Unicode Standard for Vaidika Sanskrit: Efforts & Status

The ancient holy text in India is known by the name of the *Veda*. The word *Veda* is derived from the root verb *"vid"- to know, to attain the knowledge*. There are 4 *Vedas* namely *RigVeda, YajurVeda, SaamaVeda* and *AtharvaVeda*. Composed in verse format using Sanskrit language, the contents of these *Vedas* are of spiritual, philosophical and practical nature and are to be vocalized in terms of recitation, chanting, singing, etc. The collections of the *Mantras* (the compositions in poetry and prose form) of the four *Vedas* were further divided into *Shaakhaas* (branches). The range of Vaidika texts extend to their respective *Shaakhaas, Upashaakhaas* (branches-subbranches), *Samhitaas, Braahmannas, Aarannyakas, Upanishads, Shrautasuutras, Praatishaakhyas* and *Shikshaas*. Originally, the Vaidika text (mostly in verse form) –*Samhitaas* along with their respective *Padapaathas* exist in the oral mode and were passed on from *Guru* to *Shishya* through various processes such as *Shravanna* (listening), *Anuchchaaranna* (repetition), *Arthabodha* (understanding of the meaning), *Kannthasthiikaranna* (memorisation), *Charchaa-Samvaada* (discussion) and *Prashnottara* (Questions &Answers).

The classes of reciters known as Vaidikas belonging to individual branches have retained their chanting characteristics while vocalizing the *Mantras*. At later stage such oral recitations had to be written down in order to retain the knowledge in the written mode. To write such Vaidika text with vocal nuances, various signs (*svaraankana-paddhati*) were devised and placed around syllables. These are called Vaidika accent and tonal signs.

There is special Vedic grammar, rules for each *Sakha* known as *Prati Sakhya* and phonetic rules known as *Siksha.* There are four *Upangas: Mimansa Sutra-s* (described rules for interpretation of *Vedic* text), *Nyaya & Vaisheshika sutra-s* (deal with logical aspects, ontological classification, process of human understanding), *Purana-s* (narrations of messages and teachings of Veda-s, *Dharma Sastra-s* (describe code of conduct for universal harmony). The *Shikshaas* cover the study of the terms such as *Varna* (letters of alphabet), *Svara* - (accents like *Udaatta, Anudaatta* and *Svarita*), *Maatraa* (the time measurement required for the pronunciation), *Bala* (stress - in context to *sthaana* and *prayatna*) etc. The *Praatishaakhyas* constitute studies on phonetics, which prevailed in ancient India in different schools of the *Vedas*.

It is observed that the Vaidika text and its accent marks vary in their application and presentation as per its branch, sub-branch, tradition and location. It is interesting to note that a single chant if vocalised by the *Vaidikas* of different branches, would have slight variation in terms of nasalization, intonation, continuum, stress etc. And to indicate these differences, many variants of a single sign have been developed. For example, there are 15 variations of *Anusvaara* indicated by 15 graphic signs with slight variations in their visual forms. The selective usage of these signs indicates that the particular text is from a particular branch of *Veda*.

For example:

In *RigVeda*, the *Udaatta Svara* is left unmarked. In *Rigveda* and *YajurVeda* the *Svarita Svara* is marked with short vertical line on the top of a syllable. The same sign appears as *Anudaatta* mark below a syllable in *AtharvaVeda* (*Paippalaada Shaakhaa*).

Sanskrit

Sanskrit is one of the most ancient languages of the world, which has molded the culture and the thought systems not only of India but also of many other countries in Asia. Sanskrit language is variously referred to as *Devavani, Amarvani, Girvanavani, Surbharati, Amarbharati*, etc. each expressing connoting its inherent vitality, versatility and greatness. The script in which Sanskrit texts are written is called Devanagari. Sanskrit has its own place recognized by the linguists all over the world. It is agreed that the study of Sanskrit enables us to draw freely upon our tradition, which can lead to the newworld outlook of modern man. Further, it helps to keep pace with the rapid social change, advances in modern science & technology and the process of modernization, at the same time inculcating the right type of social, moral and spiritual value through self-discipline.

Sanskrit is a living spoken language with a considerable literature of its own. Sanskrit is still spoken in some Indian families. Its vocabulary has permeated all Indian languages, and thus provides continuity with the past of our country. Panini's book on Sanskrit Grammar, named Ashtadhyayi, has been considered by eminent American linguist Bloomfield as "one of the greatest monuments of human intelligence". Panini was preceded by a long chain of grammarians, and his tradition continued even afterwards. With his 4000 sutras, each of which is usually no more than two or three words, Panini was able to explain how almost all the words used in Sanskrit of his time were formed. It is precursor of today's generative grammar. Sanskrit grammar is prescriptive, that takes phoneme as smallest unit with meaning, knowledge representation is deeper and holistic at sentence level with three necessary and sufficient conditions of Expectancy, Compatibility and Proximity whereas modern linguistics is descriptive and empirical, that takes morpheme as the smallest unit with meaning, uses word-by-word approach rather than sentence respectively.

Sanskrit is syntax-free and word-order-free natural language. Shastric Sanskrit is the Natural Language with all the desirable properties of formal artificial language, such as naturalness, expressiveness, unambiguous and no redundancy.

Unicode is the defacto industry standard for representation of multilingual text in computer/ electronic media. Major scripts of the world are included in the standard including major Indian scripts. Unicode Standard is maintained by the Unicode Consortium and is in sync with ISO10646.

TDIL Programme is the member of Unicode Consortium with the objective of adequate representation of constitutionally recognized 22 languages (using 11 different scripts) in the Unicode Standard. TDIL programme is also coordinating the activities for representation/ encoding of the Vaidika Symbols and Characters in the Unicode Standard. The Devanagari script is used for writing classical Sanskrit as well as Vedic Sanskrit. This includes the multi-tier usage of diacritic marks of complex compositions, above, below and at the sides of the base glyphs and various other symbols known as Swaradi Chihna, Visarga, Bala, Gomukha, etc.

The task of identifying all such symbols, accent markers and characters used in Vaidika literature was very exhaustive. Prof. R K Joshi, CDAC Mumabi was entrusted the task of identifying all additional symbols/ characters which were required to be encoded in the Unicode Standard. Prof. Joshi consulted many Vedic experts, studied many manuscripts, visited many institutes/ libraries across the India to find out various symbols which were needed for representation of Vedic Sanskrit in Unicode Standard. With his hard work and efforts the first draft was prepared the year 2002. This draft was based on pure consonant approach and suggested encoding of the Devanagari pure consonants also alongwith symbols/ characters needed for Vaidika. The proposal was submitted to Unicode Technical committee in the year 2003. Unicode technical committee suggested modification in the proposal as the proposed encoding of Devanagari pure consonants was not as per the Unicode Encoding Policies. Also, UTC advised to submit detailed evidences for the usage of these symbols/ characters. Prof. Joshi & TDIL Programme worked to prepare the detailed evidences which were subsequently submitted to Unicode Technical Committee.

The matter of encoding was further pursued with UTC in May 2005. Mrs. Swaran Lata, Director, DIT participated in the Unicode Technical Committee meeting and presented/ discussed the modified proposal on the encoding of the Vedic Symbols and Characters in the Unicode Standard.

Brown University USA, who are working on the Sanskrit Libraries also submitted another proposal to Unicode Consortium for the encoding of the same. The draft proposal submitted by Govt. of India and the proposal

submitted by Brown University were again discussed during UTC#113. UTC recommended that Government of India will submit a unified revised proposal for encoding of Vaidika Sanskrit.

On behalf of DIT, GOI, Prof. R K Joshi, CDAC Mumbai further worked and prepared revised proposal (as advised by UTC) in consultation with the other experts in the area. DIT organized a number of meetings of Vaidika Scholars & experts from all over the India for consultations. The proposal was also uploaded on the TDIL programme website (http://www.tdil.mit.gov.in) for feedback and comments.

DIT submitted the final proposal to Unicode Consortium in January 2008 (L2/08-042 & L2/08-043) for consideration of Unicode Technical Committee (UTC). The proposal was first discussed by the South Asia Sub-Committee and later by the UTC. The UTC recommended most of the proposed characters for inclusion in the future version of the standard.

UTC recommended encoding of various characters and symbols required for representation of Vaidika text in three different blocks i.e. Devanagari, Vedic Extensions and Devanagari Extensions.

In the Vedic Extensions block (1CD0 – 1CFF), 35 characters have been recommended for encoding. 26 characters have been recommended for encoding in the Devanagari Extended block (A8E0 – A8FF). Four characters were recommended for encoding in the Devanagari block (U+A8E0 – U+A8F9).

The UTC recommended proposal was further discussed in the 53rd Meeting of the ISO Working Group-2 "Universal Coded Set (UCS)" of JTC1/SC2 in Hong Kong, China during October 13th -17th , 2008. Mrs. Swaran Lata, Director and HOD, Human Centred Computing Division, DIT participated in the meeting as the representative of Government of India and presented/ discussed the proposal. The proposal has now been accepted by ISO also and after the April 2009 meeting of the above working group of ISO, the code set will be included in the future version of ISO10646 and Unicode Standard.

The charts shown below shows the various characters accepted for encoding for the representation of Vaidika characters and symbols as recommended by Department of Information Technology, Ministry of Communications & IT, Government of India for inclusion in the future version of ISO10646 and Unicode Standard.

1. Devanagari (U+900 – U+97F)

	090	091	092	093	094	095	096	097
0	◌ँ 0900	ऐ 0910	ठ 0920	र 0930	◌ी 0940	ॐ 0950	ॠ 0960	० 0970
1	◌ँ 0901	ऒ 0912	ड 0921	ऱ 0931	◌ु 0941	◌॑ 0951	ॡ 0961	ॱ 0971
2	◌ं 0902	ओ 0912	ढ 0922	ल 0932	◌ू 0942	◌॒ 0952	◌ॢ 0962	ॲ 0972
3	◌ः 0903	ओ 0913	ण 0923	ऴ 0933	◌ृ 0943	◌॓ 0953	◌ॣ 0963	▨
4	ऄ 0904	औ 0914	त 0924	ऴ 0934	◌ॄ 0944	◌॔ 0954	। 0964	▨
5	अ 0905	क 0915	थ 0925	व 0935	◌ॅ 0945	◌ॕ 0955	॥ 0965	▨
6	आ 0906	ख 0916	द 0926	श 0936	◌ॆ 0946	▨	० 0966	▨
7	इ 0907	ग 0917	ध 0927	ष 0937	◌े 0947	▨	१ 0967	▨
8	ई 0908	घ 0918	न 0928	स 0938	◌ै 0948	क़ 0958	२ 0968	▨
9	उ 0909	ङ 0919	ऩ 0929	ह 0939	◌ॉ 0949	ख़ 0959	३ 0969	ज़ 0979
A	ऊ 090A	च 091A	प 092A	▨	◌ॊ 094A	ग़ 095A	४ 096A	य 097A
B	ऋ 090B	छ 091B	फ 092B	▨	◌ो 094B	ज़ 095B	५ 096B	ग 097B
C	ऌ 090C	ज 091C	ब 092C	◌़ 093C	◌ौ 094C	ड़ 095C	६ 096C	ज 097C
D	ऍ 090D	झ 091D	भ 092E	ऽ 093D	◌् 094D	ढ़ 095D	७ 096D	? 097D
E	ऎ 090E	ञ 091E	म 092F	◌ा 093E	◌ॎ 094E	फ़ 095E	८ 096E	ड 097E
F	ए 090F	ट 091F	य 0920	ि◌ 093F	▨	य़ 095F	९ 096F	ब 097F

2. Vedic Extensions (U+1CD0 – U+1CFX)

	1CD	1CE	1CF
0	1CD0	1CE0	1CF0
1	1CD1	1CE1	1CF1
2	1CD2	1CE2	1CF2
3	1CD3	1CE3	
4	1CD4	1CE4	
5	1CD5	1CE5	
6	1CD6	1CE6	
7	1CD7	1CE7	
8	1CD8	1CE8	
9	1CD9	1CE9	
A	1CDA	1CEA	
B	1CDB	1CEB	
C	1CDC	1CEC	
D	1CDD	1CED	
E	1CDE	1CEE	
F	1CDF	1CEF	

3. Devanagari Extended (U+A8E0 – U+A8FX)

	A8E	A8F
0	◌॒ (A8E0)	वि (A8F0)
1	◌॑ (A8E1)	ऽ (A8F1)
2	◌॒ (A8E2)	◌ (A8F2)
3	◌ (A8E3)	◌ (A8F3)
4	◌ (A8E4)	◌ (A8F4)
5	◌ (A8E5)	◌ (A8F5)
6	◌ (A8E6)	◌ (A8F6)
7	◌ (A8E7)	◌ (A8F7)
8	◌ (A8E8)	�49 (A8F8)
9	◌ (A8E9)	**ᴐ (A8F9)**
A	अ (A8EA)	◌ (A8FA)
B	उ (A8EB)	**— (A8FB)**
C	क (A8EC)	
D	न (A8ED)	
E	प (A8EE)	
F	ज (A8EF)	

(blue)	Recently proposed characters

DEVANAGARI BLOCK
Additions in Devanagari Block:

0900	◌̐	DEVANAGARI SIGN INVERTED CANDRABINDU =*vaidika adhomukha candrabindu*
0955	◌̆	DEVANAGARI VOWEL SIGN CANDRA LONG E • *used in transliteration of Avestan*

Other Consonants

0979	ज़	DEVANAGARI LETTER ZHA • *used in transliteration of Avestan*
097A	य	DEVANAGARI LETTER HEAVY YA • *used for an affricate glide JJYA*

Prishthamatra Orthography Vowel Signs

094E	ꠁ◌	DEVANAGARI VOWEL SIGN PRISHTHAMATRA E •*Character has historic use only*

VEDIC EXTENSIONS (U+1CD0 – U+1CFX)
Tone marks for the Samaveda

UTC Code	Glyph	Name of character/ symbol
1CD0	◌̂	VEDIC TONE KARSHANA = *vaidika saamasvara karshanna*
1CD1	◌̑	VEDIC TONE SHARA = *vaidika svarita uurdhva shara*
1CD2	◌̄	VEDIC TONE PRENKHA = *vaidika saamasvara prenkha* • *indicates vibrato*
1CD3	⸽	VEDIC SIGN NIHSHVASA = *vaidika saamagaana yogakaala*

Signs for Yajurvedic

1CD4	◌꠴	VEDIC SIGN MID-CHARACTER SVARITA = *vaidika svarita urasi rekhaa*
1CD5	◌̣	VEDIC TONE YAJURVEDIC AGGRAVATED INDEPENDENT SVARITA = *vaidika svarita adho nyubja*
1CD6	◌̣	VEDIC TONE YAJURVEDIC INDEPENDENT SVARITA = *vaidika svarita adhah konna*
1CD7	◌̣	VEDIC TONE YAJURVEDIC KATHAKA INDEPENDENT SVARITA = *vaidika svarita adho vakra rekhaa*
1CD8	◌̣	VEDIC TONE CANDRA BELOW = *vaidika svarita adho ardha vakra*
1CD9	◌̣	VEDIC TONE YAJURVEDIC KATHAKA INDEPENDENT SVARITA SCHROEDER =*vaidika svarita adho samyukta rekhaa*
1CDA	◌̈	VEDIC TONE DOUBLE SVARITA = *vaidika svarita uurdhva dvi rekha*
1CDB	◌⃛	VEDIC TONE TRIPLE SVARITA = *vaidika svarita uurdhva tri rekhaa*
1CDC	◌̗	VEDIC TONE KATHAKA ANUDATTA = *vaidika svarita adho rekhaa*
1CDD	◌̣	VEDIC TONE DOT BELOW = *vaidika svarita adho bindu*

Tone marks for the Satapathabrahmana

1CDE	◌̣	VEDIC TONE TWO DOTS BELOW *= vaidika svarita adho dvi bindu*
1CDF	◌̤	VEDIC TONE THREE DOTS BELOW *= vaidika svarita adho tri bindu*

Tone mark for the Rigveda

|		
1CE0	◌	VEDIC TONE RIGVEDIC KASHMIRI INDEPENDENT SVARITA *=vaidika uurdhva vakra rekhaa*

Tone mark for the Atharvaveda

1CE1	◌	VEDIC TONE ATHARVAVEDIC INDEPENDENT SVARITA *= vaidika svarita dvi vakra khannda*

Diacritics for Visarga

1CE2	◌	VEDIC SIGN VISARGA SVARITA *=vaidika madhyarekhaa*
1CE3	◌	VEDIC SIGN VISARGA UDATTA *= vaidika visarga dakshinnatah uurdhvaga*
1CE4	◌	VEDIC SIGN REVERSED VISARGA UDATTA *= vaidika visarga vaamatah uurdhvaga*
1CE5	◌	VEDIC SIGN VISARGA ANUDATTA *= vaidika visarga vaamatah adhoga*
1CE6	◌	VEDIC SIGN REVERSED VISARGA ANUDATTA *= vaidika visarga dakshinnatah adhoga*
1CE7	◌	VEDIC SIGN VISARGA UDATTA WITH TAIL *= vaidika visarga dakshinnatah uurdhva vakra*
1CE8	◌	VEDIC SIGN VISARGA ANUDATTA WITH TAIL *= vaidika visarga vaamatah adho vakra*

Marks of nasalization

1CE9	೮	VEDIC SIGN ANUSVARA ANTARGOMUKHA *= vaidika anusvaara antarmukha*
1CEA	೮	VEDIC SIGN ANUSVARA BAHIRGOMUKHA *= vaidika anusvaara naagaphanna*
1CEB	೮	VEDIC SIGN ANUSVARA VAMAGOMUKHA *= vaidika anusvaara vaamamukha*
1CEC	೮	VEDIC SIGN ANUSVARA VAMAGOMUKHA WITH TAIL *= vaidika anusvaara vaamagomukha sa-vakra*

1CED	◌	VEDIC SIGN TIRYAK *=vaidika tiryak*
1CEE	೮	VEDIC SIGN HEXIFORM LONG ANUSVARA *=vaidika anusvaara anugamii*
1CEF	೮	VEDIC SIGN LONG ANUSVARA *=vaidika anusvaara dakshinnamukha*
1CF0	೮	VEDIC SIGN RTHANG LONG ANUSVARA *=vaidika anusvaara ttha-sadrisha*

1CF1	೮	VEDIC SIGN ANUSVARA UBHAYATOMUKHA *=vaidika anusvaara ubhayatomukha*

Ardhavisarga

1CF2	◌	VEDIC SIGN ARDHAVISARGA *=vaidika jihvaamuliiya upadhamaaniiya*

DEVANAGARI EXTENSIONS (U+A8E0 – U+A8FX)
Cantillation marks for the Samaveda

A8E0	◌ॠ	COMBINING DEVANAGARI DIGIT ZERO =*vaidika saamasvara anka shuunya*
A8E1	◌	COMBINING DEVANAGARI DIGIT ONE =*vaidika saamasvara anka eka udaatta*
A8E2	◌	COMBINING DEVANAGARI DIGIT TWO =*vaidika saamasvara anka dvi svarita*
A8E3	◌	COMBINING DEVANAGARI DIGIT THREE =*vaidika saamasvara anka tri anudaatta*
A8E4	◌	COMBINING DEVANAGARI DIGIT FOUR =*vaidika saamasvara anka chatur*
A8E5	◌	COMBINING DEVANAGARI DIGIT FIVE =*vaidika saamasvara anka pancha*
A8E6	◌	COMBINING DEVANAGARI DIGIT SIX =*vaidika saamasvara anka shatt*
A8E7	◌	COMBINING DEVANAGARI DIGIT SEVEN =*vaidika saamasvara anka sapta*
A8E8	◌	COMBINING DEVANAGARI DIGIT EIGHT =*vaidika saamasvara anka ashta*
A8E9	◌	COMBINING DEVANAGARI DIGIT NINE =*vaidika saamasvara anka nava*
A8EA	◌	COMBINING DEVANAGARI LETTER A =*vaidika saamasvara abhinihita*
A8EB	◌	COMBINING DEVANAGARI LETTER U =*vaidika saamasvara u*
A8EC	◌	COMBINING DEVANAGARI LETTER KA =*vaidika saamasvara ka*
A8ED	◌	COMBINING DEVANAGARI LETTER NA = *vaidika saamasvara namana*
A8EE	◌	COMBINING DEVANAGARI LETTER PA =*vaidika saamasvara prannatam*
A8EF	◌	COMBINING DEVANAGARI LETTER RA =*vaidika saamasvara ra*
A8F0	◌	COMBINING DEVANAGARI LETTER VI =*vaidika saamasvara vinata*
A8F1	◌	COMBINING DEVANAGARI SIGN AVAGRAHA =*vaidika saamasvara diirghiibhaava*

Marks of Nasalization

A8F2	◡	DEVANAGARI SIGN SPACING CANDRABINDU =*vaidika candrabindu*
A8F3	◌	DEVANAGARI SIGN CANDRABINDU VIRAMA =*vaidika anusvaara candrabindu tiryak*
A8F4	◌	DEVANAGARI SIGN DOUBLE CANDRABINDU VIRAMA =*vaidika anusvaara dvi candrabindu tiryak*
A8F5	◌	DEVANAGARI SIGN CANDRABINDU TWO =*vaidika anusvaara candrabindu sa-dvi*
A8F6	◌	DEVANAGARI SIGN CANDRABINDU THREE =*vaidika anusvaara candrabindu sa-tri*
A8F7	◌	DEVANAGARI SIGN CANDRABINDU AVAGRAHA =*vaidika anusvaara candrabindu sa-avagraha*
A8F8	ఌ	DEVANAGARI SIGN PUSHPIKA =*vaidika pushpika* • used as a placeholder or "filler" • often flanked by double dandas

A8F9	ʘ	DEVANAGARI GAP FILLER used to indicate that a gap in a manuscript or text is not a lacuna
A8FA	˅	DEVANAGARI CARET *=vaidika trutikaa* • zero-width character centred on the point between two orthographic syllables
A8FB	‒	DEVANAGARI HEADSTROKE *=shirorekhaa* used to indicate uncertain manuscript reading

References:
1. **Vishwabharat@tdil: Language Technology Flash Vol. October 2002**
2. **www.unicode.org**

Courtesy:
Swaran Lata, Scientist 'F' & HOD
Manoj Jain, Scientist 'D'
Technology Development for Indian Languages Programme
Department of IT, New Delhi 110001