Proposed Vedic Sanskrit Coding Scheme: Some suggestions

Akshar Bharati
Amba Kulkarni
Department of Sanskrit Studies
University of Hyderabad
Hyderabad
email: apksh@uohyd.ernet.in

Abstract:

Indian languages belong to four different families. However, as far as scripts are concerned, all of them (expect for Perso-Arabic script) are derived from the Brahmi script. Indian languages are compositionally syllabic. They have a scientific phonetic base and all the syllables are derived from the phonemes compositionally. They have flexibility and can be used as alphabetic or as syllabic as the need demands. Whereas the syllabic version is suitable for writing concisely, the alphabetic is suitable for performing the linguistic operations, such as sandhi operation, morphological analysis, sorting, searching, etc.

The current Unicode for Indian Languages is neither purely phonemic nor syllabic, but a hybrid of the two. There is redundancy in the scheme regarding the vowel modifiers. The presence of vowel modifiers also hinders the public from knowing the exact nature of our scripts.

The proposed Unicode for vedic Sanskrit accounts for this issue, and is purely phonemic in agreement with our shastric texts. It also removes the redundancy in the vowel modifiers, by removing them from the chart. Further the reserved places between different consonants and vowels is also a right step towards developing an International Phonetic alphabet for all languages in the world based on Devanagari script.

There is further a need for smooth transition between the syllabic to phonemic and vice versa. It has been observed that around 1000 syllables cover almost 90% of any random text. So by assigning codes to these frequent 1000 syllables, will not only ensure the fast rendering but also offer saving in terms of storage requirement. Further viewing different Indian scripts as different fonts will enable a person to view text in any Indian language script through any other script, just by selecting appropriate fonts.

Key Words: Vedic Sanskrit, sandhi, Unicode, ISCII, UTF-8, Indian Language Scripts

Nature of Indian Scripts:

It is a well known fact that Indian Languages belong to four different language families. However, as far as scripts are concerned, all of them (except for Perso-Arabic script) are derived from the Brahmi script. Indian Language scripts are syllabic in nature. A vowel optionally preceded by one or more consonants is called a syllable. Following are some examples of syllables.

क = क्+अ
का = क्+आ
क्षे = क्+ ष्+ ए

On the right side of '=' in the above examples, is the expansion of syllables in alphabetic notation.

What we observe is the alphabetic representation of the syllables on the right side requires more space than the syllabic representation on the left side. In order to accommodate more text in less space, a concise orthography was evolved. Following are the salient features of this script.

a) 'अ' being the most frequent vowel in a syllable, the basic symbols of consonants were assumed to carry the vowel 'अ' within them. Unlike in Roman or Arabic scripts where different consonants are pronounced with different vowels (e.g. 'g' as 'gee', 'j' as 'je'), all the consonants in Indian scripts are pronounced with 'a' as in 'ka, ga, ta, etc.

b) When a vowel other than 'अ' is to be added to the consonant, then it is necessary to delete the inherent 'अ' and then add the other vowel.
e.g. कि = क्+इ = क + ◌् + इ

To represent this concisely in less space, the concept of the secondary vowel signs or 'matra' must have been evolved. There are around 12 vowels. To economize on space, then different symbols/signs representing different vowels might have been evolved. These symbols/signs are placed on the four sides of the consonant clusters viz. top, bottom, left and right.

By definition, syllable is a vowel optionally preceded by one or more consonants. Thus every syllable must have a vowel. Orthographically the symbol corresponding to a vowel may be either to the left, or right or top or bottom position of the given consonant cluster. This also answers the anomaly in 'बुद्धि' viz. the vowel indicator orthographically precedes the consonant cluster, but is pronounced later.

Further from

क्+इ = कि = क + ◌ि and
क् = क + ◌्

We get

क + ◌् + इ = क + ◌ि

or

◌ि = ◌् + इ

Thus vowel indicator is semantically equivalent to subtracting the inherent 'अ' vowel from the consonant and then adding the corresponding vowel to it. (Halanta sign represents the subtraction operation.)

c) In case of consonant cluster, the consonants are written either from top to bottom or from left to right. The vowel indicators will have the same positions. But they will appear to be with the top-most consonant (in case of indicators written on the top) or the bottom-most consonant (in case of indicators written at the bottom), and so on.

From the above discussion, what we observe is Indian language scripts are syllabic in nature, and it has been a tradition to fall back to the alphabetic expansion while analyzing the language string. Thus our scripts have flexibility and can be used as alphabetic or as syllabic as the need demands. Whereas the syllabic version is suitable for writing concisely, the alphabetic is suitable for performing the linguistic operations, such as sandhi operation, morphological analysis, sorting, searching, etc.

Further, though Indian Language scripts are syllabic in nature, the syllables are compositional. In this sense their development is post-alphabetic. These are developed after a thorough understanding of the script. The compositional syllabic nature of the scripts allows one to go from alphabetic to syllabic or vice versa smoothly. The natural choice for computation is the alphabetic expression while the syllabic expression is most suited for compact display. The text to speech package and the meter analysis in Chandas also require syllable as a basic unit.

Problems with ISCII:

ISCII has codes for both the matras (vowel indicators) as well as the vowels. Thus this scheme is neither purely phonetic nor purely syllabic, but a half-hearted combination of the two. Since matras (vowel indicators) are not independent entities, their presence in the chart leads to redundancy. Moreover, this also creates problems in the language analysis. For example, if we do not consider the inbuilt 'अ', then one needs separate sandhi rules to handle the following cases.

राम+ आलय = रामालय

देव+ आलय = देवालय

However, if we consider the underlying consonant - vowel combination, then we have

र्आम्अ + आल्अय्अ

and

द्एव्अ + आ ल्अय्अ

leading to a single rule for sandhi viz. 'अ + 'आ = 'आ

The search engines also have to do more work, to search the vowels and also for matras. People have tendency to use matras and the ignorance about the nature of our scripts leads them to write ि न to mean न िetc. Further since it is not purely syllabic, for the purpose of display as well as keyboard input, in any case we require a rendering engine.

Problems with Unicode:

Unicode is based on the ISCII-91. So Unicode also has all the problems which ISCII has. In addition it has a disadvantage of having different pages for Indian languages, leading to the diversity. ISCII had a single underlying code for all Indian Languages originating from Brahmi script. This ensured a unity among all the scripts. Further, the use of UTF-8

maps each Indian language 2 byte Unicode string to a 3 byte sequence.

Thus the transition from ISCII to UNICODE has led us to the WORST case – both in terms of unity as well as byte size.

Advantages of the proposed scheme for Vedic Sanskrit:

1. Rules for Linguistic analysis will be simple and will directly follow from the shastric texts without any modifications.
2. Redundancy of matras is gone.
3. The proposed InPa will be an important step towards evolving an IPA based on Devanagari.

Further Suggestions:

As pointed out earlier, though our scripts are syllabic in nature, the syllables are compositional, and are composed from the basic units viz. phonemes. Further we require the phonemic representation for linguistic analysis, whereas the syllabic representation is required for display, and few linguistic operations pertaining to chandas, etc. or speech synthesis that take syllable as a unit. So Indian languages require a smooth transition between the two. The compositionality of the syllables from the basic phonemes provide the basis for this smooth transition.

Following are the suggestions:

1. Treat all Indian Language scripts as different fonts with one underlying script – 'Indian Script'.

This will then allow one to change the script just by changing the fonts. One can view Indian language texts using ANY of the Indian language scripts just by changing the fonts. (This feature is the same as the one present in GIST technology.) This will help in easy accessibility of texts in Indian language to the other language speakers also. Further this will have more relevance in case of Sanskrit texts, where there is a practice of using different Indian scripts to write the Sanskrit texts.

2. Though our scripts are compositional, and theoretically it is possible to construct syllables with one or more number of consonants followed by an vowel leading to hundreds of thousands of such possible combinations, actual data suggests that there are only around 12000 syllables that are used very frequently. Of these around 1000 syllables account for the 90% coverage in almost all Indian languages. By treating all the Indian language scripts as fonts, one can pool a space of 10 pages of 128 codes each, space enough for storing these 1000 syllables together with the basic phonemes.

a) This will ensure not only quick rendering of the text, but also a fall back mechanism for those syllables which are not covered in these 1000 and need to be composed.
b) User can switch over between the phonemic and syllabic representation seamlessly.
c) While storing one can save the codes corresponding to syllables thereby saving storage space as well as transmission cost.
d) Other alternative is: Define the UTF-8 encoding based on the relative addresses

instead of absolute addresses. This will help in assigning 1 byte for Devanagari codes as well.

Conclusion:

The proposed phoneme base coding scheme is a right step towards removing the redundancy in the current scheme, which is neither purely phonemic nor syllabic.

This also is a step towards developing an Devanagari based Phonetic alphabet InPa, which may be extended further to cover all languages in the world.

The compositionally syllabic nature of our scripts allow a smooth transition from the phonemic to syllabic and vice versa transition. The orthographically our scripts being syllabic in nature, there is a need to have a fast rendering engine. Assigning codes to high frequency syllables and storing them as ready-made glyphs not only ensures fast rendering, but also save in storage space.

Bibiliography:

1. "Acharya: Multilingual Computing for Literacy and Education": http://**acharya.iitm.ac.in/**
2. Bharati, Akshar, Amba P Kulkarni, Vineet Chaitanya and Rajeev Sangal, "अनुवाद के उपकरण , संगणक तथा भाषायें" (Tools of Translation: Computer and Languages), University of Hyderabad, Distance Education Programme, Hyderabad, India, Feb 1998.
3. Devanagari (TDIL Newsletter Jan 2002) http://tdil.mit.gov.in/devanagari.pdf
4. ISCII - Indian Script Code for Information Interchange - ISCII Bureau of Indian Standards: New Delhi, 1991.
5. "ISFOC Standard for Fonts," http://www.cdac.in/html/gist/standard/isfoc.asp
6. "Microsoft Typography – Features of TrueType and OpenType", http://www.microsoft.com/typography/SpecificationsOverview.mspx
7. Vedic (TDIL Newsletter Oct 2002) http://tdil.mit.gov.in/Vedic.pdf