

Issues in Representation of Indic Scripts in Unicode

L2/05-063

Dr. Om Vikas

Government of India

Ministry of Communications & Information Technology

Department of Information Technology

Email: omvikas@mit.gov.in

UTC # 102

Unicode Technical Committee Meeting
February 7-10, 2005, Mountain View, CA, USA

Contents

- Characteristics of Indic script.
- India's initiative for Language Technology Development
- Awaiting updates in Unicode.
- Proposal for New Scripts

Characteristics of Indic Scripts

Linguistic Scenario in India

- India is Multilingual Multiscript Country
- Twenty Two constitutionally recognized Indian Languages are mentioned as follows with their scripts within parentheses: Hindi (*Devanagari*), Konkani (*Devanagari*), Marathi (*Devanagari*), Nepali (*Devanagari*), Sanskrit (*Devanagari*), Sindhi (*Devanagari/Urdu*), Kashmiri (*Devanagari/Urdu*); Assamese (*Assamese*), Manipuri (*Manipuri*), Bangla (*Bangali*), Oriya (*Oriya*), Gujarati (*Gujarati*), Punjabi (*Gurumukhi*), Telugu (*Telugu*), Kannada (*Kannada*), Tamil (*Tamil*), Malayalam (*Malayalam*) Urdu (*Urdu*), Bodo (*Devanagari, Assamese*), Dogri (*Devanagari*), Maithili (*Devanagari*), Santhali (*Ol Chiki*). There are 10 Indic Scripts in vogue.
- Less than 5 percent of people can either read & write English. Over 95 percent population is normally deprived of the benefits of English-based Information Technology.

Characteristics of Indian Languages:

- What You Speak Is What You Write (WYSIWYW)
- Script grammar describes transformation rules
- Relatively word-order-free
- Common phonetic based alphabet
- Common concept terms (from Sanskrit)

Indian languages owe their *origin to Sanskrit*, hence they have in common rich cultural heritage and treasure of knowledge. Indic scripts have originated from Brahmi script.

Script Grammar

In *linear scripts* such as Roman, Characters juxtapose – one after another. Characters do not change their shape.

In *non-linear scripts* such as Indic scripts, Characters may change shape. Indian Script Grammar describes rendition rules:

- There is a grapheme for a phoneme
- Consonants and vowels do not always juxtapose.
- Consonant is represented as a character that depicts pure consonant (C) plus embedded vowel sound /a/.

Ka	Ga	Pa	Ba	Ma
----	----	----	----	----

क	ग	प	ब	म
---	---	---	---	---

- Pure Consonant is represented with Halant (्): omission of embedded vowel /a/

क (ka) + ् = क् (k)

-Vowel (V) changes its shape when it follows a consonant (C) or consonant cluster (CC). This change in shape is represented by a Vowel sign called 'matra'.

Vowels in Devanagari:

अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ
a	aa	i	ii	u	uu	e	ai	o	au

Vowel-sign

◌	◌ा	◌ि	◌ी	◌ु	◌ू	◌े	◌ै	◌ो	◌ौ
---	----	----	----	----	----	----	----	----	----

CV:

क	का	कि	की	कु	कू	के	कै	को	कौ
ka	kaa	ki	kii	ku	kuu	ke	kai	ko	kau

- Consonant Cluster results into consonants fusion or a new conjunct.

CC:

क् + व् + आ = क्वा

k + v + aa = kvaa (fusion)

क् + ष् + अ = क्ष

K + s'h + a = ks'ha (conjunct)

प् + ऋ + थ् + व् + ई = पृथ्वी

p + r'i + th + v + ii = pr'ithvii

- Vowel sign (matra) may split on either side of a consonant in some Indic scripts.

e.g. Tamil: ொ (Tamil Vowel Sign O), ோ (Tamil Vowel Sign AU)

Bengali: ୌ (Bengali Vowel Sign O), ୊ (Bengali Vowel Sign AU)

- CCCV is also seen
In Devanagari मर्त्य स्वास्थ्य
 martya swaasthya
 CV CCCV CCV CCCV
- Consonant 'r' (र्) also change its shape differently when it precedes or succeeds consonant
 'r' precedes पूर्व (puu rva)
 'r' succeeds चक्र (cha kra)
- Nasal consonant combines only with a character in the same group.
 पन्त पम्प पङ्गु
 panta pampa panggu
- Nasal consonants may transform into anuswar (ँ) also
 पंत पंगु
 panta pangu

-There are varying ligatures for CCV combinations

kka may be written as क्क or क्क

kta may be written as क्त or क्त

- Rendering of Vowel sign may vary depending on the associated consonant /consonants cluster.

e.g रि कि मि स्थि

ri ki mi sthi

rendering of “i ” (ि) matra (vowel sign) has upper alignment: shortest, shorter, short & long respectively for र् (r), क् (k), म् (m) & स्त् (sth). Hence this has 4 different glyphs..

Devanagari alphabet has been enhanced to include characters for language-specific speech sounds.

India's initiative for Language Technology Development

TDIL Programme

Vision statement

Digital unite and knowledge for all.

Mission statement

Communicating without language barrier & moving up the knowledge chain.

INITIATIVES

- **Knowledge Resources**
(Parallel Corpora, Multilingual Libraries/Dictionaries, lexical resources)
- **Knowledge Tools**
(Language Processing Tools)
- **Translation Support Systems**
(Machine Translation, Cross Language Information Retrieval)
- **Human Machine Interface System**
(OCR, Voice Recognition Systems, Text-to-Speech System)
- **Localization**
(Localisation of Operating System, Software and Content)
- **Language Technology Human Resource Development**
(in Computational Linguistics & Knowledge Engineering)
- **Standardization**
(ISCII, Unicode, XML, INSFOC, MPEG, Terminology, etc.)

Long Term Goals

- **Speech - to - Speech translation.**
- **Human Inspiring Systems**

Towards Indic scripts representation in Unicode :

- **Consensus Building:**

Ministry of Communications & Information Technology, Government of India, organized a number of discussion meetings of Experts/ Linguistic groups/ Industry to identify the solutions to overcome the deficiencies in the existing Unicode Standard for Indian scripts.

It took pretty good time to reach to an agreement so that all the Indian languages are properly represented.

- **Language Design Guides**

In consultation with the linguists/ experts, the Language Design Guides have been prepared for all the constitutionally recognized Indian Language to ease the Language Technology Industry and user groups. These should be used for providing detailed explanation about all the Indic scripts in the Unicode Standard.

Indic Script Encoding:

- Initially phonetic based encoding was for 7-bit code-set in the single byte (8-bit) code in order to co-exist with Roman. National standard ISCII (Indian Scripts Standard Code for Information Interchange). This is base in the Unicode.
- Content creation requires glyph-coding. Propriety glyph-codes hamper spread of IT in local languages.
- Glyph encoding for Indic Scripts would require over 300 code points.
- There are discrepancies in representation of Indic scripts in Unicode.

Awaiting Update in Unicode
&

Proposal for New Scripts

Enhanced Devanagari Code Chart

Proposed changes in the Unicode Standards for Indic Scripts - Devanagari

	090	091	092	093	094	095	096	097
0		ऐ 0910	ठ 0920	र 0930	ी 0940	ॐ 0950	ऋ 0960	० 0970
1	ँ 0901	ऑ 0911	ड 0921	ॡ 0931	ु 0941	ं 0951	ॠ 0961	ॠ 0971
2	ं 0902	औ 0912	ढ 0923	ल 0932	ू 0942	॒ 0952	ॡ 0962	क्ष 0972
3	ः 0903	ऐ 0913	ण 0923	ळ 0933	ृ 0943	े 0953	ॢ 0963	ज्ञ 0973
4	ओ 0904	औ 0914	त 0924	ॠ 0934	ॢ 0944	ँ 0954	॥ 0964	श्र 0974
5	अ 0905	क 0915	थ 0925	व 0935	ँ 0945	ॡ 0955	॥ 0965	ँ 0975
6	आ 0906	ख 0916	द 0926	श 0936	ै 0946	ॢ 0956	० 0966	श 0976
7	इ 0907	ग 0917	ध 0927	ष 0937	े 0947	ॡ 0957	१ 0967	ल 0977
8	ई 0908	घ 0918	न 0928	स 0938	ै 0948	क्र 0958	२ 0968	८ 0978
9	उ 0909	ङ 0919	ॡ 0929	ह 0939	ॉ 0949	ख 0959	३ 0969	ग 0979
A	ऊ 090A	च 091A	प 092A	० 093A	ौ 094A	ग 095A	४ 096A	ज 097A
B	ऋ 090B	छ 091B	फ 092B		ो 094B	ज 095B	५ 096B	ड 097B
C	ॠ 090C	ज 091C	ब 092C	॒ 093C	ौ 094C	ड़ 095C	६ 096C	ब 097C
D	ँ 090D	झ 091D	भ 092D	ऽ 093D	॑ 094D	ढ 095D	७ 096D	
E	ऐ 090E	ञ 091E	म 092E	ा 093E		फ़ 095E	८ 096E	
F	ए 090F	ट 091F	य 092F	ि 093F		य़ 095F	९ 096F	

Encoding of Vedic Symbols & Characters

Government of India proposed to encode Vedic Symbols and characters at the Code Space U+0800 to U+08FF, but so far space has not reserved for Vedic symbols and characters. Space for these may be reserved in BMP immediately and the process of investigation for encoding these symbols and characters should be expedited. DIT is ready to provide all required support for the same.

New Proposal for Vedic Sanskrit (Submitted a year ago)

Vedic Code chart - http://www.tdil.mit.gov.in/prop_uni/Vedic.pdf

Proposed Vedic Code Set Page 1

	080	081	082	083	084	085	086	087
0	० 0800	ॠ 0810	ऐ 0820	ॡ 0830	ॢ 0840	द 0850	ॣ 0860	। 0870
1	॥ 0801	इ 0811	ॢ 0821	क 0831	झ 0841	ध 0851	ॣ 0861	ARLU 0871
2	२ 0802	ॠ 0812	ओ 0822	ॡ 0832	ॢ 0842	न 0852	ल 0862	। 0872
3	३ 0803	ई 0813	ओ 0823	ख 0833	ज 0843	ॣ 0853	ॣ 0863	। 0873
4	४ 0804	उ 0814	ॢ 0824	ॢ 0834	द 0844	ॣ 0854	व 0864	। 0874
5	५ 0805	ॢ 0815	ऑ 0825	ग 0835	उ 0845	प 0855	ॣ 0865	। 0875
6	६ 0806	ऊ 0816	औ 0826	ॢ 0836	इ 0846	फ 0856	श 0866	० 0876
7	७ 0807	ॢ 0817	ॠ 0827	घ 0837	ॢ 0847	ॢ 0857	ष 0867	। 0877
8	८ 0808	ॠ 0818	ॠ 0828	इ 0838	ॢ 0848	ब 0858	स् 0868	। 0878
9	९ 0809	ॠ 0819	ॠ 0829	च 0839	ह 0849	ॢ 0859	ह 0869	SNDH 0879
A	अ 080A	लृ 081A	ॠ 082A	ॢ 083A	ॢ 084A	भ 085A	ळ 086A	ASND 087A
B	ॢ 080B	लृ 081B	ॠ 082B	छ 083B	ण 084B	म् 085B	ॢ 086B	RIKT 087B
C	ॢ 080C	ऐ 081C	ॠ 082C	ॢ 083C	ॢ 084C	य 085C	ॢ 086C	ARRI 087C
D	ॢ 080D	ए 081D	ॠ 082D	ज 083D	त् 084D	ॢ 085D	ळह 086D	DRSK 087D
E	आ 080E	ऐ 081E	ॠ 082E	ॢ 083E	ॢ 084E	र 085E	ॢ 086E	VIST 087E
F	ॢ 080F	ॢ 081F	ॢ 082F	ॢ 083F	थ 084F	ॢ 085F	ॢ 086F	ॢ 087F

Proposed Vedic Code Set Page 2

	088	089	08A	08B	08C	08D	08E	08F
0	 0880	 0890	 08A0	 08B0	 08C0	 08D0	 08E0	 08F0
1	 0881	 0891	 08A1	 08B1	 08C1	 08D1	 08E1	 08F1
2	 0882	 0892	 08A2	 08B2	 08C2	 08D2	 08E2	 08F2
3	 0883	 0893	 08A3	 08B3	 08C3	 08D3	 08E3	 08F3
4	 0884	 0894	 08A4	 08B4	 08C4	 08D4	 08E4	 08F4
5	 0885	 0895	 08A5	 08B5	 08C5	 08D5	 08E5	 08F5
6	 0886	 0896	 08A6	 08B6	 08C6	 08D6	 08E6	 08F6
7	 0887	 0897	 08A7	 08B7	 08C7	 08D7	 08E7	 08F7
8	 0888	 0898	 08A8	 08B8	 08C8	 08D8	 08E8	 08F8
9	 0889	 0899	 08A9	 08B9	 08C9	 08D9	 08E9	 08F9
A	 088A	 089A	 08AA	 08BA	 08CA	 08DA	 08EA	 08FA
B	 088B	 089B	 08AB	 08BB	 08CB	 08DB	 08EB	 08FB
C	 088C	 089C	 08AC	 08BC	 08CC	 08DC	 08EC	 08FC
D	 088D	 089D	 08AD	 08BD	 08CD	 08DD	 08ED	 08FD
E	 088E	 089E	 08AE	 08BE	 08CE	 08DE	 08EE	 08FE
F	 088F	 089F	 08AF	 08BF	 08CF	 08DF	 08EF	 08FF

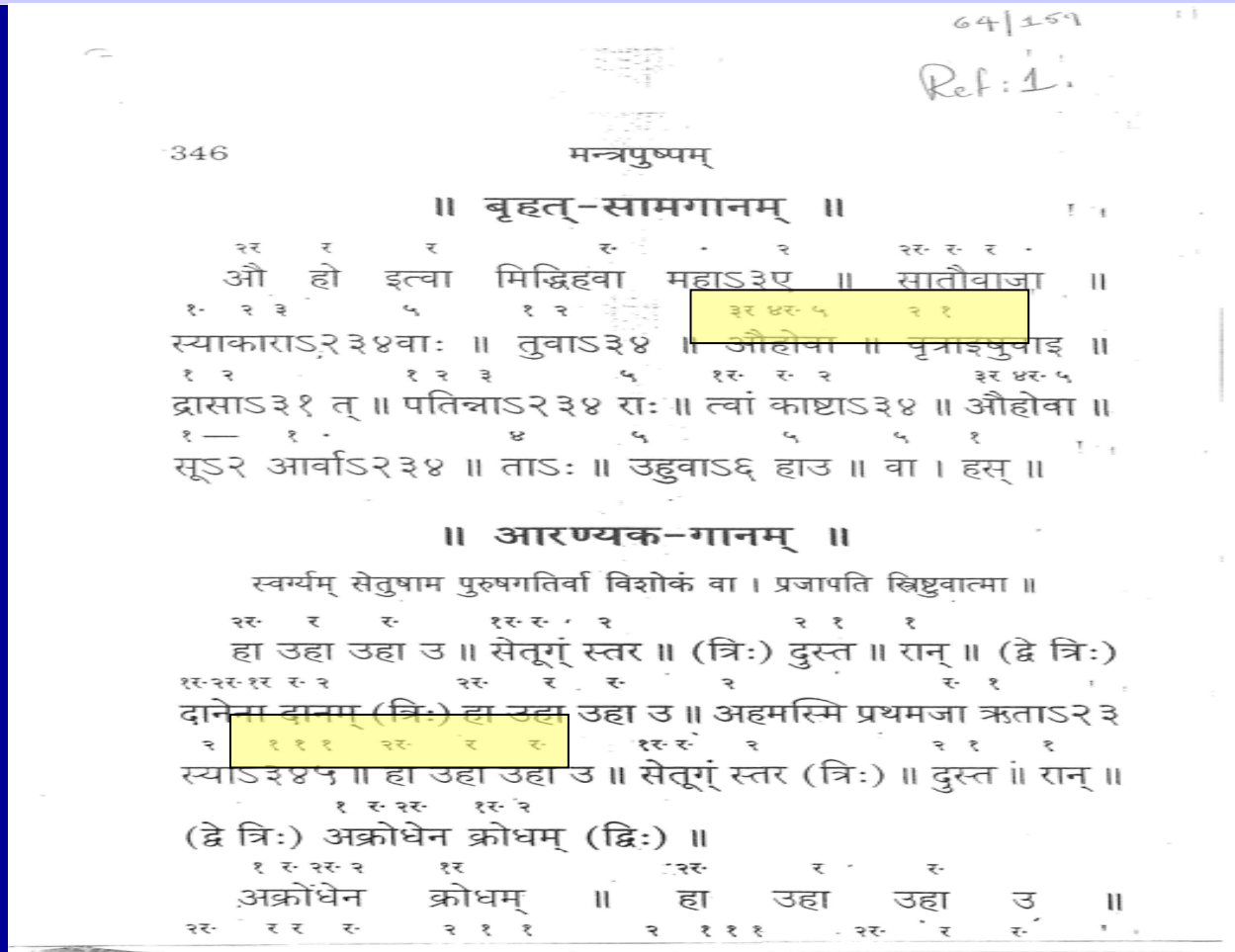
Typical Evidences for Vedic Characters and Symbols

ऋषिश्राद्धम् । ७०/१२५ (२६९) २६९।
धर्मिणो विसख्यानि विसृजामह इति च सर्वे जपेयुः विरताः
स्मेति प्रवृत्तः ॥ इत्युपिश्राद्धम् ॥
अथ विसर्जनम् । ॐ उत्तिष्ठ ब्रह्मणस्पते देवयन्तस्त्वे
महे ॥ उपप्रयन्तु मरुतः सुदानव इन्द्रं प्राशू र्भवासर्चा ॥ १ ॥
स्वस्वप्रवरान् विसर्जयित्वा जले प्रवाहयेत् । ॐ सुमुद्गच्छ-
स्वाहान्तरिक्षं च्छस्वाहा देवैः सवितारं च्छस्वाहा मित्रा-
वरुणौ च्छस्वाहा होरात्रे च्छस्वाहा छंदांसि च्छस्वाहा
द्यावापृथिवी च्छस्वाहा यज्ञं च्छस्वाहा सोमं च्छस्वाहा
दिव्यन्नभोगं च्छस्वाहा शिवैश्चानुरं च्छस्वाहा मनो मेहादि-
यं च्छदिवन्ते धूमो गच्छतु स्वर्ग्योतिः पृथिवीं भस्मना
पूणस्वाहा ॥ २ ॥ ततो द्विराचनं कुर्युः ॥ “अनेनाध्या-
योत्सर्गकर्माङ्गत्वेन कृतेन ऋषिपूजनतर्पणादिकर्मणा
श्रीभगवान् परमेश्वरः प्रीयतां न मम” ॐ तत्सद्ब्रह्मार्पण-
मस्तु ॥ इति श्रावणीप्रयोगः समाप्तः ॥
अथ रक्षावन्धनमस्यामेव पूर्णिमायां भद्रोरहितायां कार्यम् ॥
१ अयं रक्षावन्धो भद्रायां न कार्यः । भद्रायां न कर्तव्ये श्रावणी फाल्गुनी
व्या । श्रावणी नृपतिं हन्ति ग्रामं दहति फाल्गुनी ॥ इति वचनादिति विश्वनाथः ॥
अथ श्रावणीशब्देन रक्षावन्धनमेव गृह्यते नोपाकर्मणि विदुषां परामर्श इति ध्येयम् ॥

(२४८) नित्यकर्मप्रयोगमालायां- २४८।
अथैनमतस्त्वनत्येव ॥ ३९ ॥ पर्णकषाय निष्पक्वाऽएत
आपो भवन्ति ॥ ४० ॥ भुयासि हवापि भवन्ति
॥ ४१ ॥ रुक्मम्प्रतिमुञ्च विभक्तिं ॥ ४२ ॥ व्वनीवाद्देव
विभ्रदित्याहुः ॥ ४३ ॥ गार्हपत्यं चेप्यन्पलाशशाख
व्युद्धति ॥ ४४ ॥ अथातो नैर्ऋतिर्हरति ॥ ४५ ॥ चितं
दपत्यो भवति ॥ ४६ ॥ आत्मन्यग्निं हृत्ते चेप्यन् ॥ ४
कूर्ममुपदधाति ॥ ४८ ॥ प्राणभृत उपदधाति ॥ ४९ ॥ हि
यां चितिमुपदधाति ॥ ५० ॥ तृतीयां चितिमुपदधाति ॥ ५
चतुर्थीं चितिमुपदधाति ॥ ५२ ॥ पंचमीं चितिमुपदधाति
॥ ५३ ॥ नाक्सुद उपदधाति ॥ ५४ ॥ ऋतुव्या उपदधा
॥ ५५ ॥ अथातः शतुरुद्रियञ्जुहोति ॥ ५६ ॥ उपवस्
वेहन् प्रातरुदितऽआदित्ये ॥ ५७ ॥ अथातो वैश्वानरु
होति ॥ ५८ ॥ अथातो राष्ट्रभृतो जुहोति ॥ ५९ ॥ अथ
पयोव्रततायै ॥ ६० ॥ अग्निरेष पुरस्ताच्चीयते ॥ ६१ ॥ ३
पतितं स्वर्गलोकं मुजिगात्सत् ॥ ६२ ॥ प्राणो गायत्री ॥
प्रजापतिं विस्रस्तम् ॥ ६३ ॥ तस्य वाऽएतस्याग्नेः ॥ ६

The detail evidences for the proposed Vedic Characters and Symbols
have been submitted to UTC

....Typical Evidences for Vedic Characters and Symbols



The detail evidences for the proposed Vedic Characters and Symbols have been submitted to UTC

Proposal to Encode Lepcha in Unicode

Lepcha is second language of the Government of Sikkim, India. It has its own Grammar and it is also taught in school of Sikkim. Lepcha may be represented in the BMP at code points “1A002-1A7F”.

Proposed Code Block for Lepcha in Unicode

Draft – Unicode for LEPCHA								
	1A0	1A1	1A2	1A3	1A4	1A5	1A6	1A7
0								
	1A00	1A10	1A20	1A30	1A40	1A50	1A60	1A70
1								
	1A01	1A11	1A21	1A31	1A41	1A51	1A61	1A71
2								
	1A02	1A12	1A22	1A32	1A42	1A52	1A62	1A72
3								
	1A03	1A13	1A23	1A33	1A43	1A53	1A63	1A73
4								
	1A04	1A14	1A24	1A34	1A44	1A54	1A64	1A74
5								
	1A05	1A15	1A25	1A35	1A45	1A55	1A65	1A75
6								
	1A06	1A16	1A26	1A36	1A46	1A56	1A66	1A76
7								
	1A07	1A17	1A27	1A37	1A47	1A57	1A67	1A77
8								
	1A08	1A18	1A28	1A38	1A48	1A58	1A68	1A78
9								
	1A09	1A19	1A29	1A39	1A49	1A59	1A69	1A79
A								
	1A0A	1A1A	1A2A	1A3A	1A4A	1A5A	1A6A	1A7A
B								
	1A0B	1A1B	1A2B	1A3B	1A4B	1A5B	1A6B	1A7B
C								
	1A0C	1A1C	1A2C	1A3C	1A4C	1A5C	1A6C	1A7C
D								
	1A0D	1A1D	1A2D	1A3D	1A4D	1A5D	1A6D	1A7D
E								
	1A0E	1A1E	1A2E	1A3E	1A4E	1A5E	1A6E	1A7E
F								
	1A0F	1A1F	1A2F	1A3F	1A4F	1A5F	1A6F	1A7F

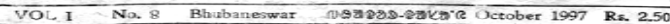
Page 1 of 5

Page 1 of 5

Unicode Representation of Santhali

Santhali has been recently added in the list of constitutionally recognized scheduled languages of India. It is spoken by approximately 5 Million people in India. Government of India proposes inclusion of the same in the Unicode standard. The proposal n2505 (Proposal submitted by Mr. Michael Everson) for encoding of Santhali is not acceptable as such to the experts in India. A committee of experts will deliberate in details and the consensus inputs will be communicated to UTC accordingly.

𐤀 LO, <i>la, la, A</i>	𐤁 OT, <i>at, at, AT</i>	𐤂 OKH, <i>ak', ag, AG</i>	𐤃 ONG, <i>an, an, ANG</i>	𐤄 OL, <i>al, al, AL</i>
𐤅 LA, <i>lā, lā, AA</i>	𐤆 AK, <i>āk, āk, AAK</i>	𐤇 ACH, <i>āc', aj, AAJ</i>	𐤈 AM, <i>ām, ām, AAM</i>	𐤉 AW, <i>āw, āw, AAW</i>
𐤊 LI, <i>li, li, I</i>	𐤋 IS, <i>is, is, IS</i>	𐤌 IH, <i>ih, ih, IH</i>	𐤍 INY, <i>iñ, iñ, INY</i>	𐤎 IR, <i>ir, ir, IR</i>
𐤏 LU, <i>lu, lu, U</i>	𐤐 UC, <i>uc, uc, UCH</i>	𐤑 UTH, <i>ut', ud, UD</i>	𐤒 UNN, <i>un, un, UNN</i>	𐤓 UY, <i>uy, uy, UY</i>
𐤔 LE, <i>le, le, E</i>	𐤕 EP, <i>ep, ep, EB</i>	𐤖 EDD, <i>ed, ed, EDD</i>	𐤗 EN, <i>en, en, EN</i>	𐤘 ERR, <i>er, er, ERR</i>
𐤙 LOO, <i>lō, lo, O</i>	𐤚 OOTT, <i>ot, ot, OTT</i>	𐤛 OOPH, <i>op', ob, OB</i>	𐤜 OOWN, <i>ow, ow, OV</i>	𐤝 OOH, <i>oh, oh, OH</i>



10039 10040 10041 10042 10043 10044 10045 10046 10047 10048 10049 10050 10051 10052 10053 10054 10055 10056 10057 10058 10059 10060 10061 10062 10063 10064 10065 10066 10067 10068 10069 10070 10071 10072 10073 10074 10075 10076 10077 10078 10079 10080 10081 10082 10083 10084 10085 10086 10087 10088 10089 10090 10091 10092 10093 10094 10095 10096 10097 10098 10099 10100 10101 10102 10103 10104 10105 10106 10107 10108 10109 10110 10111 10112 10113 10114 10115 10116 10117 10118 10119 10120 10121 10122 10123 10124 10125 10126 10127 10128 10129 10130 10131 10132 10133 10134 10135 10136 10137 10138 10139 10140 10141 10142 10143 10144 10145 10146 10147 10148 10149 10150 10151 10152 10153 10154 10155 10156 10157 10158 10159 10160 10161 10162 10163 10164 10165 10166 10167 10168 10169 10170 10171 10172 10173 10174 10175 10176 10177 10178 10179 10180 10181 10182 10183 10184 10185 10186 10187 10188 10189 10190 10191 10192 10193 10194 10195 10196 10197 10198 10199 10200 10201 10202 10203 10204 10205 10206 10207 10208 10209 10210 10211 10212 10213 10214 10215 10216 10217 10218 10219 10220 10221 10222 10223 10224 10225 10226 10227 10228 10229 10230 10231 10232 10233 10234 10235 10236 10237 10238 10239 10240 10241 10242 10243 10244 10245 10246 10247 10248 10249 10250 10251 10252 10253 10254 10255 10256 10257 10258 10259 10260 10261 10262 10263 10264 10265 10266 10267 10268 10269 10270 10271 10272 10273 10274 10275 10276 10277 10278 10279 10280 10281 10282 10283 10284 10285 10286 10287 10288 10289 10290 10291 10292 10293 10294 10295 10296 10297 10298 10299 10300 10301 10302 10303 10304 10305 10306 10307 10308 10309 10310 10311 10312 10313 10314 10315 10316 10317 10318 10319 10320 10321 10322 10323 10324 10325 10326 10327 10328 10329 10330 10331 10332 10333 10334 10335 10336 10337 10338 10339 10340 10341 10342 10343 10344 10345 10346 10347 10348 10349 10350 10351 10352 10353 10354 10355 10356 10357 10358 10359 10360 10361 10362 10363 10364 10365 10366 10367 10368 10369 10370 10371 10372 10373 10374 10375 10376 10377 10378 10379 10380 10381 10382 10383 10384 10385 10386 10387 10388 10389 10390 10391 10392 10393 10394 10395 10396 10397 10398 10399 10400 10401 10402 10403 10404 10405 10406 10407 10408 10409 10410 10411 10412 10413 10414 10415 10416 10417 10418 10419 10420 10421 10422 10423 10424 10425 10426 10427 10428 10429 10430 10431 10432 10433 10434 10435 10436 10437 10438 10439 10440 10441 10442 10443 10444 10445 10446 10447 10448 10449 10450 10451 10452 10453 10454 10455 10456 10457 10458 10459 10460 10461 10462 10463 10464 10465 10466 10467 10468 10469 10470 10471 10472 10473 10474 10475 10476 10477 10478 10479 10480 10481 10482 10483 10484 10485 10486 10487 10488 10489 10490 10491 10492 10493 10494 10495 10496 10497 10498 10499 10500 10501 10502 10503 10504 10505 10506 10507 10508 10509 10510 10511 10512 10513 10514 10515 10516 10517 10518 10519 10520 10521 10522 10523 10524 10525 10526 10527 10528 10529 10530 10531 10532 10533 10534 10535 10536 10537 10538 10539 10540 10541 10542 10543 10544 10545 10546 10547 10548 10549 10550 10551 10552 10553 10554 10555 10556 10557 10558 10559 10560 10561 10562 10563 10564 10565 10566 10567 10568 10569 10570 10571 10572 10573 10574 10575 10576 10577 10578 10579 10580 10581 10582 10583 10584 10585 10586 10587 10588 10589 10590 10591 10592 10593 10594 10595 10596 10597 10598 10599 10600 10601 10602 10603 10604 10605 10606 10607 10608 10609 10610 10611 10612 10613 10614 10615 10616 10617 10618 10619 10620 10621 10622 10623 10624 10625 10626 10627 10628 10629 10630 10631 10632 10633 10634 10635 10636 10637 10638 10639 10640 10641 10642 10643 10644 10645 10646 10647 10648 10649 10650 10651 10652 10653 10654 10655 10656 10657 10658 10659 10660 10661 10662 10663 10664 10665 10666 10667 10668 10669 10670 10671 10672 10673 10674 10675 10676 10677 10678 10679 10680 10681 10682 10683 10684 10685 10686 10687 10688 10689 10690 10691 10692 10693 10694 10695 10696 10697 10698 10699 10700 10701 10702 10703 10704 10705 10706 10707 10708 10709 10710 10711 10712 10713 10714 10715 10716 10717 10718 10719 10720

[illegible][illegible][illegible]

Enhancement of Devanagari for Dogri

Dogri has also been recently added in the list of constitutionally recognized scheduled languages of India. Dogri uses Devanagari script with two more Diacritic Marks. There are two tone markers used to write Dogri text using Devanagari script. These Tone markers need to be added in the Devanagari Code Block of the Unicode.

Dogri Tone Marker: Example

खून दा मेआर

□ शशि पठानिया

रमा नै कल्ले खंवां शा इसलै तगर रेडियो परा ते केई बारी एखबार सुनी लेई ही जे बिद्वान वस्ती च दो बमारिये किट्ठा हगला करी दित्ता ऐ। जिन्हे ओ ईमानदारी दी बमारी नै बड़े बड़ पमाने पर लोकें गो प्रभावत कीता ऐ, हूए नम्बरें पर राज बरोधी बमारी ऐ। ईमानदारी दी महामारी दा नगाना बने न कीहू, साल दी उमरी शा उपर दे लोक ते दूई बमारी दे जरासीम तेजी कर्ने जोहें साल्ले शा पट्ट आली पीछी गो मुतासर करे करे न। समाज दे जोर देने पर मैहू कमा सेहू नै सावधानी दे तोरे पर इनें लोकें गो हस्पताल च भरथी करी लेआ ऐ। समाज दे ठेकेदार जनता अग्ये अपील कीती ऐ जे ओ परहेजी पुजाऽ करिये महामारी दी रोक-धाम च उनें गो सँहयोग देन। समाज-सुधार संस्था आसेआ परहेज सरबन्धी साहित्य दे पोस्टर थां-थां कग्ये पर चमकाई दिते गे न, जिन्हे पर लिखे दा ऐ मरीजें दे बोलने पर इनें बमारिये दे जरासीम कर्ने रस्ते हूए मनुबल्ले दे शरीर च दाखल होन्दे न। इस आस्ते जनता अग्ये नवेदन ऐ जे ओ कर्ने च हं देई लैन जां सिक्का भरी लैन। सिक्के आलां मुस्ला जैदा असरदार ऐ।

रमा गो हुन इन्तजार हा जे टी० बी० परा हस्पताल दा दिशस दिखेआ जा। टी० बी० ऑन कीता, इक सरकरदा नेता जी प्रगट होए, ओ दोऐ हत्य जोड़िये जनता अग्ये अपील करे करे हे “वस्ती दे बसनीकें अग्ये ए मुखारण ऐ जे वस्ती खाली नैई करन, अपने-अपने घरे च गै रोहन। मुतासर लोकें गो हस्पताल च भरथी करी लैता गेआ ऐ। बाहुरे थमां माहिर डाक्टर दी टीम बुलाई लैती गेई ऐ मेद ऐ तीले गै डाक्टर दी टीम बी ब्याई-ज्हाजे राहें पुजदी गै होण। असें इक जुट्ट होइये इस कुदरती खबाही शा समाज की बचाना ऐ। बाकी लोकें अग्ये बी प्रार्थना ऐ जे जेकर ए लोक पर छोड़िये निकलदे न तां कोई इनें गो शरण नैई देऐ। जेहूदा शरण देण, अपने आस्ते बमारी ते स्नेहण गै दुआ कनूनी तोरे पर की कसूरबार होय।

टी० बी० आलें हस्पताल च मरीजें पर ते सरसरी कैमरा गै मारेआ इस लेई उन्दे चा-कोई नुहार बी नैई पनछोई पर इस गल्ले दा बड़ा ताज्जुब होआ जे उनें मरीजें कोल कोई बी

अटेष्ट नैई हा। बाबें साफ-सुपरे ते शांत। सारे मरीज इक-डाल मेदे हे, लेटे दे हे जां सुले हे, पर उन्दे च हरकत कोई नैई ही, शायद कोई दुआ देइये गे-होण कीते दे हे।

शायद अखबार च खो-लिखे मान कीते दा होए। इस आस्ते रमा-गो बे-सबरी कर्ने अखबार दी निहासप हो। ओ जानना चाहू दी ही जे उस वस्ती चा कु-न-कु-न हले महामारी दी सपेट च आया ऐ की जे ओदे कोलोग्ज उस्ते वस्ती च रोह दे न। ओ-की-मान उस्ते गै रोह दी होंदी जेकर उस्ती छोड़िये ओदे शा जूनियर गो पलट नैई देई दित्ता गेआ हुन्दा। अज्ज पैहूली बारी उस्ती पकीन आया हा जे भगवान जे किण करदा ऐ, ओदे च गै मनुबल्ले दी बेहूतरी होन्दी ऐ। अज्ज खबरें अखबार आले गो केहू होई गेआ हा। अट्ट बजे तगर अखबार नैई देई गेआ। रमा शा नैई बलगोआ तां उन्न नोकर बजार दुड़ाया, पर बेकार, बजार लोकल अखबार मुक्की गेदा हा। उन्न अपने असुलें गो बी ताकड़े रक्खिये नोकरे गो मुआदिये दे भेजेआ पर ओ उपुआं की हस्प फंडा गै परतोआ। अहू हूए मुआंकी लेई गेदे न। इनें चिरे च अखबारा आला बी आई गेआ। इस शा पैहूले जे ओ अखबार सुट्टिये परतोआ, रमा नै उस्ती टोकेआ “केहू गल्ल ऐ, ए टैम ऐ अखबार देने दा, लोकें दे भ्यामा छे अवे दी आई गेदी ऐ।” इसलै ओ ए भुल्ली गेदी ही जे ओ बी रोज सवेस्ता गै देई जन्दा ऐ। ओ हफे दा बोलेआ “मैन जी, मै बी सवेस्ता गै निकलेआ हा पर रस्ते च हस्पताल अग्ये गइहये दी भीई मूजब रस्ता गै ककी गेदा ऐ। मै बड़ी मुत्कले इस भीड़े च साईकिल छोड़िये आंदा ऐ।”

रमा अखबार लेइये बैठी ब अक्लें उनें मरीजें दे दर्शन ओदे नसीबें च नैई हे। ओदे बाहूदे गै दरवाजे पर बैठी करताई। रमा नै नोकरे गो दरवाजे पर दिखने लेई आखेआ। नोकरें ने दरवाजा खो-लेआ बहरें दा उनें पैरे गै परतोई आया। “मेमसाहूव, मेमसाहूव कोई आदमी ते जानाती न, दोनें कर्ने पर पट्टियां बन्नी थियां न ते अपने आप गै मिसेज एंड मिस्टर इंजनीयर दसदे न।” अखबार रमा दे हस्ता खुकी गेई। “तू दरवाजा ते बन्द करी आयां जां नैई” जा, आख, साहूव पर नैई हुन। “नोकर पही परतोई आया,” “मेमसाहूव ओ भेरी गल्ल सुनदे गै नैई अपनी गै सनाई जा करेदे न, “बिपदा दे मारे दे आं, “अपने घरे-आले गो फोन कीता। उन्न आखेआ “कनूनी तोरा पर ते सुस कोई मुक्ता कइहिये अपनी बकालत आपुं करी लेंगे ब जेकर बमारी चमकी गेई तां?”

“हां जी।”

रमा सोचें च पैई गेई जे केहू कीता जा। जेकर इनें गो अन्दर ओन दिन्नी आं तां इक्क ते कनूनी तोर पर कसूरबार होइ दुआ इन्केशन बी होई सकदी ऐ। पर दरवाजे आए दे दी मदद करना बी ते मनुबल्लता दा सकाजा ऐ। जिसलै ओ आपुं फोन फंसला नैई करी सकी तां अपने घरे-आले गो फोन कीता। उन्न आखेआ “कनूनी तोरा पर ते सुस कोई मुक्ता कइहिये अपनी बकालत आपुं करी लेंगे ब जेकर बमारी चमकी गेई तां?”

भलिये लोकें, दाज बरोधी बमारी ते असें गो चमकने शा रेही की जे अस बहोए-बरे दे आं ते रेही ईमानदारी दी गल्ल तां अपनी रजिस्ट्रार उपर मिथी पूरा भरोसा ऐ।”

रमा फोन रक्खिये गेई ते दरवाजा खो-लेआ, ओ अग्ये अपना समान रक्खिये बैठे दे हे। रमा ने बारे कर्ने उनें गो अन्दर ओन लेई आखेआ। उनें दोनें मियां-बीची जे लक्ख-लक्ख सीसां दिन्दे होई धनबाद कीता। उनें गो परोहते आला कमरा दसिये, रमा ने नोकरे गो आखेआ जे उन्दे लेई न्हारी दा सरोता करे। रमा पही उन्दे कोल जाई बैठी ते “हुन कर्ने

Encoding of Tamil Syllables

Government of India recommends the New Encoding scheme for the Tamil Syllables as proposed by the State Government of Tamil Nadu. Encoding of these may be expedited.

UNICODE: Bhutia and Limbu

- Bhutia is spoken in Sikkim, India. It uses Tibetan Script for writing. A character “Modified Tsalak” need to be added in the Tibetan code block at code point U+ 0FBD for complete representation of Bhutia.
- Limbu is also spoken in Sikkim, India. It's already represented in the Unicode standard. The Glyphs of the Vowel signs (U+1920-U+1928) need to be modified for adequate representation of Limbu.

Need to represent Minority & Historic Scripts of India

Historic Scripts:

Grantha, Sharada,

Minority Scripts:

Ol Chiki (for Santhali), Tirhuta (for Maithili) & Idumishmi (in Arunachal Pradesh).

There is need to find adequate representation of these historic & minority scripts in Unicode.

Work has been initiated and necessary inputs will be furnished to UTC for consideration.

Updates/ Suggestions accepted by Unicode

UTC has accepted the following updates/
suggestions and included in Unicode 4.0

Sincere Thanks for that !

DEVANAGARI:

- Devanagari ‘Short A’ has been added in Unicode 4.0.
- Detailed description has been provided about characteristics of Indic Scripts.
- A detailed description has been provided in Unicode 4.0 regarding the Marathi allograph SHA, LA and SOFT RA.
- Annotation has been provided for Devanagari Sign Halant, Devanagari Punctuation Marks - Deergh Virama and Poorna Virama.

BENGALI:

- 09BD (৳) BENGALI SIGN AVAGRAH
- Annotations for the four Bengali Code points i.e. Bengali Sign Hasant (U+09CD) and Bengali Currency Numerators One, Two and Three have been added in Unicode 4.0.
- Encoding of “BENGALI LETTER KHANDA TA” (09FC ঙ).

GUJARATI:

0A8C	૨	GUJARATI LETTER VOCALIC L
0AE1	૨૯	GUJARATI LETTER VOCALIC LL
0AE2	૨૯	GUJARATI VOWEL SIGN VOCALIC L
0AE3	૨૯	GUJARATI VOWEL SIGN VOCALIC LL

KANNADA:

0CBC	ೠ	KANNADA SIGN NUKTA
0CBD	ೡ	KANNADA SIGN AVAGRAH

GURMUKHI:

0A50	ੴ	GURMUKHI EK ONKAR (God is one)
0A74	ਖ਼	GURMUKHI SIGN KHANDA
0A01	ੰ	GURMUKHI TIPPI (Nasalization)
0A03	ੴ	GURMUKHI VISARG
0A3B	ੳ	GURMUKHI ADDAK (Doubles following consonant)

ORIYA:

ORIYA LETTER WA (0B35 - ଳ) and Change in Explanation and annotations are added for Oriya code points U+0B2C, U+0B5C and U+0B5D.

TAMIL:

- Tamil Special signs like day sign, month sign, year sign, debit sign, credit sign, Tamil Rupee sign have been encoded in Unicode 4.0.
- Annotation for TAMIL LETTER AYTHAM, TAMIL SIGN ANUSVARA and TAMIL SIGN VISARGA have been added in Unicode 4.0.

MALAYALAM:

- A detailed explanation has been provided in the Unicode 4.0 for formation of Malayalam *Chillu* characters (MALAYALAM LETTER NN, MALAYALAM LETTER N, MALAYALAM LETTER RR, MALAYALAM LETTER L and MALAYALAM LETTER LL).
- Annotation has been provide for the code point 0D4D (Malayalam Sign Halant) as Malayalam Sign CHANDRAKKALA.

- ARABIC LETTER ULTA PESH, ARABIC LETTER JAZM, ARABIC LETTER KHARI ZER, ARABIC LETTER HAMZA (Diacritical) have been encoded in the Unicode 4.0.
- Also the separate code points provided for Sindhi-Arabic numerals.
- Annotation have been added for the code points used in Sindhi and Urdu being used in India.

Awaiting Update in Unicode

Delay in accepting Country-level Inputs

Obstructs use of local languages in IT resulting into
Sprawling Digital Divide

Public Review Issue: Danda Dis-unification

Government of India strongly recommends the disunification of the punctuation marks Deergh Virama (double danda) and Poorna Viram (danda) for all the Indic scripts as the glyph for these signs vary in all the Indic scripts.

Awaiting Updates-

Devanagari

Additions required for Sindhi language written in Devanagari

0979	ग़	DEVENAGARI LETTER Implosive GA
097A	झ	DEVENAGARI LETTER Implosive JA
097B	ड़	DEVENAGARI LETTER Implosive DA
097C	ब़	DEVENAGARI LETTER Implosive BA

These consonants are exclusively used in Sindhi and have different Phonetic Value then GA,JA,DA,BA.

Awaiting Updates- Devanagari

Additions required for *Laukik Sanskrit* written in Devanagari

0955 ँ DEVENAGARI ANUSWARA

0956 ः DEVENAGARI SIGN YAJURVEDIC ANUSWARA

0957 ऌ DEVENAGARI JIVHYAMULIYA

These characters are commonly used in *Laukik Sanskrit shlokas* which are often used in Indian language text.

Additions for Currency Sign

0971 ₹ DEVENAGARI CURRENCY SIGN

Awaiting Updates- Devanagari

Change in Glyph

	Present	Proposed	
0929	ॢ	ॣ	DEVANAGARI LETTER NNNA For transcribing Dravidian alveolar n Fricative Consonant used in Marathi
0931	।	॥	DEVANAGARI LETTER RRA For transcribing Dravidian alveolar r Half form is represented as “Eyelash RA” Fricative Consonant used in Marathi
0934	०	ॠ	DEVANAGARI LETTER LLLA For transcribing Dravidian alveolar l Fricative Consonant used in Marathi

The Glyph of these characters need to be changed because the sound associated with these consonants are different from the consonants used with Nukta.

Awaiting Updates- Bengali & Oriya

Bengali – Additions

09FB ঔ BENGALI LETTER YA FALLA

09FD ঋ BENGALI LETTER KSHA (Used in Assamese)

Oriya - Addition

0B44 ୠ ORIYA VOWEL SIGN VOCALIC RR need to be Added
in the Unicode Standard for Oriya.

Awaiting Updates- Gujarati

Additions

0 AF6 ઠ	GUJARATI FRACTIONAL NUMERAL /PA/ 1/4
0 AF7 ઠઠ	GUJARATI FRACTIONAL NUMERAL /Addho/ 1/2
0 AF8 ઠઠઠ	GUJARATI FRACTIONAL NUMERAL /Pono/ 3/4

The fractional numerals and diacritic marks are very commonly used in Gujarati. Hence these characters should be added in the Gujarati code block.

Awaiting Updates-

Gurmukhi

Additions

0A4E	ੴ	GURMUKHI PARI – RA (Consonant modifier)
0A4F	ੵ	GURMUKHI PARI – HA (Consonant modifier)
0A76	ੴ	GURMUKHI PARI VA [Halant ੴ (0A4D) ਵ (0A35)]
0A77	ੵ	GURMUKHI PARI YA [Halant ੴ (0A4D) ਯ (0A2F)]
0A78	ਯ	GURMUKHI Half Y-YA [ਯ (0A2F) Halant ੴ (0A4D)]
0A79	ੴ	GURMUKHI PARI GU [Halant ੴ (0A4D) ਗ (0A17)]
0A7A	ੴ	GURMUKHI PARI CA [Halant ੴ (0A4D) ਚ (0A1A)]
0A7B	ੴ	GURMUKHI PARI TA [Halant ੴ (0A4D) ਤ (0A24)]
0A7C	ੴ	GURMUKHI PARI NA [Halant ੴ (0A4D) ਨ (0A2B)]

Awaiting Updates- Telugu

Additions

0C3C ◌◌ TELUGU SIGN NUKTA
Placed at bottom left corner of the letter

0C3D ఱ TELUGU SIGN AVAGRAH

There is a strong need of adding (Telugu) diacritic marks in the Unicode Standard for Telugu, as there are no independent characters in the script for representation of sounds borrowed from other languages.

TAMIL

The TAMIL user worldwide has been demanding to have the Syllable based Encoding. Tamil Scripts has minimum number of Vowels and Consonants so there are only 313 Syllables in Tamil. The proposed syllable based encoding scheme proved to be efficient for text/ data processing applications.

ARABIC - Urdu, Sindhi and Kashmiri

ARABIC LETTER BAT (ب) need to be added in the Arabic code set for Urdu.

Workshop for Unicode awareness/ Proliferation

Government of India proposes to Organize

“Unicode Workshop for Developers”

in April 2005, and solicits the Unicode Consortium support by participation of the Unicode Experts for the workshop .

Helpline

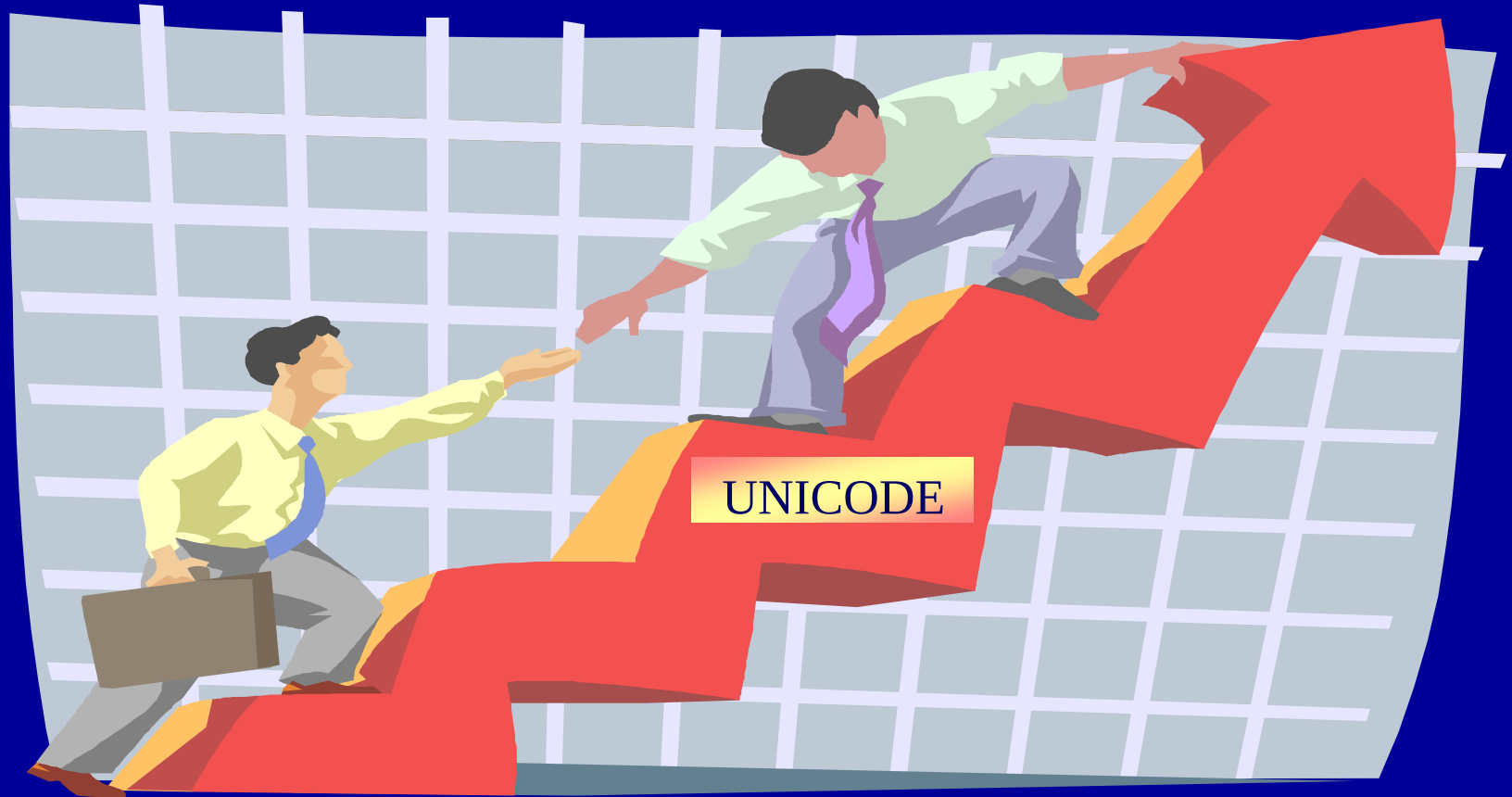
indic@unicode.org is the mail group to discuss issues related to Unicode representation of Indic scripts.

<http://www.tdil.mit.gov.in> provides updates on Indian language technologies, tools and resources .

info@tdil.mit.gov.in to answer various queries of the user.

New Order of Knowledge based Society :

- Universalisation of Content
- Rise, Raise & Race



Unicode plays an important role.

Government, Academia, Industry
together to play **globally** and to serve
locally in multilingual computing.

धन्यवाद
ਧਨ੍ਯਵਾਦ
ધન્યવાદ

Thank You