

SIGNA

S I G N O G R A P H I C S T U D I E S

The Script Encoding Initiative

Original English version of the article
published in Signa Nr. 6

1.0 – April 2004

Author:
Deborah Anderson
UC BERKELEY

This file is provided free of charge for scholarly purposes. The provider allows to copy it and hand it down to anyone in an unaltered and unabridged state. No commercial use of this file is permitted. The editor shall be grateful for any hints or comments upon the contents of this document since it may be provided in an updated version if appropriate. – This file consists of 12 pages.

DENKMALSCHMIEDE HÖFGEN
E D I T I O N W A E C H T E R P A P P E L

The Script Encoding Initiative, the Unicode Consortium, and the Character Encoding Process

Deborah Anderson

Abstract

The Script Encoding Initiative is a project at UC Berkeley that was established in part to help bring the academic community and other user groups into the Unicode character encoding standards process so their needs can be made known and, hopefully, met. A description of the project, its achievements, and future goals will be presented. The project touches on topics that are pertinent to typographers, both in Europe and throughout the world, who would like to also have their needs met by Unicode.

Introduction

Over ninety-six thousand characters are covered in the international character encoding standard Unicode, which now encompasses a wide range of symbols and ancient and modern scripts. As a result of the growth of the Unicode Standard and its widespread adoption in computer architecture and software, many computer users who are sending texts electronically – whether in email, on Web pages, or in other electronic documents – can now rely on this standard to have their documents sent without corruption of their data. While the development of the Unicode Standard represents a monumental achievement and allows increasing electronic communication amongst a wide array of groups, the work is still unfinished, for over ninety scripts are missing from the Standard (see table on p. 10/11).

In order to address the issue of missing scripts and to provide expert feedback to the Unicode Consortium on specific language and script topics, a project at UC Berkeley was begun, the Script Encoding Initiative (SEI). SEI was also established to provide information to various user communities about Unicode, groups for whom the standards process – and even the objectives of Unicode itself – may not be well known or understood. Since the work of typographers touches on character encoding and hence the Unicode Standard, the SEI project can serve as a case study on how to work with the Unicode Consortium in getting characters proposed.

Core Concepts of Unicode

In order to discuss the Unicode Standard process a brief description of the basics of character encoding is helpful.¹ A few of the basic concepts underlying Unicode are:

(1) Unicode encodes characters, not glyphs. Characters are the abstract representations of the smallest components of written language that have semantic value,² whereas glyphs are what appear on the printed page or on your monitor; they are the surface representations of abstract characters. It is the glyphs that make up your font (see figure 2).

(2) Unique, idiosyncratic, proprietary symbols, and logos are not eligible for encoding. In order to cover such symbols, users can use markup or codepoints in a special area called the Private Use Area (PUA) of Unicode, which is intended for interchange between private parties.³

a	← Unicode's domain
»LATIN SMALL LETTER A« ABSTRACT CHARACTER	
a a a a a	← Font's domain
Different GLYPHS, representing the character »a«	

2. Example of glyphs vs. character.

(3) In order for characters to be approved, users need to show that the characters occur with in-line text (i.e., they occur with words in a line and not just in diagrams) and are needed in a plain text environment. Plain text is defined as computer-encoded text that consists only of a sequence of code points from a given standard, with no other formatting or structural information. It is opposed to rich or fancy text, which is plain text with additional information added in, such as font information, color, styles, etc.⁴

The Character Encoding Process

Script proposals are used to make requests to the standards bodies for new characters and scripts. The proposal itself is composed of a chart with a representative picture (=glyph) of each character (see figure 3) and a proposed name in a names list (see figure 4). Information on each character's properties is required (i.e., whether a character is a number, a letter, a symbol, a mark of punctuation, etc.,⁵ as well as details on how characters combine typographically. Examples of the characters from printed texts should be included to demonstrate how a script works, along with the names of recent standard reference works. A general introduction for the lay

reader and computer implementer who may not be familiar with the script can be provided.⁶

Once a proposal has provided the basic information outlined above, it can be forwarded to the two standards committees, the Unicode Technical Committee (UTC) and the Working Group 2 of the ISO/IEC JTC1/SC2/WG2. The UTC meets quarterly, often in the San Francisco Bay Area, and is composed of the members of the Unicode Consortium. The ISOWG2 meets once or twice a year at different locations throughout the world. Its members are national body members (governments) who are ISO WG2 members.

Proposals need to be considered and approved by both groups before they are accepted into the official standard, which is technically the Unicode Standard and the ISO 10646 standard, which are now fully synchronized. Because the UTC meets more often, proposals are typically reviewed by this group first. The meetings consist of a short presentation of the proposal by the author (or a designated representative). The UTC members typically pose specific questions about the proposal. It usually takes at least two UTC meetings before all the questions are answered. The entire process from the initial proposal until acceptance can span from two to five years, in part because the ISO WG2 only approves the proposals for balloting, then several rounds of international balloting are needed.

Description and Background of SEI

A number of unmet needs were identified by DEBORAH ANDERSON, liaison to the Unicode Consortium for the Department of Linguistics at UC Berkeley, and RICK MCGOWAN, Unicode Vice President:

(1) Academic users (and others) have often been unclear about what the Unicode Standard is, why it is important for their work, and how to get missing characters or scripts proposed. Hence, a contact person is needed who can answer these questions and act generally as an interlocutor between scholars and the standards bodies (especially the Unicode Technical Committee).

(2) There should be more scholarly participation in the Unicode standards process, since a number of outstanding script proposals still do not have active input from specialists. Comments are sought verifying that all the characters of a script are included in a proposal and that the representative pictures and names are acceptable. Letters from scholars to the standards bodies are also needed that verify a script is needed in Unicode and/or that a particular proposal is accurate and complete.

(3) Much of the work for past proposals has relied exclusively on volunteer efforts. As a result, proposals have appeared irregularly, meaning that some proposals lie unattended for years. In order to rectify this, funding is needed that can be earmarked exclusively for the writing and researching of character proposals by script proposal authors, such as veteran proposal author MICHAEL EVERSON.⁷

3. Sample portion of a Unicode proposal codechart from the Buginese proposal by Michael Everson (<http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2588.pdf>):

	1A0	1A1
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		
A		

00	BUGINESE LETTER KA
01	BUGINESE LETTER GA
02	BUGINESE LETTER NGA
03	BUGINESE LETTER NOKA
04	BUGINESE LETTER PA
05	BUGINESE LETTER BA
06	BUGINESE LETTER MA
07	BUGINESE LETTER MPA
08	BUGINESE LETTER TA
09	BUGINESE LETTER DA
0A	BUGINESE LETTER NA
0B	BUGINESE LETTER NRA
0C	BUGINESE LETTER CA
0D	BUGINESE LETTER JA
0E	BUGINESE LETTER NYA
0F	BUGINESE LETTER NYCA
10	BUGINESE LETTER YA
11	BUGINESE LETTER RA
12	BUGINESE LETTER LA
13	BUGINESE LETTER VA
14	BUGINESE LETTER SA
15	BUGINESE LETTER A
16	BUGINESE LETTER HA
17	BUGINESE VOWEL SIGN I
18	BUGINESE VOWEL SIGN YI
19	BUGINESE VOWEL SIGN U
1A	BUGINESE VOWEL SIGN E
1B	BUGINESE VOWEL SIGN O
1C	BUGINESE VOWEL SIGN AE
1D	BUGINESE VIRAMA
1E	BUGINESE PALLAWA
1F	BUGINESE END OF SECTION

4. (o.) Names list from the Buginese proposal by Michael Everson

(4) Attendance at Unicode Technical Committee meetings is highly recommended, because the author (or his/her designate) is able to answer specific questions posed by the UTC. This cuts down the time on the revision process considerably so a proposal can be fairly quickly be re-submitted to the UTC.

In response to these identified problems, SEI was set up at UC Berkeley in 2002. It is run in collaboration with Rick McGowan, who has considerable experience with the proposal process. The project has several ongoing activities:

(1) The director of SEI, Deborah Anderson, has written articles and presented papers in various venues to help promote Unicode and to explain the standards process.

However, the most time has been devoted to working with user groups in identifying eligible characters and encouraging the use of Unicode. Also, because many scholars want to continue using their (often older) operating systems, software programs, and fonts, it has become essential to spend time explaining to users how Unicode-compliant software and fonts can help assure the longevity and integrity of their textual data, as well as how they can bring their existing data forward into Unicode.

(2) In order to keep scholars up to date on the latest proposals, SEI maintains a website listing all the unencoded scripts⁹ with links to the latest proposals.¹⁰ It also includes a webpage listing the scripts that are currently under review by the two standards groups,¹¹ with a link to questions that need to be answered by scholars.

(3) In order to maintain a stable funding base, SEI is applying for funding from foundations and governmental grant agencies through U.C. Berkeley. Because it is a part of a non-profit educational institution, it can also receive tax-free donations made to SEI from groups and individuals are tax-deductible within the U.S..¹² Money raised thus far has been able to pay for the authoring of several Unicode proposals.

(4) The director, Deborah Anderson, continues to serve as a liaison to the Unicode Con-

sortium and attends Unicode Technical Committee meetings. As such, she can notify groups about upcoming deadlines for meetings, present script proposals for those script authors who are not able to be present, and can relay any questions to the author(s) from the Unicode Technical Committee.

The Unicode Consortium and SEI

The Unicode Consortium, which maintains the Unicode Standard, is primarily made up of companies (such as IBM, Microsoft, Apple, Sun, and Hewlett-Packard), with the participation of a few other non-profit or governmental groups (including SIL International,¹³ Research Libraries Group, the Government of India and the Government of Pakistan). This industry-heavy composition of the Consortium has ensured that Unicode is able to be implemented in computers and that it is interoperable between different computer platforms.

While the business interests have been actively behind much of the character encoding effort to date, advocates for the lesser-known scripts have not had a similarly strong presence amongst the Unicode Consortium membership. This vacuum is now being filled in part by SEI. Because the groups who are interested in the encoding of these lesser-known scripts are often not able to come to the standards committee meetings, advocates for missing scripts ought to come from institutions that are interested in the study of such scripts, such as from higher education or professional societies. If budget constraints prevent certain scripts from being studied or taught at a university, the need for having such a script included in the Unicode Standard is even more pressing, for online courses can be developed.

SEI Progress and Goals

To date, SEI has raised funds for work on various script proposals and a number of proposals partially funded by SEI have been approved, including Buginese, Glagolitic, Coptic, Old Persian cuneiform, and a Sumero-Akkadian cuneiform

proposal. Work is continuing on several other script proposals.

Additional funding for other script proposals is still being sought. While publicity has garnered some media attention, relatively little in terms of actual donations has resulted. Grant-writing to foundations and governmental agencies – itself a time-consuming effort – is needed.

Locating scholars with the time and interest in reviewing proposals is time-consuming, and often involves describing the core concepts of Unicode, emphasizing its importance, and outlining the standards process. Hopefully, as Unicode becomes more accepted, less time will be needed to explain what Unicode is and why it is important to scholars, though many in the academic world are still tied to their familiar older – but often non-standard – software and fonts.

Suggestions for the Typographers

As head of SEI, it has been my experience that the Unicode Technical Committee has been very responsive to proposals on missing characters as long as (a) the proposed characters are indeed eligible and (b) the proposals have all the necessary information provided in a well-researched and clear proposal. A number of specific suggestions for typographers include:

- (1) Join the Unicode email list. The membership includes people from a wide variety of backgrounds, all with an interest in Unicode.¹⁴
- (2) Read The Unicode Standard 4.0, which is available in print and in PDF format on the Unicode Consortium website.¹⁵
- (3) If a number of characters are deemed needed by members of the typography community, send in a note to the Unicode list to see if they might be eligible, as the characters may have already been discussed or may be in a proposal currently being drafted. If no proposal is in the works, a paper that discusses the use and need for particular characters (with proper documentation) can be sent to Deborah Anderson, who can then forward this document to other members of the Unicode Technical Committee for feedback, and relay this input back to the original author.

SEI can assist in guiding proposals through the standards process if the character(s) are eligible.

(4) Join the Unicode Consortium, either as a group (i.e., as a typographers group) or as an individual member. Unicode has memberships for individuals, specialists, for-profit and non-profit entities (at both the associate and full membership levels), and liaisons.

Conclusion

The Unicode Standard is vital for users today who are working with text electronically. But the Unicode Standard would benefit from having input and participation from the scholars and other user groups who can help to complete the task of encoding the remaining scripts of the world. *

Notes

1 More details are spelled out in The Unicode Standard 4.0 (TUS 4.0), which is accessible online and in print from the Unicode Consortium website, www.unicode.org

2 TUS 4.0, p. 15.

3 For further information on how to handle symbols deemed ineligible for encoding in Unicode, please see Martin Duerst's Missing Characters and Glyphs page at www.w3.org/International/O-MissCharGlyph

4 TUS 4.0, p. 1375.

5 as described in Chapter 4 of TUS 4.0.

6 For details on how to submit a proposal, see www.unicode.org/pending/proposals.html.

7 www.evertype.com

8 Articles have appeared in Ariadne (www.ariadne.ac.uk/issue37/) and Multilingual Computing & Technology (Volume 14 Issue 6), and papers have been presented at a variety of conferences, including the Electronic Metastructure for Endangered Languages Data workshops in 2002 and 2003, the UCLA Indo-European Studies conference in 2000 and 2001, Society of Biblical Literature annual conference in 2002, the Internationalization and Unicode Conference in 2002 and 2003, and at UC Berkeley as part of the Unicode Working Group in 2002.

9 www.linguistics.berkeley.edu/~dwanders

10 www.linguistics.berkeley.edu/~dwanders/alpha-script-list.html

11 www.linguistics.berkeley.edu/~dwanders/ScriptsNeedInput.html

12 www.linguistics.berkeley.edu/~dwanders/donations.html#Online

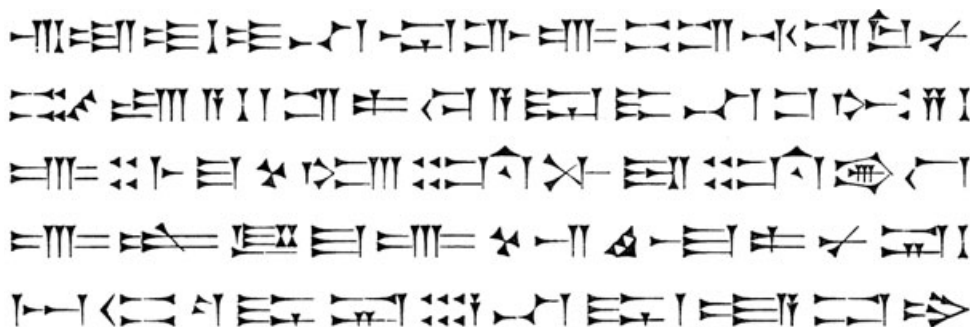
13 www.sil.org

14 Directions are included on the Unicode website at: www.unicode.org/consortium/distlist.html

15 www.unicode.org

Dr Deborah Anderson is a researcher in the Department of Linguistics at UC Berkeley. She received her Ph.D. from UCLA in Indo-European Studies, with an emphasis in linguistics. She currently edits the UCLA Indo-European Studies Bulletin and is assisting on a project to create a Unicode-compliant font for Indo-Europeanists. Additionally, she was a contributor for the etymologies in the American Heritage Dictionary, third edition, and worked as a consultant on linguistic-related projects for Houghton-Mifflin and Inso Corporation.

Examples of scripts still to be encoded



5. Cuneiform

6. Tifinagh

[illegible]

7. Javanese

[illegible]

8. Samaritan

· මත්තැන් · ත්වමත්තැන් · ත්වම · පුංඤ
· ඤාණයන් : පුංඤ · මාත්මයන් : ත්වමත්තැන්
· මාත්මයන් : ත්වමත්තැන් · ත්වමත්තැන් · ඤාණයන් : ත්වමත්තැන්
· පුංඤ · මාත්මයන් · ත්වමත්තැන් · පුංඤ · පුංඤ

9. Aramaic

ቶህህዮ ዘቐኮኝ ዘ 22 || ሊሃ ገገሃህ ቶህህ ገሃኮዘገሃ ገሃሃ ገሃቶ ገሃሃ
 ገሃኮዘ ገሃቶ ገሃቶ ገሃ ገሃሃቶ ቶህሃ ገሃሃ ገሃሃ
 ... ገሃ 22 ቶህሃ ገሃሃ ገሃሃ ገሃሃ ገሃሃ || ገሃ ገሃ ገሃ ገሃ

10. Palmyrene

[illegible][illegible]

11. Nabataean

12. Avestan

[illegible]

All figures of this spread, handlettering in original, are taken from:
Les caractères de l'Imprimerie Nationale
 (Imprimerie Nationale Éditions, [Paris], 1990).

Scripts still to be encoded

Ahom	G m	Glagolitic	A h
Alpine scripts †	A h	Grantha	A
Arabic (North-)	A	Greek (extensions)	A h
Arabic (South-)		Hatran	h
Aramaic **		Hittite Hieroglyphs	L h
Avestan **	A h	Hungarian Runic	A h
Aztec pictograms	I h	Iberian	A h
Balinese	G m	Indus Valley Script	L h
Balti	G m	Javanese **	G m
Bamum		Jurchin	I h
Bassa	m	Kaithi	G m
Batak	G m	Karian *	A h
Blissymbolics	m	Kaya Li *	G m
Brahmi	G h	Kawi	G h
Buginese	G m	Khamti	G h
Buthakukye	A h	Kharoṣṭhi-script	G h
Byblos	h	Khotanese	G h
Chakma	G m	Kitan Large Script	h
Chalukya	G h	Lahnda	G
Cham	G m	Lanna	
Chinook	h	Lepcha	G m
Chola		Linear A	L h
Coptic (additions)	A m	Lycian	A h
Cirth	a	Lydian *	A h
Cypro-Minoan	L h	Mandaic	A
Egyptian Hieroglyphs	L h	Manichaean	
Egyptian (Latin transliteration)	A h	Mangyan	h
Elbasan		Mayan Hieroglyphs	I h
Elymaic	h	Mende	m
Ethiopian (extensions)	S m	Meroitic	A h
Georgian (Nuskhuri)	m	Methei/Manipuri *	G m

Modi			Tengwar	a
Nabataean **	A	h	Tifinagh **	A m
Naxi Geba *		m	Turkestani	
Naxi Tomba		m	Uighur	A
New Tai Lue		m	Vai	m
Newari		m	Varang Kshiti	m
N'ko	A	m	Vedic, accents	G h
Numidian	A	h	Viet Thai	m
Nushu		m	Yezidi	m
Ol Chiki	G	m	Yi (extensions)	L m
Orkhon	A			
Pahawh Hmong		m		
Pahlavi	A	h		
Palmyrene **	A	h		
Permic (Old Permic)		h		
'Phags-pa	G	h		
Phoenician *	A	h		
Pollard Phonetic	S	m		
Proto-Elamite		h		
Pyu (Tircil) *	G	h		
Rejang	G	m		
Rongo Rongo	L	h		
Samaritan **	A	m		
Satavahana	G			
Saurashtra	G	m		
Siddham	G	h		
Sorang Sompeng	G	m		
Soyombo				
Sumero-Akkadian Cuneiform **	L	h		
Syloti Nagri	G	m		
Tai Lue		m		
Takri				
Tangut Ideograms	I	h		

This list includes scripts which have already passed one standardizing body, the *Unicode Technical Committee*, and await approval for balloting by the ISO Working Group 2 in June 2004.

Abbreviations: A – Alphabets and Abjads, G – Abugidas, I – Ideographic systems, L – Logosyllabic systems, S – Syllabic systems; a – artificial scripts, h – historic scripts, m – recent minority scripts.

† Alpine scripts include ancient Raetic, Venetic, Lepontic and Gallic “alphabets”.

* see figures on p. 2

** see figures on p. 8/9

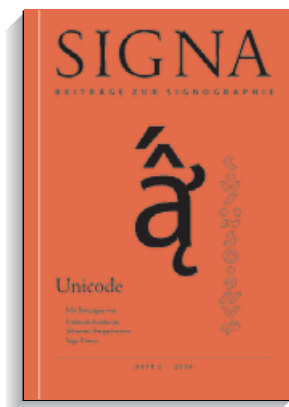
Compiled according to:
<http://linguistics.berkeley.edu/~dwanders/alpha-script-list.html>

SIGNA signographic magazin is intended to give the study of **Graphic signs** a platform of its own. Research, design and use of graphic signs are to be considered a subject in its own right.

SIGNA does provide *signographic studies* which embrace morphological, anthropological, semiological aspects alongside with those of history, design, communication and processing. The very nature of the graphical is to be focussed on, as is with general basics underlaying any graphic system or convention. SIGNA does promote such research and its professional application.

SIGNA features recent research papers, either on single signs or on whole signage complexes, about theoretical problems as upon method or terminology.

The thoroughfull edited thematic volumes appear once or twice a year. SIGNA is edited in German.



For more information about SIGNA
and signographics
visit
www.signographie.de

□ SIGNA_Anderson_SEI_1.0

🌐 www.signographie.de

April 2004

© Denkmalschmiede Höfgen gGmbH 2004

Denkmalschmiede Höfgen gGmbH
Edition Wächterpappel
Postfach 436, D-04663 Grimma
Tel. +49-3437-9877-0
Fax +49-3437-9877-10
www.hoefgen.de