Leaks in the Unicode pipeline: script, script, script…

Michael Everson, Everson Typography, www.evertype.com

Some 52 scripts are currently allocated in the Unicode Standard. This reflects an enormous amount of work on the part of a great many people. An examination of the Roadmap shows, however, that there are at present no less than 96 scripts yet to be encoded! These scripts range from large, complex and famous dead scripts like Egyptian hieroglyphs, to small, little-known but simple scripts like Old Permic. But, importantly, about a third of the scripts are living scripts which are intended to go on the BMP. Over the past few years, some implementers and standardizers alike have expressed their concern about how much work remains to be done. "When will the standard be finished?" they have asked. This talk will give a brief overview of the history of Unicode allocations, and discuss the standardization process required for newly-allocated scripts, including discussion of the kinds of procedural, political, and implementation issues which are met with in trying to get a script standardized. The different types of scripts remaining to be encoded will be discussed with regard to the ease with which they can be both encoded and implemented. Finally, a proposal for the way forward will be given.

Many of you will know me as the author of a rather large number of proposals to add various scripts and characters to the standard. One of our colleagues recently sent me an e-mail saying that he considered me to be to Unicode script proposals what the inherent vowel is to Indic consonants!

Though the title of my talk is "Leaks in the Unicode pipeline", I don't mean to imply that there are errors or faults in our encoding process – I just mean to underscore the fact that a good many scripts remain to be encoded, and that, given the current rate of demand or urgency for them, as well as the lack of resources to facilitate the work, we can expect these scripts to be added slowly, like drips out of a pipe. It will doubtless take many years before they are all encoded. Whether that is a desirable situation is a question I am raising.

History of allocations

Unicode was conceived as a solution to the chaos of formal character set standards, industrial standards, and font hacks by creating a single universal set containing, in layman's terms, all the letters of all the alphabets of all the languages of the world. It began with a set of the major writing systems of the world: European alphabets, West Asian alphabets and abjads, East Asian logographies and syllabaries, and Central and South Asian abugidas. It was believed, back in 1988, that a single 16-bit plane – the BMP (Basic Multilingual Plane) – would suffice to meet the world's encoding needs.

It quickly became clear that 65,000 code positions were not sufficient, particularly as a large number of punctuation, mathematical, technical, and general symbol systems would need to be encoded as well. With Unicode 3.1, three more planes intended for characters were admitted: the SMP (Supplementary Multilingual Plane), the SIP (Supplementary Ideographic Plane), and the SSP (Supplementary Special-purpose Plane). During this time, the list of scripts deemed acceptable for encoding grew, culminating in a paper by Joe Becker and Rick McGowan in 1993. By October 1998, I had conceived of the idea of drawing up a set of graphic roadmaps, which give the current allocations and show the empty slots into which new scripts could fit. These roadmaps are altered as each new script is encoded, or as information becomes available about the expected size of the unencoded scripts. In 2001, the roadmaps were adopted as formal, informative documents on the Unicode web site.

As of today, there are 52 scripts currently allocated in the Unicode Standard, in addition to the various symbol sets used for mathematical, technical, musical, and other purposes. The roadmaps show, however, that there are at present no less than 96 scripts which remain to be encoded – and about a third of these are intended for the BMP. It is worth asking how much work remains to be done, as some implementers and standardizers have been concerned that an unfinished standard is in some respects unstable.

Standardization process for new scripts

One way of gauging the work remaining to be done is to look at the processes required to get a script encoded. The most efficient procedure is to have experts work with experienced standardizers to prepare a preliminary proposal. This proposal is examined by the Unicode Technical Committee and ISO/IEC JTC1/SC2/WG2, and may be modified once or more than once before a final proposal is accepted for SC2 balloting. During the voting period, the proposal may undergo further revision if necessary. The more comprehensive the work done by the experts and standardizers in the initial stages, the easier the road is later on. The UTC and WG2 committees themselves do not do the work of preparing and perfecting proposals; it is participants in those committees who do, between meetings. Fortunately, we have honed our skills in script analysis and encoding, and we are better at ensuring that all the right questions are asked so that initial proposals can be quite mature.

We have established a number of criteria which assist us in determining which scripts belong on the roadmap and which do not. Chief among these criteria is the requirement of modern

users to exchange data using the scripts. Undeciphered scripts are at present not considered good candidates for encoding, as the character/glyph model cannot be applied to them, since, obviously, we can't know what the glyphs stand for. A few of these scripts (such as Indus and Rongorongo) are kept on the roadmap because we do have some idea of the apparent glyph repertoire, but it is unlikely that formal encodings will be pursued absent actual decipherment. A few scripts (such as Aymara, Paucartambo, and Woleai) have not been roadmapped because, despite their appearance in books about writing systems, we have at present no real information about them at all.

Tengwar and Cirth, two scripts created by J. R. R. Tolkien – one of the most influential writers of the twentieth century – to represent the languages he created for use in his literary universe, *are* considered to be candidates for encoding, because scholars and enthusiasts study both his published words and his manuscripts, create new texts in these scripts both in his invented languages and in modern languages, and have expressed an interest in making use of a standard for interchanging data written with them. The Klingon "alphabet", on the other hand, was rejected, because although there is a rather large community of rather enthusiastic users of the Klingon language, they invariably prefer to use the ASCII-based orthography of that language for communication and interchange, and use the Klingon font almost exclusively to create gifs for web pages. (Were this not the case, the Klingon script could well have been taken seriously. It certainly has more active users than other constructed languages, such as Volapük, have. One Bulgarian colleague undertook the task of translating Lewis Carroll's "The Hunting of the Snark" into Klingon – in a version which scans and rhymes in the same way as the original!)

We have found that a set of characters and names by itself is not enough to enable a script to be encoded. Character properties and behaviour are important for an actual implementation of a script. Such information is standardized by the UTC but not formally taken into account by WG2. However, by addressing it in the proposal it becomes possible not only to encode the characters, but to guide developers in making fonts and other resources that work properly. Synchronization between the Unicode Standard and ISO 10646, requires that such information be available to the UTC. It is therefore recommended that *all* proposals include, as explicitly as possible, information about character properties and behaviour, as well as complete multi-level ordering information. Directionality and positioning of combining characters are important and necessary for Unicode implementation. Ordering information for the UCA (Unicode Collation Algorithm) and ISO/IEC 14651 make it possible for users of scripts to get the behaviour they require.

Compatibility considerations are also brought to bear, sometimes trivially affecting encoding proposals, sometimes profoundly affecting them. As I pointed out in 1995 after the first Yi proposal, Yi ought to have been considerably smaller, since 25% or so of the encoded characters are simply existing base characters with a single diacritic. But compatibility with a Chinese Standard for Yi prompted the Chinese to request their separate encoding. Still, if we were to find additional syllables of the mid-level tone, it would require us to explicitly encode them given the accepted model – a potential disadvantage for Yi implementation.

Trivial effects of more-or-less political considerations can be seen in the Myanmar and Sinhala blocks. Representatives from Myanmar insisted that the script not be given its traditional name in English – Burmese – and required the Sanskrit-specific characters to be separated out of the normal sorting order. Similarly, the character names for Sinhala are not easily recognizable as their Brahmic *akṣara* names are not given, but instead their Sinhalese names. This helps the Sri Lankans assert their identity, but makes the identification of character by name more difficult for non-Sri Lankan implementers. There isn't much that can be done about political pressure levied on the encoding process, even when such pressure comes in after the fact, as occurs from time to time, as has been seen in recurring discussion about Arabic presentation forms and the Brahmic shaping model. But often, delays can be avoided if script experts work together with experienced standardizers, as we know many of the pitfalls, and can ask the right questions early on in order to avoid dispute later on. Syriac, Gothic, Osmanya, Limbu, and Deseret are examples of scripts for which we had good information early on. Aegean is one where we had significant scholarly input subsequent to the initial proposals.

Types of scripts

Turning to the 96 as-yet unencoded scripts, it's important to describe them. After the publication of the roadmaps, some standardizers became alarmed by what seemed to be a huge number of scripts yet to be encoded, and expressed their concern (as in SC2 N3243) about the effort it would require to encode them and the possible burden on implementers. And it will indeed take effort, and resources, to do the work. But such concerns are less well-founded than they appear at first. What I hope to do here is describe the as-yet unencoded scripts in categories, which should illustrate that a good many of them, while unique writing systems, do not differ much from already-encoded scripts. Therefore, it can be seen that the great majority of them present no particular difficulties in implementation.

21% of the unencoded scripts are simple left-to-right (LTR) alphabets and syllabaries. Of these, some of them make use of combining marks, but none present great difficulties for implementation than any LTR alphabet already encoded. In addition to Vai, Bamum, and Mende there are a number of other African syllabaries which have recently come to my attention but all but one of them would belong in this category. Here, and below, I give the names of these scripts, with an asterisk * preceding scripts which are used actively to represent modern spoken languages, and a dagger † preceding scripts which have active liturgical or other modern use.

Old Persian Cuneiform, *Hittite Hieroglyphs/Luvian*, *Cypro-Minoan*, *Lycian*, *Iberian*, †*Coptic*, †*Glagolitic*, *Old Permic*, *Elbasan*, *Büthakukye*, †*Hungarian Runic*, †*Cirth*, *Bassa*, **Vai*, *Bamum*, *Mende*, **Naxi Geba*, *Yi Extensions*, **Pollard Phonetic*, **Blissymbols.*

24% of the unencoded scripts are right-to-left (RTL) abjads and syllabaries. Some of these are similar to Hebrew, though a few of them have complex ligature shaping as Arabic does. Kharoshthi follows the Brahmic shaping model, though it is an RTL script. In January 2001 I proposed a unification of a number of Semitic scripts, reducing the number of scripts in the roadmap (WG2 N2311).

Meroitic, *Phoenician*, *Lydian*, *Carian*, †*Samaritan*, *Numidian*, **Tifinagh*, *North Arabic*, *South Arabian*, *Aramaic*, *Kharoshthi*, *Pahlavi*, *Avestan*, *Orkhon*, *Uighur*, *Balti*, *Yezidi*, **N'ko*, *Elymaic*, *Hatran*, †*Mandaic*, *Palmyrene*, *Nabataean.*

34% of the unencoded scripts are Brahmic abugidas; none are more complex than any we have encoded to date. Siddham is often written in vertical columns. Modern users of Meithei prefer a radically different sorting order than the usual Brahmic one. Some researchers have suggested that there are a great many more historical Brahmic scripts than we have identified.

Brahmi, *Turkestani*, *Soyombo*, †*Siddham*, *Chola*, *Chalukya (Box-Headed)*, *Satavahana*, **Newari*, **Siloti Nagri*, *Saurashtra*, *Takri*, *Kaithi*, *Modi*, **Meithei*, **Lepcha*, *Landa*, **Cham*, *Ahom*, *Khamti*, *Pyu*, **Chakma*, **New Tai Lü*, **Lanna*, **Viêt Thái*, *Javanese*, *Balinese*, *Rejang*, **Batak*, **Buginese*, **Kayah Li*, **Ol Cemet'*, **Sorang Sompeng*, **Varang Kshiti.*

4% of the unencoded scripts are logographic scripts. They are large, but offer no implementation difficulties.

Tangut Ideographs, *Kitan Small Script*, *Kitan Large Script*, *Jurchin.*

8% of the unencoded scripts are undeciphered scripts and true ideographic scripts. Sumerian Pictograms may be unifiable with their Sumero-Akkadian Cuneiform descendants. Proto-

Elamite has been partially deciphered. It has been suggested that scripts which have not been deciphered not be encoded at all. It is not certain that the true ideographic scripts are strictly speaking encodable, as their use as "text" is ambiguous. We know little at present about Aztec Pictograms. Naxi Tomba characters are well-defined and catalogued and a good deal of work on the "texts" has been published.

Sumerian Pictograms, *Proto-Elamite*, *Byblos*, *Indus*, *Aztec Pictograms*, **Naxi Tomba*, *Rongorongo*.

And finally, 8% of the unencoded scripts are scripts with complex features, requiring either novel rendering models or a great deal of analysis to determine what comprises the basic character set. Cuneiform is simple enough to render but it will take a long time and a lot of work to choose which signs are unifiable and which must be encoded separately. Egyptian and Mayan Hieroglyphs are both quite complex to render, and it has been suggested that markup is the best way to handle a good bit of it. These two scripts do appear, in my analysis, to have the same essential structure, and will use the same encoding model – though Mayan fonts will have to be very, very complex indeed. Egyptian is likely to be encoded in stages, the first stage being the basic Gardiner set (about 800 characters), and the second comprising a much larger set – though it may be *decades* before the compilation, analysis, and unification of that set is complete! (Not very surprising, considering that Egyptian was a living writing system for 4,300 years.) 'Phags-pa is written in vertical columns. Pahawh Hmong deserves further study as far as input methods and ordering are concerned because of the unique way it writes phonetic syllables. Chinook is based on a manual shorthand system which is likely to be quite complex to analyse. Sutton SignWriting is written in an extremely complex vertical matrix incorporating markers for handshapes, facial expressions, positions and movements. It is, however, implemented in software with a standard interchangeable text-format. A version of XML is being developed for SignWriting which is likely to be useful in rendering Unicode-encoded characters.

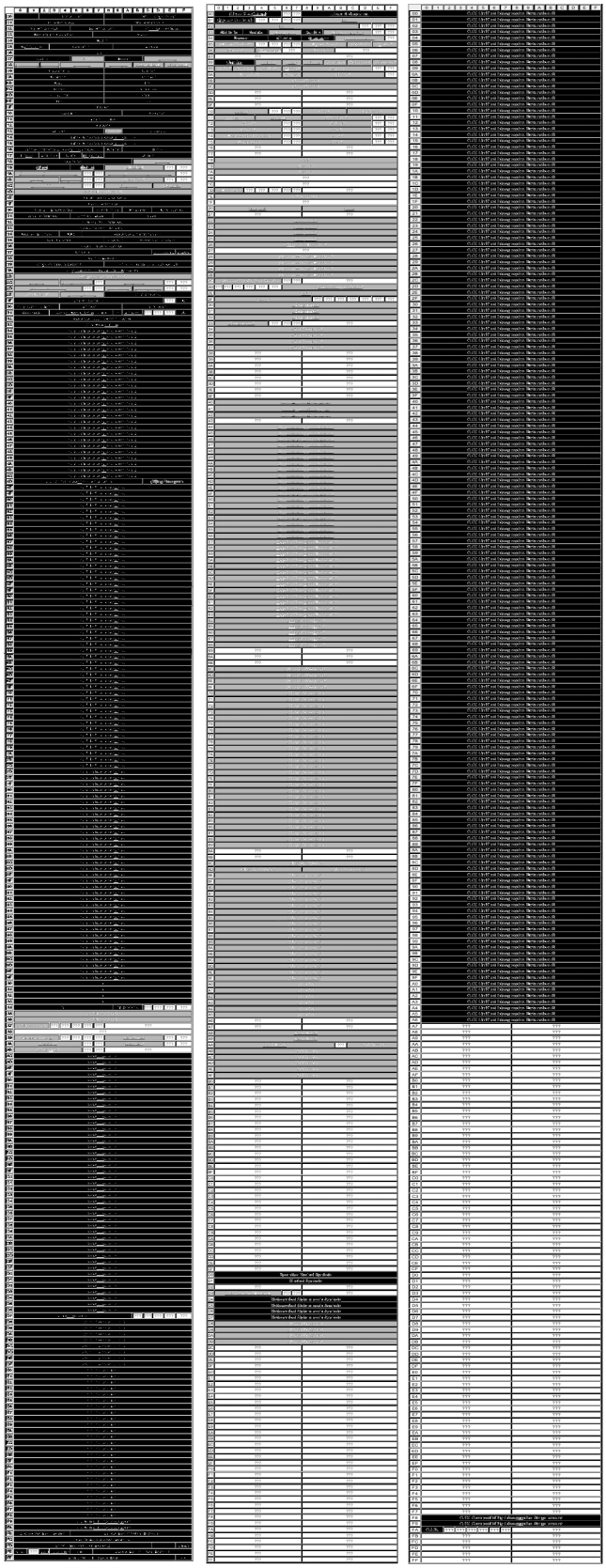
Egyptian Hieroglyphs, *Sumero-Akkadian Cuneiform*, *Mayan Hieroglyphs*, *'Phags-pa*, **Pahawh Hmong*, *Chinook*, †*Tengwar*, **Sutton SignWriting*.

The way forward

Currently the encoding process for all these scripts is initiated on more or less a first-come-first-served basis. We are endeavouring to focus on living scripts roadmapped to the BMP, but in some cases good information has been available for scripts in the SMP and it has been appropriate to serve the interested user community which helped provide information. The biggest problem we face is finding the resources to do the work of script analysis, proposal preparation,

and, in most cases, production of fonts for the charts.

A look at the roadmaps in overview shows the scale of the task we face rather dramatically. To the right the roadmaps for the BMP, SMP, and SIP are shown in their entirety, to demonstrate graphically the situation as it is at present. The blackened blocks show that the BMP is nearly full, the SMP only beginning to be filled, and the SIP more than half full but with a good bit of room to accommodate additional characters. (WG2's Ideographic Rapporteur Group is working on adding more, and is well-supported in its efforts.) The greyed blocks show the 96 scripts which remain to be encoded. About 30% of those are in the BMP.



It seems reasonable to suggest that the sooner these scripts are encoded, the happier the IT community, the JTC1 Member Bodies, and the user community will be, for the standard will be, at last, a good deal more stable, apart from the odd script, character, or symbol which will turn up from time to time. I propose that it would greatly facilitate the process if the IT community could fund the activity of experts to put in the time and effort required to achieve our goal of a complete and stable standard sooner rather than later. Doing so would certainly be in the interests of that community – as a way of plugging the leaks in the Unicode Pipeline.

Leaks in the Unicode pipeline: script, script, script…

I am happy to report that just recently a project, the Scripts Encoding Initiative, has been established through the Department of Linguistics at UC Berkeley to raise funds specifically for these purposes, that is, to oversee the creation of script proposals for missing scripts and to produce freely-available fonts for certain scripts. The project is being run in conjunction with the Unicode Vice President, with the goal that proposals will be able to get approved by the Unicode Technical Committee without much intervention on the part of the Committee. For those who would like to see long-term stability in the universal character set, this is an opportunity for you (and your company) to effectively support the effort.

Cheques (in U.S. dollars) should be made out to "UC Regents", with "Script Encoding Initiative" written on the memo line, and sent to:

Script Encoding Initiative
c/o Deborah Anderson
Department of Linguistics
1203 Dwinelle Hall #2650
University of California at Berkeley
Berkeley, CA 94720-2650
USA

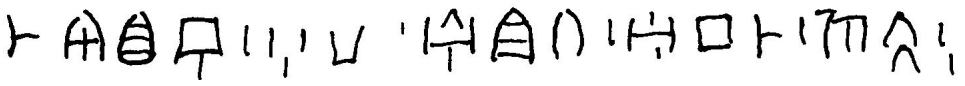
If a letter accompanies the cheque, it should specify that the money is a "gift." Donations are tax-deductible in the US within the limits as prescribed by law; 2% of donations go automatically to the campus Development Office, as is usual for gifts to the University of California at Berkeley. Questions may be directed to Deborah Anderson at the above address, or by e-mail to: dwanders@socrates.berkeley.edu

Straightforward LTR alphabets and syllabaries

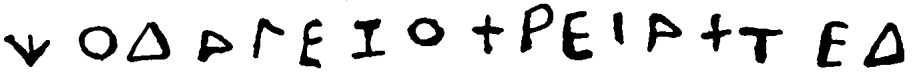
Old Persian Cuneiform

Hittite Hieroglyphs/Luvian

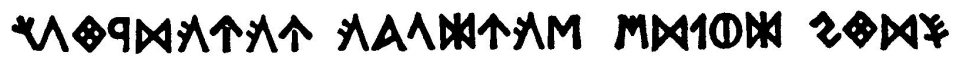
Cypro-Minoan



Lycian



Iberian

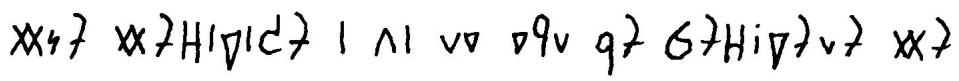


†Coptic

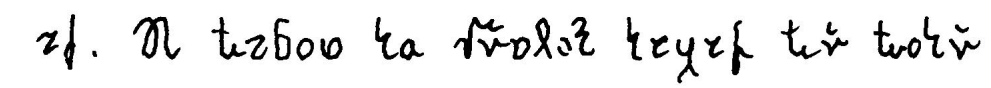
†Glagolitic

Old Permic

Elbasan



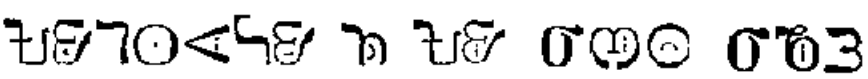
Büthakukye



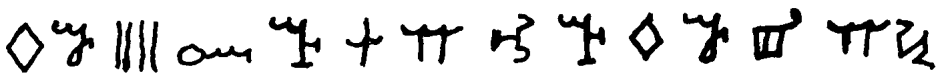
†Hungarian Runic

†Cirth

Bassa

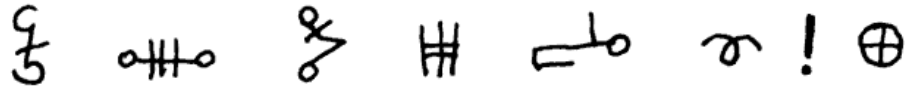


*Vai

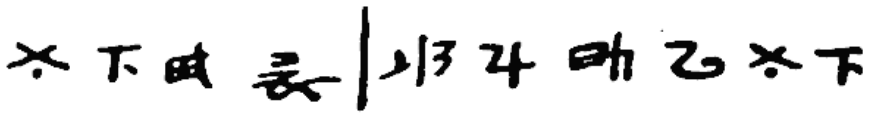


Bamum

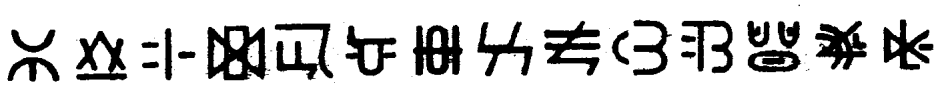
Mende



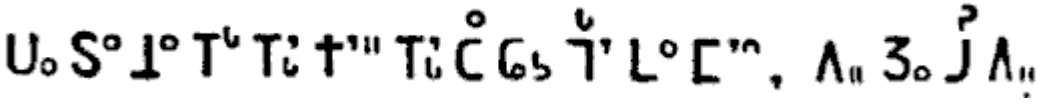
*Naxi Geba



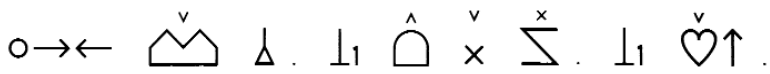
Yi Extensions



*Pollard Phonetic



*Blissymbols



Straightforward RTL abjads and syllabaries

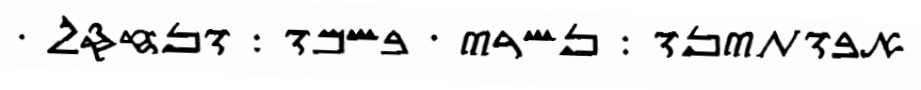
Meroitic

Phoenician

Lydian

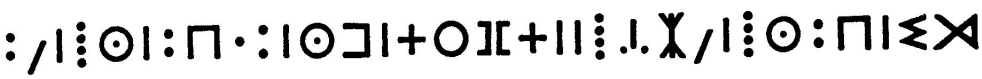
Carian

†Samaritan

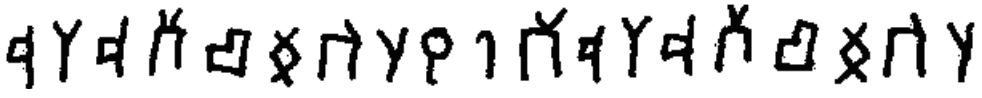


Numidian

*Tifinagh



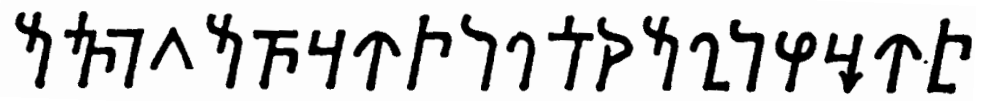
North Arabic



South Arabian

Aramaic

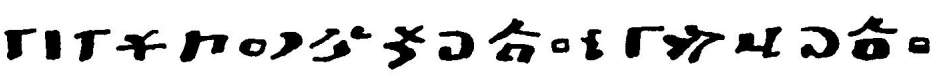
Kharoshthi



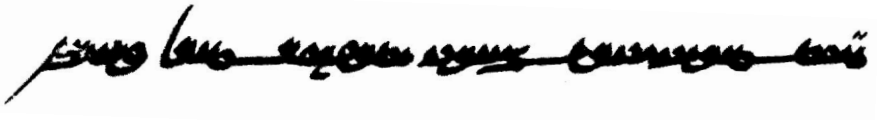
Pahlavi

Avestan

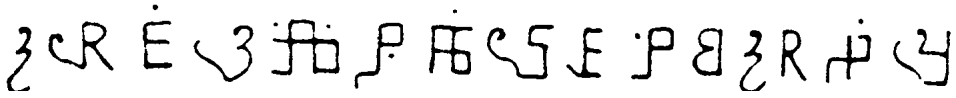
Orkhon



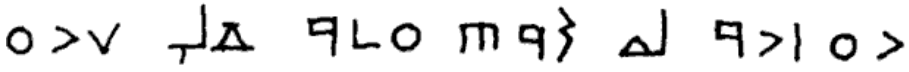
Uighur



Balti

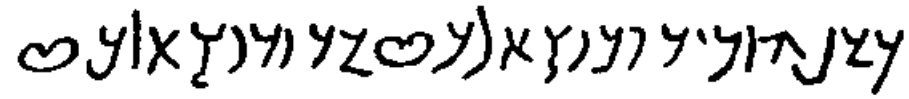


†Yezidi

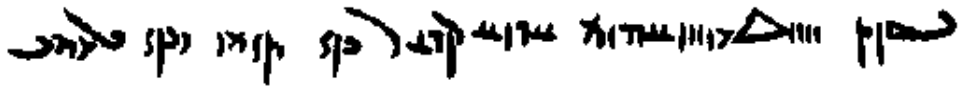


*N'ko

Elymaic

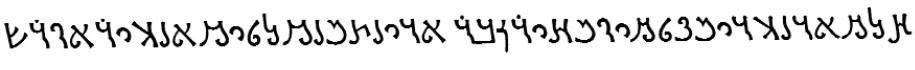


Hatran

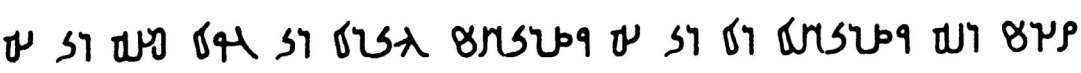


†Mandaic

Palmyrene



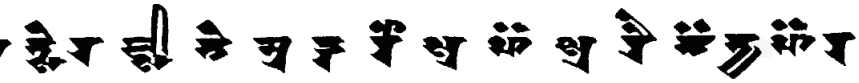
Nabataean



Straightforward Brahmic abugidas

Brahmi

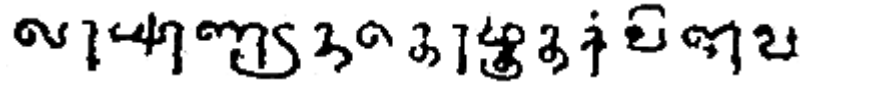
Turkestani



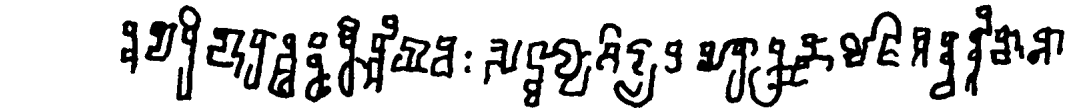
Soyombo

†Siddham

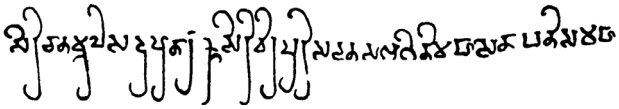
Chola



Chalukya (Box-Headed)



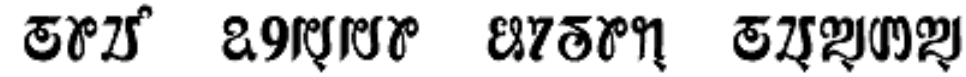
Satavahana



*Newari

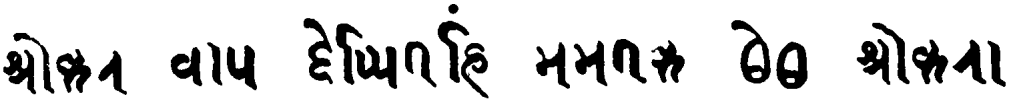
*Siloti Nagri

*Saurashtra

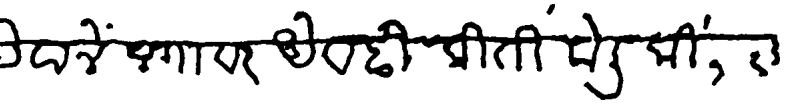


Takri

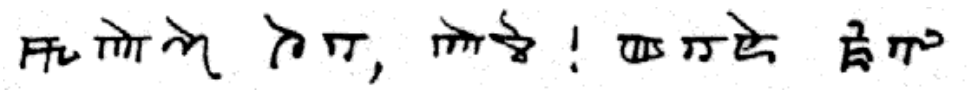
Kaithi



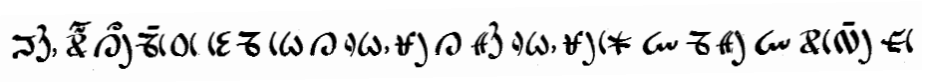
Modi



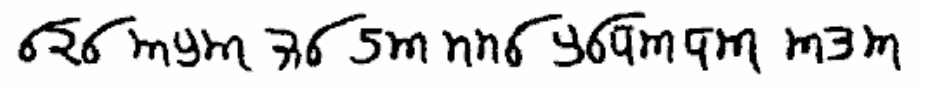
*Meithei



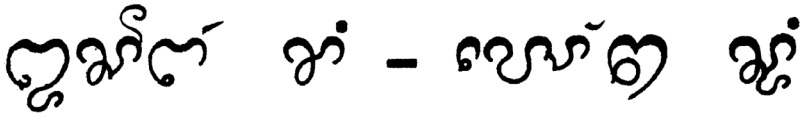
*Lepcha



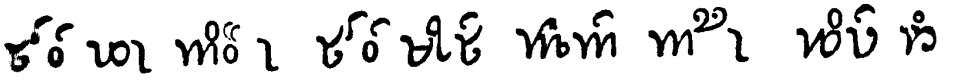
Landa



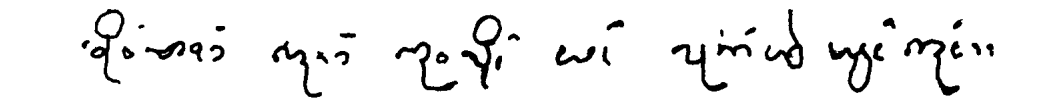
*Cham



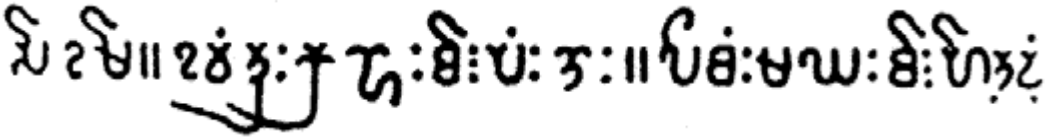
Ahom



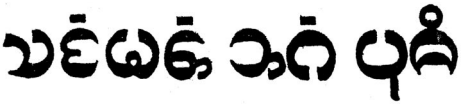
Khamti



Pyu

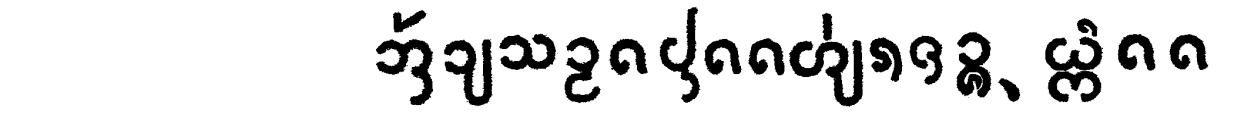


*Chakma

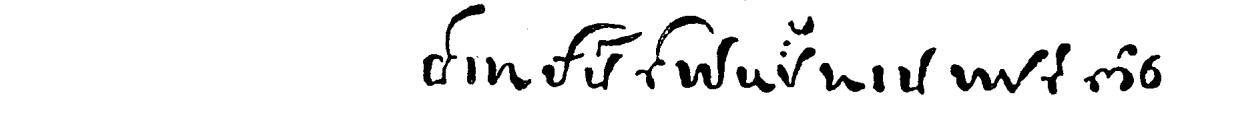


*New Tai Lü

*Lanna

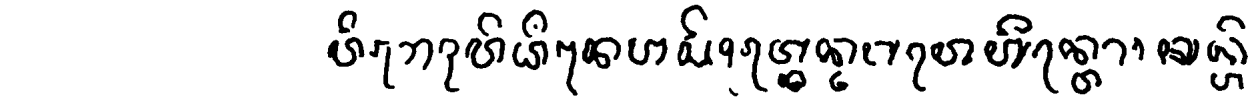


*Viêt Thái

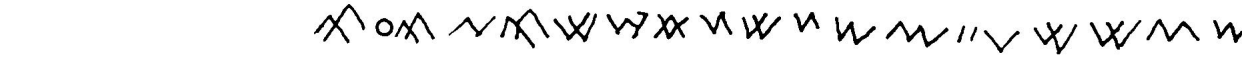


Javanese

Balinese



Rejang

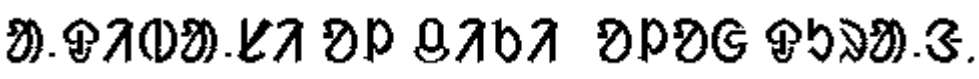


*Batak

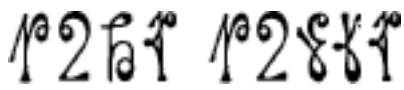
*Buginese

*Kayah Li

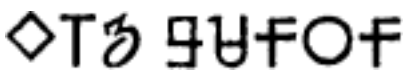
*Ol Cemet'



*Sorang Sompeng



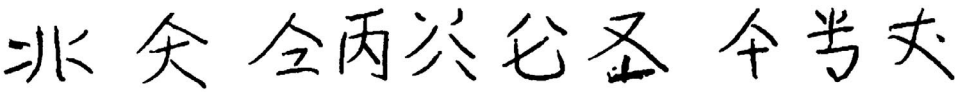
*Varang Kshiti



Straightforward logographic scripts

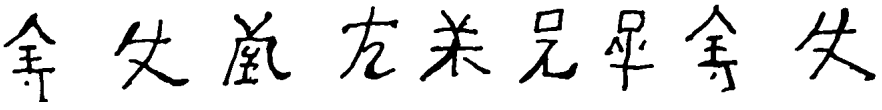
Tangut Ideographs

Kitan Small Script



Kitan Large Script

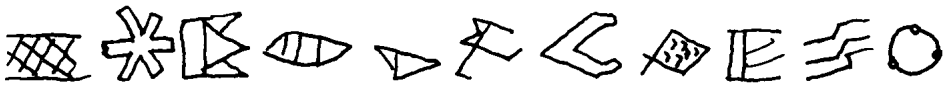
Jurchin



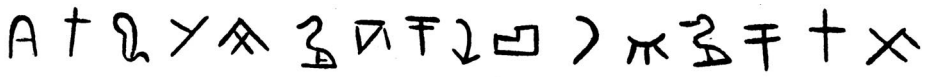
Undeciphered scripts and true ideographic scripts

Sumerian Pictograms

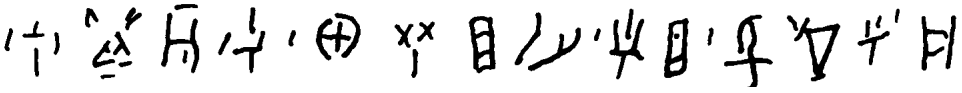
Proto-Elamite



Byblos



Linear A



Indus

Aztec Pictograms

*Naxi Tomba

Rongorongo

Scripts with complex features

Sumero-Akkadian Cuneiform

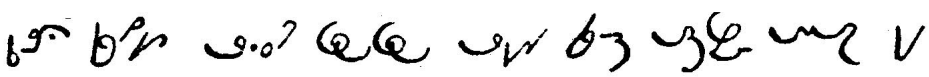
Egyptian Hieroglyphs

Mayan Hieroglyphs

’Phags-pa

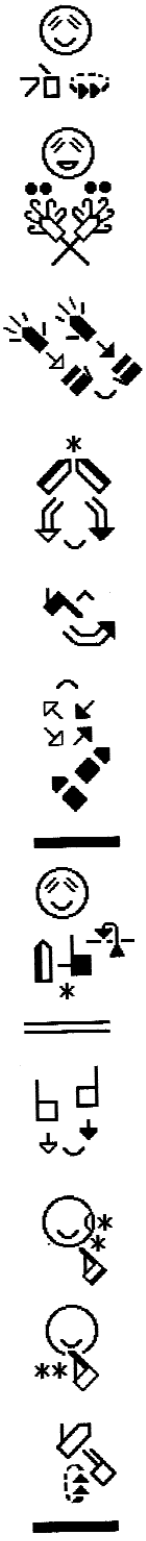
*Pahawh Hmong

Chinook



†Tengwar

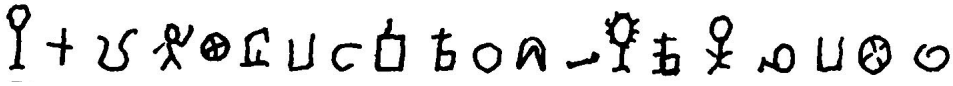
*Sutton SignWriting



Scripts not roadmapped

Aymara

Paucartambo



Woleai