Universal Scripts Project:
Statement of Significance and Impact

The Universal Scripts Project expands the capabilities of the Internet by providing digital access to text materials from a variety of modern and historical cultures whose writing systems are not currently included in the international standard for electronic representation of scripts, known as Unicode. People who write in these scripts find it difficult to use email, compose and send documents electronically, and post documents on the World Wide Web, without relying on nonstandard fonts or other cumbersome workarounds, and are therefore left out of the "technological revolution." About 66 scripts are currently included in the Unicode standard, but over 80 are not. Some 40 of these missing scripts belong to modern linguistic minorities in Africa, the Indian subcontinent, China, and other countries in Southeast Asia; about 40 are scripts of historical importance.

The project's goal for 2007–2008 is to provide the standards bodies overseeing character sets with proposals for 15 scripts to be included in the Unicode standard. The scripts selected for inclusion include 9 modern minority scripts and 6 historical scripts. The need is urgent, because the entire process, from first proposal to acceptance, typically takes from 2 to 5 years, and support among corporations and national bodies for adding more scripts to Unicode is uncertain. If the proposals are not submitted soon, these user communities will not be able to use their scripts in the near future. The scripts selected for this grant have established scholarly and user-community connections, which will help guarantee that the proposals meet the users' needs.

This project, based at the Department of Linguistics at the University of California at Berkeley, centralizes the effort to include missing scripts and to encourage the active participation of scholars. It is led by Dr. Deborah Anderson, the department's (and UC Berkeley's) representative to the Unicode Consortium, who has been active in the Unicode effort since 2000. The project will enable the single most successful author of Unicode proposals to write proposals for 15 scripts, for most of which preliminary research has already been completed. To guarantee that the proposals will proceed smoothly through the approval process, they will be carefully reviewed in advance by the Unicode vice president, who is a consultant to this project, as well as by those members of the advisory board who are on the Unicode Technical Committee. Significantly, funds will support research on further scripts, in an effort to actively involve new authors in the script-proposal process. An important part of the effort is for the project leader, assisted by a graduate student, to actively seek out specialists to review the proposals and make drafts freely available to scholars and the user communities for comment. Because of the large number of missing scripts, it is our hope that the project will gather momentum and more groups will participate in the script-proposal process. (Over the past year and a half, the project has encouraged several other groups to work on proposals.)

The results of this project are of great import to humanity. It will facilitate the learning of both modern and ancient languages via the Internet, and permit online communication among users of these scripts. A standardized encoding will facilitate collaborative work and discussions among scholars and online publication of documents written in these scripts. The project will also preserve and make accessible the documents that reflect the linguistic and cultural heritage of many groups, both ancient and modern. If these scripts are not included in the international digital standard, our record of these languages—and the people who spoke them—may be lost, or remain inaccessible to all but a very few.

Universal Scripts Project

Table of Contents

Universal Scripts Project
for the Electronic Transmission of Texts

Narrative

Universal Scripts Project
for the Electronic Transmission of Texts

1. Significance

At present, computer users can send and receive email, access online text materials, and create and read web pages in languages written in any of 66 scripts in use throughout the world—including the Latin alphabet (used for English and many other languages of the world), Cyrillic (for Russian and others), Greek, Arabic, and Chinese—because these scripts are all included in the international character-encoding standard known as **Unicode,** which is widely supported by the computer industry and national bodies. However, more than 80 scripts are missing from the standard, making it difficult to work electronically with text materials in these scripts.

Powerful and flexible as they are, computers can only communicate with each other reliably when they use exactly the same digital representation, or "encoding," for each character. For example, to virtually all personal computers the character code 0083 means the capital letter "S." (Alternate fonts can map the numeric entities to other characters: for example the Windows "Symbol" font renders 0083 as the Greek upper-case sigma [Σ].) Since the beginning of personal computers, the standard set of character encodings has been the ASCII character set, which includes the letters and numbers of the Latin alphabet and many accented and other European language characters (e.g. é, ç, ñ, ø, and ß), as well as a number of other common characters like &, ®, ¶, and §. The ASCII set was geared to the North American and European interests who originally developed computers, and served them well, but it contained only a limited number of characters, and made no provision for those who needed other characters (like the Turkish ı), or who used other alphabets like Cyrillic or Greek. These users had to create, or obtain, fonts that contained the letters and symbols they needed, and send those fonts to others with whom they wanted to exchange documents. But if the font on the recipient's computer did not precisely match that on the sender's, the text would be garbled. Unfortunately, these fonts proliferated, coming from companies like IBM and Apple, from national bodies that created their own encoding schemes, and from individual users and groups. Without a single agreed-upon standard, confusion was inevitable.

To bring some order to the growing chaos, specialists in the industry developed the concept of *Unicode,* which could expand the PC character set from the original 256 to the tens of thousands. Originally an attempt to unify the scripts important for "modern computer use," especially the disparate standards for Chinese, Japanese, and Korean (abbreviated CJK), the effort gradually expanded to include more national scripts. Eventually, Unicode was merged with the competing scheme ISO 10646, to create a single international standard. But just as the original ASCII standard had excluded users of Cyrillic, Greek, Arabic, and CJK, so it became obvious that users of many other

writing systems in Africa, Asia, and elsewhere still had no easy access to personal computing in their native scripts, and Unicode proponents soon developed the grand vision of a standard encoding for all scripts in use throughout the world. The academic world has also begun to recognize the value of a common encoding standard for the scripts used for ancient languages, as scholars strive to adapt their methods of study and collaboration to the new technology.

When there is no standard encoding for a script, users who wish to share texts written in it must rely on ad-hoc, nonstandard character encodings and homemade or proprietary fonts, often along with home-brewed software. But this approach invites confusion and error. Consider the predicament of an editor, for example, attempting to compile the electronic contributions of numerous authors who all use different, incompatible fonts. Perhaps more importantly, it demands a level of technical expertise and sophistication that is unrealistic to expect of most users. Other workarounds have also been tried—for example using graphics or SGML entities for the missing characters—but these are even more difficult and have other serious drawbacks: graphics are not searchable or often scalable, and are large data objects compared with text; and SGML entities can be ad-hoc creations that are cumbersome to keep track of. Nonstandard solutions like these cannot ensure the integrity of data transmitted electronically, or the long-term usability of the data as equipment becomes obsolete. Unicode provides a permanent, open, and industry-supported standard. As the default encoding for XML, which will be the basis for documents on the World Wide Web and many text-processing systems for some time to come, it is destined to become the universally accepted scheme.

Although great progress has been made, many important scripts are still missing from the Unicode standard. The missing scripts fall into two categories. One group consists of modern minority scripts: those used by speakers of minority languages, who are often marginalized in their own countries. Unencoded modern scripts currently total about 40, and are located in Africa, the Indian subcontinent, China, and other countries in Southeast Asia—typically in the poorest nations.[1] The second group consists of scripts of historical interest, studied by scholars and others; these also number about 40.

For modern populations that use these scripts, the inclusion of their writing systems in Unicode will enable the preservation of and access to humanities collections in electronic formats, and promote native-language instruction and literacy by allowing them to post documents and materials on the Web and to send texts electronically to other members of their community. Access to the scripts will aid foreign-language instruction in general, for they will now be usable in campus instructional programs and in distance-learning environments throughout the world. The ability to communicate electronically

1. According to the World Bank's statistics for 2004 (accessed June 2006), all the missing modern scripts are used in nations with annual per capita gross national incomes of under $825, except for China, Iran, Indonesia, and Thailand, which are classified "lower income economies" (GNI of $826–$3255). For a full listing of all the missing scripts with background information, see http://linguistics.berkeley.edu/sei/USR.html.

provides minority language groups with a voice and a presence in the modern world, as well as a means of preserving their cultural identity and heritage. For the Javanese people, for example, using the native script expresses and increases their pride in their culture in a way that writing Javanese in Latin letters cannot (see the description of Javanese in appendix A).

For scholars working with historical scripts, accurate transmission of text materials will make communication with other scholars and the wider lay audience easier. Currently, scholars employ a variety of transliteration and transcription systems (for cuneiform scripts, for example), but once the script has been accepted into the international standard, they will be able to use the original scripts directly in documents (as an example, see the letter from Brian Mubaraki on Mandaic, in appendix F). Including historical scripts in Unicode will improve online research and communication, by enabling scholars to discuss specific signs in their electronic discussions and by furthering the creation of online reference works. It will be a boon, too, to online study, which becomes particularly important when institutional funds are cut for instruction in the "dead" languages. Distributing texts on the Internet should reduce the cost of scholarly publication in these scripts, which in turn will help make materials more widely accessible.

The need for action is urgent. Because they must be approved by two standards bodies, proposals typically take from 2 to 5 years to go through the entire process. In the past 17 years, 66 scripts have been approved. At the current rate, unfunded, it would take at least 20 years for all the remaining scripts to be approved. But the door to including new scripts is already closing. Interest among corporations in encoding the remainder of the scripts is uncertain; their attention has turned toward improving computer implementations of those already encoded and security issues. Sadly, the delay in encoding has the effect of depriving the user communities even longer of access to their scripts at a time when many groups are keenly interested in having their script in Unicode (see the letters from Donny Harimurti for Javanese and Provungshu Chakma for Chakma, appendix F).

The reason for these scripts' absence up to now is largely economic: historical scripts and those used by linguistic minorities offer little economic incentive for computer companies or national bodies to pursue their inclusion, as the number of potential consumers does not justify an effort to provide access to their script. Another difficulty, particularly for those from linguistic minorities, lies in the nature of the standard-setting process itself. In order to have a script included in the international standard, a script proposal—including a listing of the characters of a script, a representative picture of each character, supporting information, and references—must be submitted to two bodies: the Unicode Consortium (via the Unicode Technical Committee, or UTC) and the International Organization for Standardization Working Group 2 (with the full title ISO/IEC JTC/SC2 WG2). Because the UTC meets in the United States and the ISO WG2 meets at various locations throughout the world, it takes time, expense, and commitment to successfully move proposals forward. While it is possible to submit a

proposal without physically being present at the meetings, it is highly advisable to attend them to answer any questions posed by the committees and to promote the proposal. Seeing a proposal through to final acceptance requires a substantial commitment of time, as well as familiarity with Unicode and the standards process, which is not widely known except to those who regularly attend and participate in the meetings. All these factors weigh against speakers of minority languages, who often lack the resources to travel and participate. The need for active regular participation is also not sufficiently recognized by academics, probably because of the time and commitment required: there is no regular voting academic member of the UTC except for the Universal Scripts Project's director, Dr. Anderson (who became UC Berkeley's representative in September 2004 and was earlier the liaison representative for the Department of Lingustics in 2002).

In the past, script proposals have been submitted by volunteers. Michael Everson, for example, the most prolific script proposal author, who will be providing scripts and basic fonts for this project, has worked in the past largely without pay and on his own time. As a result, the amount of time he could devote to proposals was limited by financial constraints. Since the remaining scripts are less well known and, in the case of historical scripts, sometimes fragmentary, preparing a full script proposal can require additional research and considerable consultation with language experts. Relying on this voluntary effort for the remaining 80+ scripts will only delay the effort—at a time when the standardizing bodies (Unicode Technical Committee and ISO WG2) are beginning to consider setting limits on the number of further updates that will be permitted.

By providing financial backing for the script-encoding effort, we can make a concerted push now that will guarantee access to a complete, stable set of scripts by linguistic minorities and scholars. But the task must be undertaken in conjunction with the university (home to linguists with an interest and expertise in these scripts), native users, Unicode officers, and the computer industry (which will be implementing the scripts in computer systems). The Universal Scripts Project aims to lead the drive to include the missing scripts in Unicode by involving all these key players.

2. Background of applicant

The University of California at Berkeley offers the ideal base from which to run a character encoding project because of its long-time connection with online text-based projects, its breadth of linguistic expertise, and its wider connection to other online projects in the UC system. UC Berkeley already hosts several internationally recognized computerized dictionaries managed by linguistics faculty members, including the *Turkish Electronic Living Lexicon* (TELL); Johanna Nichols's Ingush and Chechen dictionaries; *FrameNet*, a machine-readable lexicon that includes semantic descriptions of English vocabulary; and the *Sino-Tibetan Etymological Dictionary and Thesaurus* (STEDT). STEDT contains data on approximately 250 Tibeto-Burman languages, as well as material in Chinese. Although STEDT has already converted its non-Chinese data to Unicode, some of the other projects have not yet fully embraced it. This reflects the

situation elsewhere: many projects have not yet or are just now converting to Unicode. The Universal Scripts Project will be deemed successful if, as part of its mission, it can persuade more faculty to convert their text data in such online projects to Unicode, which in turn could help convince others in the wider academic community of the importance of adopting Unicode.

Furthermore, the Linguistics Department and other language departments at UC Berkeley include a wide range of experts in the different language families. These faculty members can serve as vital connections to user communities and to other specialists with necessary expertise in the various scripts that the Universal Scripts Project is working on. The Linguistics Department also hosts the Survey of California and Other Indian Languages, which is in the first phase of a three-year project to digitize and make accessible its collection of manuscript material (the largest university archive of linguistic documentation outside of the Smithsonian Institution, SIL International, and the American Philological Society). This project will involve displaying and searching with Unicode, and hence will serve as another important model for text projects relying on Unicode.

UC Berkeley provides a conduit to projects elsewhere in the University of California system, including UC Irvine's *Thesaurus Linguae Graecae (TLG),* a digital library that contains nearly all surviving Greek texts from Homer to A.D. 600 (and the majority of works that date to A.D. 1453), and UCLA's Cuneiform Digital Library (CDL), which includes cuneiform texts and images from ca. 3350 B.C. until the end of the pre-Christian era. Both of these groups have now become involved in the Unicode effort: The *TLG* has successfully advanced several proposals for Greek characters that occur in the *TLG*-hosted corpus but are missing from Unicode, and the CDL was actively involved in the Sumero-Akkadian Unicode proposal project. The TLG is currently working closely with Dumbarton Oaks on a proposal for Byzantine Greek symbols missing from Unicode.

Finally, UC Berkeley is also the home of another NEH-funded program: a collaboration among five universities and libraries nationwide to develop standards and practices for creating and retrieving historical sources on the Internet. Whereas that project worked to preserve digitized resources and make them more accessible, ours will complement it by enabling the inclusion of historical and cultural records from groups that have been excluded up to now.

3. History, scope, and duration

History of the project

The proposed project represents a continuation of the NEH-funded “Universal Scripts Project” awarded for the 2005–2006 time period, which is itself an outgrowth of Dr. Deborah Anderson’s original project, begun in April 2002 in the UC Berkeley Department of Linguistics, under the name “Script Encoding Initiative” (SEI). (The name was changed for the NEH grant because "Universal Scripts" was deemed more widely

understandable. Essentially, SEI and the Universal Scripts Project are the same, but USP is specifically the NEH project.)

The key players in SEI and the Universal Scripts Project all have extensive Unicode experience: Dr. Anderson, the project director, worked (with Michael Everson) on the Old Italic and Aegean script proposals, and has overseen scholarly review of proposals since 2000; Rick McGowan has been participating in the development of Unicode since 1989; and Michael Everson has been working on Unicode proposals and fonts since 1993.

In 2002, when Dr. Anderson first started the project, under the name Script Encoding Initiative, she created a web page that described the project and listed the unencoded scripts (hosted on the UC Berkeley Linguistics Department web site at www.linguistics.berkeley.edu/sei/), and sent out email announcements about the project to the Unicode and other email lists (for example LinguistList, and the Ancient Near East email list hosted by the Oriental Institute at the University of Chicago). She also presented papers at conferences promoting the effort to encode missing scripts. This publicity generated responses from potential collaborators at other institutions, as well as from interested students and experts at Berkeley and elsewhere. The exposure also raised startup funds at the outset of the project in 2002: the project received a seed-funding donation, and individuals and professional societies made small donations (see *History of Grants and Gifts,* page 100). With only a relatively small budget for three years, the project oversaw or assisted with 10 successful script proposals, including Buginese, Old Persian cuneiform, Glagolitic, Coptic, Tai Lue, and the Sumero-Akkadian cuneiform proposal. SEI assisted other groups in writing proposals that were also approved, including several by the *Thesaurus Linguae Graecae* at UC Irvine and Kharosthi, which was written by Andrew Glass and Stefan Baums at the University of Washington. The results of the initial project underlined key aspects that were adopted into the 2004 NEH grant's work plan (and the current application):

- provide funding to pay veteran Unicode authors to work on proposals and, if needed, fund travel to remote user communities
- have the project leader actively promote the project and encourage new authors to write proposals
- involve user communities and experts in the review of proposals
- have a representative participate in Unicode Technical Committee meetings to represent the historic script and modern minority script users
- work cooperatively with other projects involved in encoding proposals (i.e., SIL International) to prevent any duplication of effort[2]

2. The informal collaboration with SIL International in the past is now spelled out more clearly in a letter from Joel Lee, director of SIL International's Non-Roman Script Initiative, in appendix F.

- work with NGOs and local groups to get feedback on proposals from the user community (e.g., "Balinese Encoding Project" and United Nations Development Programme)[3]

During the past 18 months of NEH funding, the Universal Scripts Project has overseen the proposal of 12 scripts, with work on 6 still under way.[4] Thus NEH funding has significantly sped up writing and submission of proposals: from about 3 scripts per year when the project relied solely on donations, to nearly 9 per year with NEH support. The success of the project and its visibility have encouraged other script authors to try their hand at writing proposals: Dr. Elena Bashir, Lecturer in Urdu at the University of Chicago, for example, contacted Dr. Anderson for assistance on a proposal to encode 16 Perso-Arabic characters for minority languages in Pakistan (see her letter of support in appendix G). Similarly, Alexey Kryukov and the Slavonic typographic community in Russia sent in a proposal for historical Cyrillic letters. In both cases, Dr. Anderson helped review the proposals, made suggestions, and explained the standards process; the UTC approved both proposals. While NEH did not specifically pay for the authoring of such proposals, the project leader's guidance helped two of them to be successfully approved. Because the number of scripts is over 80, it is only through enlisting other groups and scholars to write proposals that the unencoded scripts will be proposed in the foreseeable future.

A second result of NEH funding is that it has enabled a veteran Unicode author to travel to meet with minority user communities, which normally would not have been possible because of the cost or political circumstances. A trip by Michael Everson to Myanmar enabled the Myanmar Computer Federation and the Myanmar Language Commission to organize a workshop on Myanmar language processing in Yangon, on 13–15 February 2006, to discuss specific issues relating to Myanmar minority languages. Over 14 people were in attendance (participants in the meeting appear in the photograph on page 4 of http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3043.pdf). In the same way, in January 2005, a "Balinese Encoding Group" (in Bali) assembled a meeting of over 20 script experts and font developers that led directly to another successful Unicode proposal.

A third outcome of NEH support has been increased participation by the university in Unicode-related meetings. A reallocation of grant funds (approved by NEH) allowed UC Berkeley to become an institutional member in the Unicode Technical Committee—the first time an academic institution has had a full vote on the committee. Because of the increased activity generated by UC Berkeley and specifically this project, UC

3. See the letter of support from Ida Bagus Adi Sudewa in appendix G as attestation of such a relationship in 2005, and letter proposing such collaboration with UNDP's representative, Helen Leake, in appendix F, under "Chakma."

4. Balinese, Vai, Lepcha, Saurashtra, Ol Chiki, Lycian, Lydian, Carian, Sundanese, Kayah Li, Rejang, and Myanmar Extensions. Still in progress are: Egyptian hieroglyphs, Lanna, Hieroglyphic Luwian, Avestan, Meitei, and Cham.

Berkeley was invited to become a liaison member of the ISO committee on coded character sets, the second standards committee that must approve proposals. In May 2006, the ISO Working Group 2 and its parent committee, Subcommittee 2, approved the liaison request. As a result, UC Berkeley can now submit contributions as a liaison member and comment on proposals, and thus represent groups that otherwise would not be heard from. Also, the increase in script proposals being submitted has led the ISO Working Group 2 to return to scheduling two meetings per year, rather than one as in 2003 and 2004.

To date, completion of the proposed tasks funded by the current grant is on schedule. The project has enlisted a number of new authors to do research or work on preliminary proposals, and this work is still ongoing: Charles Riley, for example, will be traveling to Cameroon in August 2006 to collect materials on the Bamum script. In May 2006, Richard Cook examined the Naxi manuscripts at the Harvard-Yenching Library for details on that script. (Many of the "new authors" are included in the present proposal for work on unencoded scripts.) Graduate students have been engaged to add material to the project's web pages, including script details, basic Unicode information, and a sample Unicode proposal; more work is scheduled over the summer or in the fall of 2006. Documents the project has produced so far (totaling 42) are listed in appendix H.

Based on feedback from the current project, the present proposal includes a new item in the work plan: to support software development, the software industry and open-source community are requesting that proposal authors submit *locale data* on modern scripts to the Common Locale Data Repository, a project hosted by the Unicode Consortium. Locale data are local conventions, such as standard date and time formats, that vary from one language or country to another.

In sum, we have made real progress on encoding scripts over the past 18 months, building on earlier work and relationships forged since 2002. The project continues to gain strength, and an appropriate level of funding will help continue the momentum and lead the way toward completing this important work (see *4. Methodology and standards,* page 10).

Scope and duration

For the two-year period of the proposed project, 16 scripts have been selected for work (see table 1, page 17; table 2, page 20; and appendix A). Michael Everson already has preliminary research at hand on all these scripts.[5] With the encoding of the scripts

5. For the Unified Canadian Aboriginal Syllabics, the work was begun by Chris Harvey, who has already relayed the material to Everson. For Bamum, a provisional list of names has been drafted, but Charles Riley will be collecting more material in August 2006.

outlined for this project, all scripts in the Unicode Basic Multilingual Plane will have been encoded.[6]

The concrete outcome of this project will be the approval of 15 script proposals by the SEI Advisory Board, with locale data provided on all the modern scripts, and one study on the varieties of one script, Pahawh Hmong.[7] Although 16 represents a modest number in the larger pool of over 80, we expect that NEH support now will spur other agencies, corporations, and foundations to provide longer-term support. We estimate that the effort to encode all missing scripts will take more than 10 years, even if ongoing support is found. At the rate proposed for the current project (8 scripts per year), 80 scripts might be completed within that time. But in response to publicity about the Universal Scripts Project and Unicode, we expect additional groups to take up the cause and work on further script proposals. Indeed, a number of proposals are now under way or planned at various other institutions, including:

- a proposal on Byzantine Greek symbols, being prepared by an editor at Dumbarton Oaks, in conjunction with *Thesaurus Linguae Graecae* at UC Irvine
- work on missing papyrological symbols, by the Center for Tebtunis Papyri (Todd Hickey, Donald Mastronarde) at UC Berkeley
- SIL International has expressed an interest in writing proposals on SignWriting, Fraser, and Pollard, as well as Ethiopic extensions and Arabic additions.

In order to encourage script encoding by new authors, the project has set aside a pool of funding for research and script proposal authoring (see *Budget Narrative,* page 35, for a breakdown of tentative authors and scripts).

4. Methodology and standards

The methodology for the project includes the following components:

- creating Unicode proposals
- review of proposals by scholars and the user communities
- publicity and outreach
- project evaluation

6. The only script not on the Basic Multilingual Plane "roadmap" (http://www.unicode.org/roadmaps/bmp/) is the Hangul Jamo extensions. However, South Korea has already stated that they intend to make a proposal for these characters.

7. For Pahawh Hmong, a study of the varieties needs to be conducted. It is not possible to ensure a proposal in the near future.

Creating Unicode proposals

A prime task of this project, of course, is to promote the creation of successful Unicode proposals, both by supporting the work of veteran authors and by guiding new authors through the process. Unicode is an international standard, with an established script proposal process that includes the following steps:

1. **Write the script proposal.** A proposal comprises a list of a script's characters, a representative picture of each character (called a *glyph*), a name for the character, information on the characters' properties (such as whether a character is a letter, a number, or a punctuation mark), sorting and line-breaking information, and a general introduction for the lay reader and computer implementer. The proposal typically includes a bibliography of recent or standard handbooks, and examples of the script that illustrate specific problems or characteristics and show the script in a longer text sample. (See appendix B for an example.)
2. **Submit the proposal to the Unicode Technical Committee.** The UTC, comprising primarily representatives from the computer industry, will comment on and ask questions about the proposal, particularly from the perspective of technical completeness and consistency with established models. Any problems raised by UTC members must be resolved before continuing through the process. Proposals, even when presented in great detail, typically take at least two UTC meetings to pass the scrutiny of the committee. Proposals should also have had input from scholars in the field before presentation at the UTC. A font must be provided in order to print the Unicode standard.
3. **Submit the proposal to the ISO WG2 group.** After UTC approval, proposals are usually submitted to the ISO WG2 group, either by expert contributions or a national body such as the U.S. Proposals may also be submitted to ISO WG2 before the UTC or at the same time. Typically, those presented at the ISO WG2 show up shortly thereafter at Unicode, as both groups work closely together.

Ideally, experts in individual scripts—representatives of the user community or scholars—would write script proposals and present them directly to the UTC and ISO working group. In the real world, though, this seldom happens: almost all proposals have been written and presented by experienced proposal authors. The process of writing, submitting, and advocating a proposal in the standards bodies can be daunting and confusing, and an experienced advocate is often needed to be present at the meetings, address the political issues that often arise, and stay with the proposal through the whole process, which may last several years. Several groups have remarked that the efforts of SEI and the Universal Scripts Project have been invaluable in helping them see their proposals through to acceptance (see for example the letter from Dr. Elena Bashir in appendix G). A significant part of the project's work will continue to include measures aimed at bringing user-community experts and scholars more actively

into the process, and providing them with the support of the Unicode specialists on its advisory board:

- Proposal authors, especially if new, will be encouraged to consult with Dr. Anderson regarding their proposals. She will answer basic questions and forward more difficult ones to Rick McGowan or the other advisors. Authors will be strongly encouraged to contact the relevant user communities as well, whether scholars or speakers of the languages.
- Before a proposal is submitted to the UTC, reviewers (Anderson, McGowan, and the board of advisors) will review it in order to anticipate questions and problems and suggest enhancements, thereby shortening the time to approval. In essence, the reviewers will help massage proposals into the correct format, applying the standards already set by Unicode and ISO WG2.

Review of proposals

A critical component of the Unicode proposal process involves the active participation of scholars and users who can review proposals to verify that they are correct and reflect all the needed characters, and write supporting letters to the two standards bodies.

Because Unicode is not especially well understood by the public at large (or the academic audience), new reviewers often require a basic discussion of what Unicode is, so they will be encouraged to read the introductory chapters in *The Unicode Standard* (available online on the Unicode Consortium web site), relevant Frequently Asked Questions on the Unicode web site, and information provided on the project's web site (http://linguistics.berkeley.edu/sei).

Because the missing scripts cover such a wide range of language families, it can be difficult to find reviewers who are familiar both with a given script and with Unicode. Assisted by the graduate student researcher, the project director will make an active effort to locate and recruit suitable experts, by:

- soliciting recommendations of knowledgeable experts from faculty in the Department of Linguistics and other relevant departments on campus
- sending announcements to the LinguistList and Unicode email lists and other relevant lists (e.g. AegeaNet for scripts relating to the ancient Aegean area).

Publicity and outreach

An essential aspect of this project is the need to continually publicize the effort of encoding scripts in Unicode, and to educate the relevant constituencies and the public at large about this effort. Our experience has shown that a number of questions need to be clearly addressed, including:

- what is Unicode?
- why should a script be included in the Unicode standard?
- which scripts are missing and what languages are they used for?
- why is there an urgent need to encode the scripts now?
- for new scripts in Unicode, how can I get the script incorporated into widely used fonts and keyboards?
- for languages without an orthography, how can I pick characters that will not cause problems for the user community when typing and sending text in their language?
- how can I help with this project?

The project director will continue to revise the project's web pages in order to answer these questions, with review of the content by McGowan and the board of advisors. This awareness-raising effort will serve several purposes. By making known the urgent need to include these scripts in Unicode, it will help generate donations from the general public and from professional societies and other interested groups (which can be used as part of the required "third-party donations"). Publicity about the project will also help circulate the list of unencoded scripts, so that additional experts in the field can be found who can review proposals, write letters of support, or participate actively in the proposal-writing process, either as primary authors or collaborators. Gathering and providing accurate background on unencoded scripts is especially helpful, as there is little information on these scripts available.

Project evaluation

At the end of each year we will perform an evaluation of the project, emailing a questionnaire to those who have participated in the encoding process, either as proposal authors or reviewers. Specific questions on the form will ask:

- Was information on the web site clear?
- Was guidance from the project staff helpful?
- How can the web site and assistance from staff be improved?

Participants will also be able to send comments to the project director at any time directly from the web page. Anderson and McGowan will review all feedback and make any needed structural changes to the web site and procedures. The project will be successful if, after consultation with user communities, the script proposals are approved by the two standards bodies.

5. Work plan

Year 1

1. Hire graduate student researchers

The project will employ graduate student researchers (GSRs) to assist in proposal-related tasks (as outlined in steps 2–4 below). Given the scope of the task ahead, it is important to involve grad students so they learn about Unicode now and can incorporate it in their work. Upon notification of successful funding, the project director will post an announcement on the Linguistics Department's email list, announcing positions for two GSRs. A duplicate announcement will appear electronically and on the bulletin boards of other relevant departments on campus (e.g. South and Southeast Asian Studies and the Institute of Slavic, East European, and Eurasian Studies), in case no Linguistics students are available. She will screen all applicants and forward vitae to McGowan and the advisory board, who will consult on the final selection. After the students are selected, she will train them in the required tasks (see job description, appendix D), oversee their work, and check regularly on their progress.

2. Add content and update information on the project's web site

The project maintains a "Unicode scripts research" page with detailed information on the unencoded scripts (http://linguistics.berkeley.edu/sei/USR.html). It lists the languages a given script is used for, the geographical spread of the script, dates of use (for historic scripts), the current status of the script (i.e., "proposed," "no script proposal written"), etc.

GSR1 will verify that the proposal status information on the "scripts research" page is current. This will involve updating the status of those scripts that have proceeded through the standards committees and providing links to the latest versions of the proposals. This task needs to be performed at least every four months, after the quarterly Unicode Technical Committee meetings, when scripts are approved. Also, if any new scripts are added to the list, he or she should add new entries.

A second task for GSR1 is to provide links in the script entries to the Ethnologue language entries, give an approximate number of speakers of the language (citing the source and its date), and add the ISO 639 language codes, which will help promote the use of these standardized codes. This task will be done by 1 September 2007.

The third task for GSR1 is to perform research to fill in any gaps in the "scripts research" page and verify with experts that the current content is correct. He or she will obtain names of scholars from faculty in the Linguistics Department (including the two members of the project's advisory board: James Matisoff for Sino-Tibetan and Johanna Nichols for others) and the project leader. (GSR1 will also maintain a list of such contacts as a resource list of potential experts to review proposals.) The project leader will review

any changes in the content before they are uploaded to the web site. This task will be done by 1 September 2007.

Lastly, GSR1 will improve the web site as directed by the project leader, primarily by drafting content to address questions posed by new script authors, scholars, user communities, and the public. Content will be reviewed by the project leader and McGowan before it is uploaded to the web site.

3. Assist with current Unicode proposals

GSR2 will assist the project director with tasks relating to Unicode proposals currently under way by Everson or others. These include the following:

Bibliographic assistance to proposal authors:

- locate (in books and journals), photocopy (or scan), and distribute to proposal authors samples of characters needed for proposals
- assist with documentation and other bibliographic tasks as needed

Helping with the review process:

- under the project leader's direction, locate contact information for scholars who might be available to review proposals, based on names provided by linguistics faculty, the project leader, and those names collected by GSR1 (see 2, above)
- if possible, contact any professional organizations that should be aware of the encoding effort, and see if they would like to publicize the effort (such as the Society of Biblical Literature, the Linguistic Society of America, etc.)
- send out email announcements to the LinguistList and Unicode email lists, with links to the proposals

4. Gather statistics

To aid third-party funders who are tracking the amount of material published in a given script and any change in its use once the script is encoded, GSR2 will collect statistics on the unencoded and recently encoded scripts, for example:

- number of existing pages (printed or handwritten)
- number of web pages with nonstandard fonts
- for scripts that have been recently encoded, amount of increased traffic to web sites with the newly approved scripts

Prof. Mikami's project, Language Observatory, will be able to assist with the latter two tasks (see letter of collaboration in appendix F).

5. Select specialists for work on additional proposals

An important goal of the project is to expand the cadre of qualified Unicode proposal authors. To this end, the project director, McGowan, and the board of advisors will select specialists with Unicode experience to work on additional proposals. The budget allocates $8,000 per year (a total of 400 hrs. @ $20/hr.), to be used either for creating new proposals or for additional required research on particular scripts. Potential candidates are listed in the attached budget narrative and in the tentative script lists (below). The project director will authorize payment for the work upon satisfactory completion of the proposal or a document on a topic.

Dr. Anderson will invite any brand-new specialists to submit proposals and a C.V. and list of references by 1 February. By 15 February she, McGowan, and the advisory board will select candidates to begin working on projects—preparing script proposals or background research. Anderson will periodically check on their progress and, for new proposals, will set an October deadline, allowing for proposals to be submitted at the UTC meeting in November.

6. Finalize list of scripts for year 1

Allowing for the schedule of upcoming standards-body meetings and necessary feedback, a final list of proposed scripts to be completed by Michael Everson for year 1 will be agreed upon by 1 February 2007. A tentative list appears in table 1 (page 17; for background information on these scripts, see appendix A). All of the modern scripts listed for year 1 have active user communities.

The estimated pay rate for these proposals and fonts is $50/hr., with most scripts estimated to require between 100 and 150 hours in order to complete a proposal. (New script authors, listed in the budget narrative, will receive $20/hr.)

7. Review and publish draft proposals

Upon completion of a proposal, McGowan and Anderson will review it and make comments. Once it is deemed acceptable, Anderson (or GSR1) will upload the proposal to the project's web site and provide a link to the proposal from the "scripts research" page, thus making it available for public comment. The proposal's author and Anderson (with GSR2) will seek scholarly input on it, as well as feedback from advisory board members. After comments have been relayed to the author and changes made as requested, Anderson will send a request to the chair of the UTC for the proposal to be put on the agenda for the next meeting, and the author will send it to the Unicode Technical Committee document register.

8. Present proposals at UTC meetings

Anderson will attend quarterly UTC meetings to present new proposals and report on ongoing proposal work, and will relay comments and suggested corrections and improvements to the proposals' authors. After these changes have been made and a

proposal font provided (see *9. Approve proposal fonts,* below), she will authorize payment for the proposal, in consultation with McGowan and the advisory board.

9. Approve proposal fonts

A font needs to be created in conjunction with every Unicode proposal, in order to print the standard. Fonts must receive input and approval from the user community before the project director authorizes payment.

10. Oversee entry of locale data

For modern scripts approved by the UTC, the project leader will select a tech-savvy member of the user community or a linguist who is very familiar with the script to enter locale data into the CLDR project. Questions on how to enter locale data will be sent to the project leader and Steven Loomis (see appendix F), who is on the CLDR team and developed the user interface. Locale data are important because implementers need this information for building localized software. The CLDR project makes the data freely available in a standardized format.

11. Year-end evaluation

At the end of year 1, an evaluation questionnaire will be sent to those who have participated in the Unicode proposal process, and will be reviewed by Anderson and McGowan.

Goal at the end of year 1: Complete at least seven script proposals and fonts and submit proposals to the UTC for approval.

Table 1: Tentative list of scripts, year 1

Modern scripts
Batak: Revise and complete proposal and create proposal font (170 hrs., $8,500 for Everson, $3,500 travel); collate and submit locale data (20 hrs., $400) Medium-complex Brahmic script without complex rendering behavior. The difficulty is in unifying the different versions used by different speech communities. Travel to Indonesia is budgeted in order to meet directly with the user community. (Travel funds are partly listed under Batak, partly under Javanese, for one trip to Indonesia.) A letter of collaboration with Prof. Uli Kozok appears in appendix F.
Javanese: revise and complete proposal (150 hrs, $7,500 for Everson, $2,500 travel); collate and submit locale data (20 hrs., $400) Complex Brahmic script with many glyphs and complex rendering behavior. Related to Balinese script. Everson wrote a document investigating these points in July 2005, including preliminary code charts and names. Travel to Indonesia is budgeted to meet and work directly with user community. (Travel funds are partly listed under Batak, partly under Javanese, for one trip to Indonesia.) See appendix F for letters of collaboration from I. Supriyanto, W. van der Molen, and Donny Harimurti.

Pahawh Hmong: perform a study of the problems involved in reconciling the several varieties of written Hmong script (100 hrs., $5,000 for Everson, $1,000 travel); submit preliminary locale data (20 hrs., $400)

Complex ordering and inputting. Everson wrote a document investigating these points in July 1999, including preliminary code charts and names. A letter of support from Vang Tzianeng appears in appendix F. Travel is included for a trip from Ireland to the U.S., where Vang is organizing a meeting on the Hmong language.

Unified Canadian Aboriginal Syllabics: Write and complete proposal and create proposal font (150 hrs., $7,500 for Everson; 50 hrs., $1,000 for specialist Harvey); collate and submit locale data (20 hrs., $400)

Straightforward syllabary with simple glyph representation. A fairly large number of characters (though less than 200) needs to be added for communities whose needs were not met by the original encoding. Preliminary work by Chris Harvey has been communicated to Everson. Work with local communities will probably necessitate a modular approach to completion. Chris Harvey has written a letter of collaboration (appendix F).

Unified Tai (Viet Thai): Review proposal, create proposal font (85 hrs., $4,250 for Everson; 75 hrs., $1,500 for Brase); collate and submit locale data (20 hrs., $400)

Complex Brahmic script with many glyphs and complex rendering behavior. Related to Thai and Lao scripts. Some preliminary work was done by Everson in 2000, and more has been done by Ngo Trung Viet and Jim Brase in 2006. Everson will review the proposal by Ngo Trung Viet, who is working with representatives in the Vietnamese government and other user communities, and Jim Brase, SIL. SIL International will provide the font.

Historical scripts

Hungarian Runic (Old Hungarian, Szekely runic script): Revise and complete proposal and create proposal font (100 hours; $5,000 for Everson)

This is a simple right-to-left script with a large set of ligatures, mostly optional, but some fairly obligatory. Everson wrote a preliminary proposal and supplementary discussion paper in 1998. See appendix F for letters from specialists Árpád Berta and Sandor Klara.

Mandaic: Revise and complete proposal, and create proposal font (100 hrs., $5,000 for Everson)

A complex right-to-left script with joining behavior and diacritics. Everson has completed a preliminary proposed code table and names list, dated 2001. Appendix F includes letters of collaboration from Dr. Brian Mubaraki, who is a member of the user community; Dr. Erica Hunter at Oxford; and William Clocksin, who has been working on OCR tools for Mandaic.

Meroitic: Revise and complete proposal and create proposal font (100 hrs., $5,000 for Everson)

This is a simple alphabetic script (actually an abugida, as there is an inherent vowel), which will be encoded left-to-right like Egyptian, although right-to-left and top-to-bottom are also found (as in Egyptian). Everson wrote a preliminary proposal in 1999. A letter from Michael Zach appears in appendix F.

Year 2

1. Hire new GSRs, if necessary

Project director will hire new GSRs, if necessary. The GSRs will update and improve the project web site in response to comments. GSR1 will update the alphabetical list on the "script research" page at regular intervals and do additional research; GSR2 will assist with current Unicode proposals and collect and update statistics on unencoded scripts as described above.

2. Select outside specialists to work on unencoded scripts

As described above for year 1. Some research may carry over from year 1, depending on the complexity of the script or difficulty in obtaining information.

3. Finalize list of scripts for year 2

A tentative list of scripts for Everson's work for year 2 appears in table 2 (page 20).

4. Review and publish proposals

Anderson and McGowan will review each proposal and make it available for general comment and criticism to the board of advisors and eventually to the UTC meeting.

5. Attend UTC meetings and present proposals

(as above for year 1)

6. Font services

(as above for year 1)

7. Oversee entry of locale data

(as above for year 1)

8. Seek additional funding

The project director will continue to research additional funding, beginning in April 2008, so that funding will be in place by January 2009.

9. Year-end evaluation

(as above for year 1)

Goal at the end of year 2: Complete eight script proposals and fonts, and submit proposals to UTC for approval.

Table 2: Tentative list of scripts, year 2

Modern scripts:
Bamum: Write and complete proposal and create proposal font (175 hrs., $8,750 for Everson to assist on proposal and create a font, $1000 travel; 100 hrs., $2,000 for specialist Riley to work on script proposal); collate and submit locale data (20 hrs., $400) Although Bamum is a simple right-to-left script, the size of the repertoire is quite uncertain; the current estimate is about 550 characters. Everson has a very preliminary proposed names list completed, but it is extremely provisional. Charles Riley will collect additional data in Cameroon in August 2006 and share it with Everson. Everson will prepare the font, and Charles Riley will work on the script proposal. Travel is included for Everson in Ireland to visit Riley in the U.S. and work on the script proposal. A letter of collaboration by Riley is included in appendix F.
Chakma: Revise and complete proposal and create proposal font (150 hrs., $7,500 for Everson); collate and submit locale data (20 hrs., $400) A complex Brahmic script with diacritic vowels and subjoined consonant clustering. Not well known, but Everson has developed a preliminary code chart and list of names. It has an active user community in India. Letters offering assistance on Chakma from Uttamalankar Chowdhury of the Parbatya Bouddha Mission, Provungshu Chakma, and Helen Leake of the Regional Indigenous Peoples Programme, UNDP, are included in appendix F.
Newari: Revise and complete proposal and create proposal font (150 hrs., $7,500 for Everson); collate and submit locale data (20 hrs., $400) Complex Brahmic script with many glyphs and complex rendering behavior; related to the Devanagari script. Everson wrote a document investigating these points in 2000, including preliminary code charts and names. A letter of collaboration from Allen Bailochan Tuladhar is included in appendix F.
Sorang Sompeng: Revise and complete proposal and create proposal font (100 hrs., $5,000 for Everson); collate and submit locale data (20 hrs., $400) Everson wrote a preliminary proposal in 1999 for this simple left-to-right abugida script. (Greg Anderson and K. David Harrison will assist in locating experts on Sorang Sompeng.)
Varang Kshiti: Revise and complete proposal and create proposal font (100 hrs., $5,000 for Everson); collate and submit locale data (20 hrs., $400) A simple left-to-right abugida. Everson wrote a preliminary proposal in 1999. A letter from linguists Greg Anderson and K. David Harrison, who are working with the user community of Ho language speakers, is included in appendix F.
Ancient scripts:
Manichaean: Revise and complete proposal and create proposal font (100 hrs., $5,000 for Everson) This right-to-left script has character joining behavior that needs to be described. Everson wrote a proposal in 2002 with Desmond Durkin-Meisterernst. P. Oktor Skjærvø, Aga Khan Professor of Iranian, has agreed to assist on the proposal (private communication).

Samaritan: Revise and complete proposal and create proposal font (100 hrs., $5,000 for Everson) Right-to-left script with combining diacritical marks, some of which have problematic display properties. Preliminary code table made by Everson in 2001. Two letters from collaborators, Mark Shoulson and Benny Tsedaka, are included in appendix F.
Vedic accents: Review and collaborate with scholars both within and outside India on a proposal (170 hrs., $8,500 for Everson) Everson wrote a preliminary document on the topic in 2000. The government of India has submitted a preliminary proposal, but significant questions remain on a number of aspects. The question of how many accents is extremely difficult due to lack of accurate information. For example, even if there are 11 different kinds of Ayurvedic visargas used with Devanagari, the question remains, which (if any) are used in Bengali, Gujarati, Malayalam, Telugu, and the rest? Can they all be unified or are they all script-specific? Access to experts with texts is needed to move forward with this. The government of India will provide the proposal font. Madhav Deshpande has offered his assistance in a letter in appendix F.

6. Project staff

Core personnel

The project's core participants include the project director, a general consultant, and an expert on fonts and the creation of Unicode script proposals. A pool of new script authors will also participate.

Project Director: Dr. Deborah Anderson

Deborah Anderson, a researcher in the Department of Linguistics at UC Berkeley, will spend 50% of her time on this project. She currently leads the Universal Scripts Project at Berkeley and headed its progenitor, the Script Encoding Initiative. Under her leadership, numerous scripts have proceeded through the standards process. She has also established ongoing collaboration on script proposals with other institutions and organizations (SIL International, the Language Observatory Project, and the Asian Language Resource Network; see letters of collaboration, appendix F), which is expected to continue to be fruitful. As Project Director for the Universal Scripts Project, her tasks include the following:

- hire and train graduate students to maintain the project's web pages and develop content as needed, assist with script proposals, and solicit scholarly input on proposals
- supervise graduate students and outside specialists to work on unencoded scripts
- relay any feedback received during the proposal review process to the proposal authors

- attend UTC meetings in order to present script proposals and report on upcoming proposals
- follow up on proposals so that UTC-requested changes are made and proposals are resubmitted for consideration at the next UTC meeting
- attend conferences to promote the project
- seek collaboration from scholars, users, and institutions
- raise funds for the Universal Scripts Project for future years

Consultant: Rick McGowan

Rick McGowan (Vice President of Unicode) will serve as a consultant and key collaborator. He will review new proposals, comment on potential problems that need to be addressed before full presentation at the UTC, and check script proposals and fonts before funding is approved. He will also participate in the hiring of specialists and advise on the graduate student selection process. His long-standing participation in Unicode (since 1989; he became technical director in 1993 and has been vice president since 2000) means that he brings with him an intimate familiarity with proposals and experts, problems that have been encountered with Unicode proposals and their implementation by industry, and where specific future needs lie.

Proposal and font services: Michael Everson and other specialists

Michael Everson will write proposals and create fonts as part of the budget line item *Script proposal services.* Mr. Everson is the Irish national representative to the ISO WG2 committee and is the single largest contributor by far of Unicode/WG2 proposals (see appendix C for a résumé). He has authored or coauthored over 178 standardization documents, most of which were proposals for the addition of scripts and characters to ISO WG2 and Unicode. He is also one of the authors of Unicode 5.0. Thanks to his experience and familiarity with the process and its requirements, his rate of approval of script proposals is over 95%.

Several other "new authors" will be selected, after consultation with McGowan and the advisory board, from a pool of specialists with Unicode experience. They will either work on new proposals or do background research on more difficult scripts. One goal of employing outside specialists (other than Everson) is to begin filling the ranks of potential Unicode proposal authors. (Tentative assignments are listed in the Budget Narrative.)

Advisory board

This group includes industry professionals, members of standards groups, academicians, and librarians, all groups for whom this project is important. Members of the advisory board will assist the project director in selecting specialists who will work

on proposals and in hiring two graduate students, and in general act as information resources in their respective fields of expertise. Those who have served (or currently serve) on the UTC (Whistler, Constable, Mansour, Collins) will review proposals.

Lee Collins is Manager of the Operating Systems Engineering Asia group in Apple's Software Engineering division. He is one of the original founders of Unicode and has made significant contributions to the East and South Asian portions of Unicode. As a result, he provides valuable insight into the early encoding decisions for the scripts of Asia, as well as current software development for scripts in this area. Most recently he has been participating in the proposal for Egyptian hieroglyphs and actively participated in discussions on the proposal for Myanmar extensions.

Peter Constable, a program manager for new scripts at Microsoft, was formerly the representative to Unicode from SIL International, a nonprofit organization that documents languages and scripts throughout the world. Peter has authored numerous Unicode script proposals, especially for Asian scripts and linguistic symbols. Because of his current work in developing implementation of Unicode at Microsoft, he provides key insights into the needs of users as well as industry's perspective. He holds a master's degree in linguistics and has done work as a field linguist in Mexico and Thailand.

Kamal Mansour is the manager of non-Latin products with the font company Monotype Imaging. His particular strength is in Middle Eastern scripts, especially Arabic and Hebrew; Greek; and the Latin alphabet for minority languages. His expertise in font development is especially useful, since this is an important component of work for Unicode. He is on the Unicode Technical Committee.

James Matisoff, Ph.D. is a professor in the Department of Linguistics at UC Berkeley, whose specialty is the languages of Southeast Asia, where a large percentage of unencoded scripts are used. His expertise and contacts in the field will be extremely helpful for any proposals documenting scripts in this area, particularly those from the Sino-Tibetan family.

Mike McKenna is an architect for the Advanced Technologies Group of the California Digital Library (University of California, Office of the President). He comes to the project with over fifteen years of industry experience. He will be able to provide input on any potential problems new scripts might have on current computer architecture and software. He will also act as a valuable resource for contacts within the digital library community.

Johanna Nichols, Ph.D. is a professor of Slavic languages and literatures at UC Berkeley and an affiliate professor of the UCB Linguistics Department, whose specialties are language typology and Caucasian and Slavic languages. She has extensive contacts with specialists in many language families because of her current cross-linguistic database project, AUTOTYP (http://www.uni-leipzig.de/ ~autotyp/).

Ken Whistler, Ph.D. is a Unicode technical director and managing editor of *The Unicode Standard,* U.S. representative to the ISO WG2, and an engineer at Sybase, with a degree

in linguistics. Dr. Whistler has been associated with Unicode since 1989 and has been an active supporter of the effort to get missing scripts included in Unicode and ISO 10646, both within the UTC and WG2.

7. Dissemination

The tangible results of this project will be free online access to the Unicode proposals, as they progress through the standards process, on the Universal Scripts Project's web site.[8] (The project budget includes funds for copying, faxing, and postage expenses, to permit sending hard copies to those who request them.) The web site will also make available a list with background information on the remaining unencoded scripts.

The long-term goal is the acceptance and adoption of the script proposals by the two standards bodies, thereby ensuring that the scripts will be supported in software and fonts. Because Unicode is an open standard, no intellectual property issues are involved.

8. Approved proposals that have already been adopted into the standard are posted publicly on the Unicode web site (http://www.unicode.org), and are printed by the Unicode Consortium in *The Unicode Standard* (Addison Wesley, 1991-2003), which will reappear with Unicode 5.0.

Universal Scripts Project
for the Electronic Transmission of Texts

Appendices
and Supporting Materials

Appendix A:
Brief background information on scripts

Project-supported scripts

Bamum

The Bamum language is spoken by about 215,000 people in Cameroon. The script, dating from before the arrival of the Europeans in 1902, went through several versions, from ideographic to syllabic, and although it is a simple right-to-left script, the size of the repertoire is uncertain. The current estimate is about 550 characters. There are over 7,000 documents in the script. The script is still taught but very few can read and write it fluently. However, there is very strong community interest in the script and reviving its use.

Sample of Bamum (source: Michael Everson):

Batak

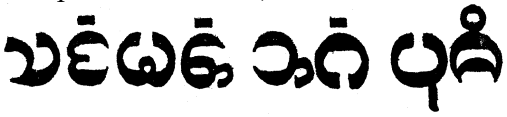
The Batak script is used by the Batak speakers on the northern part of Sumatra, Indonesia. There are several related Batak languages, with the total number of speakers coming to 5.8 million. Different versions of the script are used by different Batak user communities. The script originates from the Brahmi script, and is written left to right but sometimes printed vertically on bamboo.

Sample of Batak (source: Michael Everson):

Chakma

This script is used to write the language of the same name, which belongs to the Indo-Iranian branch of the Indo-European family and is used by 560,000 speakers in India and Bangladesh. In India, the Chakma people have been classified a scheduled tribe, which means they are eligible for special provisions in terms of education, economic interest, and protection against exploitation. Bengali and Chakma scripts are used to write Chakma.

Sample of Chakma (source: Michael Everson):



Hungarian Runic (Old Hungarian, Szekely runic script)

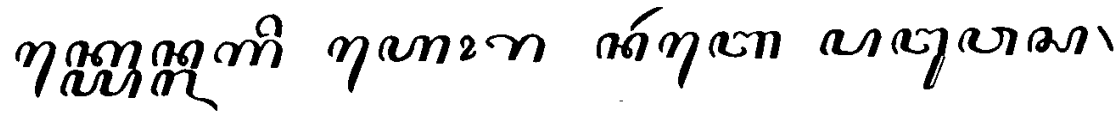
The Hungarian Runic script is descended from the Kök Turki script used in Central Asia. It was used by the Székler Magyars in Hungary before A.D. 1000, but most specimens of the script date to later periods. It is an alphabetic script with right-to-left directionality, and forms numerous ligatures.

Sample of Hungarian Runic (source: Michael Everson):

Javanese

There are some 75 million speakers of Javanese, an Austronesian language spoken in Indonesia, Malaysia, and Singapore. The Javanese script has much less currency than Latin does, but there are enthusiastic users. There is a great deal of traditional literature in Javanese. One informant reports that Javanese is still taught in a few primary schools in Java, but few students gain fluency. Being able to publish in the script (with a Unicode encoding) could help encourage its survival, particularly among the younger generation.

Sample of Javanese (source: Michael Everson):



Mandaic

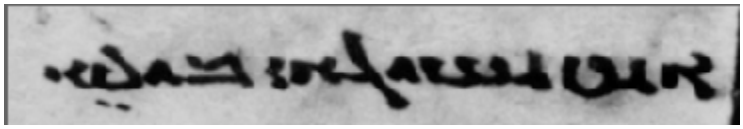
Mandaic is the liturgical language of the Mandaean religion; a vernacular form is still spoken by a small community in Iran around Ahwaz. Mandaeans also live in the U.S., Australia, and other countries.

Sample of Mandaic (source: Michael Everson):

Manichaean

Manichaean is a right-to-left script evolved from the Syriac Estrangelo script. It is used in liturgical texts for the Manichaean religion, founded by Mani in the third century A.D., who some believe devised the script. The religion flourished for several centuries before eventually dying out in the 14th century, but did become the official state religion in the Uighur kingdom in Central Asia (A.D. 762 – 840). The most signifcant Manichaean texts in the East were found in the Turfan oasis on the Silk Road in Central Asia. These texts, written in the Manichaean script, were used for the Iranian languages Middle Persian, Parthian, Sogdian, and the Turkic language Uighur.

Sample of Manichaean (source: Michael Everson):



Meroitic

Meroitic is an abugida based on Egyptian hieroglyphic and Demotic scripts, and was used to write the Meroitic language of the Kingdom of Meroë by at least c. 200 B.C.. It may have been used to write the Nubian language, since three of its letters were borrowed into the Coptic script used to write Nubian. There are, strictly speaking, two Meroitic scripts; it remains to be seen whether the Hieroglyphic and the Demotic versions should be unified or be encoded separately.

Sample of Meroitic (source: Michael Everson):

Newari:

Newari, or Ranjana script, is used alongside Devanagari for the Newari language of Nepal, a Tibeto-Burman language with about 825,000 speakers. The script itself derives from Brahmi. (Note: A secondary script, Nepali, is also used and may require encoding.)

Sample of Newari (source: Michael Everson):

Pahawh Hmong

The Pahawh Hmong script was devised by Shong Lue Yang in 1959 for Hmong, an Austro-Tai language spoken by about 5.5 million people in China, Vietnam, Thailand, Laos, and the U.S. The language is written in different scripts: Chinese (in China), Thai

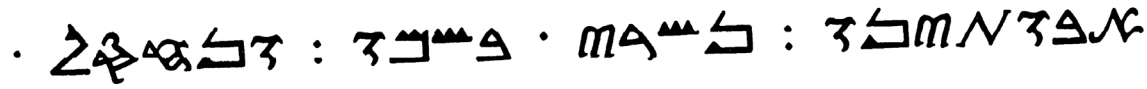
(in Thailand), and various other scripts that have been devised or adopted for Hmong. Special pride has been felt for Pahawh Hmong because it was invented by a Hmong.

Sample of Pahawh Hmong (source: Michael Everson):

Samaritan

Samaritan is used today to write Samaritan Aramaic and Hebrew. The script is preserved in Biblical scrolls, mezuzahs, amulets, and even a biweekly newspaper. The Samaritans are descendants of the ancient Israelites who broke from Judaism about 2,200 years ago. They number today about 700, approximately half of whom live in the West Bank and the other half in the Israeli town of Holon near Tel Aviv.

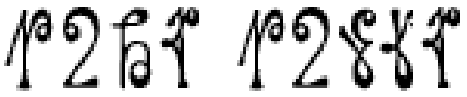
Sample of Samaritan (source: Michael Everson):



Sorang Sompeng

The Sorang Sompeng script is used to write the Sora language, a member of the Munda family. The Sora people, about 288,000 in number, live between the Oriya- and Telugu-speaking populations in what is now the Orissa-Andhra border area. Sorang Sompeng was devised by Mangei Gomango, son of the charismatic community leader Malia Gomango, as part of a comprehensive cultural program, and was offered as an improvement over scripts used by Christian missionary linguists. Sorang Sompeng is used in primary and adult education, and is published in a variety of printed materials. Simple alphabetic script.

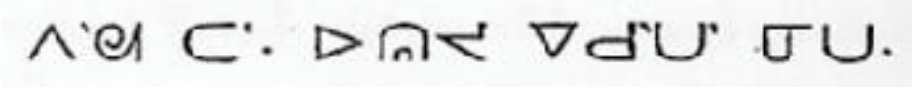
Sample of Sorang Sompeng (source: Michael Everson):



Unified Canadian Aboriginal Syllabics Extensions

The Unified Canadian Aboriginal Syllabics have been encoded in the Unicode standard since Unicode 3.0 (2000). The block comprises various local syllabaries of Canada that were unified based on their appearance. The syllabics were originally invented for the Algonquian languages in the 1830s by James Evans. About 200 syllabics are known to be missing from the UCAS block, however, impeding use by some users. The script communities involved for these extensions include the Naskapi, Blackfoot, Carrier Dene, Chipewyan Dene, Slavey Dene, Hare Dene, Beaver Dene, Ojibway, and Cree.

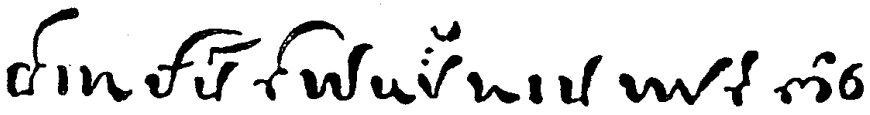
Sample:



Unified Tai (Viet Thai)

The script is used for a number of related languages: Tai Daeng (population 165,000), Tai Dam (764,000), Tai Dón (490,000), and Thai Song (32,000). The Tai languages are spoken in northwestern Vietnam, with populations in Australia, China, France and the United States. The script is related to the Thai and Lao scripts. Use of the script varies, depending upon the community, but there is some discussion of introducing it into the formal education system in Vietnam.

Sample of Unified Tai (source: Michael Everson):



Varang Kshiti

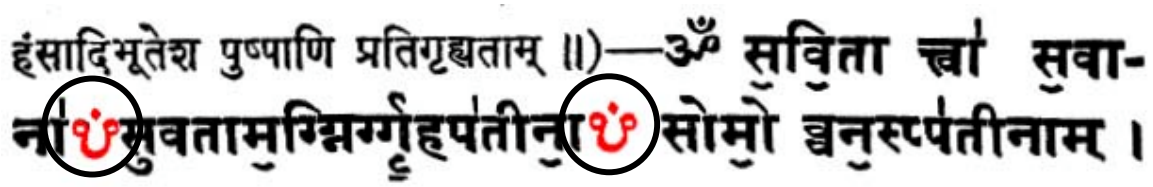
Varang Kshiti is used to write Ho, a Munda language of the Austroasiatic language family spoken primarily in India by about 1,077,000 people. Ho speakers are found living fairly evenly split between Hindi-speaking provinces and the Oriya-speaking ones, which each use a different script (Devanagari and Oriya). Use of a single script, Varang Kshiti, would help unify the Ho language speakers. The script is an abugida. It is used in educational publication materials and could be used for other smaller Munda languages.

Sample of Varang Kshiti (source: Michael Everson):

Vedic Accents

Accents for classical Sanskrit are already included in Unicode, but several are missing that are needed for the Vedic Sanskrit texts.

Sample of Vedic Accents (with missing accents circled) (source: http://www.omkarananda-ashram.org/Sanskrit/vedicaccents.htm):

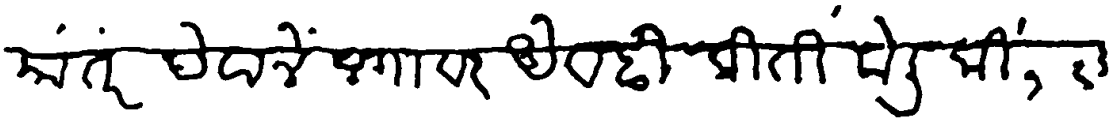


Other script research (tentative):

Various Indic scripts (Sharada, Modi)

Sharada and Modi are both "near modern" scripts from India; a number of other historical and "near modern" scripts from India are also eligible for research and proposals.

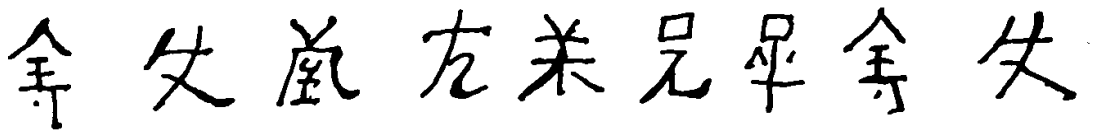
Sample of Modi (source: Michael Everson):



Jurchen

This script was used in China from the 12th until the 16th century. It was created by Wanyan Xiyin in 1120. The repertoire is reported to have about 720 characters.

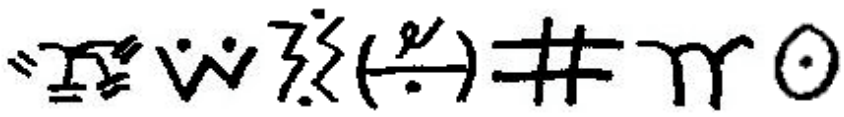
Sample of Jurchen (source: Michael Everson):



Loma

This script was developed in the 1930s by Wido Zobo for the Loma and Toma people of Liberia and Guinea. The script is historical; the last known use dates to the mid-1980s, but there has been interest in reviving its use from within the community. Some published script charts exist that document the syllabic values of its character repertoire, but known samples of use of the script in primary sources are scarce.

Sample of script (from http://www.ed.arizona.edu/loma/):

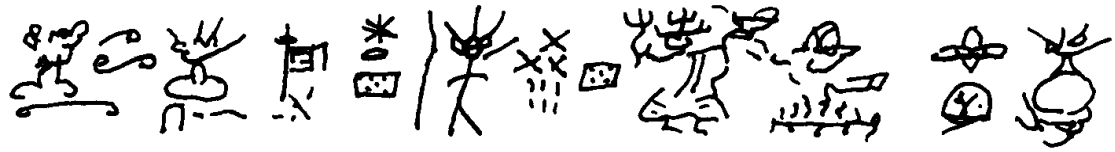


Naxi Pictographic (Naxi Tomba) and Syllabographic (Naxi Geba) Scripts

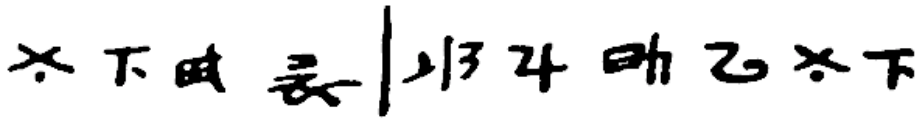
According to J. F. Rock's *Naxi-English Encyclopedic Dictionary* (1963), the Naxi script has a pictographic and syllabic component. The pictographic syllabary is said to have been invented between A.D. 1200 and 1253, though it may be much older; the syllabary is considered to be ancient. The script is used to write Naxi, a Tibeto-Burman (TB)

language, related to Moso (a.k.a. W. Naxi) and Loloish (Yi) languages. Rock's dictionary identifies a total of 3,414 signs, which should be used as a basis for the initial encoding.

Sample of Naxi Tomba (source: Michael Everson):



Sample of Naxi Geba (source: Michael Everson):



South Arabian

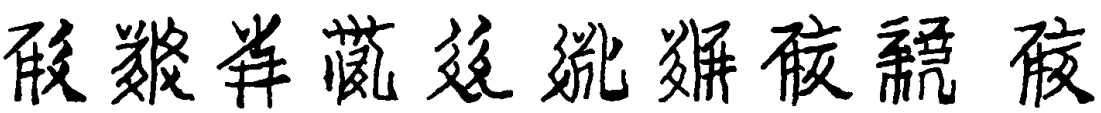
Over 10,000 inscriptions in Old South Arabian script have been found. The script runs right to left, and dates from 600 B.C. to A.D. 600. It was used to write several extinct Semitic languages that were spoken in southern Arabia.

Sample of South Arabian (source: Michael Everson):

Tangut

Tangut, also known as Xixia, is a Sino-Tibetan language formerly spoken in northwestern China, but which has been extinct since perhaps the 16th century. The Tangut script, modeled on Chinese and Khitan, has been in use since the early 11th century. Research on this script has been of keen interest to Sino-Tibetan linguists attempting to situate the language within the language family. There are over 7,000 Tangut characters.

Sample of Tangut (source: Michael Everson):



Appendix B:
Example Unicode proposal

The following pages represent a very short sample Unicode proposal. Proposals are typically much longer, depending upon the complexity of the script. This Rejang script proposal was written by Michael Everson in April 2006 as part of the Universal Scripts Project.

Components of a typical script proposal include:

(a) The proposal proper (pp. 1–2)

Provides basic background information on the script and its user community and discusses the script itself (its structure, ordering, naming, punctuation and digits, linebreaking, Unicode character properties, and any processing issues). Such descriptions of how a script works are needed for software implementers and font designers. The proposal includes a bibliography of standard reference works.

(b) Samples of the script as found in printed works (pp. 3–4)

These depict the script as it appears in print, both in tables and in running text. Several well-selected examples help to show a range of features and variations. In this proposal, Figure 1 lists the characters from a table found in a handbook. Figure 2 depicts the script as it appears in running text, with a caption referring to specific characters in the figure.

(c) A chart with the representative pictures of the characters (p. 5)

The layout refers to tentative assignments to specific Unicode codepoints.

(d) A list of Unicode names for each character (p. 6)

(e) ISO/IEC JTC1/SC2/WG2 forms (pp. 7–8)

These administrative forms need to be filed before final submission. Note that submissions may be made to the two standards groups, Unicode and ISO WG2, at the same time. The two groups work closely together.

The lengthy set of initials (abbreviated to "ISO WG2") stand for:

> International Organization for Standardization (ISO),
> International Electrotechnical Commission (IEC),
> Joint ISO/IEC Technical Committee (JTC, the group concerned with information technology),
> SubCommittee 2 (of a Technical Committee, concerned with Coded Character Sets),
> Working Group 2 (WG2, working on Multiple Octet Codes).

ISO/IEC JTC1/SC2/WG2 N3096
L2/06-139
2006-04-24

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Proposal for encoding the Rejang script in the BMP of the UCS**
Source: **UC Berkeley Script Encoding Initiative (Universal Scripts Project)**
Author: **Michael Everson**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2006-04-24**

Rejang is spoken by about 200,000 people living in Indonesia on the island of Sumatra in the southwest highlands, north Bengkulu Province, around Argamakmur, Muaraaman, Curup, and Kepahiang, and also in the Rawas area of South Sumatra Province, near Muara Kulam. There are five major dialects of Rejang: Lebong, Musi, Kebanagung, Pesisir (all in Bengkulu Province), and Rawas (in South Sumatra Province). Most of its users live in fairly remote rural areas, of whom slightly less than half are literate. The traditional Rejang corpus consists chiefly of ritual texts, medical incantations, and poetry.

Origin
The Rejang script is of the Brahmic type, and is related to other scripts of the region, like Batak, Buginese, and Kerinci. The script was in use prior to the introduction of Islam to the Rejang area; the earliest attested document appears to date from the mid-18th century CE.

Structure
Vowel signs are used in a manner similar to that employed by other Brahmi-derived scripts. Consonants have an inherent /a/ vowel sound. Consonant conjuncts are not formed. Syllable structure is C(V)(F) consonant followed by optional vowel sign and/or optional final consonant or virama.

Ordering
The arrangement of the consonants is basically Brahmic and turns up in numerous sources. No strong evidence has been found for any strong preference with regard to the relative order of the vowel signs and of the final consonants; indeed most secondary sources give contradictory evidence in their charts. A generic Brahmic relative ordering for these characters is used in the code chart, with the vowel signs following the consonants, and the final consonant signs following the vowel signs.

Naming
Character names use the usual UCS conventions for Brahmic scripts.

Digits and punctuation
Unique Rejang digits are unknown; Jaspan presents a letter written to him in Rejang by Ali Akbar which uses both Roman numerals in an ordered list and European digits in a date. Ali Akbar uses comma, full stop, and colon, as well as the unique REJANG SECTION MARK which he uses both at the beginning and end of paragraphs.

Linebreaking

Traditional texts tend not to use spacing, but Ali Akbar's letter to Jaspan does; NON-BREAKING SPACE can be used in *scriptio continua* and SPACE otherwise. Hyphenation has not been observed, but could only occur after an orthographic syllable.

Unicode Character Properties

```
A930;REJANG LETTER KA;Lo;0;L;;;;;N;;;;;
A931;REJANG LETTER GA;Lo;0;L;;;;;N;;;;;
A932;REJANG LETTER NGA;Lo;0;L;;;;;N;;;;;
A933;REJANG LETTER TA;Lo;0;L;;;;;N;;;;;
A934;REJANG LETTER DA;Lo;0;L;;;;;N;;;;;
A935;REJANG LETTER NA;Lo;0;L;;;;;N;;;;;
A936;REJANG LETTER PA;Lo;0;L;;;;;N;;;;;
A937;REJANG LETTER BA;Lo;0;L;;;;;N;;;;;
A938;REJANG LETTER MA;Lo;0;L;;;;;N;;;;;
A939;REJANG LETTER CA;Lo;0;L;;;;;N;;;;;
A93A;REJANG LETTER JA;Lo;0;L;;;;;N;;;;;
A93B;REJANG LETTER NYA;Lo;0;L;;;;;N;;;;;
A93C;REJANG LETTER SA;Lo;0;L;;;;;N;;;;;
A93D;REJANG LETTER RA;Lo;0;L;;;;;N;;;;;
A93E;REJANG LETTER LA;Lo;0;L;;;;;N;;;;;
A93F;REJANG LETTER YA;Lo;0;L;;;;;N;;;;;
A940;REJANG LETTER WA;Lo;0;L;;;;;N;;;;;
A941;REJANG LETTER HA;Lo;0;L;;;;;N;;;;;
A942;REJANG LETTER MBA;Lo;0;L;;;;;N;;;;;
A943;REJANG LETTER NGGA;Lo;0;L;;;;;N;;;;;
A944;REJANG LETTER NDA;Lo;0;L;;;;;N;;;;;
A945;REJANG LETTER NYJA;Lo;0;L;;;;;N;;;;;
A946;REJANG LETTER A;Lo;0;L;;;;;N;;;;;
A947;REJANG VOWEL SIGN I;Mn;0;NSM;;;;;N;;kaluan;;;
A948;REJANG VOWEL SIGN U;Mn;0;NSM;;;;;N;;kamitan;;;
A949;REJANG VOWEL SIGN E;Mn;0;NSM;;;;;N;;kamica;;;
A94A;REJANG VOWEL SIGN AI;Mn;0;NSM;;;;;N;;katiling;;;
A94B;REJANG VOWEL SIGN O;Mn;0;NSM;;;;;N;;;;;
A94C;REJANG VOWEL SIGN AU;Mn;0;NSM;;;;;N;;katulung;;;
A94D;REJANG VOWEL SIGN EU;Mn;0;NSM;;;;;N;;;;;
A94E;REJANG VOWEL SIGN EA;Mn;0;NSM;;;;;N;;kajina;;;
A94F;REJANG CONSONANT SIGN NG;Mn;0;NSM;;;;;N;;katulang;;;
A950;REJANG CONSONANT SIGN N;Mn;0;NSM;;;;;N;;duo deatas;;;
A951;REJANG CONSONANT SIGN R;Mn;0;NSM;;;;;N;;kajunjung;;;
A952;REJANG CONSONANT SIGN H;Mc;0;L;;;;;N;;;;;
A953;REJANG CONSONANT SIGN VIRAMA;Mc;9;L;;;;;N;;;;;
A95F;REJANG SECTION MARK;Po;0;L;;;;;N;;;;;
```

Bibliography

Jaspan, M. A. 1964. *Folk literature of South Sumatra: Redjang Ka-Ga-Nga texts*. Canberra: Australian National University.

Acknowledgements

This project was made possible in part by a grant from the U.S. National Endowment for the Humanities, which funded the Universal Scripts Project (part of the Script Encoding Initiative at UC Berkeley).

Figures

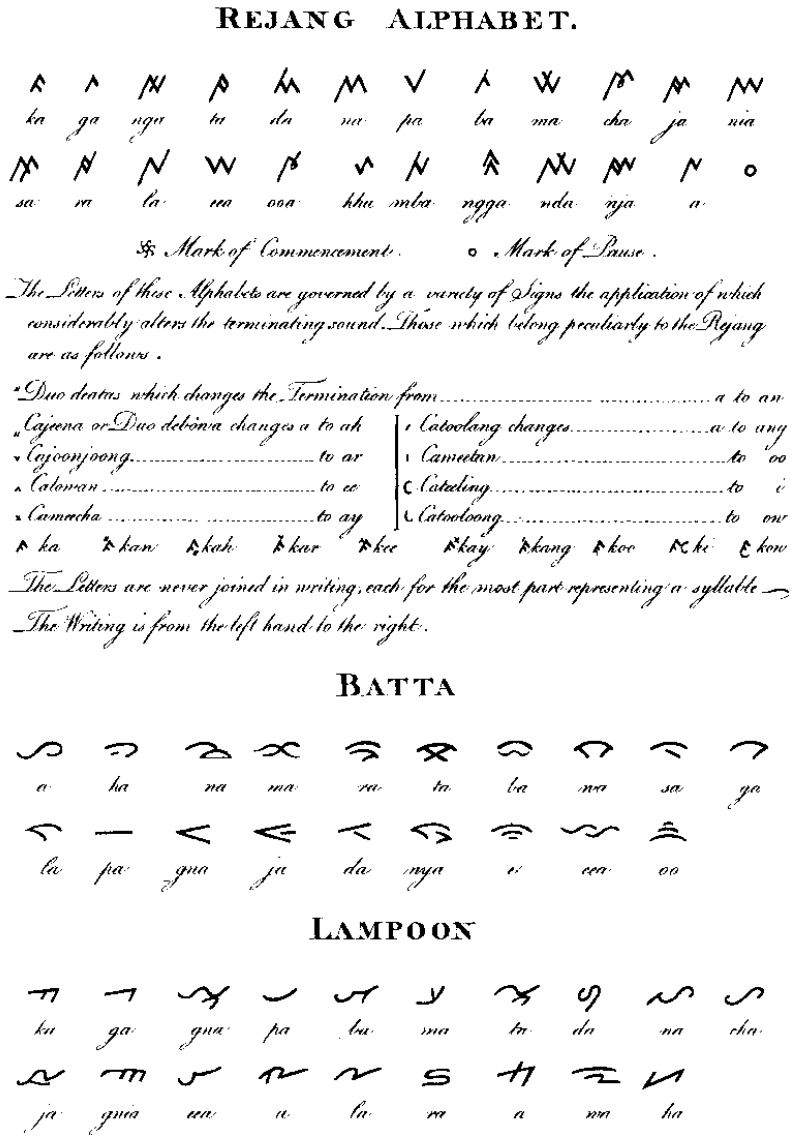


Plate 2 - Marsden's 1783 'Sumatran Alphabets'

Figure 1. Table of Rejang characters from Marsden's 1783 book *Sumatran alphabets*, as presented in Jaspan 1964. Marsden gives traditional names for the vowel signs, which Jaspan notes are no longer current with modern users of Rejang.

Text H - ALI AKBAR'S LETTER TO JASPAN [on paper]

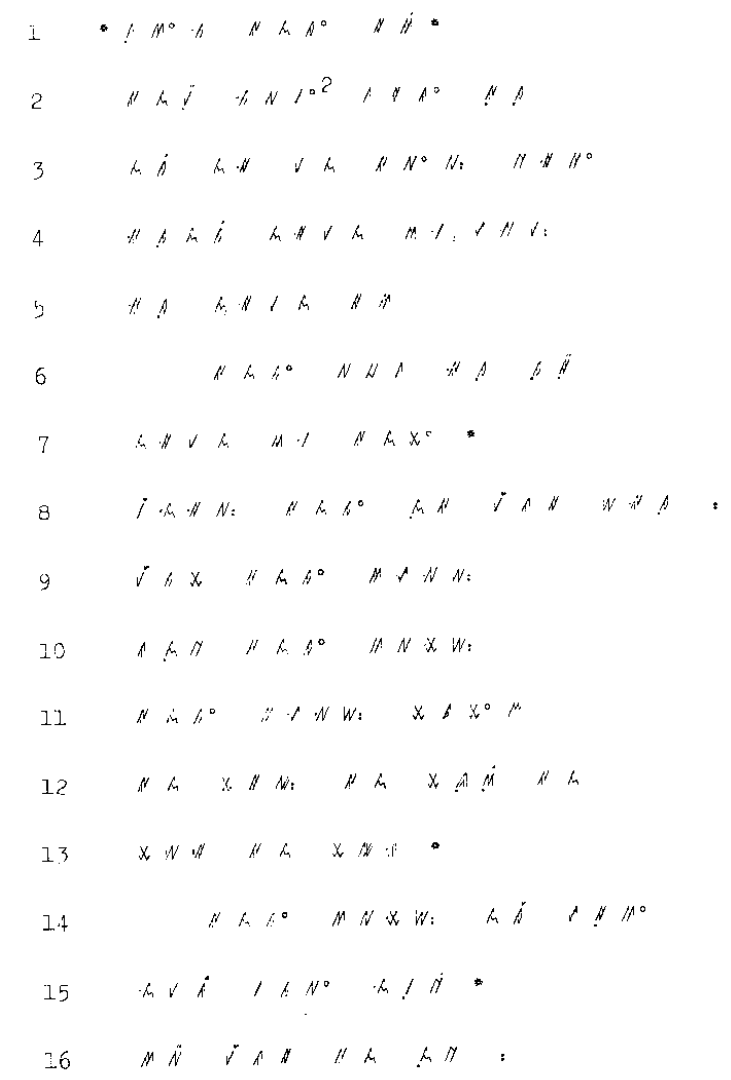


Figure 2. The beginning of Ali Akbar's letter to Jaspan, typeset and reproduced in Jaspan 1964. The REJANG SECTION MARK can be seen in lines 1, 7, 13, and 15.

Proposal for encoding the Rejang script in the UCS Michael Everson

TABLE XX - Row A9: REJANG

	A93	A94	A95
0			
1			
2			
3			
4			
5			
6			
7			
8			
9			
A			
B			
C			
D			
E			
F			

G = 00
P = 00

5

Proposal for encoding the Rejang script in the UCS Michael Everson

TABLE XX - Row A9: REJANG

hex	Name
30	REJANG LETTER KA
31	REJANG LETTER GA
32	REJANG LETTER NGA
33	REJANG LETTER TA
34	REJANG LETTER DA
35	REJANG LETTER NA
36	REJANG LETTER PA
37	REJANG LETTER BA
38	REJANG LETTER MA
39	REJANG LETTER CA
3A	REJANG LETTER JA
3B	REJANG LETTER NYA
3C	REJANG LETTER SA
3D	REJANG LETTER RA
3E	REJANG LETTER LA
3F	REJANG LETTER YA
40	REJANG LETTER WA
41	REJANG LETTER HA
42	REJANG LETTER MBA
43	REJANG LETTER NGGA
44	REJANG LETTER NDA
45	REJANG LETTER NYJA
46	REJANG LETTER A
47	REJANG VOWEL SIGN I (kaluan)
48	REJANG VOWEL SIGN U (kamitan)
49	REJANG VOWEL SIGN E (kamica)
4A	REJANG VOWEL SIGN AI (katiling)
4B	REJANG VOWEL SIGN O
4C	REJANG VOWEL SIGN AU (katulung)
4D	REJANG VOWEL SIGN EU
4E	REJANG VOWEL SIGN EA (kajina)
4F	REJANG CONSONANT SIGN NG (katulang)
50	REJANG CONSONANT SIGN N (duo deatas)
51	REJANG CONSONANT SIGN R (kajunjung)
52	REJANG CONSONANT SIGN H
53	REJANG VIRAMA
54	(This position shall not be used)
55	(This position shall not be used)
56	(This position shall not be used)
57	(This position shall not be used)
58	(This position shall not be used)
59	(This position shall not be used)
5A	(This position shall not be used)
5B	(This position shall not be used)
5C	(This position shall not be used)
5D	(This position shall not be used)
5E	(This position shall not be used)
5F	REJANG SECTION MARK

hex	Name

Group 00 **Plane 00** **Row A9**

6

A. Administrative

1. Title
Proposal for encoding the Rejang script in the BMP of the UCS.
2. Requester's name
UC Berkeley Script Encoding Initiative (Universal Scripts Project)
3. Requester type (Member body/Liaison/Individual contribution)
Individual contribution.
4. Submission date
2006-04-24
5. Requester's reference (if applicable)
6. Choose one of the following:
6a. This is a complete proposal
Yes.
6b. More information will be provided later
No.

B. Technical – General

1. Choose one of the following:
1a. This proposal is for a new script (set of characters)
Yes.
Proposed name of script
Rejang.
1b. The proposal is for addition of character(s) to an existing block
No.
1c. Name of the existing block
2. Number of characters in proposal
37
3. Proposed category (A-Contemporary; B.1-Specialized (small collection); B.2-Specialized (large collection); C-Major extinct; D-Attested extinct; E-Minor extinct; F-Archaic Hieroglyphic or Ideographic; G-Obscure or questionable usage symbols)
Category A.
4a. Proposed Level of Implementation (1, 2 or 3)
Level 2
4b. Is a rationale provided for the choice?
Yes.
4c. If YES, reference
Rejang uses Brahmic vowelsigns.
5a. Is a repertoire including character names provided?
Yes.
5b. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document?
Yes.
5c. Are the character shapes attached in a legible form suitable for review?
Yes.
6a. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard?
Michael Everson.
6b. If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:
Michael Everson, Fontographer.
7a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?
Yes.
7b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?
Yes.
8. Special encoding issues: Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?
Yes.
9. Additional Information: Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script.
See above.

C. Technical – Justification

1. Has this proposal for addition of character(s) been submitted before? If YES, explain.
Yes, a preliminary proposal was submitted in N3023.
2a. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?
Yes.

2b. If YES, with whom?
Richard McGinn, Associate Professor Emeritus of Linguistics and Southeast Asian Studies, Department of Linguistics, Ohio University.
2c. If YES, available relevant documents
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?
Rejang is used on the island of Sumatra in Indonesia.
4a. The context of use for the proposed characters (type of use; common or rare)
Used to write the Rejang language.
4b. Reference
5a. Are the proposed characters in current use by the user community?
Yes.
5b. If YES, where?
In Sumatra.
6a. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?
Yes. Positions A930-A95F are proposed.
6b. If YES, is a rationale provided?
Yes.
6c. If YES, reference
Contemporary use and accordance with the Roadmap.
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?
Yes.
8a. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?
No.
8b. If YES, is a rationale for its inclusion provided?
8c. If YES, reference
9a. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?
No.
9b. If YES, is a rationale for its inclusion provided?
9c. If YES, reference
10a. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?
No.
10b. If YES, is a rationale for its inclusion provided?
10c. If YES, reference
11a. Does the proposal include use of combining characters and/or use of composite sequences?
Yes.
11b. If YES, is a rationale for such use provided?
Yes.
11c. If YES, reference
Vowel signs and consonant signs.
11d. Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
No.
11e. If YES, reference
12a. Does the proposal contain characters with any special properties such as control function or similar semantics?
No.
12b. If YES, describe in detail (include attachment if necessary)
13a. Does the proposal contain any Ideographic compatibility character(s)?
No.
13b. If YES, is the equivalent corresponding unified ideographic character(s) identified?