Unicode Encoding Proposal for the Indus Script – Early Draft

Devajyoti Sarkar

April 3, 2023

Tokyo, Japan

DOI: https://doi.org/10.5281/zenodo.7794001

Author Note

Devajyoti Sarkar: https://orcid.org/0009-0007-3432-2082

Mastodon: @dsarkar@universeodon.com

Correspondence concerning this article should be addressed to Devajyoti Sarkar. Email: customercare@vamra-vaikhanasa.com

Abstract

The Indus script is not deciphered yet. There is no consensus on what is and what is not a character. It is likely that many new characters will emerge in future excavations. Despite its aim of universality, ***the Unicode Standard considers writing systems for which insufficient information is available to enable reliable encoding of characters to be out of its scope***. We are not at the stage where we can make a proposal for Unicode Encoding for the Indus script. We may not get there for decades to come. In the meantime, we still need some kind of a common encoding for people to record and share text. This section will outline a proposal for an encoding in the *Private Use Area (PUA)* of the *Unicode Basic Multilingual Plane (BMP)*. The hope is that in the future this can be leveraged to an RFC to the community that defines what Unicode calls a "*Private Agreement*". This will allow the community to agree upon a shared encoding and enable tools and technology like fonts and IMEs. This is an ***Early Draft*** of the proposal. Much of it will only make sense when one can read the words and texts of the Indus script. I will publish a decipherment separately. In the absence of a bilingual inscription, we have no authoritative source to determine the validity of an interpretation. This paper describes a general method to do so based on the fit to data. The upcoming decipherment meets these criteria. The following proposal and discussion are based on this decipherment and may be somewhat cryptic to read until it is available. I humbly ask your patience in this regard.

Keywords: indus script, unicode encoding, proposal, semantics, bidi, collation, akhyats, compounds, phono-semantic compounds, canonical sequence, canonical order

Quick Summary

This proposal defines the range U+E8FF to U+F1AF in the Unicode BMP for the Indus script.

- a proposed Character Encoding Table (Appendix A)
- size of range: 2,225
- number of Characters defined: 638
- number of Characters reserved: 1587

This proposal attempts to conform to or enable the following *Design Principles of Unicode*:

- **Characters, not glyphs**: It specifies character encodings, not glyph forms.
- **Stability**: Once assigned, cannot be reassigned and key properties are immutable.
- **Logical order**: Articulating the default for memory representation.
- **Specification of Bi-directional (Bidi) text handling**
- **Efficiency**: Focusses on making Unicode text simple to parse and process.
- **Plain text**: It represents plain text not rich or structured text.
- **Universal repertoire**: The Unicode Standard provides a single, universal repertoire.

This proposal does not address or considers **out of scope** the following:

- Semantics: so that characters have well-defined properties.
- Unification: unifying duplicate characters within scripts across languages (e.g. Egyptian Hieroglyphs)
- Dynamic composition: accented forms to be dynamically composed.
- Convertibility: guaranteeing convertibility with other standards.
- Unicode Character Database and formal semantics
- Tailorings for Unicode Common Locale Data Respository (CLDR)
- Specification of Collation methods or weights.
- Emoji

Background

The Indus Script has survived to us in the form of seals with relatively short inscriptions of 4-5 characters. The earliest seals discovered so far come from about 3,500BC but the majority are from the Mature Harappan Period of 2,600-1,900BC. Only a few thousand seals have been found tiil date, out of which 40% of the seals have an illustration apart from text, of those more than half are that of a Unicorn bull. There is no long-form script available to us.

This relative scarcity of information has made it difficult to decipher the script. Unlike Egyptian Hieroglyphs which had the Rosetta stone and Cuneiform which had the Behistun inscription, there has been no bilingual inscription discovered so far. In its absence, we have no way

to learn from someone who knew. Any attempt to decipher the script, including the current one, is a guess that may or may not be correct. There is no logical way to eliminate possibilities down to a single one.

Farmer et al. suggest that the Indus script was not a script (Farmer et al., 2004). Among a number of arguments, they make two statistical arguments to support this. They say that the inscriptions are too small compared to other known languages and that there were too many "singleton" characters (characters that occur just once in the whole corpus).

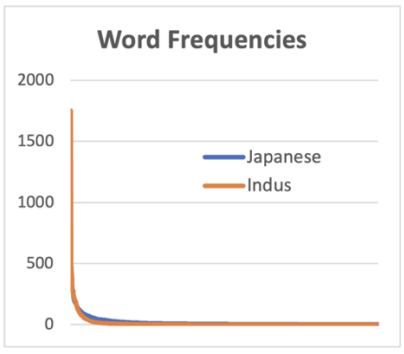
It is important to point out that these points are neither necessary nor sufficient to conclude that this is not a script. In many ways, the Indus seals are quite similar to Japanese hankos or personal seals. Hankos are used to sign contracts or authorize bank transactions. Most hankos range between 2-4 characters and are enough to serve an economy of the scale of Japan.

Figure 1

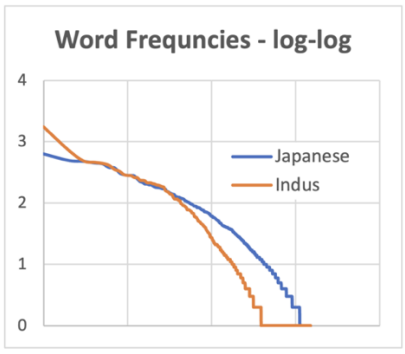
A comparison of Japanese and Indus scripts

Description	Indus	Japanese
Inscriptions	4,615	8,850
Corpus Size	16,678	26,040
Unique Characters	585	1,507
Average Length	**3.61**	**2.94**
Singletons	**33.85%**	**25.95%**

(a)



(b)



(c)

As an example, if we were to take the train station names across Japan and make "seals" of them, we would get about 8,850 names as in Figure 1(a). This is somewhat comparable to the 4,615 Indus seal inscriptions in the Indus corpus. We find that each seal in Japan would, on the average,

have about 3 characters, whereas the corresponding number is closer to 4 in the Indus case. The modern script is actually more terse. This is not surprising because it has a greater number of unique characters. Both have a large number of singletons, although the Indus script has a slightly higher number at 33.85% vs. 25.95%. It is important to note that even in the Japanese case, ***for random samples of 4,615 names we get an average of 31.65% singletons***. As we see in Figure 1(b) and 1(c), their relative character distributions also have similar long tails.

The Japanese corpus has 1,507 unique Kanjis versus the 586 unique characters in the Indus. While the Japanese number is larger, it does represent the full range of place names of one of the largest economies, more than an order of magnitude greater in population compared to the Indus and a much later stage in the evolution of the script - i.e. the literature available in Japanese today and the complexity of what it represents is considerably greater than the Indus script would have had in its time. Even though such metrics cannot represent the complexity of either script, but all things considered, they are surprisingly comparable.

Characters, not glyphs

Unicode considers characters to be the smallest component of a written language that has semantic value. As an example, a given character or piece of text may be written in multiple fonts, each of which can have very different appearances, but they all represent the same text, the same characters. So when you search for a word, you don't need to remember or specify what font you wrote it with, just the word because the text is represented in characters, not glyphs. It is the character that receives a Unicode code.

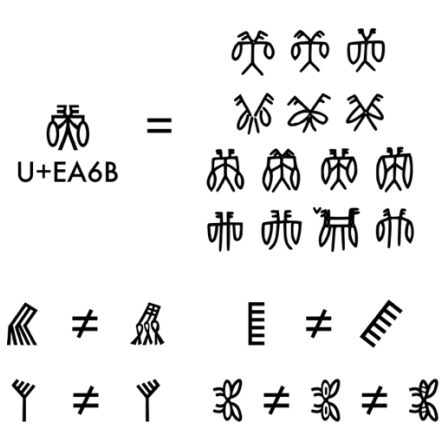
There are two major kinds of scripts in broad use - phonetic scripts and ideographic scripts. Phonetic scripts, like the alphabet, represent sounds and we write a word by how we hear it. In a logographic or Ideographic script, like the Han characters of CJKV (Chinese-Japanese-Korean-Vietnamese) scripts, a character represents a meaning. They generally have one root meaning and retain their semantic content across linguistic boundaries. Instead of understanding what text means by hearing it, we can recognize it by seeing it. It gives communication a visual metaphor.

The Indus script is Ideographic. The characters encode meaning. It is peculiar that even as far as today we do not have a name for these characters. I will call them **Akhyats**. “Akhyat” in Sanskrit means "that which brings to cognition" or “announced, named”. The more familiar “Vikhyat” means “famous or easily recognized” and comes from the same root.

Akhyats are in many ways like CJKV ideographs, and like them, they were most likely used to encode multiple languages. In fact, this is also why Indus inscriptions are terse. They need fewer characters. Unlike a phonetic script like the alphabet where two characters can at most represent 26x26 (676) combinations, in an ideographic script like the Indus, two characters could potentially represent (638x638) more than 400,000. In a CJKV script this number is many orders of magnitude larger. Like any script, not all combinations are valid, but there is no doubt that one needs fewer ideographs to specify the same word compared to phonetic characters.

Figure 2

Different Akhyats versus different glyphs



(a)

Scholars	Script size
Langdon (1931)	288
Hunter (1932)	149
Von Meriggi (1934)	270
Dani (1963)	537
Koskenniemi and Parpola (1982)	396
Mahandevan (1977)	417
Fairservis (1992)	419
Bryan Wells (1998)	587

(b)

One of the primary tasks to create a Uncode encoding of Akhyats is to define what an Akhyat is. When do two glyphs mean the same thing and when do they not. In Figure 2 (a), the top shows the same Akhyat in different glyphs or styles found across the corpus. In the bottom we see different Akhyats that look similar but are different. Akhyats can often look quite different without any change in its semantic content and yet can look very similar and mean something different. This

can sometimes be reasonably obvious to see, but many times it is not. In Figure 2(b) we see the many widely varying Akhyat counts from different studies because of this (Possehl, 2002).

The only way to reliably define the set of Akhyats in the Indus script is by reading them. One needs a decipherment. In the absence of a bilingual inscription, in the absence of an authoritative source telling us what the Akhyats meant, any attempt at listing them is an interpretation. It is a guess, and guesses can be wrong. We need some way to validate whether one is right.

There is a way to do this. We create a dictionary definition for all Akhyats at once. We then see if we can meaningfully read all the seals and inscriptions given this definition. If the Akhyat's glyph drawing also happens to convey the same meaning, it is further confirmation, but oftentimes the match is not specific. Han characters evolved into abstract strokes that are unambiguous rather than remain pictograms with many potential interpretations.

Although such a dictionary definition is not guaranteed to be unique or correct, in practice it is incredibly difficult to do. It is not easy to "fit" a set of meanings so that one can read all the text. When it occurs, even in part, such an assignment is quite robust and changes very slowly. If the set of seals that one can read crosses 50% of the corpus, this stability is not likely to be a coincidence. If it further provides reasonable explanations for known facts and gives verifiable and/or verified predictions of things not known, it becomes ***a generic methodology to validate a decipherment even when there is no authoritative source***.

In its essence, this method is inspired by the way a machine learns in AI. We are, in effect, trying to fit a model to the data. It is more like a scientific theory than a "logical" argument. And unlike many arguments that selectively focus on some elements while ignoring the rest, this addresses all of them, all the time. It can give surprisingly good results when it finds a good fit.

But it can only speak to validity, not truth.

Using this, if two glyphs occur in roughly the same contexts and mean roughly the same, they may be the same Akhyat, even if they are written in different glyphs. Similarly, if the same

Akhyat reads differently across too many contexts, it is possible that it corresponds to more than one Akhyat and we might be missing something subtle.

The above method does not prescribe how one finds a reasonable mapping in the first place, only the means to evaluate it. I have spent over 10 years attempting to decipher the script and for the past 4 years I have been able to read more than 80% of the Indus characters and published seals. This interpretation represents a large-scale attempt to do the above.

The definition of Akhyats in this proposal is based on this "decipherment". But it is by no means normative. Although the bulk has not changed for years, every time I go through the corpus, I discover things I missed or got wrong. And there is always to chance of a breakthrough needing large scale changes. One of the Design Principles of Unicode is **Stability**, that once a character is assigned it cannot be reassigned, its definition must be immutable (*Submitting Character Proposals*, 2016). This is a high bar. We are not there yet, and we are not going to get there any time soon. This proposal is a Private Agreement and can afford to change. The hope is that over time we will get to a definition stable enough to consider a formal Unicode proposal.

We are choosing to use the PUA of the Unicode Basic Multilingual Plane for this. We have a choice of three possible PUAs, one in the BMP and two larger areas in Plane 15 and 16 respectively. We are going with BMP because we do not need the larger space and BMP is more likely to have support in older e-readers and devices/software. We are reserving a range of 2,225 codepoints and is about the same order as Joyo Kanji (2,136) which is a normative list from the Japanese Ministry of Education to define the minimum requirement to get into college (*Joyo Kanji*, 2023). We have assigned 638 so far, but this has been done with enough headroom to accommodate future Akhyats that we discover as well as establish areas for legacy characters to provide compatibility with or deprecate. While one would like to create a safe playpen to "*Fail fast, fail often, learn and try again*", the truth is that any change may be disruptive to texts that have already been written in a previous version. This needs to be managed in terms of *Versioning and Deprecation* and *Backward Compatibility* where possible. Ideally, one wants a slow-moving target.

Direction of Writing and Unicode BiDi

Figure 3

The Indus Script is strongly Right-to-Left.

Word	Freq	Reverse	Freq
"𐫂	174	𐫂"	5
"𐫄	90	𐫄"	4
𐫃𐫅	145	𐫅𐫃	3
𐫆𐫃	40	𐫃𐫆	1
𐫇𐫈	45	𐫈𐫇	3
𐫉𐫊	14	𐫊𐫉	0

(a)

(b)

(c)

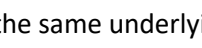
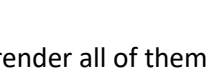
Note: The Seal Images are taken from (Marshall, 1931)

One of the aspects of the Indus script where there is relative consensus among scholars is the direction of writing (Mahadevan, 1977). It is Right-to-Left for the most part. This is apparent in some seals where the sculptor ran out of space as they went to the left, as we can see in the seal of Figure 3(b). It is even more evident in Figure 3(c) where the text wraps Right-to-Left, Top-to-Bottom, and then Left-to-Right; where the bottom line is written upside down and is the only place in all the corpus where these Akhyats are upside down. The table shows some of the most frequent words in the script and there is an overwhelming preference for one direction. The last line in the table shows the first two Akhyats of the seal of Figure 3(c). Thus means these characters were read Right-to-Left. The third line, also in the seal, is also in the same Right-to-left direction. The Right-to-Left direction

quickly extrapolates across the entire corpus from the strong direction preference and the reading direction in this seal.

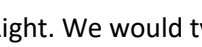
There are actually four separate possibilities for the direction of writing in the corpus:

- Right-to-Left: e.g. 
- Left-to-Right (Mirrored): e.g. 
- Left-to-Right (Symmetric): e.g. 
- Left-to-Right (Shuffled): e.g. 

It is important to note that all four directions are merely different visual representations of the same underlying text.  ,  , are the same sequence of characters. We can render all of them within standard Unicode through the proper choice of direction and mirror forms. This should cover all currently available inscriptions.

Right-to-Left is the most common pattern but we find Left-to-Right as well as mirror or not mirrored forms.

1. Right-to-Left: 94.2%
2. Left-to-Right: 5.8%
    - Mirrored: 2.7%
    - Symmetric: 2.7%
    - Shuffle: 0.4%

Unicode standardizes the Logical order of Characters in the text. This is the sequence of the characters in memory. Essentially this is how we store our text in a file, how algorithms like search match text and many other things. The specification says that we should store the characters similar to how we would enter them on a keyboard. In the case of the Indus script this would be Left-to-Right. We would type  as ① ⓪ ② ‖ ③ . is the sequence that would be stored on file as well as made available to all programs. The text rendering directionality would be implemented by Unicode's standard BiDi algorithms.

In our case, since we are assigning the Akhyats to code points in the PUA, we cannot leverage the standard Unicode BiDi algorithm. We cannot specify directionality within each Character's properties as we cannot define them. But, we can use standard css styles and Unicode directional formatting characters. Please note that these directional characters may have varying levels of support in real software. Many ereaders and others may strip them out. They can also be an attack vector for malware and social engineering plays in some cases (Corfield, 2021). As an example, although not a representative one, a file called “name-txt.exe” can look like “name-exe.txt” if one places a right-to-left override where the hyphen is. It may get somebody to execute it by thinking it was a document. Varying softwares may intentionally highlight Bidi characters as a control character by displaying a box prominently around. This may make it awkward to read the text. The Unicode Bidirectional Algorithm formatting characters can be equivalently represented by stylesheets or markup. Conflicts can arise if markup and explicit formatting characters are used together. Where available, markup should be used instead of the explicit formatting characters. As an example:

- Right-to-Left:<span style='direction: rtl; unicode-bidi:isolate-override;’>𑀀𑀁𑀂

  gives 𑀂𑀁𑀀.

- Left-to-Right:<span style='direction: ltr; unicode-bidi:isolate-override;’> 𑀀𑀁𑀂

  gives 𑀀𑀁𑀂.

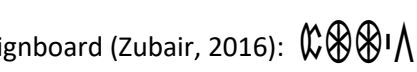
Since the PUA cannot leverage standard Unicode mechanims for handling mirror forms, the **ida font** (Sarkar, 2023), for example, comes in two varieties – Ida-Regular and Ida-Left-To-Right. The same character can be reflected simply by changing the font.

There are many limitations in the proposed scheme due to it using the PUA which has no formal support from Unicode. But it is a starting point. Please note that we need to encode text in the standard Unicode Logical order (left-to-right) so as to have uniform text representation without which software and tools like search cannot function. It will be impossible to change this along the

way. ***We must follow Unicode standard right from the start*** regardless of the teething difficulties we may have with limited support for the PUA encoding.

Canonical Decomposition

An Akhyat corresponds closest to an Ideograph in the CJKV sense but there are punctuation characters, digits and digit-like characters, and perhaps most importantly compounds, sometimes similar to the phono-semantic compounds in CJKV and sometimes not.

First of all, there are rare instances where the Indus script is used phonetically where each Akhyat represents a single phonetic character. Perhaps the most famous of these is the Dholavira signboard (Zubair, 2016):  which is an example of phonetic usage. Unlike abjads like the original Phoenician script, here both vowels and consonants are represented as well as long and short vowels are distinguished. The semantic characteristic of the Akhyat is not completely lost in such usage. The Akhyats chosen are “poetically related” in meaning with relation to the word but they primarily represent sound. The same name in Akhyats would be written with two Akhyats. In this inscription, it is interesting to observe that the Akhyats are in their proper Right-to-left direction even though the text itself is written from Left-to-Right. It is tempting to think the Akhyats face you as you read them and in 99% of the inscriptions they do, but they don’t in this one.

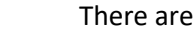
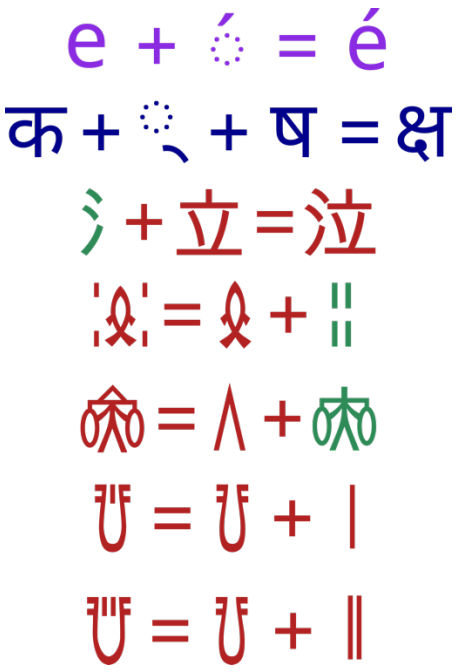
There are some other examples as like  , and others. But less than 1% of the inscriptions discovered so far are used in this phonetic way. For the rest of this proposal we will ignore them and focus on outlining Akhyats that are ideographs.

Figure 4

Compounds in the Indus Script



Unicode distinguishes between base characters and those that are derived from the combination of other characters. As an example, in Figure 4, we find an accented e which has a different Unicode code than e or its accent but is essentially a "precomposed" character. Unicode defines a canonical decomposition into its constituents such that two strings may be compared reliably. These normalized decompositions are defined for all such precomputed characters. There is another kind that is found in Indic and complex scripts where two characters combine to form a new glyph but this precomposed form is not assigned a Unicode codepoint. Instead it is handled by the font and rendered on the fly. There are well-defined rules for such composition that typically come from the language itself and encoding them with codepoints would unnecessarily use up many thousands of slots.

In CJKV scripts base forms combine to form a new character that is assigned a separate Unicode codepoint. In the third line of Figure 4 we have the Han character for "tear", like a tear drop, which is composed from the characters of "water" and "to stand". While the composition is poetically formed from the other two characters there is no doubt it is a separate meaning. One character contributes the core meaning from which the final character is derived and the other

specifies the sound. This is called a Phono-Semantic Compound. The character in green, water, specifies the meaning water and is called the radical of the phono-semantic compound. It serves as the key to looking up the word in a dictionary. In fact, more than 90% of the Chinese characters come from this same generative process (*Chinese character classification*, 2023). Each such compound is assigned a separate Unicode codepoint because it has a different meaning.

Indus Compounds and Decomposition

Indus script also has phono-semantic compounds although more often than not the compounds are not phono-semantic. In the 4th and 5th lines of Figure 4, we can decompose the Akhyat according to phonetic and radical components in exactly the same fashion as a Han compound. The resulting ideograph can and should be assigned a Unicode codepoint separate from the constituent Akhyats because it is not an obvious composition of the two, and has a meaning different from both. But the majority of the compounds in the Indus script are not phono-semantic. They are often closer to compound nouns like "computer science" where both characters participate in the specification of meaning. Examples of this are in lines 6 and 7. A basic breakdown of the types of Indus compounds is as follows:

1. **Base Akhyats** - 265
2. **Compound Akhyats** - 315
3. **Phono Semantic Akhyats** - 55
4. **Punctuation**: 3 Akhyats - 𒑰 𒑱 𒑲

So how does one handle these compounds? Can they be precomposed forms like the accented "e" or can they be rendered as a font artifact like the ligatures of a complex script. In order to answer this we need to note a fundamental different between all ideographic scripts and phonetic ones, one deals with sounds and the other with meanings. In the case of sound we can compose adjacent characters in a sequence because sound itself composes locally, that is, there is no long range correlation between a sound and one that occurred some time ago. The same is not true with ideographs as the composition of two adjacent characters is dependent on context that could be defined much earlier in the sequence.

Let us take the bottom two lines in Figure 4. 𑋟=𑋞+𑋠 can we canonically break this compound down into those two Akhyats whenever we encounter them? Are they canonically equal to the compound Akhyat? Interestingly for this Akhyat it may well be true. This is generally not the case. There are only 2 inscriptions where we encounter the two constituent Akhyats separately. 206 times they are precomposed into the compound, so effectively, the Indus scribes had already precomposed them.

The same is not true of the second example 𑋡=𑋞+𑋢. The compound occurs only 41 times in the corpus whereas the consitituents occur separately 68 times like in 𑋞𑋢𑋪. In actual fact, it is not possible to combine the two Akhyats into the compound in this case because it would completely change the meaning. A rough analogy of this can be seen in English. "computer science" is different from "science computer" so the sequence order matters; this is guaranteed by the Logical order of Unicode. But, is "(computer science) department" the same as "computer (science department)" - most probably not. This combination is not merely dependent on sequence adjacency but has a longer range correlations to a broader context. As an “algebra” it is neither "commutative" nor "associative".

Indus compounds cannot be canonically decomposed, in the Unicode sense, to its constituent base Akhyats. ***They must be assigned separate codepoints.*** I suspect this is generally true for all Ideographic scripts.

This is not to say that having a representative decomposition is not useful. As an example, a search engine may want to enrich indus text to match components, like a search for 𑋠 should match both 𑋟 and 𑋞𑋠. But when we encounter compounds in the Indus script, like the phono-semantic case, they have a meaning that is different from the mere composition of its base Akhyats, and need a separate Unicode codepoint to represent them. “computer science” means something specific and different from an “intersection” of “computer” and “science”. From this perspective they are similar to Han characters and we can treat them exactly in the same way. In fact, it almost feels like the

phono-semantic nature of Han characters is a natural and later stage of evolution from what we see in the Indus.

A Road to Canonical Order

How do you sort Indus text? This is a matter of great significance. Without the "alphabetical order" it would be hard for us to search for books in a library or even look up a word in a book's index. One could argue that the "Qwerty" system is more efficient in someways but having a generally accepted scheme has a value independent of the value of the scheme itself. Indeed, there can be multiple competing orders but it is important to have one widely adopted, a universal order. It is likely that the ordering of the Akhyats in the Unicode encoding will be the defacto universal order.

Unlike the ordering in a script with a few dozen characters where any order can do, a script with several hundred or thousands needs to be well-thought through. In the Indus case this has immediate relevance because in the absence of IMEs that can covert phonetic input to the corresponding character, we have no option but scrolling through a list of all of them every single time we want to input one.

In the 2nd century AD, Xu Shen organized the first etymological dictionary he called Shuowen Jiezi by selecting 540 radicals, the semantic component of a phono-semantic compound (*Shuowen Jiezi*, 2023). This served as the basis for partitioning the dictionary. As an example, the character for "tear" in the Figure 4 would be listed only under the "water" radical.

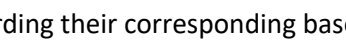
Mei Yingzuo's 1615 dictionary Zihui made two further innovations. He reduced the list of radicals to 214, and arranged characters under each radical in increasing order of the number of additional strokes – the "radical-and-stroke-count". This method is still used in most present-day dictionaries.

While the number of Akhyats discovered so far is much less, in the hundreds rather than tens of thousands for Han characters, it is still enough to make search non-trivial and memorizing a linear order impractical. One cannot really use the "radical" method as the use of phono-semantic

compounds were not as wide spread. Also, there is much more visual information in the Indus glyph that was abstracted away in the later Han characters as that script grew more complex. But the idea of classes or grouping based on related meaning can find a similar expression in the Indus case.

Etymology Classes and Systematic Polysemy

This proposal partitions the space of Indus Akhyats into **108 Etymology Classes as** in Appendix B. The classes are derived from commonality based on meaning. There are a few large blocks, grouping the classes themselves. The **U block** is essentially the "tomb of the unknown character". In any inscription, where the character is missing or damaged or we simply cannot read it, we use this character to specify that the character is not known. The **N block** is organized based on numbers although the Indus used numbers for various purposes in the script. The **P Block** corresponds to Akhyats related to peoples in the Indus civilization. The **G Block** organizes Akhyats related to places or geographies. The **PxG Block** is a set of place names that are associated to a people as France is to the French. The **O Block** contains various other groupings. Each of these classes have considerable headroom to incorporate new Akhyats while maintaining the same overall order.

It is hard to describe Etymology classes without explaining what each means. A full explanation of this section will need to wait for the decipherment book to be published. But the logic being used may be visible just from the glyph forms to a limited extent. As an example, the four Akhyats  in block P-7 are obviously related. All the classes in the N-Block (N1 to N18) have a similar relationship based on the numbers they represent. In general, various compounds are grouped according their corresponding base form like  in G-22. Phono-semantic compounds like  are grouped according to the semantic component rather than the sound component  . This has the added benefit of distributing them over a number of blocks instead of clustering them in one. It is important to realize that unlike the Han dictionaries phono-semantic compounds do not have the same level of either systematic or standardized use in

the Indus script. Block P-6 clusters phono-semantic compounds based on their sound rather than radical and I suspect this is indeed how the Indus viewed them because the most important Akhyat in their glyphs is the sound component. In the case of noun compounds, like , we organize it by the narrower component, in this case by instead of . This is like cataloging "Brooklyn Bridge" under "Brooklyn" rather than "Bridge" because it would easier to find in things related to Brooklyn rather than all things "bridges". It is important to note that the narrower category is defined semantically rather than by frequency of usage. is found in 447 inscriptions rather than which is found only in 147. Appendix C has a breakdown of the Base Forms, Compound Forms and Phono-Semantic Compound forms in the Indus script.

I realize that this is a cryptic and incomplete description. I will come back to this once my decipherment book is published and explain each element in detail. But this is to give a flavor of the kind of considerations went into the design of the Etymology classes. There is no one right answer. Most of it comes from what kind of organization that was useful in finding Akhyats to input in texts.

There is another form of natural organization that is present in the Indus. This is Systematic Polysemy. Systematic Polysemy occurs in various forms in many languages. In English the word "rabbit" can mean both the animal as well as its meat. This is used systematically to generate similar related meanings across a number of words related to animals like chicken, fish, rabit, lamb, etc. Similarly, the Indus script uses various glyph transforms to generate a form of systematic polysemy. This does not really have a counterpart in the more abstract Han characters. There are the following types - **Chiral Forms**, **Bent Forms** and **Black-and-White Forms**. You can find a listing of each type in Appendix D.

Let us take versus . The central Akhyat of each is a reflected image of the other. These are called Chiral forms. They are different Akhyats but very closely related to one another. They really differ in just one aspect. In a sort, we would ideally like to have these two right next to

one another. We achieve this by placing them in the same Etymology Class as well as right next to one another. This allows for the standard unicode encoding itself to sort them together. So everywhere we encounter Chiral forms we place them together. Everywhere there is a reasonable possibility that a chiral form may emerge in the future, we leave a gap in the encoding to accommodate it. The same is true of Bent Forms like 𒐖 𒐗 and of Black-and-White forms like 𒍅 𒍆.

Naturally, the proposed encoding will allow one to sort Indus text in a Universal Canonical order based on these Etymology classes and this order will be available to all software by default. In effect, it is the same as the alphabetical order in English. But it is important to remember that this is just one order. Just to end out the section on a Canonical Order, the following is taken from the Unicode standard for collation and applies just as well to the Indus case.

- Collation is one of the most performance-critical features in a system.
- Collation is not code point (binary) order.
- Collation is not aligned with character sets or repertoires of characters.
- Collation is not a property of strings.
- Stability is a property of a sort algorithm, not of a collation sequence.
- Collation order is not preserved under concatenation or substring operations, in general.

Versioning, Deprecation and Backward Compatibility

As our understanding of the Indus script evolves this encoding will need to evolve as well. This is an Early Stage Draft and will likely change in the coming weeks and months. Also, while this proposal is my best effort to create a reasonable encoding, it can be wrong. As I engage with the community and they have had chance to reflect their needs into this, newer and better methods of organization may emerge. We need a method to incorporate change until the requirements for change begin to subside. To put some method in the madness, I propose the follow versioning scheme.

Currently everything is pre-release meaning it does not have a version. It is given "as-is, where is" and everything can and will change. I am bootstrapping this over my various projects so as

to get to a reasonable initial state. Even after release of my decipherment and for a time that the community feels necessary this will continue in this pre-release state until people have had a chance to study it and engage with it.

Once released, versioning will start from 1.0.0 and go monotonically upwards. All versions will be retained in a central site. Each version should ideally release a corresponding font (where necessary) so that the community can see the change. Versioning is divided into ***major.minor.point*** releases.

A **major release** is one where we have breaking changes, which means that an existing character is changed - moved, deleted, etc. Every such character will be moved to a well-defined code in the **Deprecation Block O-6** and people can "search-replace" such characters in existing texts so that they continue to be readable. People may also be able continue using a previous version by embedding the corresponding font.

A **minor release** is one where new characters are added without affecting any existing character. This means that all existing texts can continue without change and any software may upgrade to the new characters, if necessary.

A **point release** means that there is no change in the character assignments but something else has changed and should not affect any existing documents or applications.

Since the versioning explicitly allows for breaking changes, backward compatibility can only be offered to a limited extent through the Deprecation block. Tools and even texts can note what version they require. People should consider embedding an equivalent font so that a text written will continue to be readable where appropriate.

This proposal is merely a suggestion to provide a starting point for discussion. Please feel free to get in touch and suggest improvements or changes.

References

Farmer, S., Sproat, R., & Witzel, M. (2004). The collapse of the Indus-script thesis: The myth of a literate harappan civilization. *Electronic Journal of Vedic Studies*, *11*(2), 19–58. https://doi.org/10.11588/ejvs.2004.2.620

Submitting Character Proposals. (2016, April 2). Unicode.org. https://www.unicode.org/pending/proposals.html

Marshall, S. J. (1931). *Mohenjo-daro and the Indus Civilization*. Arthur Probsthain.

Possehl, G. L. (2002). *The Indus civilization: A contemporary perspective*. AltaMira Press.

Joyo Kanji. (2023, March 10). Wikipedia. https://en.wikipedia.org/wiki/J%C5%8Dy%C5%8D_kanji

Sarkar, D. (2023). *The Ida Font*. Vamra Vaikhanasa Publishing. https://www.vamra-vaikhanasa.com/ida-font/

Mahadevan, I. (1977). *The Indus Script - Texts, Concordance and Tables*. Archaeological Survey of India.

Corfield, G. (2021, November 1). Trojan Source attack: Code that says one thing to humans tells your compiler something very different, warn academics. *The Register*. https://www.theregister.com/2021/11/01/trojan_source_language_reversal_unicode/

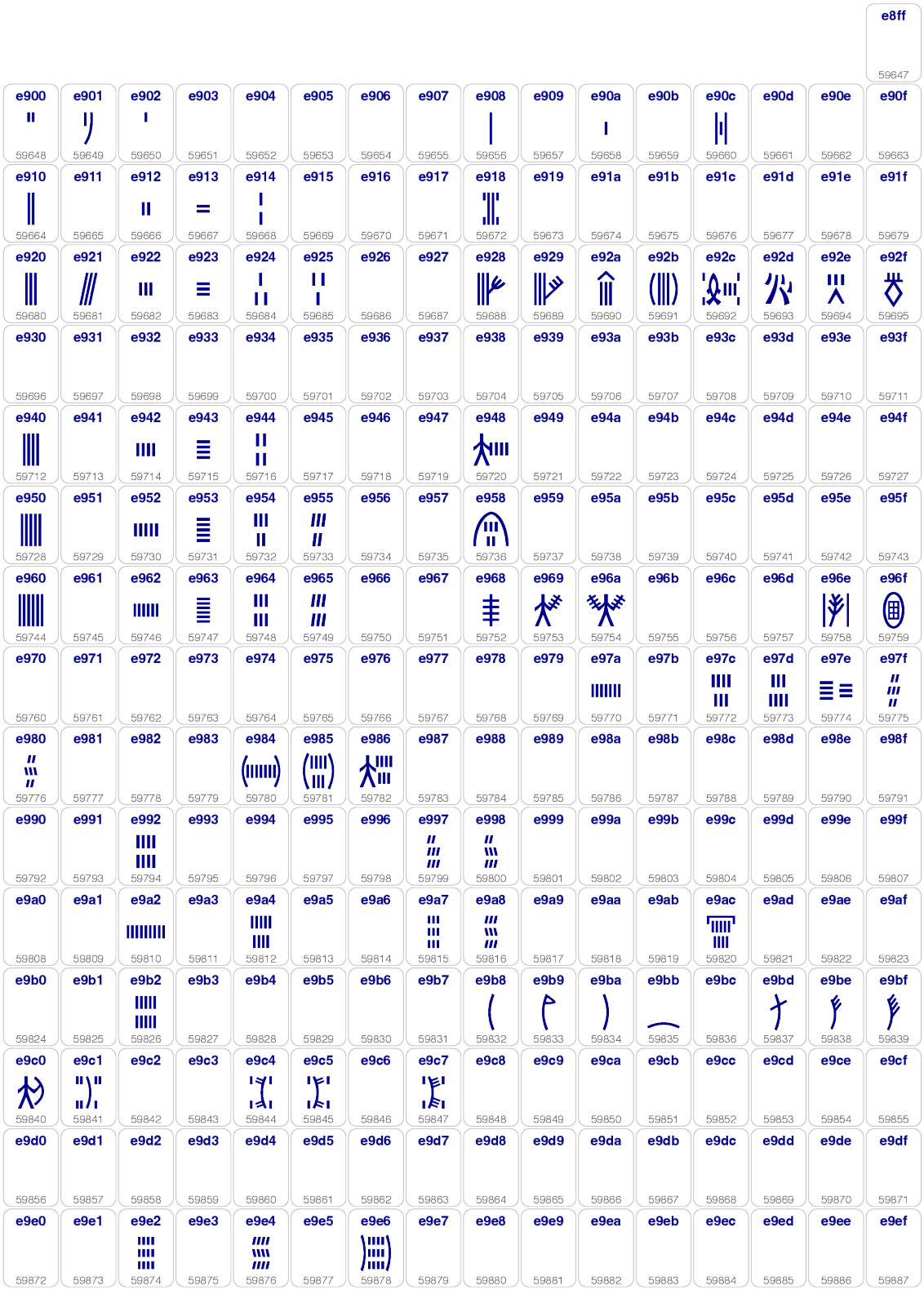
Zubair, N. (2016, November 4). Dholavira and Its Mysterious Sign Board. *Harappa.com*. https://www.harappa.com/blog/dholavira-and-its-mysterious-sign-board

Chinese character classification. (2023, February 7). Wikipedia. https://en.wikipedia.org/wiki/Chinese_character_classification

Shuowen Jiezi. (2023, March 22). Wikipedia. https://en.wikipedia.org/wiki/Shuowen_Jiezi

Appendix A

Unicode Charcater Table




e9f0 59888
e9f1 59889
e9f2 59890
e9f3 59891
e9f4 59892
e9f5 59893
e9f6 59894
e9f7 59895
e9f8 59896
e9f9 59897
e9fa 59898
e9fb 59899
e9fc 59900
e9fd 59901
e9fe 59902
e9ff 59903
ea00 59904
ea01 59905
ea02 59906
ea03 59907
ea04 59908
ea05 59909
ea06 59910
ea07 59911
ea08 59912
ea09 59913
ea0a 59914
ea0b 59915
ea0c 59916
ea0d 59917
ea0e 59918
ea0f 59919
ea10 59920
ea11 59921
ea12 59922
ea13 59923
ea14 59924
ea15 59925
ea16 59926
ea17 59927
ea18 59928
ea19 59929
ea1a 59930
ea1b 59931
ea1c 59932
ea1d 59933
ea1e 59934
ea1f 59935
ea20 59936
ea21 59937
ea22 59938
ea23 59939
ea24 59940
ea25 59941
ea26 59942
ea27 59943
ea28 59944
ea29 59945
ea2a 59946
ea2b 59947
ea2c 59948
ea2d 59949
ea2e 59950
ea2f 59951
ea30 59952
ea31 59953
ea32 59954
ea33 59955
ea34 59956
ea35 59957
ea36 59958
ea37 59959
ea38 59960
ea39 59961
ea3a 59962
ea3b 59963
ea3c 59964
ea3d 59965
ea3e 59966
ea3f 59967
ea40 59968
ea41 59969
ea42 59970
ea43 59971
ea44 59972
ea45 59973
ea46 59974
ea47 59975
ea48 59976
ea49 59977
ea4a 59978
ea4b 59979
ea4c 59980
ea4d 59981
ea4e 59982
ea4f 59983
ea50 59984
ea51 59985
ea52 59986
ea53 59987
ea54 59988
ea55 59989
ea56 59990
ea57 59991
ea58 59992
ea59 59993
ea5a 59994
ea5b 59995
ea5c 59996
ea5d 59997
ea5e 59998
ea5f 59999
ea60 60000
ea61 60001
ea62 60002
ea63 60003
ea64 60004
ea65 60005
ea66 60006
ea67 60007
ea68 60008
ea69 60009
ea6a 60010
ea6b 60011
ea6c 60012
ea6d 60013
ea6e 60014
ea6f 60015
ea70 60016
ea71 60017
ea72 60018
ea73 60019
ea74 60020
ea75 60021
ea76 60022
ea77 60023
ea78 60024
ea79 60025
ea7a 60026
ea7b 60027
ea7c 60028
ea7d 60029
ea7e 60030
ea7f 60031
ea80 60032
ea81 60033
ea82 60034
ea83 60035
ea84 60036
ea85 60037
ea86 60038
ea87 60039
ea88 60040
ea89 60041
ea8a 60042
ea8b 60043
ea8c 60044
ea8d 60045
ea8e 60046
ea8f 60047
ea90 60048
ea91 60049
ea92 60050
ea93 60051
ea94 60052
ea95 60053
ea96 60054
ea97 60055
ea98 60056
ea99 60057
ea9a 60058
ea9b 60059
ea9c 60060
ea9d 60061
ea9e 60062
ea9f 60063
eaa0 60064
eaa1 60065
eaa2 60066
eaa3 60067
eaa4 60068
eaa5 60069
eaa6 60070
eaa7 60071
eaa8 60072
eaa9 60073
eaaa 60074
eaab 60075
eaac 60076
eaad 60077
eaae 60078
eaaf 60079
eab0 60080
eab1 60081
eab2 60082
eab3 60083
eab4 60084
eab5 60085
eab6 60086
eab7 60087
eab8 60088
eab9 60089
eaba 60090
eabb 60091
eabc 60092
eabd 60093
eabe 60094
eabf 60095
eac0 60096
eac1 60097
eac2 60098
eac3 60099
eac4 60100
eac5 60101
eac6 60102
eac7 60103
eac8 60104
eac9 60105
eaca 60106
eacb 60107
eacc 60108
eacd 60109
eace 60110
eacf 60111
ead0 60112
ead1 60113
ead2 60114
ead3 60115
ead4 60116
ead5 60117
ead6 60118
ead7 60119
ead8 60120
ead9 60121
eada 60122
eadb 60123
eadc 60124
eadd 60125
eade 60126
eadf 60127
eae0 60128
eae1 60129
eae2 60130
eae3 60131
eae4 60132
eae5 60133
eae6 60134
eae7 60135
eae8 60136
eae9 60137
eaea 60138
eaeb 60139
eaec 60140
eaed 60141
eaee 60142
eaef 60143
eaf0
eaf1
eaf2
eaf3
eaf4
eaf5
eaf6
eaf7
eaf8
eaf9
eafa
eafb
eafc
eafd
eafe
eaff


60144 60145 60146 60147 60148 60149 60150 60151 60152 60153 60154 60155 60156 60157 60158 60159
eb00 eb01 eb02 eb03 eb04 eb05 eb06 eb07 eb08 eb09 eb0a eb0b eb0c eb0d eb0e eb0f
60160 60161 60162 60163 60164 60165 60166 60167 60168 60169 60170 60171 60172 60173 60174 60175
eb10 eb11 eb12 eb13 eb14 eb15 eb16 eb17 eb18 eb19 eb1a eb1b eb1c eb1d eb1e eb1f
60176 60177 60178 60179 60180 60181 60182 60183 60184 60185 60186 60187 60188 60189 60190 60191
eb20 eb21 eb22 eb23 eb24 eb25 eb26 eb27 eb28 eb29 eb2a eb2b eb2c eb2d eb2e eb2f
60192 60193 60194 60195 60196 60197 60198 60199 60200 60201 60202 60203 60204 60205 60206 60207
eb30 eb31 eb32 eb33 eb34 eb35 eb36 eb37 eb38 eb39 eb3a eb3b eb3c eb3d eb3e eb3f
60208 60209 60210 60211 60212 60213 60214 60215 60216 60217 60218 60219 60220 60221 60222 60223
eb40 eb41 eb42 eb43 eb44 eb45 eb46 eb47 eb48 eb49 eb4a eb4b eb4c eb4d eb4e eb4f
60224 60225 60226 60227 60228 60229 60230 60231 60232 60233 60234 60235 60236 60237 60238 60239
eb50 eb51 eb52 eb53 eb54 eb55 eb56 eb57 eb58 eb59 eb5a eb5b eb5c eb5d eb5e eb5f
60240 60241 60242 60243 60244 60245 60246 60247 60248 60249 60250 60251 60252 60253 60254 60255
eb60 eb61 eb62 eb63 eb64 eb65 eb66 eb67 eb68 eb69 eb6a eb6b eb6c eb6d eb6e eb6f
60256 60257 60258 60259 60260 60261 60262 60263 60264 60265 60266 60267 60268 60269 60270 60271
eb70 eb71 eb72 eb73 eb74 eb75 eb76 eb77 eb78 eb79 eb7a eb7b eb7c eb7d eb7e eb7f
60272 60273 60274 60275 60276 60277 60278 60279 60280 60281 60282 60283 60284 60285 60286 60287
eb80 eb81 eb82 eb83 eb84 eb85 eb86 eb87 eb88 eb89 eb8a eb8b eb8c eb8d eb8e eb8f
60288 60289 60290 60291 60292 60293 60294 60295 60296 60297 60298 60299 60300 60301 60302 60303
eb90 eb91 eb92 eb93 eb94 eb95 eb96 eb97 eb98 eb99 eb9a eb9b eb9c eb9d eb9e eb9f
60304 60305 60306 60307 60308 60309 60310 60311 60312 60313 60314 60315 60316 60317 60318 60319
eba0 eba1 eba2 eba3 eba4 eba5 eba6 eba7 eba8 eba9 ebaa ebab ebac ebad ebae ebaf
60320 60321 60322 60323 60324 60325 60326 60327 60328 60329 60330 60331 60332 60333 60334 60335
ebb0 ebb1 ebb2 ebb3 ebb4 ebb5 ebb6 ebb7 ebb8 ebb9 ebba ebbb ebbc ebbd ebbe ebbf
60336 60337 60338 60339 60340 60341 60342 60343 60344 60345 60346 60347 60348 60349 60350 60351
ebc0 ebc1 ebc2 ebc3 ebc4 ebc5 ebc6 ebc7 ebc8 ebc9 ebca ebcb ebcc ebcd ebce ebcf
60352 60353 60354 60355 60356 60357 60358 60359 60360 60361 60362 60363 60364 60365 60366 60367
ebd0 ebd1 ebd2 ebd3 ebd4 ebd5 ebd6 ebd7 ebd8 ebd9 ebda ebdb ebdc ebdd ebde ebdf
60368 60369 60370 60371 60372 60373 60374 60375 60376 60377 60378 60379 60380 60381 60382 60383
ebe0 ebe1 ebe2 ebe3 ebe4 ebe5 ebe6 ebe7 ebe8 ebe9 ebea ebeb ebec ebed ebee ebef
60384 60385 60386 60387 60388 60389 60390 60391 60392 60393 60394 60395 60396 60397 60398 60399
ebf0 ebf1 ebf2 ebf3 ebf4 ebf5 ebf6 ebf7 ebf8 ebf9 ebfa ebfb ebfc ebfd ebfe ebff


ec00 60416
ec01 60417
ec02 60418
ec03 60419
ec04 60420
ec05 60421
ec06 60422
ec07 60423
ec08 60424
ec09 60425
ec0a 60426
ec0b 60427
ec0c 60428
ec0d 60429
ec0e 60430
ec0f 60431
ec10 60432
ec11 60433
ec12 60434
ec13 60435
ec14 60436
ec15 60437
ec16 60438
ec17 60439
ec18 60440
ec19 60441
ec1a 60442
ec1b 60443
ec1c 60444
ec1d 60445
ec1e 60446
ec1f 60447
ec20 60448
ec21 60449
ec22 60450
ec23 60451
ec24 60452
ec25 60453
ec26 60454
ec27 60455
ec28 60456
ec29 60457
ec2a 60458
ec2b 60459
ec2c 60460
ec2d 60461
ec2e 60462
ec2f 60463
ec30 60464
ec31 60465
ec32 60466
ec33 60467
ec34 60468
ec35 60469
ec36 60470
ec37 60471
ec38 60472
ec39 60473
ec3a 60474
ec3b 60475
ec3c 60476
ec3d 60477
ec3e 60478
ec3f 60479
ec40 60480
ec41 60481
ec42 60482
ec43 60483
ec44 60484
ec45 60485
ec46 60486
ec47 60487
ec48 60488
ec49 60489
ec4a 60490
ec4b 60491
ec4c 60492
ec4d 60493
ec4e 60494
ec4f 60495
ec50 60496
ec51 60497
ec52 60498
ec53 60499
ec54 60500
ec55 60501
ec56 60502
ec57 60503
ec58 60504
ec59 60505
ec5a 60506
ec5b 60507
ec5c 60508
ec5d 60509
ec5e 60510
ec5f 60511
ec60 60512
ec61 60513
ec62 60514
ec63 60515
ec64 60516
ec65 60517
ec66 60518
ec67 60519
ec68 60520
ec69 60521
ec6a 60522
ec6b 60523
ec6c 60524
ec6d 60525
ec6e 60526
ec6f 60527
ec70 60528
ec71 60529
ec72 60530
ec73 60531
ec74 60532
ec75 60533
ec76 60534
ec77 60535
ec78 60536
ec79 60537
ec7a 60538
ec7b 60539
ec7c 60540
ec7d 60541
ec7e 60542
ec7f 60543
ec80 60544
ec81 60545
ec82 60546
ec83 60547
ec84 60548
ec85 60549
ec86 60550
ec87 60551
ec88 60552
ec89 60553
ec8a 60554
ec8b 60555
ec8c 60556
ec8d 60557
ec8e 60558
ec8f 60559
ec90 60560
ec91 60561
ec92 60562
ec93 60563
ec94 60564
ec95 60565
ec96 60566
ec97 60567
ec98 60568
ec99 60569
ec9a 60570
ec9b 60571
ec9c 60572
ec9d 60573
ec9e 60574
ec9f 60575
eca0 60576
eca1 60577
eca2 60578
eca3 60579
eca4 60580
eca5 60581
eca6 60582
eca7 60583
eca8 60584
eca9 60585
ecaa 60586
ecab 60587
ecac 60588
ecad 60589
ecae 60590
ecaf 60591
ecb0 60592
ecb1 60593
ecb2 60594
ecb3 60595
ecb4 60596
ecb5 60597
ecb6 60598
ecb7 60599
ecb8 60600
ecb9 60601
ecba 60602
ecbb 60603
ecbc 60604
ecbd 60605
ecbe 60606
ecbf 60607
ecc0 60608
ecc1 60609
ecc2 60610
ecc3 60611
ecc4 60612
ecc5 60613
ecc6 60614
ecc7 60615
ecc8 60616
ecc9 60617
ecca 60618
eccb 60619
eccc 60620
eccd 60621
ecce 60622
eccf 60623
ecd0 60624
ecd1 60625
ecd2 60626
ecd3 60627
ecd4 60628
ecd5 60629
ecd6 60630
ecd7 60631
ecd8 60632
ecd9 60633
ecda 60634
ecdb 60635
ecdc 60636
ecdd 60637
ecde 60638
ecdf 60639
ece0 60640
ece1 60641
ece2 60642
ece3 60643
ece4 60644
ece5 60645
ece6 60646
ece7 60647
ece8 60648
ece9 60649
ecea 60650
eceb 60651
ecec 60652
eced 60653
ecee 60654
ecef 60655
ecf0
ecf1
ecf2
ecf3
ecf4
ecf5
ecf6
ecf7
ecf8
ecf9
ecfa
ecfb
ecfc
ecfd
ecfe
ecff


ed00 60672
ed01 60673
ed02 60674
ed03 60675
ed04 60676
ed05 60677
ed06 60678
ed07 60679
ed08 60680
ed09 60681
ed0a 60682
ed0b 60683
ed0c 60684
ed0d 60685
ed0e 60686
ed0f 60687
ed10 60688
ed11 60689
ed12 60690
ed13 60691
ed14 60692
ed15 60693
ed16 60694
ed17 60695
ed18 60696
ed19 60697
ed1a 60698
ed1b 60699
ed1c 60700
ed1d 60701
ed1e 60702
ed1f 60703
ed20 60704
ed21 60705
ed22 60706
ed23 60707
ed24 60708
ed25 60709
ed26 60710
ed27 60711
ed28 60712
ed29 60713
ed2a 60714
ed2b 60715
ed2c 60716
ed2d 60717
ed2e 60718
ed2f 60719
ed30 60720
ed31 60721
ed32 60722
ed33 60723
ed34 60724
ed35 60725
ed36 60726
ed37 60727
ed38 60728
ed39 60729
ed3a 60730
ed3b 60731
ed3c 60732
ed3d 60733
ed3e 60734
ed3f 60735
ed40 60736
ed41 60737
ed42 60738
ed43 60739
ed44 60740
ed45 60741
ed46 60742
ed47 60743
ed48 60744
ed49 60745
ed4a 60746
ed4b 60747
ed4c 60748
ed4d 60749
ed4e 60750
ed4f 60751
ed50 60752
ed51 60753
ed52 60754
ed53 60755
ed54 60756
ed55 60757
ed56 60758
ed57 60759
ed58 60760
ed59 60761
ed5a 60762
ed5b 60763
ed5c 60764
ed5d 60765
ed5e 60766
ed5f 60767
ed60 60768
ed61 60769
ed62 60770
ed63 60771
ed64 60772
ed65 60773
ed66 60774
ed67 60775
ed68 60776
ed69 60777
ed6a 60778
ed6b 60779
ed6c 60780
ed6d 60781
ed6e 60782
ed6f 60783
ed70 60784
ed71 60785
ed72 60786
ed73 60787
ed74 60788
ed75 60789
ed76 60790
ed77 60791
ed78 60792
ed79 60793
ed7a 60794
ed7b 60795
ed7c 60796
ed7d 60797
ed7e 60798
ed7f 60799
ed80 60800
ed81 60801
ed82 60802
ed83 60803
ed84 60804
ed85 60805
ed86 60806
ed87 60807
ed88 60808
ed89 60809
ed8a 60810
ed8b 60811
ed8c 60812
ed8d 60813
ed8e 60814
ed8f 60815
ed90 60816
ed91 60817
ed92 60818
ed93 60819
ed94 60820
ed95 60821
ed96 60822
ed97 60823
ed98 60824
ed99 60825
ed9a 60826
ed9b 60827
ed9c 60828
ed9d 60829
ed9e 60830
ed9f 60831
eda0 60832
eda1 60833
eda2 60834
eda3 60835
eda4 60836
eda5 60837
eda6 60838
eda7 60839
eda8 60840
eda9 60841
edaa 60842
edab 60843
edac 60844
edad 60845
edae 60846
edaf 60847
edb0 60848
edb1 60849
edb2 60850
edb3 60851
edb4 60852
edb5 60853
edb6 60854
edb7 60855
edb8 60856
edb9 60857
edba 60858
edbb 60859
edbc 60860
edbd 60861
edbe 60862
edbf 60863
edc0 60864
edc1 60865
edc2 60866
edc3 60867
edc4 60868
edc5 60869
edc6 60870
edc7 60871
edc8 60872
edc9 60873
edca 60874
edcb 60875
edcc 60876
edcd 60877
edce 60878
edcf 60879
edd0 60880
edd1 60881
edd2 60882
edd3 60883
edd4 60884
edd5 60885
edd6 60886
edd7 60887
edd8 60888
edd9 60889
edda 60890
eddb 60891
eddc 60892
eddd 60893
edde 60894
eddf 60895
ede0 60896
ede1 60897
ede2 60898
ede3 60899
ede4 60900
ede5 60901
ede6 60902
ede7 60903
ede8 60904
ede9 60905
edea 60906
edeb 60907
edec 60908
eded 60909
edee 60910
edef 60911
edf0 60912
edf1 60913
edf2 60914
edf3 60915
edf4 60916
edf5 60917
edf6 60918
edf7 60919
edf8 60920
edf9 60921
edfa 60922
edfb 60923
edfc 60924
edfd 60925
edfe 60926
edff 60927


ee00 60928
ee01 60929
ee02 60930
ee03 60931
ee04 60932
ee05 60933
ee06 60934
ee07 60935
ee08 60936
ee09 60937
ee0a 60938
ee0b 60939
ee0c 60940
ee0d 60941
ee0e 60942
ee0f 60943
ee10 60944
ee11 60945
ee12 60946
ee13 60947
ee14 60948
ee15 60949
ee16 60950
ee17 60951
ee18 60952
ee19 60953
ee1a 60954
ee1b 60955
ee1c 60956
ee1d 60957
ee1e 60958
ee1f 60959
ee20 60960
ee21 60961
ee22 60962
ee23 60963
ee24 60964
ee25 60965
ee26 60966
ee27 60967
ee28 60968
ee29 60969
ee2a 60970
ee2b 60971
ee2c 60972
ee2d 60973
ee2e 60974
ee2f 60975
ee30 60976
ee31 60977
ee32 60978
ee33 60979
ee34 60980
ee35 60981
ee36 60982
ee37 60983
ee38 60984
ee39 60985
ee3a 60986
ee3b 60987
ee3c 60988
ee3d 60989
ee3e 60990
ee3f 60991
ee40 60992
ee41 60993
ee42 60994
ee43 60995
ee44 60996
ee45 60997
ee46 60998
ee47 60999
ee48 61000
ee49 61001
ee4a 61002
ee4b 61003
ee4c 61004
ee4d 61005
ee4e 61006
ee4f 61007
ee50 61008
ee51 61009
ee52 61010
ee53 61011
ee54 61012
ee55 61013
ee56 61014
ee57 61015
ee58 61016
ee59 61017
ee5a 61018
ee5b 61019
ee5c 61020
ee5d 61021
ee5e 61022
ee5f 61023
ee60 61024
ee61 61025
ee62 61026
ee63 61027
ee64 61028
ee65 61029
ee66 61030
ee67 61031
ee68 61032
ee69 61033
ee6a 61034
ee6b 61035
ee6c 61036
ee6d 61037
ee6e 61038
ee6f 61039
ee70 61040
ee71 61041
ee72 61042
ee73 61043
ee74 61044
ee75 61045
ee76 61046
ee77 61047
ee78 61048
ee79 61049
ee7a 61050
ee7b 61051
ee7c 61052
ee7d 61053
ee7e 61054
ee7f 61055
ee80 61056
ee81 61057
ee82 61058
ee83 61059
ee84 61060
ee85 61061
ee86 61062
ee87 61063
ee88 61064
ee89 61065
ee8a 61066
ee8b 61067
ee8c 61068
ee8d 61069
ee8e 61070
ee8f 61071
ee90 61072
ee91 61073
ee92 61074
ee93 61075
ee94 61076
ee95 61077
ee96 61078
ee97 61079
ee98 61080
ee99 61081
ee9a 61082
ee9b 61083
ee9c 61084
ee9d 61085
ee9e 61086
ee9f 61087
eea0 61088
eea1 61089
eea2 61090
eea3 61091
eea4 61092
eea5 61093
eea6 61094
eea7 61095
eea8 61096
eea9 61097
eeaa 61098
eeab 61099
eeac 61100
eead 61101
eeae 61102
eeaf 61103
eeb0 61104
eeb1 61105
eeb2 61106
eeb3 61107
eeb4 61108
eeb5 61109
eeb6 61110
eeb7 61111
eeb8 61112
eeb9 61113
eeba 61114
eebb 61115
eebc 61116
eebd 61117
eebe 61118
eebf 61119
eec0 61120
eec1 61121
eec2 61122
eec3 61123
eec4 61124
eec5 61125
eec6 61126
eec7 61127
eec8 61128
eec9 61129
eeca 61130
eecb 61131
eecc 61132
eecd 61133
eece 61134
eecf 61135
eed0 61136
eed1 61137
eed2 61138
eed3 61139
eed4 61140
eed5 61141
eed6 61142
eed7 61143
eed8 61144
eed9 61145
eeda 61146
eedb 61147
eedc 61148
eedd 61149
eede 61150
eedf 61151
eee0 61152
eee1 61153
eee2 61154
eee3 61155
eee4 61156
eee5 61157
eee6 61158
eee7 61159
eee8 61160
eee9 61161
eeea 61162
eeeb 61163
eeec 61164
eeed 61165
eeee 61166
eeef 61167
eef0 61168
eef1 61169
eef2 61170
eef3 61171
eef4 61172
eef5 61173
eef6 61174
eef7 61175
eef8 61176
eef9 61177
eefa 61178
eefb 61179
eefc 61180
eefd 61181
eefe 61182
eeff 61183
ef00
ef01
ef02
ef03
ef04
ef05
ef06
ef07
ef08
ef09
ef0a
ef0b
ef0c
ef0d
ef0e
ef0f

61184	61185	61186	61187	61188	61189	61190	61191	61192	61193	61194	61195	61196	61197	61198	61199
ef10 61200	ef11 61201	ef12 61202	ef13 61203	ef14 61204	ef15 61205	ef16 61206	ef17 61207	ef18 61208	ef19 61209	ef1a 61210	ef1b 61211	ef1c 61212	ef1d 61213	ef1e 61214	ef1f 61215
ef20 61216	ef21 61217	ef22 61218	ef23 61219	ef24 61220	ef25 61221	ef26 61222	ef27 61223	ef28 61224	ef29 61225	ef2a 61226	ef2b 61227	ef2c 61228	ef2d 61229	ef2e 61230	ef2f 61231
ef30 61232	ef31 61233	ef32 61234	ef33 61235	ef34 61236	ef35 61237	ef36 61238	ef37 61239	ef38 61240	ef39 61241	ef3a 61242	ef3b 61243	ef3c 61244	ef3d 61245	ef3e 61246	ef3f 61247
ef40 61248	ef41 61249	ef42 61250	ef43 61251	ef44 61252	ef45 61253	ef46 61254	ef47 61255	ef48 61256	ef49 61257	ef4a 61258	ef4b 61259	ef4c 61260	ef4d 61261	ef4e 61262	ef4f 61263
ef50 61264	ef51 61265	ef52 61266	ef53 61267	ef54 61268	ef55 61269	ef56 61270	ef57 61271	ef58 61272	ef59 61273	ef5a 61274	ef5b 61275	ef5c 61276	ef5d 61277	ef5e 61278	ef5f 61279
ef60 61280	ef61 61281	ef62 61282	ef63 61283	ef64 61284	ef65 61285	ef66 61286	ef67 61287	ef68 61288	ef69 61289	ef6a 61290	ef6b 61291	ef6c 61292	ef6d 61293	ef6e 61294	ef6f 61295
ef70 61296	ef71 61297	ef72 61298	ef73 61299	ef74 61300	ef75 61301	ef76 61302	ef77 61303	ef78 61304	ef79 61305	ef7a 61306	ef7b 61307	ef7c 61308	ef7d 61309	ef7e 61310	ef7f 61311
ef80 61312	ef81 61313	ef82 61314	ef83 61315	ef84 61316	ef85 61317	ef86 61318	ef87 61319	ef88 61320	ef89 61321	ef8a 61322	ef8b 61323	ef8c 61324	ef8d 61325	ef8e 61326	ef8f 61327
ef90 61328	ef91 61329	ef92 61330	ef93 61331	ef94 61332	ef95 61333	ef96 61334	ef97 61335	ef98 61336	ef99 61337	ef9a 61338	ef9b 61339	ef9c 61340	ef9d 61341	ef9e 61342	ef9f 61343
efa0 61344	efa1 61345	efa2 61346	efa3 61347	efa4 61348	efa5 61349	efa6 61350	efa7 61351	efa8 61352	efa9 61353	efaa 61354	efab 61355	efac 61356	efad 61357	efae 61358	efaf 61359
efb0 61360	efb1 61361	efb2 61362	efb3 61363	efb4 61364	efb5 61365	efb6 61366	efb7 61367	efb8 61368	efb9 61369	efba 61370	efbb 61371	efbc 61372	efbd 61373	efbe 61374	efbf 61375
efc0 61376	efc1 61377	efc2 61378	efc3 61379	efc4 61380	efc5 61381	efc6 61382	efc7 61383	efc8 61384	efc9 61385	efca 61386	efcb 61387	efcc 61388	efcd 61389	efce 61390	efcf 61391
efd0 61392	efd1 61393	efd2 61394	efd3 61395	efd4 61396	efd5 61397	efd6 61398	efd7 61399	efd8 61400	efd9 61401	efda 61402	efdb 61403	efdc 61404	efdd 61405	efde 61406	efdf 61407
efe0 61408	efe1 61409	efe2 61410	efe3 61411	efe4 61412	efe5 61413	efe6 61414	efe7 61415	efe8 61416	efe9 61417	efea 61418	efeb 61419	efec 61420	efed 61421	efee 61422	efef 61423
eff0 61424	eff1 61425	eff2 61426	eff3 61427	eff4 61428	eff5 61429	eff6 61430	eff7 61431	eff8 61432	eff9 61433	effa 61434	effb 61435	effc 61436	effd 61437	effe 61438	efff 61439
f000	f001	f002	f003	f004	f005	f006	f007	f008	f009	f00a	f00b	f00c	f00d	f00e	f00f

f010 61456	f011 61457	f012 61458	f013 61459	f014 61460	f015 61461	f016 61462	f017 61463	f018 61464	f019 61465	f01a 61466	f01b 61467	f01c 61468	f01d 61469	f01e 61470	f01f 61471
f020 61472	f021 61473	f022 61474	f023 61475	f024 61476	f025 61477	f026 61478	f027 61479	f028 61480	f029 61481	f02a 61482	f02b 61483	f02c 61484	f02d 61485	f02e 61486	f02f 61487
f030 61488	f031 61489	f032 61490	f033 61491	f034 61492	f035 61493	f036 61494	f037 61495	f038 61496	f039 61497	f03a 61498	f03b 61499	f03c 61500	f03d 61501	f03e 61502	f03f 61503
f040 61504	f041 61505	f042 61506	f043 61507	f044 61508	f045 61509	f046 61510	f047 61511	f048 61512	f049 61513	f04a 61514	f04b 61515	f04c 61516	f04d 61517	f04e 61518	f04f 61519
f050 61520	f051 61521	f052 61522	f053 61523	f054 61524	f055 61525	f056 61526	f057 61527	f058 61528	f059 61529	f05a 61530	f05b 61531	f05c 61532	f05d 61533	f05e 61534	f05f 61535
f060 61536	f061 61537	f062 61538	f063 61539	f064 61540	f065 61541	f066 61542	f067 61543	f068 61544	f069 61545	f06a 61546	f06b 61547	f06c 61548	f06d 61549	f06e 61550	f06f 61551
f070 61552	f071 61553	f072 61554	f073 61555	f074 61556	f075 61557	f076 61558	f077 61559	f078 61560	f079 61561	f07a 61562	f07b 61563	f07c 61564	f07d 61565	f07e 61566	f07f 61567
f080 61568	f081 61569	f082 61570	f083 61571	f084 61572	f085 61573	f086 61574	f087 61575	f088 61576	f089 61577	f08a 61578	f08b 61579	f08c 61580	f08d 61581	f08e 61582	f08f 61583
f090 61584	f091 61585	f092 61586	f093 61587	f094 61588	f095 61589	f096 61590	f097 61591	f098 61592	f099 61593	f09a 61594	f09b 61595	f09c 61596	f09d 61597	f09e 61598	f09f 61599
f0a0 61600	f0a1 61601	f0a2 61602	f0a3 61603	f0a4 61604	f0a5 61605	f0a6 61606	f0a7 61607	f0a8 61608	f0a9 61609	f0aa 61610	f0ab 61611	f0ac 61612	f0ad 61613	f0ae 61614	f0af 61615
f0b0 61616	f0b1 61617	f0b2 61618	f0b3 61619	f0b4 61620	f0b5 61621	f0b6 61622	f0b7 61623	f0b8 61624	f0b9 61625	f0ba 61626	f0bb 61627	f0bc 61628	f0bd 61629	f0be 61630	f0bf 61631
f0c0 61632	f0c1 61633	f0c2 61634	f0c3 61635	f0c4 61636	f0c5 61637	f0c6 61638	f0c7 61639	f0c8 61640	f0c9 61641	f0ca 61642	f0cb 61643	f0cc 61644	f0cd 61645	f0ce 61646	f0cf 61647
f0d0 61648	f0d1 61649	f0d2 61650	f0d3 61651	f0d4 61652	f0d5 61653	f0d6 61654	f0d7 61655	f0d8 61656	f0d9 61657	f0da 61658	f0db 61659	f0dc 61660	f0dd 61661	f0de 61662	f0df 61663
f0e0 61664	f0e1 61665	f0e2 61666	f0e3 61667	f0e4 61668	f0e5 61669	f0e6 61670	f0e7 61671	f0e8 61672	f0e9 61673	f0ea 61674	f0eb 61675	f0ec 61676	f0ed 61677	f0ee 61678	f0ef 61679
f0f0 61680	f0f1 61681	f0f2 61682	f0f3 61683	f0f4 61684	f0f5 61685	f0f6 61686	f0f7 61687	f0f8 61688	f0f9 61689	f0fa 61690	f0fb 61691	f0fc 61692	f0fd 61693	f0fe 61694	f0ff 61695
f100	f101	f102	f103	f104	f105	f106	f107	f108	f109	f10a	f10b	f10c	f10d	f10e	f10f

f110	f111	f112	f113	f114	f115	f116	f117	f118	f119	f11a	f11b	f11c	f11d	f11e	f11f
61712	61713	61714	61715	61716	61717	61718	61719	61720	61721	61722	61723	61724	61725	61726	61727
f120	f121	f122	f123	f124	f125	f126	f127	f128	f129	f12a	f12b	f12c	f12d	f12e	f12f
61728	61729	61730	61731	61732	61733	61734	61735	61736	61737	61738	61739	61740	61741	61742	61743
f130	f131	f132	f133	f134	f135	f136	f137	f138	f139	f13a	f13b	f13c	f13d	f13e	f13f
61744	61745	61746	61747	61748	61749	61750	61751	61752	61753	61754	61755	61756	61757	61758	61759
f140	f141	f142	f143	f144	f145	f146	f147	f148	f149	f14a	f14b	f14c	f14d	f14e	f14f
61760	61761	61762	61763	61764	61765	61766	61767	61768	61769	61770	61771	61772	61773	61774	61775
f150	f151	f152	f153	f154	f155	f156	f157	f158	f159	f15a	f15b	f15c	f15d	f15e	f15f
61776	61777	61778	61779	61780	61781	61782	61783	61784	61785	61786	61787	61788	61789	61790	61791
f160	f161	f162	f163	f164	f165	f166	f167	f168	f169	f16a	f16b	f16c	f16d	f16e	f16f
61792	61793	61794	61795	61796	61797	61798	61799	61800	61801	61802	61803	61804	61805	61806	61807
f170	f171	f172	f173	f174	f175	f176	f177	f178	f179	f17a	f17b	f17c	f17d	f17e	f17f
61808	61809	61810	61811	61812	61813	61814	61815	61816	61817	61818	61819	61820	61821	61822	61823
f180	f181	f182	f183	f184	f185	f186	f187	f188	f189	f18a	f18b	f18c	f18d	f18e	f18f
61824	61825	61826	61827	61828	61829	61830	61831	61832	61833	61834	61835	61836	61837	61838	61839
f190	f191	f192	f193	f194	f195	f196	f197	f198	f199	f19a	f19b	f19c	f19d	f19e	f19f
61840	61841	61842	61843	61844	61845	61846	61847	61848	61849	61850	61851	61852	61853	61854	61855
f1a0	f1a1	f1a2	f1a3	f1a4	f1a5	f1a6	f1a7	f1a8	f1a9	f1aa	f1ab	f1ac	f1ad	f1ae	f1af
61856	61857	61858	61859	61860	61861	61862	61863	61864	61865	61866	61867	61868	61869	61870	61871

Appendix B

Etymology Classes and Blocks

P-6 16

P-7 8

P-8 8

P-9 24

P-10 8

P-11 24

P-12 8

P-13 16

P-14 8

P-15 16

P-16 16

P-17 16

P-18 48

P-19 16

P-20 8

P-21 48

P-22 24

P-23 8

P-24 24

P-25 8

P-26 8

P-27 8

P-28 8

P-29 16

P-30 8

P-31 24

P-32 8

P-33 8

P-34 8

P-35 8

P-36 16

P-37 16

P-38 16

P-39 16

G-1 8

G-2 8

G-3 16

G-4 16

G-5 16

G-6 16

G-7 16

G-8 16

G-9 16

G-10 8

G-11 8

G-12 8

G-13 8

G-14 8

G-15 16

G-16 24

G-17 16

G-18 8

G-19 8

G-20 8

G-21 8

G-22 16

G-23 16

G-24 8

G-25 8

G-26 16

G-27 16

G-28 16

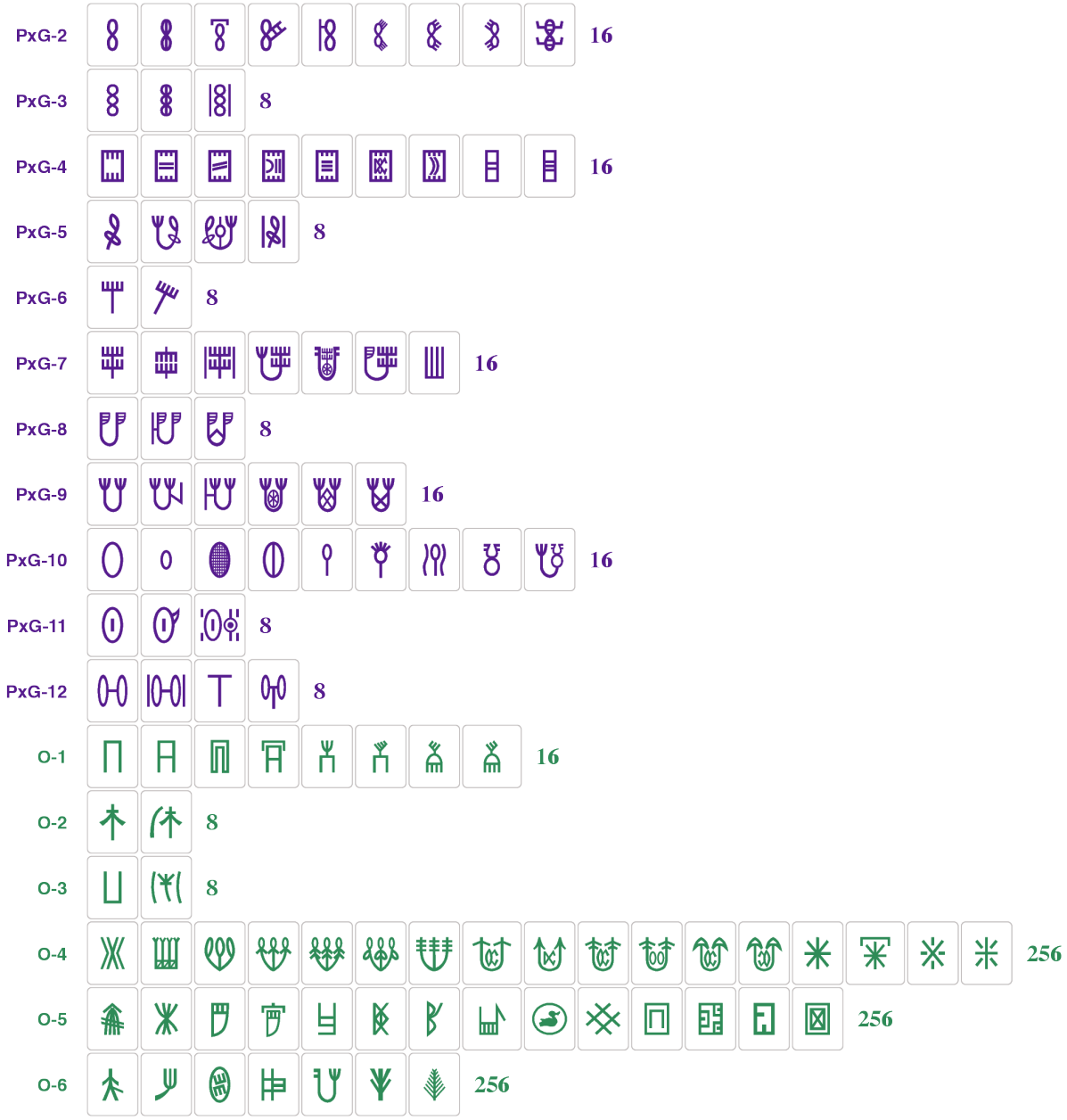
G-29 16

G-30 16

G-31 16

G-32 8

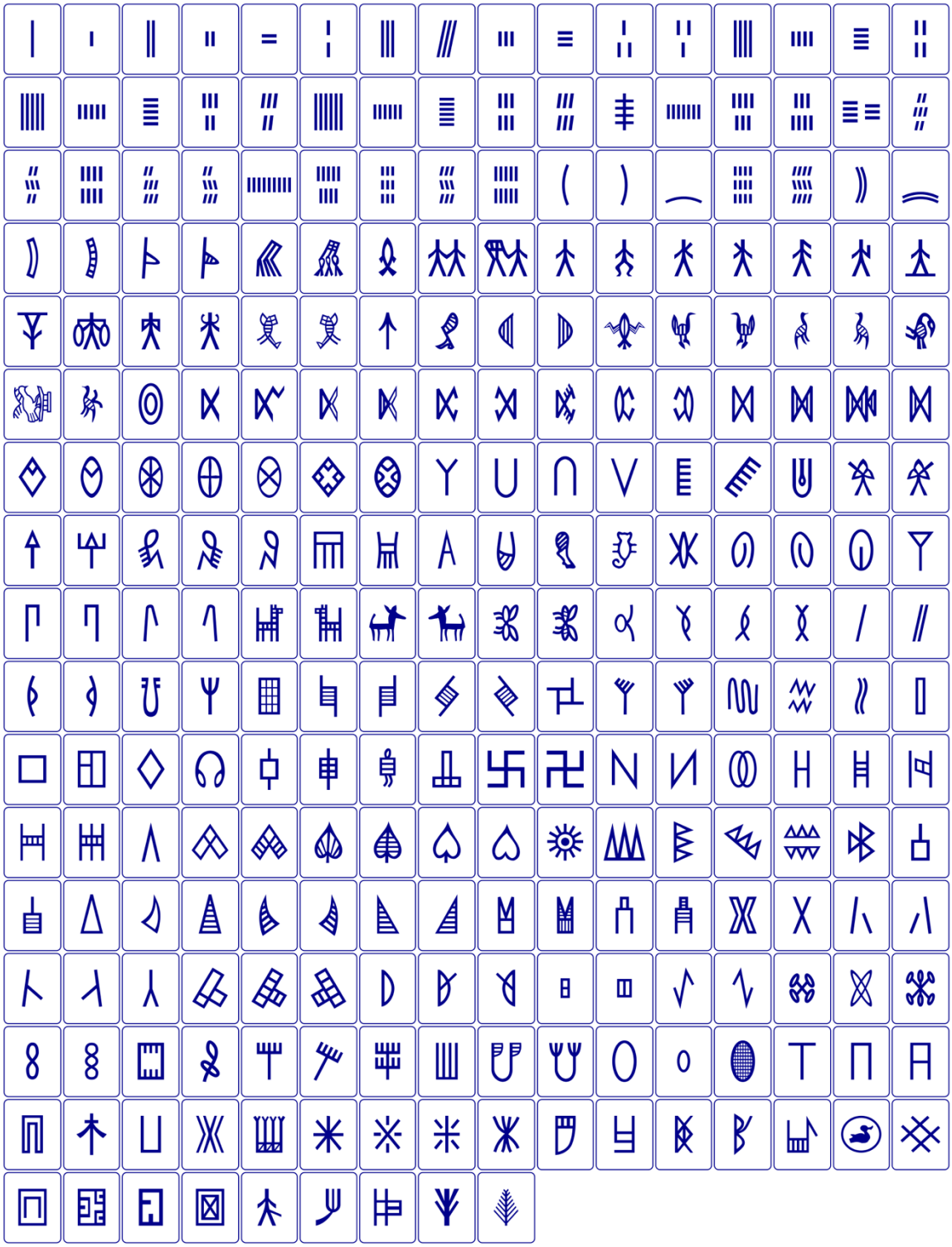
PxG-1 8


PxG-2 16
PxG-3 8
PxG-4 16
PxG-5 8
PxG-6 8
PxG-7 16
PxG-8 8
PxG-9 16
PxG-10 16
PxG-11 8
PxG-12 8
O-1 16
O-2 8
O-3 8
O-4 256
O-5 256
O-6 256

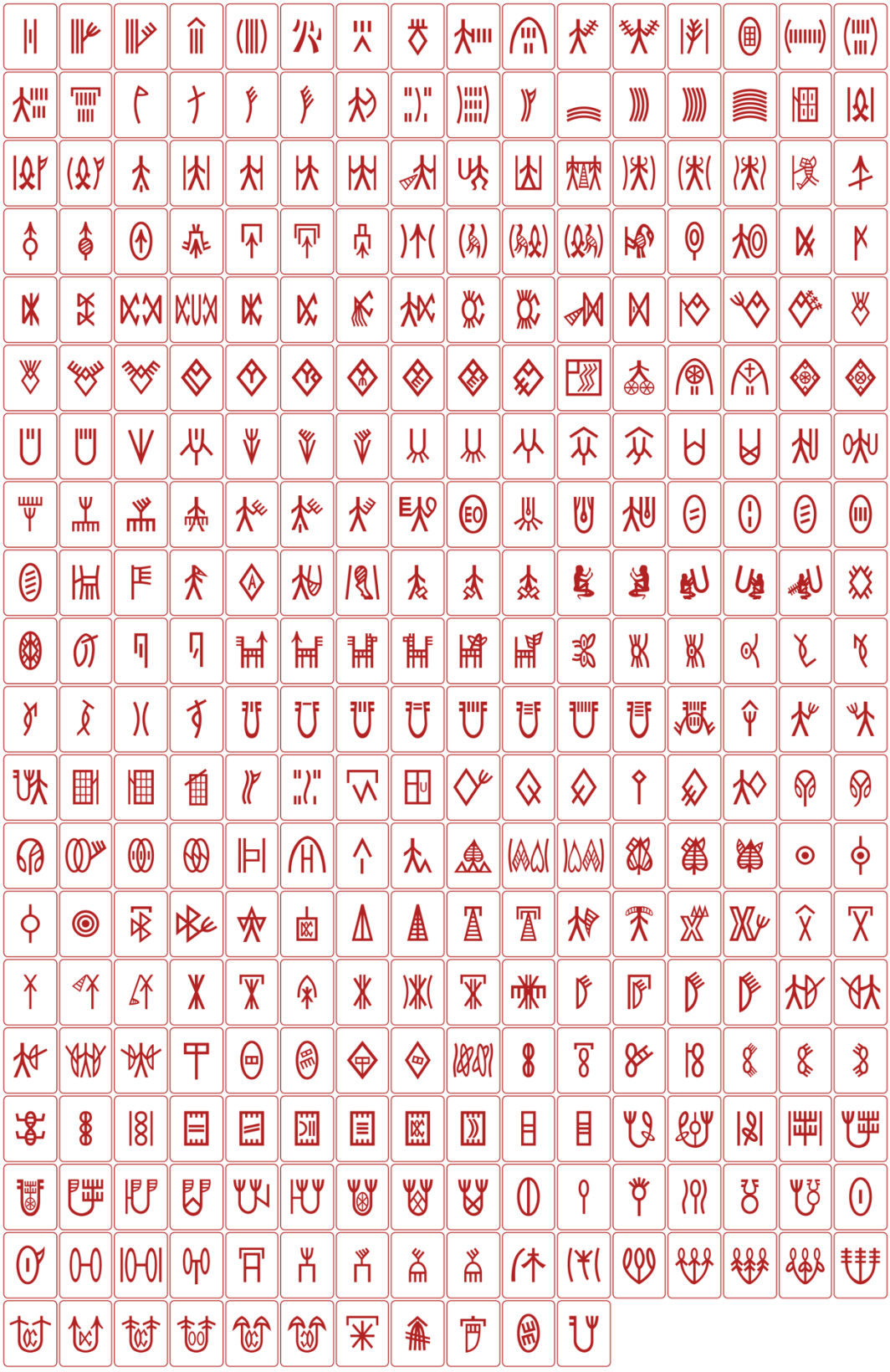
Appendix C

Base and Compound Akhyats

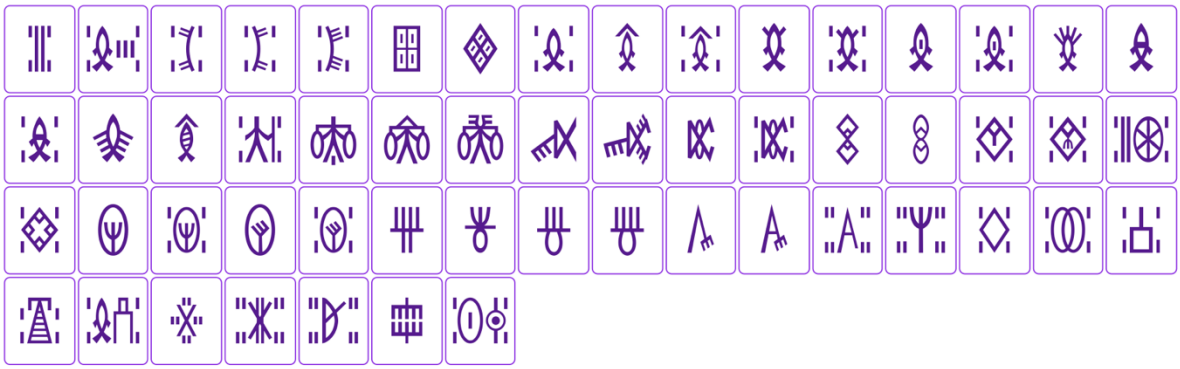
Base Akhyats



Compound Akhyats



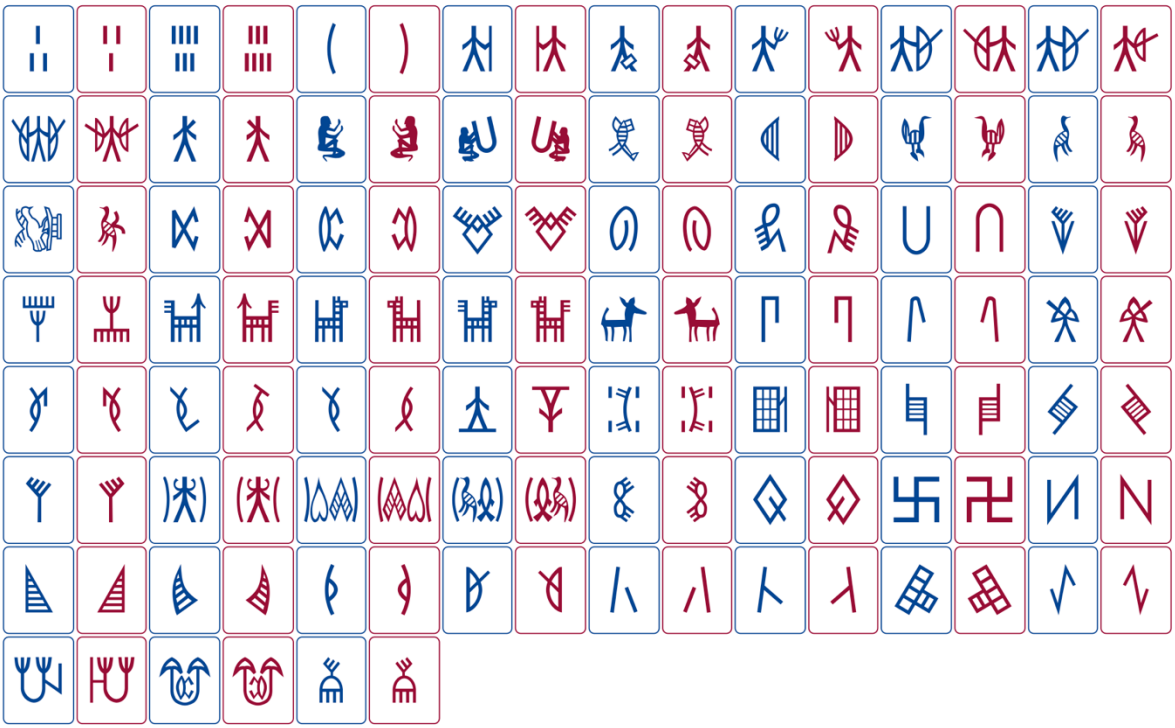
Phono-Semantic Akhyats



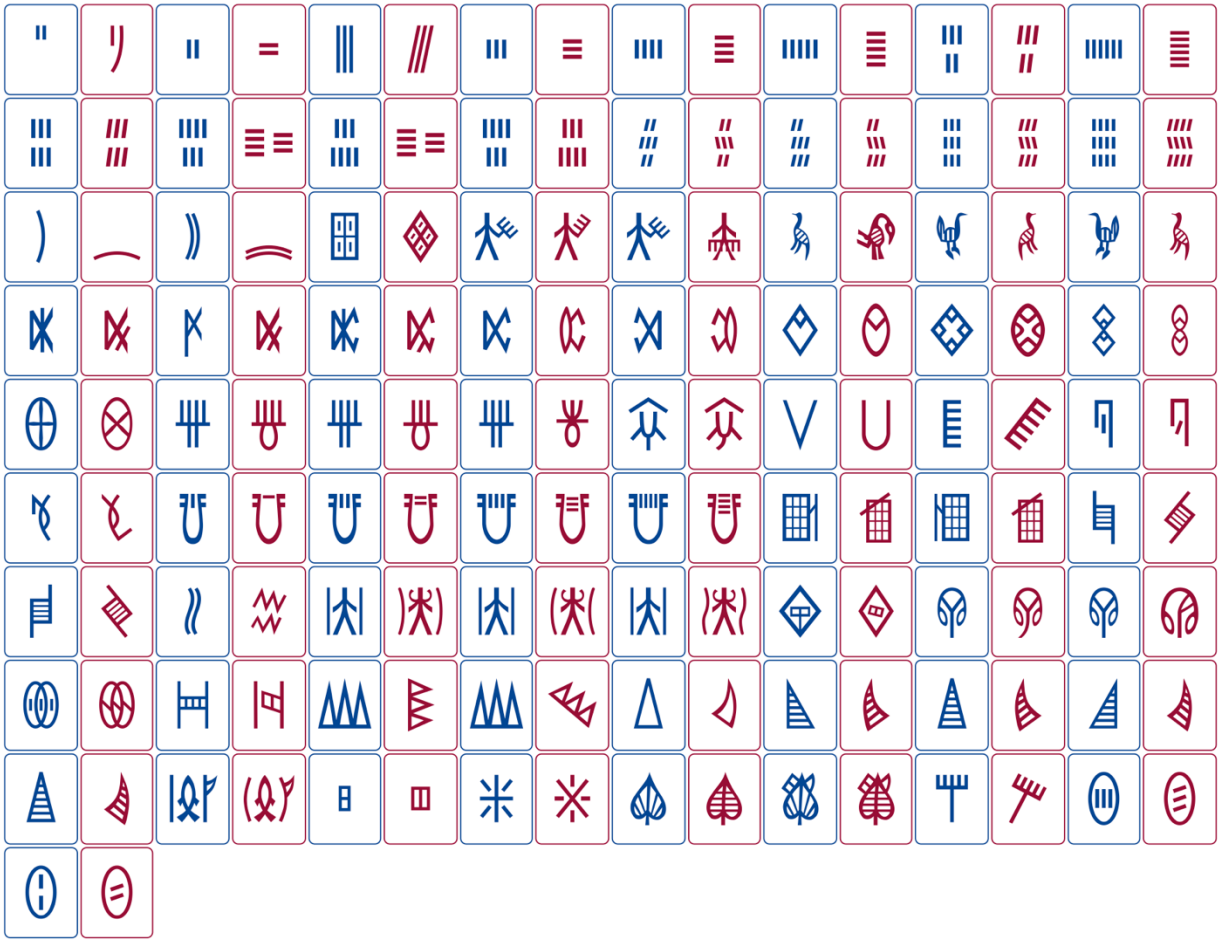
Appendix D

Systematic Polysemy

Chiral Forms



Bent Forms



Black and White Forms