3. Workshop Reports

3.1 National Workshop on Unicode : September 24-26, 2003

Background :

TDIL programme has been focusing on the Technology Development for Indian Languages, making it feasible to take IT to the people in their local languages.

Department of IT (then Dept of Electronics) evolved ISCII Standards for use of Indian languages in Computers. These recommendations were accepted by BIS as the National standards for information Interchange ISCII:1991 (Indian Script Code for Information Interchange-ISCII- IS13194:1991).

Unicode encoding scheme has proven to be very useful to create and store Multilingual Data. Unicode Standard includes standard code sets for Indian scripts based on ISCII-1988 document whereas present National standard is ISCII:1991. Department of Information Technology, Ministry of Communications & Information Technology is the voting member of the Unicode Consortium. Some modifications were necessary in the existing Unicode Standard for adequate representation of Indian Scripts. Department of Information Technology organized a number of meetings with State Government IT Departments, Linguists and members of Indian IT industry to finalize the changes in the existing Unicode Standards for adequate representation of Indian Scripts. After detailed discussion with Linguists, respective State Government and Indian IT Industry the proposed changes were finalized. These recommendations were submitted to Unicode Technical Committee for inclusion in the Unicode Standards. Some of the proposed changes are included in the Unicode Standards 4.0.

Dr. Om Vikas, as the representative of Govt. of India participated in Unicode Technical Committee meeting held in California in March 2003. Various issues related to Indic scripts were discussed in the meeting. It was proposed in the meeting to hold a Unicode Workshop in India to discuss and sort out various issues related to Indic scripts computing.

A "National Workshop on Unicode" was organised by DIT in New Delhi from September 24-26, 2003 with the following objectives : To create awareness and clarity about Unicode representations system and enable a wider cross-section of linguists, language experts and technology developers of Indian Languages to interact with the International Unicode experts.

To address all Unicode-related apprehensions (of linguists and developers) via sessions in the workshop. To focus on solution to various problems being faced by the industry in implementing the Unicode Standard for application development in Indic languages.

The workshop was inaugurated by Prof. M.G.K. Menon, Chairman, Board of Governors, IIT-Delhi by lighting the lamp. Dr. Om Vikas, Senior Director and Head, TDIL Mission welcomed the delegates. Prof. Menon delivered the inaugural address. Dr. Mark Davis, President Unicode Consortium presented the keynote address. Shri K.K. Jaswal, Secretary, Department of Information Technology addressed the gathering during the Welcome Dinner.

In his address Secretary DIT highlighted the need of Localisation and the importance of Universal Encoding Standards to support multilingual content creation. He added,... Today 80% of the web contents are in English. The real benefit of Information Technology can reach to the masses only when the content is available in their own language. As Unicode encoding has proved to be very useful to create the contents in Local languages, its use will spur the content creation and web enabling activities for Indian Languages and the Localization efforts will get a boost.

Dr. Monon said that he had closely watched the development of the remarkable field i.e. IT over the past 50 years and have seen many of these developments at close quarters; and known personally many of the pioneers. Every individual in this world is, in some way a participant in the innovations, the interactions, new ways of doing things, the diverse areas of application and their impact.

He added that ICT Revolution has been fuelled by parallel revolutionary advances in many fields, that have produced an interaction of a symbiotic and synergistic nature: foundational thinking (such as of Babbage, Turing and Von Neumann) software development (languages, Internet, www-through the work of Tim Berners-Lee); microelectronics (from the transistor of Bardeen, Brattain and Shockley to the integrated circuits of Kilby and Noyce, to integration as of Intel); space technology; radio and wide spectrum communication capabilities. These have all been based on the great advances in mathematics, physics and chemistry; and now in biology, particularly through improved understanding of the functioning of the human brain.

One can see that the de facto standard language on Internet is English. The result of this is that the knowledge gap will grow and widen, with those who know English rapidly increasing their knowledge capabilities whilst those who do not know English will be unable to access this huge and growing knowledge base, which is predominantly in English. There are, therefore, two issues in the existing nature of the digital divide that will have to be overcome: (i) how can one make computation and the use of computers for knowledge acquisition and utilization low in cost; also how can one make access e.g. communication systems inexpensive; (ii) how can one ensure that language will not be a barrier over the future to accessing global knowledge. He said that language is intrinsically connected with perception, cognition and understanding of concepts. One must preserve and grow the mother tongues of what are truly huge populations – since these relate to rich diverse cultures and content.

Prof. Menon concluded that the ICT will create wholly new opportunities in the world, through which the quality of life for all will be improved, preserving and continuing to grow the rich heritage of cultures and creative activities of diverse civilizations & peoples, and thus truly establish unity in diversity, or what is the basic Indian concept of 'Vasudeva Kutumbakam', meaning that all the world is one family.

Session-2: Unicode – Architecture, Design and Encoding Fonts

The first technical session in the workshop was presided over by Mr. Lalit Yagnik, IBM software group and addressed by Dr. Ken Whistler. Dr. Whistler presented technical issues of unicode. Various projects mentioned by Dr. Whistler are reproduced here.

i. Scope of Unicode:

Where it fits:

a. Simplest definition of Unicode: It is the Universal Character set.

b. The character set of Unicode is greater than the set of characters in different languages and scripts of the world. This is because some 'characters' in Unicode are used to specify the behaviour of script-elements of different languages.

c. Unicode is universal in more than one sense. Not only does it have a single code for each character in a script, it is intended to also address all the languages of the world.

d. Unicode is usable by all, especially on the web, because it is widely supported by software vendors for its relative-simplicity of design.

Implications for development of the Unicode Standard:

a. To function as a Universal Character set, any missing characters must be added.

b. To facilitate interoperability with the existing standard some specific characters may need to be added for additional clarity of use.

c. Spoken languages have scripts whose script is based on some functional model. When adding these new scripts to Unicode, the difference of the functional models of these scripts with that of Unicode may require some 'strange' characters to be added. But the resulting encoding will have Universal functionality.

Items relevant to Indic-specific languages for Unicode:

a. All official languages of India are covered by the Unicode specification.

b. Universality of Unicode implies that even the minority languages of India are addressed by it.

c. It is important to specify the language-specific behaviour of these characters to complement its use in the Unicode Standard.

d. The use of Unicode for implementing Indic language support in software will expose Indian languages to international users.

e. Any historical characters which are not in use presently but did have functions earlier also need to be encoded.

ii. Architecture of Unicode

What is a character? :

a. To encode elements of scripts, abstract characters are mapped to a number.

b. This causes doubt because people don't know what the abstract character to be encoded is.

c. Example: Are Latin characters which may each exist as an upper case or lower case separate character? This was resolved before Unicode was created by assigning different numbers for each case for every character.

d. Different solutions for character elements of different scripts imply different handling decisions, to resolve the functional model for each script to that of Unicode.

e. Indian example:

Evaluation of implementation of Indic scripts in ISCII shows that independent vowels and matras (or dependent vowels/ vowel modifiers) were used in implementation as against pure-consonant implementation.

The Unicode Standard is based completely on ISCII-1988 implementation.

The encoding was made consistent with Unicode design principles without any change. Consistency will be maintained in the future.

f. Punctuation marks are handled only once, hence commas and other marks are not assigned separately in the code-space for each Indic script.

iii. Character-glyph model:

a. A glyph is the graphical representation of an abstract character.

b. Example :

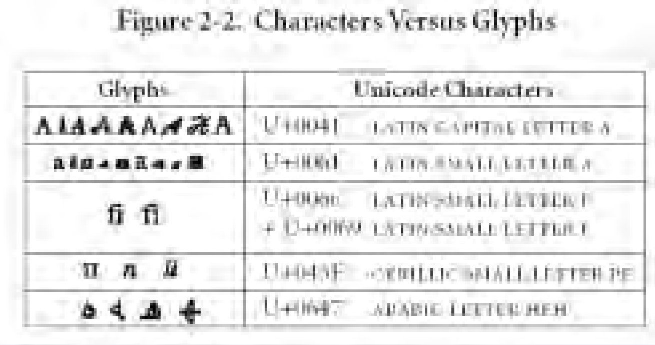


Figure 2-2. Characters Versus Glyphs

Glyphs	Unicode Characters
A A A A A A A A A	U+0041 LATIN CAPITAL LETTER A
a a a a a a a a a	U+0061 LATIN SMALL LETTER A
fi fi	U+0066 LATIN SMALL LETTER F + U+0069 LATIN SMALL LETTER I
п п п	U+043F CYRILLIC SMALL LETTER PE
ه ه ه ه	U+0647 ARABIC LETTER HEH

Each is different graphical representations of the same character

c. Sometimes context dependent choices may be necessary to show which glyph is shown when a character appears at a particular location.

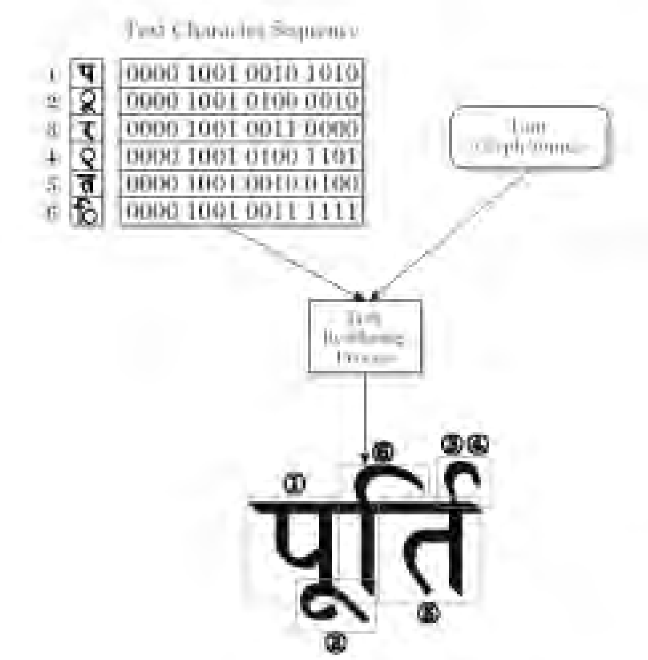
d. This could raise the question: Would you encode the character or the glyph? Each choice has its advantage and disadvantage. For example there are choices in sorting (Sorting glyphs are not inherently logical), displaying (Displaying using glyph-based encoding is much simpler), etc.

e. The preferred choice is based on the design principles of Unicode.

f. For Indic languages, ISCII, and therefore Unicode Indic encoding, encodes the character only and not the glyph. This means conjunct-characters are not individually encoded and characters are not repeated among separate scripts (e.g. Hindi, Marathi, Bengali, and Assamese).

g. Following is an example of how the rendering is handled for Devanagari.

Figure 2-3. Unicode Character Code to Rendered Glyphs



iv. Allocation choice:

Unicode is fundamentally a BIG FLAT uniformly accessible encoding space.

There are almost a million characters.

Each character has a single number or code-point.

For Indic scripts each block has 128 code points.

a. One to one for each language was attempted for simplicity but some languages were merged into closer blocks (e.g. Assamese and

Bengali even though there are implemented separately in ISCII.)

b. Ordering of code-points for characters is **not equivalent** to their linguistic sorting. This is anyway impossible since multiple languages may be ordered within the same block. The sorting may be addressd with collation algorithms.

v. Encoding forms

Unicode encoding has 3 main variant forms to represent the numbers assigned.

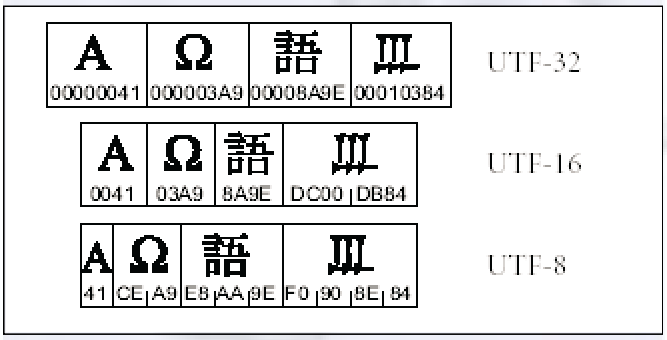
Character encoding is not the same as Encoding form. The encoding forms vary to facilitate their use in different environments.

These forms are UTF-8, UTF-16 and UTF-32

All three forms can store any character in the Unicode Standard. They can be mutually converted into other forms without any loss of data.

Any of these encoding forms are conformant for encoding characters using Unicode. There is no distinction between the validity of any of the forms. They may be used according to the context and use of the data.

Figure 2.4 Unicode Encoding Forms



UTF-8 is used for byte-oriented or ASCII-based systems. It is a variable-width encoding form to preserve transparency of 8-bit code units. It is the encoding form used for HTML and other Internet protocols. It is useful for migration and as with other forms there is no overlap of byte-values.

UTF-16 is a fixed-width encoding for the Basic Multilingual Plane (BMP) which addresses the vast majority of the languages in common-use. UTF-16 is the current form of the earliest form of Unicode which was originally designed to use a fixed-width 16-bit encoding. Java, ICU and Windows are some of its application domains.

UTF-32 is the simplest implementation of Unicode encoding where each Unicode code point is represented uniquely in directly-identifiable form by a 32-bit value. This is used in Linux UNIX implementations.

v. Encoding Schemes:

Even though computers store data in digital form, the exact ordering of this data may differ among computers of different architecture. When Unicode data is exchanged between computers in such a situation, a process called serialization is used to arrange the data into well defined sequences of bytes.

The ordering of which end of a byte to start reading from (i.e. least significant bit or most significant bit) to decode the data, are called 'little-endian' and 'big-endian.'

Serialization must include the 'endian-ness' of the system to make the data interchangeable. Byte Order Marks are used to differentiate endian-qualities of data.

vi. Observation on naming schemes:

The policies of the Unicode specification do not allow renaming of any characters in the Standard. However, annotation may be added.

Annotations are used to clarify or correct the names of characters if they are wrong.

viii. Canonical Equivalence:

When implementations of new languages are added to the Unicode Specification, the complexity of how these languages have been implemented in legacy-systems may cause some design principles of the Unicode Specification to be compromised.

As the following example shows, the two forms of A and LJ can be created in two different ways. This would not be recommended in an ideal situation. However, these two different sequences represent the same end-form and are called Canonical Equivalents.

Figure 2.5 Equivalent Sequences

B + Ä → B + A + ◌̈

LJ + A → L + J + A

Canonical equivalents are achieved by a process called Normalization.

In general duplicate encoding is a bad practice and discouraged and done only out of compulsion. Some examples are Hangul and some examples of Latin.

Canonical forms relevant for Indic languages are few. This should be maintained to avoid unnecessary problems. (Example: Some nukta-based characters for Hindi and Bengali which can also be displayed independently : U095C = U0921 + U093C)

In a question by Prof. J.N. Shah from VJTI Mumbai about the design-bases of multiple encoding forms, Dr. Whistler replied that Unicode Standard 4.0 may be referred.

Session-3: Unicode – Conformance issues and Character properties: Collation

The third session in the workshop was addressed by Ms. Cathy Wissink, Microsoft Corporation, USA and chaired by Mr. Narendra Bhandari, Asia Regional Manager, Intel Corporation.

Ms Wissink spoke about the need to manage Indian language computing through Unicode, and how people should understand clearly the various issues involved, which are:

- Sorting & collation
- Font support
- Keyboards support

Ms Wissink also delineated about the forthcoming Office 2003 Hindi, which is expected to have all menus and dialog boxes fully localized into Hindi, and also showed some screen shots of the same

Ms Wissink also discussed how Unicode based Indian language fonts would work better with IE, as the browser would recognize these fonts as specifically being those of Indian languages.

Session-4: Unicode - Implementation Guidelines

The fourth session in the workshop was chaired by Mr. S. Regunathan, Principal Secretary- IT, Government of Delhi and addressed by Dr. Mark Davis, President, Unicode Consortium, USA.

Dr. Davis presented a detailed overview of Unicode Architecture, Advantages, Collation, Normalization, Unicode Stability Policy etc. The main points of his presentation are reproduced here:

Dr. Davis highlighted the advantages of Unicode. He said Unicode and ISO 10646 cover all major scripts used in India today. Unicode is now part of many vendors' solutions. There are many benefits of creating a single version of product which can be used worldwide, and using Unicode many of those benefits are available. The characters in an Indic Unicode script block are a superset of the ISCII character sets.

There are now three different forms of Unicode UTF-8, UTF-16, UTF-32 to allow for implementing Unicode depending on the systems capability to support 8,16 or 32 bit data.

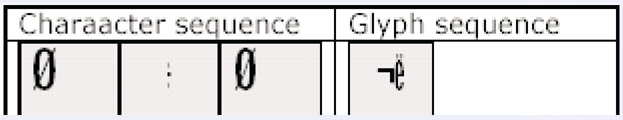
Scripts and not languages

Unicode Standard defines character set for *scripts and not languages*. Devanagari, Marathi, Nepali, Sanskrit are all represented by Devanagari range

0900-097F. Some of the Unicode code points are used to represent multiple characters like (. U-002E) is used to represent a Period, Decimal and full stop.

Characters vs. Glyphs

In response to a question related to Dr. Davis said that there is difference between *Character and glyphs*. Characters are code points used for storage while glyphs are visual representation of one or more code points.



Charaacter sequence | Glyph sequence

Fonts

Font is a collection of glyphs for visual representation. OpenType is a new cross-platform font file format developed jointly by Adobe and Microsoft. Opentype fonts contain different tables to support advanced typographic functions like

Character to glyph mapping (CMAP)

Positioning Information (GPOS)

Glyph definition data (GDEF)

Glyph substitution data (GSUB)

Locale data

Systems now support locale related information specific to user's language. Applications can use this data by using a programming interface layer of the system.

Date Time formats

Number formats

Currency formats

Input Methods

Sorting orders

Storage vs. performance

Another issue raised often related to storage taken by Unicode as compared to ISCII. One must evaluate advantages of Unicode in the implementations.

1. Most file formats like MS Word doc have much less storage for text data as compared to private binary data.
2. Most RDBMS today support Unicode collations for various Indic scripts.
3. Encryption schemes are available today which can compress small strings down to 1,2 bytes.

   ZIP

   SCSU

   BOCU (Binary ordered compression for Unicode)
4. Applications that support Unicode can be deployed Worldwide.
5. In new systems where the code units match the machine word size, access can be faster.

Properties and behavior of characters

The Unicode standard explains properties of Indic characters and their rendering behaviors.

Properties

Every character defined in Unicode has various character attribute values like Name, character classes, collation weights etc. These properties are used by application developers.

Unicode standard defines attributes for character properties. Implementations that confirm to Unicode Standard and use a character property must follow the specification as defined in this standard.

Behavior

Implementations must take care of script/language related behavior as specified in the Standard

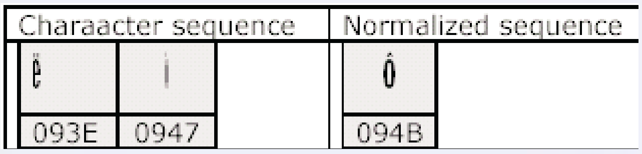
a. Bi-directional Algorithm (Arabic, Hebrew and Roman)

   To handle Right to left and left to right scripts

b. Line Break, User Character, Word to identify character, word and line boundaries.

c. Selection using Arrow keys and mouse

d. Normalization



Charaacter sequence		Normalized sequence
093E	0947	094B

A process of removing alternate representations of equivalent sequences from textual data, to convert the data into a form, which can be binary-compared for equivalence. In the Unicode Standard, normalization refers specifically to processing to ensure that canonically equivalent (and/or compatibility-equivalent) strings have unique representations.

Some of the usage of normalization are

Comparison, matching of Unicode streams

Conversion between different encoding schemes.

Collation

International Domain names

W3C Character set

Network File system

e. Collation

f. Regular Expression

For text and data pattern matching and manipulation.

g. Programming interfaces

Most platform software provide a layer for interaction with development environment through programming interface

Unicode standard defines Collation as: The process of ordering units of textual information. Collation is usually specific to a particular language. Also known as *alphabetizing* or *alphabetic sorting*. Unicode Technical Report #10, "Unicode Collation Algorithm," defines a complete, unambiguous, specified ordering for all characters in the Unicode Standard.

As Unicode defines encoding for script and not languages, implementations must implement collation rules applicable to the actual language represented by the character set.

There are differences in sorting order for Hindi and Marathi both of which use Devanagari character set.

क्ष *Devanagari Ksha (U+0915+U+094D+U+0937)* is treated as 35th consonant in Marathi

ज्ञ *Devanagari Jnya (U+091C+U+094D+U+091E)* is treated as 36th consonant in Marathi

Transcoding

Conversion of character data between different character sets. Applications must support conversion from Legacy, National encodings (ISCII) or other encodings to Unicode and vice versa.

ISCII -> UNICODE -> ISCII

Transliteration

Conversion of character data between different languages

Hindi -> Gujarati -> Hindi

Transliteration can be implemented using various methods.

1. Round Trip

This ensures transliteration from one language to another and back is always same. This may not be the correct transliteration.

Unique source -> Unique target

ç -> sá (sha)

è -> sà (sha)

2. Best Method

This ensures transliteration from one language to another is correct. This may not be ensure roundtrip transliteration.

ç -> sha (sá)

è -> sha (sà)

3. Keyboard Transliteration

Linked to QUERTY keys

ç -> sha

è -> Sha

4. Indic to Indic

It is often perceived that transliteration between Indian scripts is straight forward. Many Indic scripts have unused consonants and vowels(holes) which does not ensure round trip transliteration. Like in Tamil character set code points 0B96-0B98 are not used.

UNICODE stability

The Unicode Standard is continually expanding, new characters are added to meet a variety of uses. Character properties are also expanded to meet implementation requirements.

The Unicode standard stability policy ensures that implementations can depend on the fact that each version of the standard is a superset of previous versions.

Some of the policies are listed below :

1. Encoding (When encoded codepoints never changed)
2. Removal (No removal of existing character)
3. Name (When encoded names never change)
4. Normalization (No change in Canonical equivalences)
5. Property value (Structure of certain property values does not change)

Session-5 : Unicode & ISCII – Co-existence Paradigm

The fifth session in the workshop was divided into two part. In this session Dr. Peri Bhaskar, Tokyo University, Japan presented his views on “Unicode in the context of Indian Scripts”. This session was chaired by Dr. Omvikas, Senior Director, MCIT. Dr. M.N. Cooper, Modular Infortech, Pune, Ms Sharbani Bhonsle, CDAC, Pune, Mr. N. Anbarasan, Applesoft, Bangalore presented their views on the subject.

Ms Sharbani started her presentation by explaining the merits of the present 8 Bit National Standard ISCII. She mentioned that ISCII based software products can perform all the required functionality. She favored that ISCII should remain in existence for the few more years. She also pointed out that use of Unicode encoding scheme will require more data storage and processing time, apart from new set of hardware and software as most of system are windows 95/98 based. She gave some examples in support of her statement. It was also mentioned that the immediately switching to Unicode will require lot of expenditure in terms of hardware and software, to convert existing data into Unicode.

Dr Cooper, in his presentation mentioned that its transition period from ISCII to Unicode. He said that though the future encoding will be Unicode, we should not hurry up to convert our data into Unicode. He favored that ISCII should remain in existence for the few more years. In the answer to a question asked by Prof. J. Shah, VJTI, Mumbai, he mentioned that cost of converting data into Unicode will be more if we don’t start this activity at the earliest.

Mr Anbarasan of Applesoft advocated the existence of ISCII for the coming years. Also he mentioned that Applesoft has developed software patches so that older operating systems such as Windows 95 and 98 can support 16 bit Unicode encoding. He demonstrated a application in Tamil for the same. He also requested that Government of India should take up the matter with Microsoft to provide assistance to develop software to Unicode in Windows 95/98.

Dr Mark Davis from Unicode Consortium clarified that data storage requirement will be approximately doubled if Unicode encoding is used but it will not be 20 times as mentioned by Ms Sharbani Bhonsle in her presentation.

Dr Om Vikas in his concluding remarks mentioned that though Unicode is popularizing very fast there is a need that ISCII should not be phased out as there are already many applications, legacy data which has been stored in the ISCII encoding.

Session-6 : Unicode issues relating to Indo-Aryan Group 1.

The sixth session in the workshop was dedicated to "Unicode issues relating to Indo-Aryan Group 1: Devanagari, Gujarati & Gurmukhi". The session was chaired by Shri R.K. Arora, Eexecutive Director, CDAC, Pune. In this session Dr Mark Davis, Unicode Consortium & Prof. R.K. Joshi, C-DAC, Mumbai presented their views.

Mr. Arora introduced the speakers and also highlighted the importance of Unicode in present multilingual computing scenario.

Prof. Joshi presented the various characteristic of Indic scripts. He mentioned the peculiar features of Devanagari script. He advised that separate code points for Marathi should be added in the Unicode. He also explained the pure consonant model for Devanagari and its advantages. He mentioned that some Sanskrit characters which are very commonly used should be included in the Devanagari code chart.

Dr. Mark Davis updated the audience about the latest addition in Devanagari, Gurmukhi and Gujarati code charts. He also said any more missing characters may be added after discussion and proposal may be submitted to Unicode Consortium.

Session-7 : Unicode issues relating to Indo-Aryan Grou 2.

The seventh session in the workshop was on the "Unicode issues relating to Indo-Aryan Group 2: Assamese, Bengali, Manipuri & Oriya". The session was chaired by Dr K. Subramaniam, DDG, NIC. In this session Prof. B.B. Chaudhury, Indian Statistical Institute presented his views related to the Unicode for Bengali. Prof S. Mohanty, Utkal University, Bhuwaneshwar pointed out various issues related to Oriya. Dr. K. Whistler addressed various issues raised by speakers and the audience.

Dr B B Chaudhary was of the opinion that the punctuation mark, Viram, must not be borrowed from Devanagari. He said that this is not the basic purpose of Unicode and that the standard should have compartments for different languages.

Dr Chaudhary said that Viram being a very important punctuation mark must certainly not be shared. Instead it should be included separately for Bengali in one of the 40 vacant spaces.

Dr. Chaudhury pointed out some of the descripancies in the Unicode for Bangla. Dr. Ken Whistler assured of incorporating these suggestions.

Dr Chaudhary said that some important proposals must be considered for inclusion in the standard:

- Khanda ta is a very important character for Bengali and Assamese. This is by all means a character and must be written in its entirety.
- "Ja fala" should be included otherwise it is likely to create problems when used with reph as in ISCII coding schemes. "Ja fala" is not really a character in Bengali, but a separate coding space should be useful.
- Consonant modifier form, reph should also get a place instead of borrowing it from other scripts, because it is very widely used. Dr. Chaudhury justified this by pointing out that coding is not only for representation, it can also be used for document data compression where these issues must be considered. Dr Chaudhury further reinforced his point by saying that like "aa matra" which is not borrowed from Devanagari, we must also not borrow other forms.

Dr S Mohanty started her presentation by saying that the Oriya script is based on Devanagari script and influenced by the Dravidian writing style. The Oriya language grammar is based on Sanskrit. Oriya has influences of Pali and Prakrit. The Oriya script is non linear while the speech is linear. Dr Mohanty said that about 40 million people worldwide speak Oriya and hence a character set of vowels, consonants, conjuncts, punctuations etc. must have justified postions on the Unicode chart. She also pointed out that so far little work in this regard had happened.

The points made by Dr Mohanty during her presentation were as follows:

The Oriya language is purely phonetic. All normal consonants have the vowel 'a' merged in it. A pure consonant is *ardhamatra*, while *hraswa* vowels are *purnamatra*, *deergha vowels* are *dwi matra* and some other conjuncts are called *pluta* or *treematra*.

The combinations of consonants and matras to forms conjuncts carries with it the information of the duration of utterance. Dr Mohanty gave examples for each of the *ardhamatra*, *purnamatras*, *dwimatras* and *treematras*.

Describing the character set for Oriya, Dr Mohanty gave information on the vowels, the dipthongs, *bargya barnas*, the *abargya barnas* and the special barnas. Dr Mohanty said that Oriya used both *Purna Viram* (fullstop) and *Deergha Viram* (end of para) which are included in the Unicode chart. Other punctuation marks used by modern Oriya are also used in other languages, appropriate cross referencing for Oriya language may be provided for this.

Taking the discussion on to script specific issues and peculiarities, Dr. Mohanty said that the character set designed by Unicode is not complete.

Some observations and possible improvements as pointed out by Dr Mohanty were as follows:

- Form at 0B57 is incorrect. Oriya Script does not use a form as encoded at 0B57.
- Recent adoption by Unicode standards, of deergha R matra at 0B44, ae matra at 0B48 and aou matra as 0B4C are acceptable.
- Incorporation of all possible conjuncts in form of glyphs must be made. These glyphs are not available even in some prevalant Indian language software. These glyphs need to be created.
- Some glyphs found in Unicode compatible word processors are improper from a phonetic representation point of view. The Unicode standard needs to encode these correctly. In all about 80 glyphs for conjuncts need to be created.
- Matras need special positioning in cases like ku and kru. Also consonants when combined with i kAra and kAra form a different glyph. Rules for ku and kru must therefore be added.
- The conjuncts amka and anka mentioned on page 239 of the Unicode manuscript stands no relevance and should be avoided. It does not satisfy grammar rules.
- The representation of wa in Oriya needs discussion. Currently Oriya has two ba. One corresponds to the Devanagari ba (that follows pha) and the other is for wa (after la). Since Oriya uses mostly Devanagari style fonts, Dr. Mohanty opined that it would be better to keep just the ba as it is and use a nukta below the ba to represent wa. Dr Mohanty suggested that aum is encoded instead of wa at 0B71. The wa that is currently placed at 0B35 is incorrect since it does not follow grammar rules.
- Special characters like aum, avagraha, anuswar, visarga, chandrabindu, iswar, purna viram and deergha viram signs are frequently used in Oriya. The form referred to by the Unicode manuscript as virama on page 239 must actually be called 'halant'.
- Halant is frequently used in Oriya for marking the end of a pure consonant.

Dr Ken Whistler said that most of the issues pointed out during the session seemed to be general encoding issues and the most appropriate way to address these would be to follow up on the collection of the exact issues with the committee, formed for this purpose, and then analyze and follow up on these issues with the Unicode consortium. The consortium will then understand and resolve these issues. He said that ra hasant was added in Bengali and there has been considerable discussion on kanda ta and jafala on the mailing list and these need to be resolved.

Dr Mark Davis said that Unicode consortium has a number of Public issues updated periodically and one of these is with respect to khanda ta. He urged everyone to read these publications and contribute.

Session-8: Unicode Issues for Vedic Sanskrit, Kashmiri (Sharda Script), Tibetan & Phonetic Representation of Tribal Languages

Prof O N Koul from IILS spoke on Kashmiri (Sharda script) while Prof R K Joshi and Dr Alka Irani from C-DAC Mumbai (Formerly National Centre for Software development – NCST) presented their proposal on Vedic Coding scheme.

Prof Subbarao introduced the speakers on the dais. Dr Peri Bhaskar Rao while speaking about Sanskrit and Sanskrit-based languages commented that Paninian Grammar was precise and highly algebraic. He mentioned that Tibetian encoding for Unicode is adequate while Manipuri at present uses Bengali script but may need a few more characters. He informed the audience that the original script - Maithili – belongs to the Gupta period and there is a debate about whether it should have 36 characters or 28 characters. According to him one (with 28 characters) can be called basic script and another (with 36 characters) extended script.

Prof Omkar N Koul spoke about the script - Sharda as a script of historic importance and informed that during 9th and 10th Century, Kashmiri literature was in this script and it was a favourite script for the Kings in Kashmir. Sharda script is found only in handwritten form. He distributed a handout with Sharda vowels, consonant, vowel signs and a few conjunct letters. He mentioned that the peculiarity of the script Sharda is that conjuncts are formed by placing characters vertically one below the other. He also proposed that Devanagari should be considered as an additional script for Kashmiri.

Prof R K Joshi and Dr Alka Irani presented VagVarna - Vedic Sanskrit Coding Scheme which was already published in the TDIL newsletter, VishwaBharat@tdil in Language Technology Flash, October 2002.

Prof Joshi emphasized that Vedic Sanskrit includes multi-tier usages of diacritic marks of complex compositions, above, below and at sides of the characters. Vedic Sanskrit text needs adequate range of characters to compose exhaustive syllabic range of consonants and conjuncts. Such a concern according to them is not adequately handled in the present Unicode standard for Devanagari. The total set of characters is divided into Chart 1 and 2. Chart 1 follows uniformly additive approach for composing all possible syllabic compositions for text and speech. The proposed scheme is evolved to reflect the traditional phonological concern of Sanskrit language. The range of characters in the proposed Sanskrit Chart includes Swara Varnas, Swara Bhedak Chinhas, Vyanjan Varnas, Vyanjan Bhedak Chinhas, various other chinhas to show nuances of spoken language like kaal (time duration), bala (stress), kamp (vibration) as well as specific vedic chinhas such as Vedic Anuswaras, Visargas, Swaritas and Sammvedic intonation chinhas. The Chart 2 includes about 100 Vedic Chinhas needed for Rugveda, Yajurveda, Atharvaveda and includes Saamvedic intonation marks of various schools.

While speaking about the importance of this scheme, Prof R K Joshi told that such a scheme would facilitate the phonetic break-up of words from multi-lingual dictionaries as part of written and spoken corpora. This will help to render words with correct orthography as well as equally correct pronunciation while explicit break up of vowels would help linguistically in the areas of Sanskrit Shabdasandhi and Samasa. This scheme was modified and improved through various meetings with representatives of regional and national institutes for Sanskrit studies, Govt. of India officials and linguistic scholars and was endorsed by them.

During question-answer session Prof J N Shah from VJTI, realising the merit of the scheme, asked a question as whether Unicode will be flexible enough to make necessary changes in existing scheme to incorporate this model. Dr Alka Irani commented on this saying that Unicode's policy of never to change any character's name and coding may not allow it to incorporate these changes. However, we can evolve a national standard base on this scheme, which is based on Indian Language characteristics and peculiarities.

Session-9: Unicode Issues for Dravidian Group: Kannada, Malayalam, Tamil & Telugu

Dr V S Umamaheswaran mentioned about the features added in Unicode 4.0.

He updated that 2 Kannada characters, 8 Tamil symbols like day, month, year, credit, debit etc. have been added.

With respect to the text and descriptive material. Some text added about Dravidian characters, Handling of Danda and Double danda, Use of the name halant and virama – what is preferred and what is not.

He mentioned the points of framework for discussion.

Main point is that one should respect the stability policy. Hence

- No removal of existing characters
- No relocation
- No changes in Names
- No changes in canonical equivalence / normalization
- No new multiple spelling
- No new encoding model

He mentioned the suggestions that can be entertained by Unicode Technical Committee.

- Text for FAQs, Technical notes for better understanding
- Possible new sequences
- Annotations where appropriate
- New character only with evidence
- Deprecation with strong justification

He also addressed the Proposal for FAQ, Proposal for Unicode technical notes, Proposal for Explanatory text and Proposal for annotation, ISO and Unicode relationship.

Prof K Narayana Murthy explained about the character, Aksharas and script grammar and overall user experience, …

Akshara – C*V , C+

This is not the same as syllable. There are, theoretically speaking, infinite number of Aksharas. Up to 5 levels of combinations (e.g. kartsnya) and 10 billion possibilities. However, based on corpus analysis, about 20,000 of these are used, of which 99% use 5000 syllables or aksharas.

He also talked about

- Grammar of syllables & script grammar
- Require FSM to handle /recognize valid/invalid combinations, That must be a part of any character-encoding standard.
- Ungrammatical – script grammar must be enforced
- Examples based on ISCII grammar.
- One gylph per Akshara is not feasible. So glyphs must be composed. Issues are completeness, robustness etc.
- Different way of breaking up is required.

He also highlighted the issues on

- Keyboard layout, Fonts, Rendering, Text processing, Planned effort for total processing

He suggested that allographs / allophones should not be encoded.

He also mentioned some problems in ISCII as it is Superset theoretically not acceptable, Script grammar could have been different, Attempts to encode shape variation.

Dr. Narayana Murthy pointed that Unicode will be widely supported. There is no escape to it.

- Space not the biggest problem with the use of unicode.

Shri K.S. Lakshimarayana, MD, ELCOT and Dr. V. Krishnamoorthy presented the requirement of Tamil computing on behalf of Government of Tamil Nadu. He mentioned that present encoding model for Tamil is not suitable to Tamil requirement and suggested new model for encoding of the Tamil in Unicode Standard. He advocated the all character encoding model for Tamil as with this scheme all possible combinations of vowels and consonants require only 313 code space.

He mentioned about the Tamil people spread over the variety of countries like Singapore, Malaysia, Australia and their enthusiasm for Tamil computing, leading to TamilNet conferences from 1998.

In the 1999 conference, Standardized keyboard layout "Tamil net 99" and the font encoding - TAM and TAB.

Dr V Krishnamoorthy presented a comparison of the following three coding schemes for Tamil.

- Present Unicode standard
- Vowel and Consonant
- All character encoding

He mentioned that Tamil is different from other Indian languages as it not complex script, Has 313 characters and One to one relation between them is possible. Dr. Krishnamoorthy said that "pulli" is not a character. That is a Gylph. Consonants are represented as a combination of two code points in present Unidode. This is not natural and Current Unicode is not a character-encoding scheme.

He highlighted the second encoding scheme as Pure consonant & vowels encoding scheme.

He said that the All Character Encoding scheme is Comprehensive encoding scheme.

All 313 character depicted as they are in this scheme.

Study of the 3 encoding schemes based on typical text processing functions, and evaluated based on file size, search time etc. Text processing functions considered were, Find and replace, Creation of Database file and Morphological analysis.

On all counts, the III scheme is better than the others by 20-75%. Therefore, the recommendation is for the all-character encoding scheme. It is simple and natural in all the ways.

In response to the new encoding model for Tamil, Dr. Mark Davis, President , Unicode Consortium said that two encoding model for one script should be encouraged. It will lead to the multiple spelling for one character.

Session-10: Unicode relationship to ISO/IEC 10646

The tenth session in the workshop was on "Unicode relationship to ISO/IEC 10646" . The session was chaired by Dr. P.K. Chaturvedi, Director, Department of IT. Dr V S Umamaheswaran, IBM Toronto Lab delivered a presentation of the subject

Dr Umamaheswaran described the Unicode relationship to ISO/IEC 10646 in detail. He highlighted the various procedures adopted by ISO and Unicode for Standardizations related activities such as submitting the proposal for addition of new characters in the Unicode Standard. He emphasized that before submitting a proposal one has to do all the *homework* in terms of required information/ clarification/ example/ documentary proofs etc.

He informed that normally it takes one year to encode a character in the Unicode as this period is required for doing all the necessary formalities. He specifically mentioned the Unicode Stability policy and said that if a character is encoded it cannot be deleted, shifted or name changed. Hence it is necessary to be sure about the use/ nomenclature/ position in the code chart, about any character before submitting the proposal.

Dr Umamaheswaran also said the National Standardisation body should participate in the ISO standardization activities.

Session-11: Unicode – Generic problems faced in Implementation on various platforms - I

Session 11 & 12 in the workshop were on "Unicode - Generic problems faced in implementation on various platforms". The session was chaired by Shri Sanjay Singh, Secretary IT, Government of Madhya Pradesh. The prominent speakers in the session were Dr. Mark Davis, Unicode Consortium, Ms. Cathy Wissink, Microsoft Corporation, USA, Prof. Jitendra Shah, VJTI, Mumbai and Mr. Satish Mohan, Red Hat India.

Dr Om Vikas insisted the need of the Unicode fonts for all Indian languages. He said that for acceptability and understanding, pilot projects should be developed. He also emphasized the need of co-existence of ISCII & Unicode and also the need of arriving a consensus by removing discrepancies.

Dr Mark Davis from Unicode Consortium and Mr Satish Mohan from RedHat India answered queries of participants.

Mr Satish Mohan started his presentation with the background of Hindi spoken population and other Indian languages. The figures indicated the presence of a huge population speaking Hindi, Urdu and other Indian languages. He further discussed the enabling methodology, which covered UNICODE standards, common foundation for localization and rendering using *Pango*. The goal of the *Pango* project is to provide an open-source framework for the layout and rendering of internationalized text. *Pango* is an offshoot of the GTK+ and GNOME projects, and the initial focus is operation in those environments, however there is nothing fundamentally GTK+ or GNOME specific about *Pango*. *Pango* uses Unicode for all of its encoding, and will eventually support output in all the world's major languages.

He further specified the needs for locale information like Reading, Keyboard mappings, input methods, translation of terminologies & phrases and GUI. Collecting information for different languages like character sets, collation sequence, Day, Currency, messaging and cataloguing etc.

He also pointed out that localized versions are coming across. Different tools can facilitate this localization process. These include using Pango, Li18nux, gettext, KBabel etc.

He explained the use of UNICODE character set in programming with the help of famous 'Hello World' example. There is a need of developing Open Type fonts, Input methods, rendering and printing etc.

He stressed the need for localizing Browsers, Full GUI, E-mail clients, Multi-media tools, GNOME. The work on building locale, font installation, setting up input method, installing translations, setup multiple input methods, voice recognition, multiple keyboard overlays and inclusion of new language packs. The ultimate aim is to develop an environment from where different languages are supported from one single work space.

Dr Jitendra Shah from VJTI pointed out that in Linux printing is being done through PS/Ghostscript, which should be in push button way for easier way of printing. He also emphasized that locales should also be provided for minority languages also. And there should be mechanism for testing process, standardization in translation of terms and information and need of glossaries in other Indian languages for uniformity and standardization.

Another query raised was that why we need specific keyboard drivers, and answer to this was the need of developing a generic Keyboard.

The representative from TVS asked what type of print support is needed from developer of Printers and mentioned that print facility is there but it has to be made more user friendly and printing layer enablement is to be provided.

Dr Mark Davis emphasized that the policy of Unicode was to encode different scripts not different languages, hence languages having same scripts can not be provided separate space in Unicode, however if certain additions are needed, they are welcome provided they fit at the free blocks in Unicode.

Regarding the co-existence Dr Davis said that a path is to be defined in terms of few years not for a very long time in the future as Unicode will be prevailing everywhere from 5 years from now.

Shri Sanjay Singh introduced the speakers on the dais. He also mentioned the requirement of Open Source Software.

Prof. Jitendra Shah demonstrated the application "**Gnubharati**" developed by his group on Linux. He mentioned the problem faced by him in implementation of Unicode on Linux platform. He also developed Unicode Open Type font "Padma" and "Gargi" for which he invited audience to give feedback.

Ms Cathy Wissink assured all possible support for Unicode related application development.

3.2 Sanskrit Informatics Workshop

Venue: Indira Gandhi National Centre for Arts(IGNCA), 3 Rajendra Prasad Road New Delhi -110001

A Report on the Proceedings

Day 1 : August 7, 2003 (Thursday)

Session 1 : Inagural Session

The following participants addressed the morning session

Dr. G.C.Tripathi, IGNCA

Dr Unruh Gerd

Prof. R.K.Saxena, Rector, JNU

Dr. Om Vikas, Sr. Director, DIT, MC&IT

Prof. G.V.Singh (gvs10@hotmail.com)

The session started with a welcome address by Prof. G.V.Singh. Prof. Singh stressed the fact that Sanskrit is not just one of the languages but *the language* that linguists and we computer scientists are talking about. It is structure of this language that interests we all.

Prof. G.C. Tripathi

In his keynote address, Prof. Tripathi said that India is a treasure of time proven scientific knowledge. All this knowledge is available in the form of Sanskrit literature which is not accessible to the scholars due to various reasons. So we need to morph this knowledge so that it is available to users in localized forms. We are moving towards a knowledge society and need to develop tools and technologies that have knowledge processing capabilities.

Prof. Gerd Unruh

Prof. Gerd said that in order to develop high-end technologies we have to combine and speed up effort to justify expenditure. We need to market these knowledge systems and tools to countries in Europe and America. We need to integrate the traditional and holistic knowledge of the east with the modern and fragmented knowledge of the west.

Prof. R.K. Saxena

Prof. R.K.Saxena, Rector JNU stressed the need for such initiatives and also lauded the efforts of Prof G.V. Singh and his team in this direction. He also emphasized on 'marrying the nascent Sanskrit department in JNU to the School of Computer and System sciences.

Dr. Om Vikas (omvikas@mit.gov.in)

Dr. Om Vikas said that it heartening to know that Sanskrit Informatics is no more a national cause but it has become an international cause. More and more of international scholars from France, Germany, USA, are showing interest in our heritage knowledge. He stressed that our heritage knowledge has a holistic and integrated approach. Sanskrit language has a mathematical framework. Therefore, abstraction of this knowledge is possible. He also talked about the new initiative taken by DIT, Govt. of India, i.e., the KUNDALINI project.

Session 2 : Development of Basic Information Processing Kit (BIPK)

Chair Dr. V.V.S. Sarma (vvs@csa.iisc.ernet.in)

Speakers Mr. M.D. Kulkarni, Mr. P.Ramanujam, and Prof. G.V.Singh

Mr. M.D. Kulkarni from CDAC Pune introduced various linguistic and computional resources that are being developed at C-DAC Pune. He talked about the limitations that are coming in the way of tools and technologies being developed. He talked about the requirement of a **Basic Information Processing Kit for Sanskrit (Vedic)**. After giving the background information of the project, he listed the requirements of a **Vedic Sanskrit** BPIK.

Mr Ramanujam of C-DAC Banglore, started with a need for '**Heritage Computing**'. He showed presentation of **Desika** system. He underlined the several problems that are being faced in converting *Desika* from a DOS based system to a Windows based system. He also presented his '**Rg Veda Ratnakar**' and '**Gita Reader**' projects