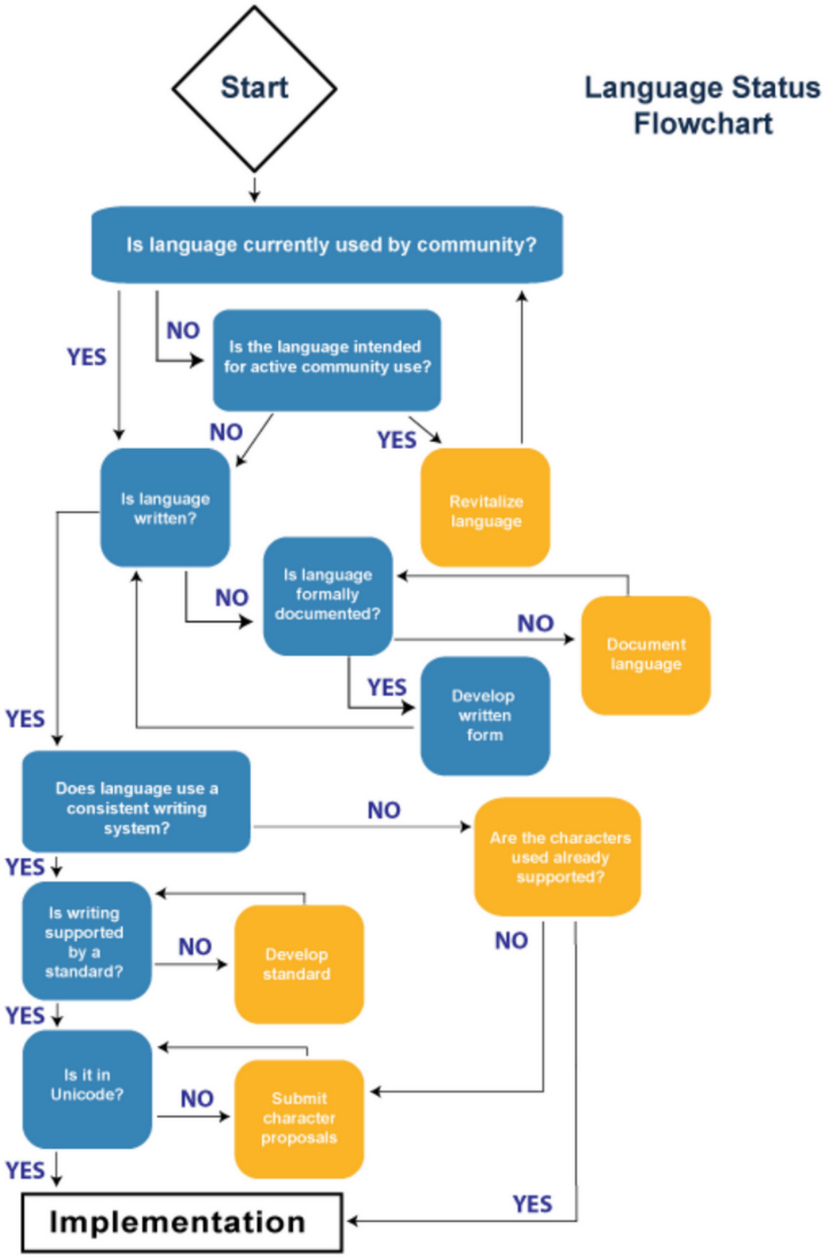
In other words, given that the limitation to 65,536 character codes genuinely does satisfy all the world's modern communication needs with a safety factor of about four, then one can decide up-front that preserving a pure 16-bit architecture has a higher design priority than publicly encoding every extinct or obscure character form. Then the sufficiency of 16 bits for the writing technology of the future becomes a matter of our active intention, rather than passive victimization by writing systems of the past.

Excerpt from "Unicode 88"

Script encoding process

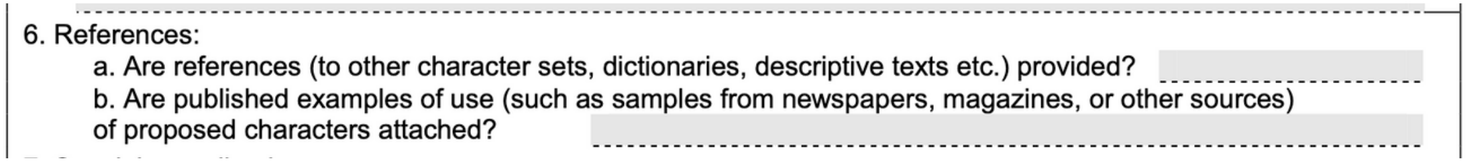
Script encoding process

1. Ensure writing system is in active use (modern scripts) and well-documented (modern and historic scripts)



Script encoding process

2. Gather evidence to to submit alongside character proposal


6. References:
a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?

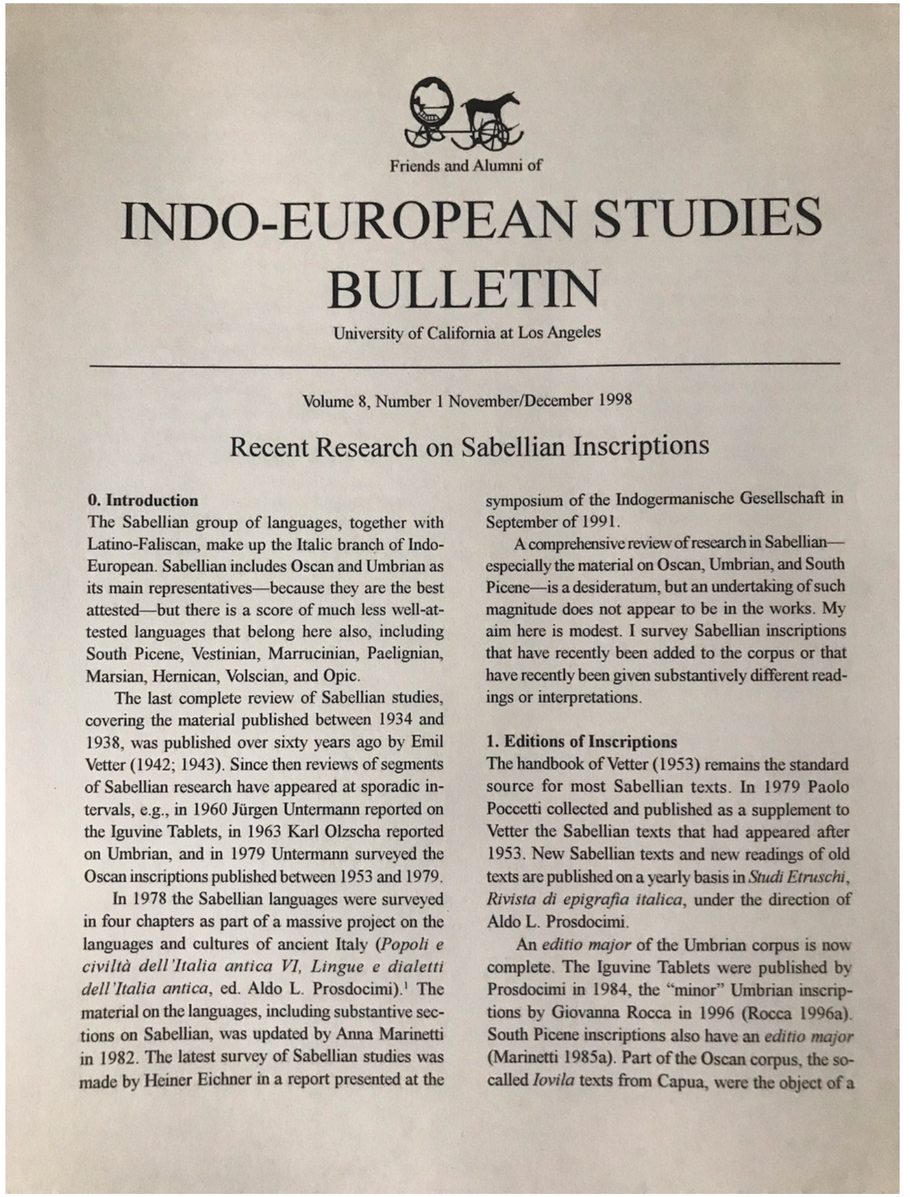
Script encoding process

Adlam meeting, 2014

Mongolian meeting, 2017

How SEI got started

"Scratching your own itch"



Friends and Alumni of

INDO-EUROPEAN STUDIES BULLETIN

University of California at Los Angeles

Volume 8, Number 1 November/December 1998

Recent Research on Sabellian Inscriptions

0. Introduction

The Sabellian group of languages, together with Latino-Faliscan, make up the Italic branch of Indo-European. Sabellian includes Oscan and Umbrian as its main representatives—because they are the best attested—but there is a score of much less well-attested languages that belong here also, including South Picene, Vestinian, Marrucinian, Paelignian, Marsian, Hernican, Volscian, and Opic.

The last complete review of Sabellian studies, covering the material published between 1934 and 1938, was published over sixty years ago by Emil Vetter (1942; 1943). Since then reviews of segments of Sabellian research have appeared at sporadic intervals, e.g., in 1960 Jürgen Untermann reported on the Iguvine Tablets, in 1963 Karl Olzscha reported on Umbrian, and in 1979 Untermann surveyed the Oscan inscriptions published between 1953 and 1979.

In 1978 the Sabellian languages were surveyed in four chapters as part of a massive project on the languages and cultures of ancient Italy (*Popoli e civiltà dell'Italia antica VI, Lingue e dialetti dell'Italia antica*, ed. Aldo L. Prosdocimi).[1] The material on the languages, including substantive sections on Sabellian, was updated by Anna Marinetti in 1982. The latest survey of Sabellian studies was made by Heiner Eichner in a report presented at the symposium of the Indogermanische Gesellschaft in September of 1991.

A comprehensive review of research in Sabellian—especially the material on Oscan, Umbrian, and South Picene—is a desideratum, but an undertaking of such magnitude does not appear to be in the works. My aim here is modest. I survey Sabellian inscriptions that have recently been added to the corpus or that have recently been given substantively different readings or interpretations.

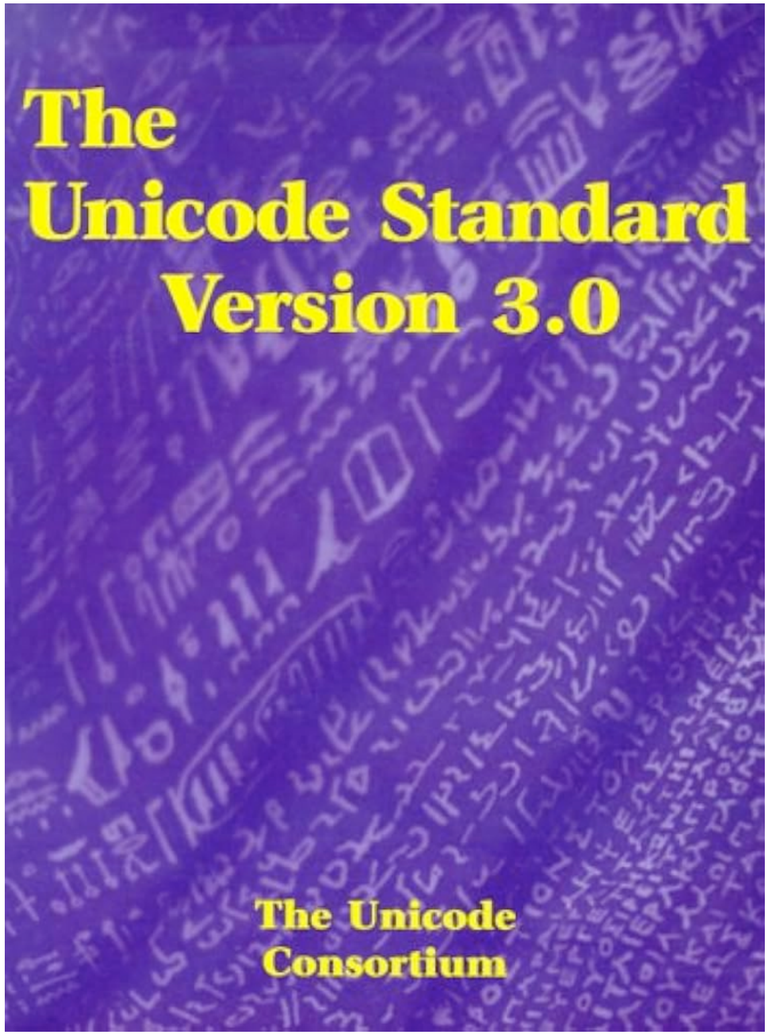
1. Editions of Inscriptions

The handbook of Vetter (1953) remains the standard source for most Sabellian texts. In 1979 Paolo Poccetti collected and published as a supplement to Vetter the Sabellian texts that had appeared after 1953. New Sabellian texts and new readings of old texts are published on a yearly basis in *Studi Etruschi*, *Rivista di epigrafia italica*, under the direction of Aldo L. Prosdocimi.

An *editio major* of the Umbrian corpus is now complete. The Iguvine Tablets were published by Prosdocimi in 1984, the "minor" Umbrian inscriptions by Giovanna Rocca in 1996 (Rocca 1996a). South Picene inscriptions also have an *editio major* (Marinetti 1985a). Part of the Oscan corpus, the so-called *Iovila* texts from Capua, were the object of a

How SEI got started

Some choice meetings...



How SEI got started

...leading to regular attendance at UTC meetings and eventually membership in the Unicode Consortium

How SEI got started

Expansion from historic
Indo-European scripts to modern
minority langauges

Old Italic

Aegean scripts (Linear B, Cypriot Syllabary)

Ol Chiki

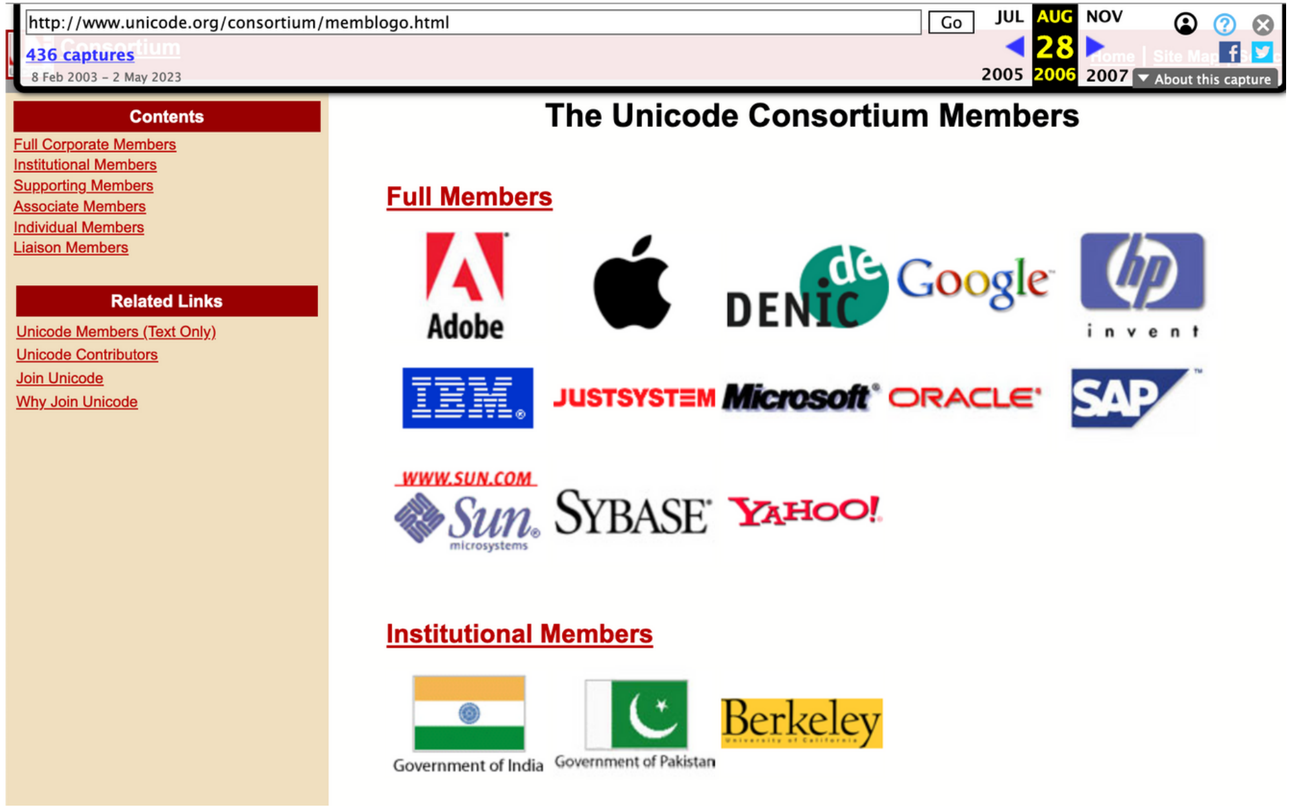
Balinese

N’Ko

...

How SEI got started

SEI (as UC Berkeley) has been the only university on Unicode Consortium roster since it joined in 2004



Managing opposing structures

Finding funding

Private donors, UNESCO, NEH, Google grant

Finding funding Private donors, UNESCO, NEH, Google grant

Getting university support for Unicode membership

Eventually paid from grants

Finding funding Private donors, UNESCO, NEH, Google grant

Getting university support for Unicode membership

Eventually paid from grants

Convincing scholars about importance of Unicode

Shifted slowly as Unicode (and the Internet) took off

Finding funding Private donors, UNESCO, NEH, Google grant

Getting university support for Unicode membership

Eventually paid from grants

Convincing scholars about importance of Unicode

Shifted slowly as Unicode (and the Internet) took off

Convincing modern language communities of Unicode's utility

Relayed effectively through the success of ADLaM and N'Ko

Doing lots with little; ongoing challenges

Finding and supporting experts and community members can be difficult

- Conferences and ISO meetings helpful
- Need to stay informed on ~40 scripts at a time that are somewhere in the pipeline
  - More examples needed?
  - Proposal vetted?
  - Outstanding issues resolved?
  - User community involved?
- Different parties working at different paces

Identifying long-term funding and full-stack support

- Providing assistance in getting script on devices and in software
- Ongoing goal to raise funds for this work, ideally disbursed over longer time horizons

How to address shifts in understanding and in Unicode expectations?

Takeaways

Keep engaging with communities already in Unicode

Better explain the steps after Unicode

Keep working with contacts to get information on unencoded scripts

Lobby companies to support SEI’s work and work on minority scripts

Encourage educational institutions to join Unicode and have a voice!

SEI, looking forward:

Help replenish ranks of Unicoders via coursework and research opportunities through SEI

Share past script communities' experiences with digitization in a *collective archive*

Translate SEI's story for public audiences through public writing and presentations

*“I found while attending Unicode Technical Committee meetings in the early 2000s that I had **found my tribe**, though it took some time to follow the Unicode lingo and to understand the encoding process”*

Thank you!

Image courtesy of www.worldswritingsystems.org

SEI

Script Encoding Initiative
Department of Linguistics
University of California, Berkeley

linguistics.berkeley.edu/sei

Debbie Anderson
dwanders@berkeley.edu

Anushah Hossain
ahossain@stanford.edu