

MOST URGENT -

By SPEED Post!

Sectt. 37 GPM. 3/225/97. 15.5 Lakhs.

L2/07-013



GOVERNMENT OF KERALA
INFORMATION TECHNOLOGY DEPARTMENT

No: 10676IT 1/2001/ITD.

Thiruvananthapuram,
Dated: 19.09.2001.

From

The Secretary to Government.

To

Ms Swaran Lata,
Additional Director,
Ministry of Information Technology,
Electronics Niketan,
6, C.G.O. Complex,
New Delhi, 110003.

Sir,

Sub:- Revision of The Unicode Standard Version 3.0 Compilation of
Proposed amendments with regard to Indian Scripts - reg.

Ref:- 1. Your letter No. 13 (3) CDD-/2000, dated 28.08.2001.
2. Letter No. 113/SECY//01/IT dated 06.09.2001.

I am directed to invite your attention to the reference cited and to forward herewith the duly filled up proposal Summary Form to accompany submission for Additions to the Repertoire of ISO/IEC 10646. The Report of the Committee on Standardisation of Malayalam Key Board and Character encoding, copy of Malayala Thanima, a publication of Kerala Institute of languages and pages of Malayalam Magazines are also forwarded herewith.

Yours faithfully

ISHITA ROY
ADDITIONAL SECRETARY
For Secretary to Government

8.9
25/9

550(MJ)

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646

Please fill Sections A, and C below. Section D will be filled by SC2/WG 2.

For instructions and guidance for filling in the form please see the document
"Principles and procedures for Allocation of New Characters and Scripts"
(<http://www.dkuug.dk/JTC1/SC2/WG2/prot>)

A. Administrative

1. **Title:**
Modification in code allocation for Malayalam in Unicode
2. **Requester's name:**
Ministry of Information Technology, Govt of India
3. **Requester type** (Member body/Liaison/Individual contribution):-
4. **Submission date:**
17.09.2001
5. **Requester's reference** (if applicable):
6. **(Choose one of the following:)**
This is a complete proposal: ; or,
More information will be provided later:

Complete proposal

B. Technical – General

1. (Choose one of the following
The proposal is for addition of character(s) to an existing block:
Name of the existing block: Malayalam (code space: 0D00 hex to 0D7F hex)
2. **Number of characters in proposal:**
Addition: 5, Shape change: 1 deletion/reserve: 14, description change: 1
3. **Proposed category** (see section II, Character Categories):

Category A

4. Proposed Level of Implementation (see clause 15, ISCO/IEC 10646 -1):

Is a rationale provided for the choice?

If Yes, reference:

5. Is a repertoire including character names provided?:

Yes

a. If YES, are the names in accordance with the character naming guidelines in annex K of ISO/IEC 10646 - 1?

Yes

b. Are the character shapes attached in a reviewable form?

Yes

c. Who will provide the appropriate computerized font (ordered preference(True Type, Post Script or 96x96 bit-mapped format) for publishing the standard?

If available now, identify source(S) for the font (include address, email, ftp-site, etc.) and indicate the tools used:

Yes, ER&DCI, Trivandrum, India

7. References:

a. Are references (to other character sets, descriptive texts etc.) provided?

Yes – Report of the Committee for Standardization of Malayalam Keyboard layout and Character Encoding

b. Are published examples (such as samples from newspapers, magazines, or other sources) of use of proposed characters attached?

Yes.(Samples from magazines attached.)

8. Special encoding issued

Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration et:

Yes, Report of the Committee for standardization of Malayalam keyboard layout and character encoding
(Enclosed Report of the Committee)

C. Technical – Justification

1. Has this proposal for addition of character(s) been submitted before?

No

2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc)?

If YES, with whom?

If YES, available relevant documents?

Yes – A committee of experts was constituted by the Government of Kerala to recommend the standards for keyboard layout and character encoding in Malayalam. Report of the committee attached.

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?

People of Kerala, a state in India. Size – 40 million. The proposed characters are used in all publications in Malayalam. Reference:

4. The context of use for the proposed characters (type of use; common or rare) Common use.

Copy of the report published by State Institute of Languages attached ('Malayalathanima')

5. Are the proposed characters in current use by the user community?

If YES, where? Reference

Yes- in all Malayalam documents. Report of the state Institute of Languages attached.

6. After giving due considerations to the principles in N 1352 must the proposed characters be entirely in the BMP?

IF YES, is a rationale provided?

If YES, reference:

Yes – Malayalam characters are entirely in BMP

7. Should the proposed characters be kept together in a continuous range (rather than being scattered?)

Yes

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?

If YES, is a rationale for its inclusion provided?

No

9. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?

The proposed chillu character (MALAYALAM LETTER N) is similar in appearance to MALAYALAM DIGIT NINE (code 0D6F).

The proposed chillu character (MALAYALAM LETTER RR) is similar in appearance to MALAYALAM DIGIT FOUR (code 0D6A).

Even though similar, functionally they are entirely different. Also, Malayalam digits are not used now

10. Does the proposal include use of combining characters and / or use of composite sequences (see clause 4.11 and 4.13 in ISO/IEC 10646 -1)?

If YES, is a rationale for such use provided?

If YES, reference:

The proposed character do not combine with other characters to form conjuncts.

11. Does the proposal contain characters with any special properties such as control function or similar semantics?

If YES, describe in detail (include attachment if necessary)

No

D. SC 2/WG 2 Administrative (To be completed by Sc 2/WG 2)

1. Relevant Sc 2/WG 2 document numbers:

KITSS

2. Status (list of meting number and corresponding action or disposition):
3. Additional contact to user communities, liaison organisations etc:
4. Assigned category and assigned priority/time frame:

Enclosures:

1. Report of the Committee on Standardisation of Malayalam Keyboard and Character Encoding
2. Copy of Malayala Thanima
3. Samples from Magazines

109

**REPORT OF THE COMMITTEE
ON
MALAYALAM CHARACTER
ENCODING
AND
KEYBOARD LAYOUT
STANDARDISATION**

108

CONTENTS

Section	Description	Page
	Executive Summary	3
1	Introduction	9
1.1	Character encoding	10
1.1.1	Character encoding process	10
1.1.2	Character Vs Glyph	11
1.2	ASCII	12
1.3	UNICODE	12
1.4	National Initiatives	14
1.5	Malayalam keyboard	15
2.	Malayalam character set	16
3.	Malayalam keyboard layout	19
3.1	Modifications in Inscript layout	21
3.2	Typing sequence	21
3.3	Conjuncts	24
3.4	Intra glyph positioning	27
	Malayalam Keyboard layout	27.a
4.	Malayalam character encoding	28
4.1	Modifications required in Unicode	29
5.	Lexicographic ordering of Malayalam characters	33
5.1	Searching and sorting	33
6.	Other recommendations	39
6.1	Visual representation of characters	39
6.2	Malayalam Research Centre	39
6.3	Official Malayalam dictionary	40
6.4	Government initiatives required	40
Annexure 1	Code table of ISCII	43
Annexure 2	Members of the Standardisation Committee	46

EXECUTIVE SUMMARY

To take the breath-taking developments in Information Technology to the common people in Kerala, the computer applications should be available in Malayalam, or at least with front end in Malayalam. A large number of applications are to be developed in Malayalam mainly text processing, publishing, database applications, web applications which can be ported to any other platform and which can be executed in any platform without any patches. This calls for two minimum requirements to be achieved - standardization of Malayalam keyboard layout and Malayalam character encoding. In the absence of these standards, the application developers will use their own schemes for keyboard layout and character encoding with the result that one document/ application developed with one package cannot be read/ executed in another platform/ environment.

The Government of Kerala had constituted a Committee for Standardisation of Malayalam Keyboard and Character Encoding with Shri Govinda Pillai as Chairman and Smt. Aruna Sundararajan IAS Secretary (IT) to the Government of Kerala as the Convener.

The Committee has taken the following decisions with regard to Keyboard layout and Character Encoding:



- I. Accept In script keyboard with modifications
(Details on page no. 21-27.a)**
- II. Accept UNICODE as the standard for Character Encoding for representation of Malayalam in computer processing.**


II. a. Modifications suggested to UNICODE


The Committee also decided that while accepting UNICODE as the standard for Character -encoding, the following modifications are to be incorporated with regard to character encoding in the existing version of UNICODE, which supercede all the earlier modifications.

1. The following chillu characters are to be included in the Malayalam character set:

<i>Suggested code Position</i>	<i>Character</i>	<i>Description</i>
0D58	ൽ	Malayalam Letter L
0D59	ൾ	Malayalam Letter LL
0D5A	റ	Malayalam Letter RR
0D5B	ൻ	Malayalam Letter NN
0D5C	ൻ	Malayalam Letter N














2. The correct glyph for Malayalam vowel sign AU (code position 0D4C) . This is wrongly shown as  in the Unicode specification. The correct description is

<i>Code Position</i>	<i>Character</i>	<i>Description</i>
0D4C		Malayalam vowel sign AU


3. Malayalam AU Length Mark (code position 0D57) is not used in Malayalam. The glyph for this character is wrongly shown as , which is actually the Malayalam vowel sign AU (as explained in para 2). So, this code position may be reserved:

<i>Code Position</i>	<i>Character</i>	<i>Description</i>
0D57		Reserved

4. The following characters, which are now provided in Unicode, may be deleted/reserved, since these are no more used in Malayalam:

	<i>Remarks</i>
Malayalam Letter Vocalic L	 (0D0C)
Malayalam Letter Vocalic LL	 (0D61)
Malayalam Letter Vocalic RR	 (0D60)
Malayalam Digit ZERO	 (0D66)
Malayalam Digit ONE	 (0D67)
Malayalam Digit TWO	 (0D68)
Malayalam Digit THREE	 (0D69)
Malayalam Digit FOUR	 (0D6A)
Malayalam Digit FIVE	 (0D6B)
Malayalam Digit SIX	 (0D6C)
Malayalam Digit SEVEN	 (0D6D)
Malayalam Digit EIGHT	 (0D6E)
Malayalam Digit NINE	 (0D6F)

5. The description of the character (code position 0D34) is to be corrected as MALAYALAM LETTER ZHA. This is wrongly specified as MALAYALAM LETTER LLLA in Unicode standard.

<i>Code Position</i>	<i>Character</i>	<i>Description</i>
0D34		Malayalam letter ZHA

OTHER RECOMMENDATIONS MADE BY THE COMMITTEE**1. LEXICOGRAPHIC ORDERING OF MALAYALAM CHARACTERS**

The following order of characters is recommended for the lexicographic ordering of Malayalam words.

അ	ആ	ഇ	ഈ	ഃ (സംവൃത ഉകാരം)
ഉ	ഊ	ഋ	ൠ	എ
ഐ	ഒ	ഓ	ഔ	
ക	ഖ	ഗ	ഘ	ങ
ച	ഛ	ജ	ഝ	ഞ
ട	ഠ	ഡ	ഢ	(ൺ ണ)
ത	ഥ	ദ	ധ	(ൻ ന)
പ	ഫ	ബ	ഭ	(ം മ)
യ	ര	(ർ റ) (ൽ ല)	വ	
ശ	ഷ	സ	(ഃ ഹ) (ശ്ശ ഷ്ശ)	

2. VISUAL REPRESENTATION OF CHARACTERS

Unicode, the coding scheme we have recommended with modifications, do not specify the visual representations of characters. It completely de-links the codes from the displayed fonts. The same text can be seen in different font styles by using a different font composition routine. A word can be displayed in a variety of styles depending on the conjunct repertoire used.

These standards define how the characters are interpreted, not how glyphs are rendered. The software or hardware-rendering engine of a computer is responsible for the appearance of a character on the screen. The standards do not specify the shape, size or orientation of the characters on the screen or on the paper.

3. MALAYALAM RESEARCH CENTRE

The Committee recommends to the Government to set up a Malayalam Research Centre with the following broad objectives:

- To be a repository of Malayalam language tools
- To develop tools and technologies in Malayalam
- To exploit the emerging technologies for use in Malayalam computing
- Assist Government in technology dissemination

4. OFFICIAL MALAYALAM DICTIONARY

The Committee recommends to the Government to publish an official Malayalam dictionary. The dictionary shall be prepared in such a way as to be used as a reference for the linguists, the software developers and the common man alike. The dictionary shall aid the standardising process of the script of Malayalam words and words will be arranged in the

CONFIDENTIAL REPORT 109

order mentioned in this recommendation. This will help in the software development activities like electronic dictionary, thesaurus, sorting engine, search engine, spell check etc.

5. GOVERNMENT INITIATIVES REQUIRED FOR IMPLEMENTATION OF THE STANDARD

If the standard recommended in this document is to be widely used, Government may implement the following plan of action.

- Official declaration of the standard to be published
- A committee to be formed to evaluate any feedback and amend the standard, if required
- Approach BIS to amend the keyboard and character encoding standards in Malayalam
- Approach Unicode consortium through MIT, to modify Unicode specifications for Malayalam
- CDAC to be approached to make amendments in the applications developed by them
- Take initiative for the development of at least one Malayalam font, which will work in all major platforms, especially Windows, Unix, Linux and Mac systems and distribute it freely on Internet.
- All software developers in Malayalam to be intimated about the new standard.
- Make it mandatory for all the Malayalam software to conform to the standard within six months
- Any training in Malayalam software (DTP) to be recognized only if the package conforms to the standard.
- Government to procure only those Malayalam software packages, which conforms to the standard

TERMINOLOGY

Alphabet: A set of letters used in writing.

Aspirated consonant: A consonant, which is pronounced with an extra puff of air coming out at the time of release of oral obstruction. This has a sound of an extra 'h'.

Basic alphabet: The minimal set of letters, which can be used for uniquely encoding every word of a language
Bit: Binary digit. It can have only two values: 0 and 1.

Byte: Group of eight bits (Also referred to as octet).

Character: A symbol which can represent a letter, a numeral, a punctuation mark, a special symbol or even a control function.

Character code: Position in the code table of the character.

Character set: A set of characters grouped together for a purpose, like that of representing a script.

Code table: A table showing the positions allotted to individual characters from a character set.

Conjunct: A letter, which is a combination of two or more letters.

Consonant: A letter representing a speech sound in which breath is at least partly obstructed, and which has to be combined with a vowel to form a syllable

Display rendering: The process by which a string of characters is displayed (or printed).

Font: A collection of glyph images that share the basic design.

Glyph: The exact shape by which the character is represented on paper or screen is called a glyph.

Letter: A character representing one or more of the simple or compound sounds used in speech. It can be any of the alphabetic symbols.

Nasal consonant: A consonant pronounced with the breath passing through the nose.

Phonetic alphabet: An alphabet, which has direct correspondence between letters and sounds.

Pure consonant: A consonant, which does not have any vowel implicitly, associated with it.

Script: A distinctive and complete set of characters used for the written form of language.

Syllable: A unit of pronunciation uttered without interruption, forming whole or part of a word, and usually having one vowel sound optionally surrounded by one or more consonants.

Vowel: A letter representing a speech sound made with the vibration of vocal cords, but without audible obstruction.

Vowel sign: A graphic character associated with a letter, to indicate a vowel to be associated with that character.

Abbreviations

ASCII	American Standard Code for Information Interchange
BIS	Bureau of Indian Standards
CDAC	Centre for Development of Advanced Computing
DBMS	Database Management System
EDCDIC	Extended Binary Coded Decimal Interchange Code
HTML	Hyper Text Markup Language
ISCII	Indian Script Code for Information Interchange
ISO	International Standards Organisation
MIT	Ministry of Information Technology
OCR	Optical Character Recognition System
OS	Operating System
TDIL	Technology Development in Indian Languages
UTF	Unicode Transformation Format
WWW	World Wide Web

1. INTRODUCTION

Information Technology has become so pervasive that it is difficult to identify any area of human concern, which is not influenced by it. One of the major area where the advancement of Information Technology is extensively used is text processing, storage and distribution. Preparing and processing of documents, including desk top publishing, is increasingly being done on computers and in the days to come, documents prepared by hand will be very rare except those prepared for very personal use.

Gone are the days where the documents were tied down to the computer in which it was prepared. With the advent of Internet and World Wide Web, people can access any document in any other computer anywhere in the world. Documents are, in general, meant for somebody else to read. If a second person has to read a document, then it has to be written in a format unambiguously understood by him. That is, there has to be a unique way of preparing the document, which is known both to the creator and the user of the text. Thus, if there is a common format in which documents are prepared, anybody who adheres to the standard will be able to read the document in any other computer.

Most of the information available on the web today is in English. There is clearly defined standard in English for inputting text through computer keyboard, for storing and retrieving data and in rendering (displaying) the text on the screen or paper. All the software developers, computer developers and users adhere to this standard. So, as far as English is concerned there is no need to worry about who created the document in what software package and what platform etc. Receiving and interpreting data in English, from whichever source it has come, are straightforward, thanks to the standard keyboard and character encoding accepted by all.

If the fruits of the breath-taking developments in Information Technology are to reach the common people in Kerala, the computer applications should be available in Malayalam, or at least with front end in Malayalam. We need to have a lot of applications in Malayalam – be it text processing, publishing, database applications, web applications or anything else – which can be ported to any other platform and which can be executed in any platform without any patches. This calls for two minimum requirements to be achieved - standardization of Malayalam keyboard layout and Malayalam character encoding. In the absence of these standards, the application developers will use their own schemes for keyboard layout and character encoding with the result that one document/application developed with one package cannot be read/ executed in another platform/ environment.

Today, there are many software packages available in Malayalam, mainly for word processing. The developers of these software packages have used different keyboard layouts and different character encoding schemes. As a result of this, a document prepared using one software package cannot be read using another package. Malayalam is also not supported in most of the widely used platforms, operating systems or application packages, because there is no standardised keyboard layout and character encoding. For the same reason, Malayalam content in the World Wide Web is also very few. The current situation is that one has to download and install several Malayalam fonts, each with its unique mapping scheme, to be able to read Malayalam contents available on the Internet.

Government of Kerala has initiated various schemes aimed at bringing the Government to the people. Computerisation efforts of various Government departments and local bodies are in full swing. Content generation in Malayalam is a main component of all these projects. This makes it all the more important and urgent to define a standard keyboard layout and character-encoding scheme for Malayalam, which is accepted by all.

On this background, Government of Kerala constituted a Committee as per GO (Rt) No.93/2000/ITD dated 2.6.2000 to recommend to the Government a Malayalam computer keyboard layout and Malayalam character encoding standard. The Committee was formed with Shri P.Govinda Pillai, Chairman, CDIT as the Chairman and Smt Aruna Sundararajan, Secretary, Department of Information Technology, Government of Kerala as Convener. Subsequently, some more experts on IT and linguistics members were co-opted to the Committee. The names of the members of the Committee are given in Annexure 1.

1.1 Character encoding

When a word processor user presses a key labeled “അ” on the keyboard, he expects the character “A” to appear on the screen. He also expects “അ” to appear in paper when he prints the text. He expects the character to be stored as “അ” when he saves the text, and somebody else using a different computer to see this character as “അ”, when he sends this text to that computer (over network).

Computer can operate only on numbers. So, any other input to the computer, including text, has to be converted to numbers before computer processing. When a word processor user types a key in the computer keyboard, the computer's system software receives a message that the user pressed a key (or a combination of keys) for “അ”, which it encodes as a number. The word processor stores this number in memory, and also passes it on to the display software responsible for putting the character on the screen. The display software, which may be a window manager or part of the word processor itself, uses the number as an index to find the image of “അ”, which it draws on the monitor screen. The process continues as the user types in more characters.

To be of any use in computers, especially in computer communications and in particular on the World wide Web, characters must be encoded. In fact, much of the information processed by computers over the past few decades has been encoded text, exceptions being images, video and numeric data. Encoding can be loosely defined as mappings between the character sequences that users manipulate and the sequences of bits that computers manipulate.

1.1.1 Character encoding process

A character set encoding has to satisfy the primary requirement, that of the unambiguous representation of the content of the written text. To make it possible to successfully encode, process and interpret text a character set must:

- Define the smallest useful elements of text to be encoded. The units of encoding, the characters, are pragmatically chosen as appropriate to express text and allow various text processes. This set is called a repertoire.

- Assign a unique code to each element of the repertoire.
- Provide basic rules for encoding and interpreting text so that programs can successfully read and process text

In some cases, the whole encoding process can be collapsed to a single step, a trivial one-to-one mapping from characters to bytes.

Because encoded text cannot be interpreted and processed without knowing the encoding, it is vitally important that the character encoding is known at all times and places where text is exchanged or stored. Mandating a unique encoding has strong virtues of simplicity, efficiency and robustness. Since the unique encoding is known implicitly from usage of such protocol or data format, the protocol or data format does not have to carry character-encoding tags.

1.1.2 Character Vs glyph

A writing system's alphabet, numbers, punctuation and other writing marks consist of characters. A character is the symbolic representation of an element of a writing system - a letter, symbol or number. As soon as we write a character, it is no longer abstract but concrete. The exact shape by which the character is represented on screen and paper is called a glyph. They are the components used to generate the visible representation of characters. Glyphs need not correspond one-to-one with characters. A single character can be represented by multiple glyphs and a single glyph can represent multiple characters. Just like characters are the basic unit of organization of encoded text, glyphs are the basic unit of organization of the visual rendering text.

The character sets that are stored in the computer are independent of the display format. The character sets and the font/glyph are defined separately and there need not be a one-to-one relationship between them.

A font is a collection of glyphs, all of similar design, that constitute one way to represent the characters of the language. Different fonts for the same language will typically have different glyphs to represent the same character. The code positions of the glyphs are not important, as are positions of characters in a characters set. For example, Times New Roman font may have the general style on Macintosh or IBM-compatible PC, but the code positions of the underlying characters are significantly different on these systems after the first 128 characters. The character set differences are taken care of by the underlying system.

But in the Internet environment, this may not be applicable always. HTML encoded documents may be created on any computer system and read by any other. It is in this cross platform environment that differences between the basic character sets of computers show through.

Most HTML content does not identify the encoding used in creating the document, nor do most HTTP servers transmit this information when it is available. As a result, browsers need to either guess at the encoding used or use a default value. Even if the encoding is specified and transmitted, browsers

are limited in what they can do to present the document correctly. Encoding translation tables and suitable font resources are necessary for proper interpretation and display of documents.

Unambiguous communication of text is much simpler when some recognized standard is followed within a document. If the character encoding and interpreting scheme is unique and widely accepted, then this scheme can be made available in any computer and any document which uses this scheme can be correctly read and processed.

1.2 ASCII

ASCII is the most common format used for encoding text files in computers and on the Internet. The most common keyboard used is QWERTY keyboard, which most of us are familiar with. The ASCII-QWERTY combination has been the de-facto standard so far. In ASCII a 7-bit number represents each of alphabetic and numeric characters or special characters like punctuation marks. Thus a total of 128 ($= 2^7$) characters are defined. Capital letter A has a code value 65, B has a code value 66, numeric 1 has a code value of 49 and so on.

ASCII encoding scheme lets a computer store a document as a series of numbers and also lets the computer share such documents with other computers that use ASCII system of coding. These stored characters can be processed using any text editor, word processor and other applications including database and web. They can be also be sent through any email system on the Internet.

Since it can define a maximum of only 128 characters, the 7-bit ASCII standard severely limits the repertoire of characters available for multilingual applications. EDCDIC (Extended Binary Coded Interchange Code) is an 8-bit character set, which can represent 256 characters. In some encoding schemes for other languages, the first 128 bits of 8 bit coded characters are used for English characters and the next 128 characters are used for the other language, so that both English and another language can be encoded in the same scheme.

1.3 Unicode

The Unicode Standard is the universal character-encoding standard used for representation of text for computer processing. It is fully compatible with the International Standard ISO/IEC 10646-1; 1993, and contains all the same characters and encoding points as ISO/IEC 10646. Unicode provides a consistent way of encoding multilingual plain text and brings order to a chaotic state of affairs that has made it difficult to exchange text files internationally.

The Unicode Consortium is an organization of major computer corporations, software developers, database vendors, international agencies and various user groups and was formed in 1991. Ministry of Information Technology, Government of India is a full member of Unicode Consortium. Unicode and International Standards Organisation (ISO) joined hands to bring out the international character code standard in 1992, known as Unicode standard Version 1.0. Unicode version 1.0 used somewhat different names for some characters than ISO 10646. In Unicode version 2.0, the names were made

the same as in ISO 10646. Version 3.0 was published in February 2000. Unicode is now a 16-bit form of the larger ISO10646, 32-bit standard.

The design of Unicode is based on the simplicity and consistency of ASCII, but goes far beyond ASCII's limited ability to encode only the Latin alphabet. The Unicode Standard provides the capacity to encode all of the characters used for the written languages of the world. It uses a 16-bit encoding that provides code points for more than 65,000 characters. To keep character coding simple and efficient, the Unicode Standard assigns each character a unique 16-bit value, and does not use complex modes or escape codes.

While 65,000 characters are sufficient for encoding most of the many thousands of characters used in major languages of the world, the Unicode standard and ISO 10646 provide an extension mechanism called UTF-16 that allows for encoding as many as a million more characters, without use of escape codes. This is sufficient for all known character encoding requirements, including full coverage of all historic scripts of the world.

The Unicode Standard defines codes for characters used in the major languages written today. Scripts include Latin, Greek, Cyrillic, Armenian, Hebrew, Arabic, Devanagari, Bengali, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, Malayalam, Thai, Lao, Georgian, Tibetan, Japanese Kana, the complete set of modern Korean Hangul, and a unified set of Chinese/Japanese/Korean (CJK) ideographs. Many more scripts and characters are to be added shortly, including Ethiopic, Canadian Syllabics, Cherokee, additional rare ideographs, Sinhala, Syriac, Burmese, Khmer, and Braille.

In all, the Unicode Standard, Version 3.0 provides codes for 49,194 characters from the world's alphabets, ideograph sets, and symbol collections. These all fit into the first 64K characters, an area of the code space that is called basic multilingual plane, or BMP for short.

Unicode also reserves some code values for private use, which the software and hardware developers can assign internally for their own characters and symbols.

The Unicode standard directly addresses only the encoding and decoding of the text elements, that is, it defines how characters are interpreted. The character identified by a Unicode code value is an abstract entity, such as "MALAYALAM CONSONANT PA ". The mark made on the screen or paper, called a glyph, is the visual representation of the character. Unicode standard does not define glyph images, that is how the character should appear on the screen or paper, which is the responsibility of the hardware or software-rendering engine.

The Unicode Standard defines three encoding forms that allow the same data to be transmitted in a byte, word or double word oriented format (i.e. in 8, 16 or 32-bits per code unit). All three encoding forms encode the same common character repertoire and can be efficiently transformed into one another without loss of data. The Unicode Consortium fully endorses the use of any of these encoding forms as a conformant way of implementing the Unicode Standard.

UTF-8 is popular for HTML and similar protocols. UTF-8 is a way of transforming all Unicode characters into a variable length encoding of bytes. It has the advantages that the Unicode characters corresponding to the familiar ASCII set have the same byte values as ASCII, and that Unicode

characters transformed into UTF-8 can be used with much existing software without extensive software rewrites.

UTF-16 is popular in many environments that need to balance efficient access to characters with economical use of storage. It is reasonably compact and all the heavily used characters fit into a single 16-bit code unit, while all other characters are accessible via pairs of 16-bit code units.

UTF-32 is popular where memory space is no concern, but fixed width, single code unit access to characters is desired. Each Unicode character is encoded in a single 32-bit code unit when using UTF-32.

All three encoding forms need at most 4 bytes (or 32-bits) of data for each character.

Code spaces U+0900 to U+0D7F is allotted to Indian scripts in Unicode. The allotment of code space for Indian scripts in Unicode is as follows:

Language	Code space
Devanagari	U+0900 to U+097F
Bengali	U+0980 to U+09FF
Gurumukh	U+0A00 to U+0A7F
Gujarati	U+0A80 to U+0AFF
Oriya	U+0B00 to U+0B7F
Tamil	U+0B80 to U+0BFF
Telugu	U+0C00 to U+0C7F
Kannada	U+0C80 to U+0CFF
Malayalam	U+0D00 to U+0D7F

Malayalam is allotted 128 character positions, code space from U+0D00 to U+0D7F. The table showing the Malayalam character codes of Unicode is given in section 4.1.

Unicode has been hailed by many in the computing communities as an ideal solution to the problems of multiplatform internationalization. It is destined to replace ASCII and other single and multibyte character sets currently in existence. Majority of the software developers the world over have declared conformance to Unicode. They include IBM, Microsoft, Oracle, Sybase, Unisys, Apple, Bell Labs, Compaq, GNU/Linux, Sun, SCO, Hewlett Packard, Netscape, Ericsson and Novell. More and more applications are becoming Unicode compliant. It is expected that Unicode will become the de facto standard in the multilingual world, especially with the spread of Internet.

1.4 National Initiatives

The demand for computer systems capable of providing input/output facilities in Indian languages started in 1970s. Department of Electronics (now Ministry of Information Technology), Government of India took the initiative in 1983 for standardization of Indian language keyboard and character

encoding. The code thus arrived at was ISCII-83, an 8-bit code, which complied with ISO 8-bit recommendations. ISCII stands for Indian Script Code for Information Interchange.

The fact that all Indian scripts, with the exception of Perso-Arabic scripts, have evolved from the ancient Brahmi script and they have a common phonetic structure, formed the basis of ISCII. A common platform for all Indian script was envisaged and DOE recommended the 8-bit ISCII code in 1986. ISCII was tried out in various implementations and refined in 1986 and again in 1988.

ISCII is defined in such a way that all Indian languages can use a single character encoding scheme. 8 bits used for ISCII coding can represent a maximum of 256 characters. Out of this, Roman characters and punctuation marks as defined in ASCII is allotted the first 128 character slots. The next 128 character positions are allotted for the Indian languages. The standard provides a unique code for the vowel, consonant and modifier characters of Indian languages and leaves the screen rendering process to the hardware or software.

The revision of this standard in 1991 made the code chart more compact. ISCII was accepted as the encoding standard for Indian languages in 1991 by Bureau of Indian Standards (BIS) as IS 13194:1991. ISCII standard was reaffirmed by BIS in 1997. The code table for ISCII is given in Annexure 1.

1.5 Malayalam Keyboard

Malayalam script contains hundreds of characters, including conjuncts. It is widely accepted that as the number of characters become large, it becomes unwieldy and inefficient for typing and printing. After detailed study, discussions and deliberations, characters of Malayalam script were modified in 1968 to make it easier to handle in typewriters. Along with that, Malayalam typewriter keyboard layout was also standardized.

The manufacturers of the typewriter keyboards added their own modifications to this standard. The typewriter keyboard layout was designed for mechanical typewriters and the criteria is slightly different for computer keyboards.

Department of Electronics also defined a keyboard layout called Inscript for all Indian languages in 1986. Inscript keyboard is based on phonetic layout and it matches with ISCII in terms of the characters provided on the keyboard and the codes provided for characters. Bureau of Indian Standards accepted Inscript as National standard along with ISCII.

2. MALAYALAM CHARACTER SET

94

The first step in evolving a keyboard layout as well as character-encoding scheme is to define the character set in Malayalam. The character set shall contain all the smallest useful elements of text to be encoded. The characters are to be pragmatically chosen as appropriate to express text and allow various text processes.

State Institute of Languages, Thiruvananthapuram has conducted a detailed study and recently has arrived at the basic character elements required for Malayalam. The basic character elements have been finalised on the basis of the following works:

- The discussions and debates by linguistic experts in Malayalam
- The workshop organised by State Institute of Languages and Central Institute of Indian Languages (Mysore) from 17th to 27th of August 1998, with the support of experts in Department of Official Languages, Kerala Sahitya Academy, NCERT, Linguistics Department (University of Kerala) and Sarva Vijnana Kosam.
- Seminar organised by Lexicon Department (University of Kerala), at Senate Hall on 19th and 20th of March 1999.
- Combined discussions of the Committees for Malayalam script standardisation, under the Chairmanship of Sri P.Govinda Pillai and Dr.V.R.Prabhodhachandran Nayar.
- Evaluation of the feedbacks received on the recommendations of the Committee.

The decisions evolved after the series of these activities are given in the Style Book (Malayalam Achatyum Ezhuthum) published by the State Institute of Languages. The Committee accepted the basic Malayalam character set as given in the Style Book, except for the consonant signs. Thus, Malayalam keyboard shall have the following characters:

Vowels (13)

അ	ആ	ഇ	ഈ	ഉ	ഊ	ഋ
എ	ഏ	ഐ	ഒ	ഓ	ഔ	

Consonants (36)

ക	ഖ	ഗ	ഘ	ങ
ച	ഛ	ജ	ഝ	ഞ
ട	ഠ	ഡ	ഢ	ണ
ത	ഥ	ദ	ധ	ന
പ	ഫ	ബ	ഭ	മ
യ	ര	ല	വ	
ശ	ഷ	സ	ഹ	
ള	ഴ	റ		

Anuswaram, Visargam, Chandrakkala (3)

o 8 ~

chillu (5)

നീ തീ റീ ശീ ണീ

vowel signs (12)

ഓ ഐ ഐ ഐ ഐ ഐ
 ഐ ഐ ഐ ഐ ഐ ഐ

Thus there are 69 basic characters required in Malayalam keyboard.

The character set consist of the following character types:

- vowels
- consonants
- Anuswaram, Visargam and chandrakkala
- chillu
- vowel signs

Consonants

The consonant letter represent a single consonantal sound, but also an inherent vowel A. In the presence of a dependent vowel, however, the inherent vowel associated with the consonant letter is overridden with the dependent vowel.

Vowels

The independent vowel letters stand on their own. The writing system treats independent vowels as (Consonant and Vowel) syllables, in which the consonant is null. The independent vowels are used to write syllables , which start with a vowel.

Vowel signs

The dependent vowels do not stand on their own, but they are depicted in combination with a consonant or consonant cluster. Explicit appearance of a dependent vowel in a syllable overrides the inherent vowel of a single consonant character. The positioning of the dependent vowel during rendering may be to the left, to the right, or both to the left and right of the consonant or conjunct, as shown below, depending on the vowel sign being attached.

Vowel sign	attached to left/right Consonant/conjunct	of the	example (vowel sign attached to ക)
ഓ	right		കാ
ഐ	right		കി

റ	right	കീ
ഡ	right	കു
ല	right	കു
പ	right	കു
ഒ	left	കെ
ഔ	left	കേ
ഈ	left	കൈ
ഓ	left and right	കൊ
ഔ	left and right	കോ
ൗ	right	കൗ

Anuswaram (ഓ) is the pure consonant form of മ. Anuswaram cannot stand independently, but is always attached to a vowel, consonant or a conjunct.

Ex: അം കം തും കതം

Visargam (ഃ) is the pure consonant form of ഹ. Visargam cannot stand independently, but is always attached to a vowel, consonant, or conjunct.

Ex: അഃ ദുഃ നഃ

Chandrakkala (~) is interpreted depending on its position in a word. If chandrakkala is the last character in a word, it is treated as the central vowel ് with spread lips.

Ex: അത് അവൻ കടവ്

If chandrakkala is not the last character of a word (it is followed by a consonant or conjunct), it is treated as the vowel omission sign. This serves to cancel the inherent vowel of the consonant to which it is applied. It functions as a combining character (to form a conjunct).

Ex: ക് ത = കത
 ഗ് ന = ഗന
 ഷ് ട് റ = ഷ്ട്ര

3. MALAYALAM KEYBOARD LAYOUT

As shown above, there are 73 basic characters that should be provided in Malayalam keyboard. This is after leaving the still larger number of conjuncts and Malayalam numerals. Since the numerals and the punctuations are required in Malayalam also, it is clear that in order to accommodate all the Malayalam characters in the keyboard, there has to be some compromises.

Any keyboard layout, which cannot be mapped to the standard QWERTY keyboard, will not be commercially sustainable. The QWERTY keyboard has provision for 26 keys for English letters (same keys used for lower case (normal position) and upper case (shift position)), 10 keys for numerals and 11 keys for punctuation marks. Thus, altogether there are 47 keys, excluding the control keys.

The optimum Malayalam keyboard layout should be such that

- ✓ It should be capable of accommodating all Malayalam characters being used
- ✓ It should be usable in all computing platforms available
- ✓ The typing efficiency should be high

The two keyboard standards available in Malayalam, which are not proprietary, are the Inscript layout and Typewriter layout.

The typewriter keyboard has been designed for the mechanical medium and is not very efficient for computer keyboards. The keyboard design does not take advantages of the flexibility of computers in rendering the conjunct letters, even if the conjuncts require more than one key-press. Also, the typing sequence of words in the typewriter keyboard is based on the actual appearance of the characters in the word. This is not very efficient, if any processing on the text, like sorting, searching, spell-check, grammar-check etc are to be done on the text.

The Inscript overlay can be used in any QWERTY keyboard. Malayalam script legends are shown on the right hand side of the key and English legends, on the left hand side. English or Malayalam overlay can be selected through the Scroll Lock key (each key press toggles the overlay) or by software selection. The keyboard is based on phonetic layout. Consonants and vowels are given keystrokes and the consonant-vowel combinations are formed by typing the consonant key followed by the vowel key. Similarly, conjuncts are formed by typing the constituent consonants with a chandrakkala typed in between.

Inscript overlay has taken into account the logical structure of script alphabet, derived from the phonetic properties. The vowels have been laid out on the left hand side of the keyboard and the consonants on the right hand side.

The vowels are allotted the shift positions of the corresponding vowel signs. This is due to the fact that the frequency of occurrence of vowel signs is more than the corresponding vowels in Malayalam. The vowel omission sign, chandrakkala, is given in the un-shifted position of A, since A does not have a corresponding vowel sign. Alternate hand action occurs while typing conjuncts, since

chandrakkala, is in the left hand side and most of the consonants are on the right hand side of the overlay (chandrakkala is used as the link to form conjuncts and is to be typed in between the constituent consonants of the conjunct).

On the consonant side, all the primary characters of the five vargs (ക ച ട ത പ) are included in the home row. The aspirated consonants are kept in the shift positions of the corresponding unaspirated counterparts.

In the Inscript scheme, vowel sign is to be typed after the consonant/conjunct to override the inherent A vowel of the consonant/conjunct. In the case of hand-written text, the vowel signs are to be put before the consonants in some cases (for example, in the case of ഐ, ഔ, ഞ signs), and in some other cases part of the vowel signs are to be put before the consonant and part after the consonant. (For example, in the case of ഐ, ഔ). For the processing of the document like sorting, searching, spell check etc. it is better that the vowel signs are typed after the consonants, since these processes are based on the construct of the words/syllables. The syllables are, in turn, built around consonants/conjuncts and modified by vowel signs.

Thus, the options available for Malayalam keyboard standard are:

- ✓ Accept Inscript layout for Malayalam keyboard, recommended by Department of Electronics, Government of India (now Ministry of Information Technology) and accepted by Bureau of Indian Standards.
- ✓ Accept Inscript layout as the base and then work on the modifications required in this specification, so as to make it a more acceptable one.
- ✓ Accept the typewriter keyboard layout
- ✓ Go for a fresh standard, based on the frequency study of occurrence of characters in Malayalam documents and the study of the existing keyboard schemes.
- ✓ Accept more than one keyboard layout standard.

The Committee decided to go for the second option, that is, accept the Inscript keyboard as the standard and address the shortcomings of the layout, for the following reasons:

- It is reported that more than 90% of the word processing packages used in Malayalam uses Inscript keyboard. So, the acceptability of a totally new standard may not be very high and may take time, even if the typing efficiency is more with the new layout. Also, the numerous documents already generated using Inscript will have to be converted to the new standard.
- The frequency study conducted by the Kerala State Institute of Languages has found that the layout of Inscript keyboard has been done scientifically. The keys of most frequent use are allotted strong key positions and those of less frequent use need two key presses (shift + character key).

ir

There are some shortcomings in the Inscript layout, and the Committee recommends the following modifications in Inscript keyboard layout for making it a standard keyboard layout.

1. Inclusion of chillu

In the Inscript layout, chillu characters are not provided on the keyboard, in spite of the fact that they are extensively used in Malayalam. Users have to type three keys to get chillu. It is recommended that five chillu characters be given place on the keyboard. The keys suggested, based on the frequency of occurrence of the chillu and the availability of keys, are as shown below:

൯൦	shift position of v
൯൧	shift position of . (decimal point)
൯൨	\ (backslash)
൯൩	shift position of 8
൯൪	shift position of x

2. It should be possible to type the vowel signs and consonant signs like

၁ ၂ ၃ ၄ ၅ ၆ ၇ ၈ ၉ ၁၀ ၁၁ ၁၂ ၁၃ ၁၄ ၁၅ ၁၆ ၁၇ ၁၈ ၁၉ ၂၀ ၂၁ ၂၂ ၂၃ ၂၄ ၂၅ ၂၆ ၂၇ ၂၈ ၂၉ ၃၀ ၃၁ ၃၂ ၃၃ ၃၄ ၃၅ ၃၆ ၃၇ ၃၈ ၃၉ ၄၀ ၄၁ ၄၂ ၄၃ ၄၄ ၄၅ ၄၆ ၄၇ ၄၈ ၄၉ ၅၀ ၅၁ ၅၂ ၅၃ ၅၄ ၅၅ ၅၆ ၅၇ ၅၈ ၅၉ ၆၀ ၆၁ ၆၂ ၆၃ ၆၄ ၆၅ ၆၆ ၆၇ ၆၈ ၆၉ ၇၀ ၇၁ ၇၂ ၇၃ ၇၄ ၇၅ ၇၆ ၇၇ ၇၈ ၇၉ ၈၀ ၈၁ ၈၂ ၈၃ ၈၄ ၈၅ ၈၆ ၈၇ ၈၈ ၈၉ ၉၀ ၉၁ ၉၂ ၉၃ ၉၄ ၉၅ ၉၆ ၉၇ ၉၈ ၉၉ ၁၀၀

without being attached to consonants and they should also be displayed as such. The word processing software may indicate (on the screen) that there is a spelling mistake in those cases where these signs are typed without attaching to a consonant/ conjuncts.

3. NUKTA is not used in Malayalam and can be deleted from the keyboard overlay.

4. കയ has been given a key in the Inscript layout. Since കയ is not a basic Malayalam character, it may be deleted from the keyboard overlay.

5. ് has been given a key in the Inscript layout. Since ് is not a basic Malayalam character, this may be deleted from the keyboard overlay.

The layout of the recommended Malayalam keyboard layout is given in figure 1.

3.2 Typing sequence

There has to be a unique typing sequence for each of the Malayalam word. For typing a word, it has to be broken into a sequence of separate keystrokes.

a) vowels and consonants

The vowels and consonants are to be typed in the sequence of pronunciation

ex:	<i>word</i>	<i>keystroke sequence</i>
	വടകര	വ ട ക ര
	മല	മ ല
	ആന	ആ ന
	ഇര	ഇ ര
	ഇറ	ഇ റ
	ഉമ	ഉ മ
	ഏല	ഏ ല
	ഓമന	ഓ മ ന
	ഔഷധ	ഔ ഷ ധ

b) vowel signs attached to consonants

The vowel sign to be attached to a consonant is to be typed immediately after the consonant/conjunct.

മാല	മ ാ ല
തിര	ത ി ര
ശീല	ശ ി ല
കൂട	ക ൂ ട
ചൂര	ച ൂ ര
കൃതി	ക ൃ ത ി
ചെവി	ച േ വ ി
വേല	വ േ ല
മൈന	മ ൈ ന
തൊലി	ത ൊ ല ി
കോടി	ക ോ ട ി
സൗമിനി	സ ൌ മ ി ന ി

c) anuswaram

Anuswaram to be attached to a consonant or vowel is to be typed immediately after the consonant/ vowel

പണം	പ ണ ണ
അംശം	അ ണ ശ ണ

If vowel sign is present, anuswaram will follow it (as is the order in which the syllable is pronounced)

വരാം	വ ര ണ
താബൂലം	ത ണ ബൂ ല ണ

d) visargam

Visargam is used to indicate an aspiration sound (h) and is to be typed immediately after the consonant/ vowel to which it is attached. If vowel sign is present, anuswaram will follow it.

ദുഃഖം	ദു ഃ ഖ ണ
-------	----------

e) Chandrakkala

Chandrakkala is typed after a consonant/ conjunct to indicate omission of inherent A from any non-final consonant or the addition of the central vowel with spread lips to the utterance-final consonant.

സ്ഥരണ	സ് ഹ ര ണ
പുഷ്പം	പു ഷ് പ ണ
കടവ്	ക ട വ്

Chandrakkala cannot be attached to vowels, vowel signs, anuswaram and visargam.

ഇ് ക്ക് സ്ക് ക്ക് are not valid syllables.

f) Attaching consonant sign ള

To attach consonant sign ള to consonants/conjuncts, type യ after consonant and chandrakkala, as shown:

സത്യം	സ ത് യ ണ
വ്യത്യാസം	വ് യ ത് യ ണ സ ണ
അന്യോന്യം	അ ന് യ ണ ന് യ ണ

However, the characters which are typed will be stored as such and ധ is used only for the visual representation.

g) Attaching consonant sign റ

To get consonant sign റ attached to consonants/conjuncts, type റ after consonant and chandrakkala, as shown:

തത്വം	ത ത് വ ണ
ശ്വാസം	ശ് വ ണ സ ണ

However, the characters, which are typed, will be stored as such and റ is used only for the visual representation.

h) Attaching consonant sign {

The attachment of ് to consonant/ conjuncts in most of the cases results in appending റ sound to the pure consonant/ conjunct. But in some cases, ് sign denotes the attachment of റ sound to the pure consonant/conjunct. So, the visual representation of the conjunct obtained by attaching ് റ and ് റ will be ് as shown in the following examples:

ചക്രം	ച ക്ക റ റ
കോയം	ക ് റ റാ യ റ
പാവ്	പ ് റ റ വ ്
ആന്യ	ആ ന്ന ് യ റ
ഗ്രാമം	ഗ്ഗ റ റ മ റ
ശ്രദ്ധ	ശ്ശ റ റ ദ്ധ
ധുവം	ധ്ധ റ റ വ റ
രാഷ്ട്രം	ര റ ഷ്ശ ് റ റ

However, the characters which are typed will be stored as such and { is used only for the visual representation.

i) Consonant sign

The attachment of ള to consonant/ conjuncts in most of the cases results in appending ള sound to the pure consonant/ conjunct, as shown above. But in some cases, ള sign denotes the attachment of ല sound to the pure consonant/conjunct. So, the visual representation of the conjunct obtained by attaching ് ള and ് ല will be ് as shown in the following examples:

ക്ലാസ്	ക ് ള റ സ ്
ക്ലോകം	ക്ല ് ല റാ ക റ

However, the characters which are typed will be stored as such and is used only for the visual representation.

3.3 Conjuncts

Malayalam has a large number of consonant conjunct forms, which serve as orthographic abbreviations of two or more adjacent letter forms. A consonant cluster is depicted with a conjunct glyph, if such a glyph is available in the current glyph. In the absence of a conjunct glyph, the conjunct is depicted with the nominal consonant forms with chandrakkala in between.

The split up (and the key sequence) of the conjuncts need to be unambiguously defined for the following reasons:

1. The key combinations and the order in which they are to be typed to get a conjunct letter is a basic need for the user.

The following points may be noted:

For uniqueness in the split up, the first of these are accepted, since the user will find it comfortable to map these split up with the visual representation of the conjuncts.

4. Conjuncts can also be shown with chandrakkala in between the constituent consonants, without formation of the conjunct. This can be achieved by typing chandrakkala twice.

ശക്തി	ശ ക്ക ി
ശക്തി	ശ ക്ക ി
യുദ്ധം	യു ങ്ങ ള്ള
യുദ്ധം	യു ങ്ങ ള്ള

5. Wherever there is no glyph available for the conjuncts, the constituent consonants can be displayed with chandrakkala in between

സ്വപ്നം	സ്വ പ്ന ള്ള
ശുഷ്കം	ശു ഷ്ക ള്ള
കാഴ്ച	കാ ങ്ങ ച

6. Typing sequence of some of the conjuncts with more than two consonants:

സ്റ്റ	സ് റ്റ റ്	(സ്റ്റേറ്റ്)
ത്സ	ത് സ ന്	(ജോത്സ)
ക്ഷ്മ	ക്ഷ മ്	(ലക്ഷ്മണൻ)
യ്ക്ക	യ് ക്ക	(നായ്ക്കൾ)
ന്ത്യ	ന് ത് റ് റ് യ	(സ്വാതന്ത്ര്യം)

3.4 Intra glyph positioning of vowel signs

When the conjunct letter is presented as a single physical entity (the constituent consonants joined together), the attachment of vowel signs to the conjunct letter is similar to attaching vowel signs to consonants. But when the conjunct letter does not have a glyph of its own, it is presented as consonants separated by chandrakkala. In such cases, confusion may arise on the position of the vowel sign with reference to the constituent consonants of the conjunct letter. The simple rule in this regard is that the vowel sign is to be attached with the second consonant of the conjunct, which is in agreement with the pronunciation of the syllable. Some of the examples are shown below, which shows the positioning of the vowel signs as well as the key sequence for the words formed.

സ്നേഹം	സ് ന്നേ ഹ ള്ള
തലശ്ശേരി	ത ല ശ്ശ ശ്ശേ റി
സ്തോത്രം	സ് ത്താ ത്ത റ് ള്ള
സ്ത്രൈണം	സ് ത്ത റ് റ് ഞ ണ ള്ള
വിദ്യോപദേശം	വി ദ്വ റ് റ് യാ പ ദ ശ്ശേ ഷ
പുഷ്പോൽസവം	പു ഷ് പ് പാ ത്ത സ വ ള്ള
രാഷ്ട്രീയം	രാ ഷ് ട് രീ യ ള്ള

FIG. 1

~ ൧ !	@	#	\$	%	^	&	* ശ	()	- ൦	+ ൧	
൧ ൦൦	2	3	4	5	6	7	8	9	0	-	= ൦	BS
TAB	Q ൦൦	W ൦൦	E ൦൦	R ൦൦	T ൦൦	Y ൦൦	U ൦൦	I ൦൦	O ൦൦	P ൦൦	{ ൦൦ } ൦൦	൦൦
Caps	A ൦൦	S ൦൦	D ൦൦	F ൦൦	G ൦൦	H ൦൦	J ൦൦	K ൦൦	L ൦൦	: ൦൦	" ൦൦	
Lock	൦൦	൦൦	൦൦	൦൦	൦൦	൦൦	൦൦	൦൦	൦൦	൦൦	൦൦	ENTER
SHIFT		Z ൦൦	X ൦൦	C ൦൦	V ൦൦	B ൦൦	N ൦൦	M ൦൦	< ൦൦	> ൦൦	? ൦൦	
CNTRL	ALT	SPACE									ALT	CNTRL

4. MALAYALAM CHARACTER ENCODING

The purpose of encoding is to represent the text elements in a language in a unique way so that applications could implement a variety of text processes in the desired language. The different operations performed on text in a computer, including input, rendering (display), searching and sorting, have different preferences for the manner in which the text is encoded. Choosing one encoding scheme may make one operation more efficient, but it may make other operations less efficient. The only possible solution is to have a compromise that can adequately satisfy all text processing needs, even though this may be result in lesser efficiency.

The character encoding scheme for Malayalam should be such that

- ✓ It should accommodate all Malayalam characters being used
- ✓ Malayalam characters should be recognized and represented uniformly in any system
- ✓ It should be usable in all widely used platforms, without any patches

There are two encoding standards for Malayalam in use today - Unicode and ISCII. Unicode is the standard developed by Unicode Consortium in consultation with ISO and caters for all languages in the world. ISCII is the standard developed by Department of Electronics, Government of India and accepted by BIS. ISCII is an 8-bit bilingual coding standard, in which 128 code spaces are available for English and the remaining 128 code spaces being multiplexed by the Indian languages.

The options available for a standard character encoding in Malayalam are:

- Accept ISCII with the necessary modifications
- Accept Unicode with necessary modifications
- Develop a new encoding standard

The Committee recommends Unicode to be the standard for character encoding in Malayalam. The shortcomings in the allotment of character codes for Malayalam can be taken up with Unicode consortium through Ministry of Information Technology, Government of India, which is a full member of the consortium. This recommendation is based on the following reasons:

- It is predicted that majority of the content available on the web and other electronic media is going to be non-English, within a short span of time. This means that content available on the web will be in various languages and it is necessary to identify, read and process correctly the content in a particular language, including Malayalam. No coding scheme available today anywhere in the world has the capability to do so, except Unicode. It is the only standard which caters to the multilingual world. It has provided for unique codes for all the characters of all the major languages used today. Unicode is destined to replace ASCII and other single and multibyte character sets currently in existence
- Most of the major software developers the world over have committed support to Unicode. They include IBM, Microsoft, Oracle, Sybase, Unisys, Apple, Bell Labs, Compaq, GNU/Linux, Sun, SCO, Hewlett Packard, Netscape, Ericsson and Novell. More and more operating

systems, Database Management Systems (DBMS), web and other applications are becoming Unicode compliant. It is expected that Unicode will become the de facto standard in the multilingual world, especially with the spread of Internet.

However, not many word processing/ application packages are available in Malayalam which conforms to Unicode, as on today. So, a period of one year may be granted for the developers of Malayalam software to conform their products to Unicode.

4.1 Modifications required to be carried out in the Unicode representation of Malayalam characters

1. The following chillu characters are to be included in the Malayalam character set:

<i>Suggested code position</i>	<i>Character</i>	<i>Description</i>
0D58	ലർ	Malayalam Letter L
0D59	ൾ	Malayalam Letter LL
0D5A	റർ	Malayalam Letter RR
0D5B	ൻ	Malayalam Letter NN
0D5C	ൻ	Malayalam Letter N

2. The correct glyph for Malayalam vowel sign AU (code position 0D4C) ു. This is wrongly shown as ു, in the Unicode specification. The correct description is

<i>Code position</i>	<i>Character</i>	<i>Description</i>
0D4C	ു	Malayalam vowel sign AU

3. Malayalam AU Length Mark (code position 0D57) is not used in Malayalam. The glyph for this character is wrongly shown as ു, which is actually the Malayalam vowel sign AU (as explained in para 2). So, this code position may be reserved:

<i>Code position</i>	<i>Character</i>	<i>Description</i>
0D57	ു	Reserved

4. The following characters which are now provided in Unicode may be deleted/reserved, since these are no more used in Malayalam:

		<i>Remarks</i>
Malayalam Letter Vocalic L	ൌ	(0D0C)
Malayalam Letter Vocalic LL	ൎ	(0D61)

Malayalam Letter Vocalic RR	൪൪	(0D60)
Malayalam Digit ZERO	൪൪	(0D66)
Malayalam Digit ONE	൪൪	(0D67)
Malayalam Digit TWO	൪൪	(0D68)
Malayalam Digit THREE	൪൪	(0D69)
Malayalam Digit FOUR	൪൪	(0D6A)
Malayalam Digit FIVE	൪൪	(0D6B)
Malayalam Digit SIX	൪൪	(0D6C)
Malayalam Digit SEVEN	൪൪	(0D6D)
Malayalam Digit EIGHT	൪൪	(0D6E)
Malayalam Digit NINE	൪൪	(0D6F)

5. The description of the character (code position 0D34) is to be corrected as MALAYALAM LETTER ZHA. This is wrongly specified as MALAYALAM LETTER LLLA in Unicode standard.

Code position	Character	Description
0D34	൪൪	Malayalam letter ZHA

The Unicode code table for Malayalam is given below. The suggested modifications (to be taken up with Unicode Consortium) are shown in the table.

Code	Character	Description	Remarks
<i>Space</i>			
<i>Various signs</i>			
0D02	൪൪	Malayalam sign Anuswaram	
0D03	൪൪	Malayalam sign Visargam	
<i>Independent vowels</i>			
0D05	൪൪	Malayalam letter A	
0D06	൪൪	Malayalam letter AA	
0D07	൪൪	Malayalam letter I	
0D08	൪൪	Malayalam letter II	
0D09	൪൪	Malayalam letter U	
0D0A	൪൪	Malayalam letter UU	
0D0B	൪൪	Malayalam letter vocalic R	
0D0C	൪൪	Malayalam letter vocalic L	To be deleted/reserved
0D0E	൪൪	Malayalam letter E	
0D0F	൪൪	Malayalam letter EE	
0D10	൪൪	Malayalam letter AI	
0D12	൪൪	Malayalam letter O	
0D13	൪൪	Malayalam letter OO	
0D14	൪൪	Malayalam letter AU	

80

Consonants

0D15	ക	Malayalam letter KA
0D16	ഖ	Malayalam letter KHA
0D17	ഗ	Malayalam letter GA
0D18	ഘ	Malayalam letter GHA
0D19	ങ	Malayalam letter NGA
0D1A	ച	Malayalam letter CA
0D1B	ഛ	Malayalam letter CHA
0D1C	ജ	Malayalam letter JA
0D1D	ഝ	Malayalam letter JHA
0D1E	ഞ	Malayalam letter NYA
0D1F	ട	Malayalam letter TTA
0D20	ഠ	Malayalam letter TTHA
0D21	ഡ	Malayalam letter DDA
0D22	ഢ	Malayalam letter DDHA
0D23	ണ	Malayalam letter NNA
0D24	ത	Malayalam letter TA
0D25	ഥ	Malayalam letter THA
0D26	ദ	Malayalam letter DA
0D27	ധ	Malayalam letter DHA
0D28	ന	Malayalam letter NA
0D2A	പ	Malayalam letter PA
0D2B	ഫ	Malayalam letter PHA
0D2C	ബ	Malayalam letter BA
0D2D	ഭ	Malayalam letter BHA
0D2E	മ	Malayalam letter MA
0D2F	യ	Malayalam letter YA
0D30	ര	Malayalam letter RA
0D31	റ	Malayalam letter RRA
0D32	ല	Malayalam letter LA
0D33	ള	Malayalam letter LLA
0D34	ഴ	Malayalam letter ZHA
0D35	വ	Malayalam letter VA
0D36	ശ	Malayalam letter SHA
0D37	ഷ	Malayalam letter SSA
0D38	സ	Malayalam letter SA
0D39	ഹ	Malayalam letter HA

Specified as LLLA in standard.

To be modified as ZHA

Dependent vowel signs

0D3E	ഏ	Malayalam vowel sign AA
0D3F	ഐ	Malayalam vowel sign I
0D40	ഓ	Malayalam vowel sign II
0D41	ഔ	Malayalam vowel sign U
0D42	ഘൃ	Malayalam vowel sign UU

0D43	ു	Malayalam vowel sign vocalic R
0D46	െ	Malayalam vowel sign E
0D47	ഈ	Malayalam vowel sign EE
0D48	ൈ	Malayalam vowel sign AI
0D4A	ഒ	Malayalam vowel sign O
0D4B	ഓ	Malayalam vowel sign OO
0D4C	ൗ	Malayalam vowel sign AU

Malayalam vowel sign AU is ഊ and not ഐ, as was specified earlier. To be corrected.

Various signs

0D4D	~	Malayalam sign Chandrakkala
------	---	-----------------------------

Specified as Virama in standard. To be corrected as chandrakkala

0D57

Specified as AU length mark.

Not used in Malayalam. To be deleted/reserved.

Chillu

0D58	ല	Malayalam letter L	To be added
0D59	ല	Malayalam letter LL	To be added
0D5A	ര	Malayalam letter RR	To be added
0D5B	ന്ന	Malayalam letter NN	To be added
0D5C	ന	Malayalam letter N	To be added

Generic additions

0D60	റ	Malayalam letter vocalic RR	Not used. To be deleted/reserved
0D61		Malayalam letter vocalic LL	Not used. To be deleted/reserved

Digits

0D66		Malayalam digit ZERO	Not used. To be deleted/reserved
0D67		Malayalam digit ONE	Not used. To be deleted/reserved
0D68		Malayalam digit TWO	Not used. To be deleted/reserved
0D69		Malayalam digit THREE	Not used. To be deleted/reserved
0D6A		Malayalam digit FOUR	Not used. To be deleted/reserved
0D6B		Malayalam digit FIVE	Not used. To be deleted/reserved
0D6C		Malayalam digit SIX	Not used. To be deleted/reserved
0D6D		Malayalam digit SEVEN	Not used. To be deleted/reserved
0D6E		Malayalam digit EIGHT	Not used. To be deleted/reserved
0D6F		Malayalam digit NINE	Not used. To be deleted/reserved

5. LEXICOGRAPHIC ORDERING OF MALAYALAM CHARACTERS

5.1 Searching and Sorting

Two utilities of main major use in any document, database or Internet applications are Searching and Sorting. Any character encoding scheme should address both these utilities.

Searching can be for a character, word or combination of words. Searching becomes complicated when the same word can be written in different ways, as is practiced in Malayalam. Many initiatives have been taken at the Government and other levels to standardize the script in Malayalam. One of the major initiatives taken by the Government of Kerala recently is known as Malayalathanima.

Malayalathanima organized various seminars and workshops to arrive at a suitable standard for writing and printing of Malayalam script. The activities were coordinated by State Institute of Languages and the expertise of Central Institute of Indian Languages, Department of Official Languages (Kerala), Kerala Sahitya Academy, NCERT, Press Academy, Department of Linguistics (University of Kerala) and various other organizations were made use of in the process. The State Institute of Languages came out with a recommendation, *a style book*, which was evolved from the seminars and workshops. These recommendations were discussed and debated in various forums and some of the suggestions were also included in the latest edition of the book.

Sorting is arranging words in the lexicographic ordering of characters. Many a times the linear order based on the codes of the characters cannot be used as the basis for sorting.

Sorting is one of the major utilities of any database applications. Be it the telephone directory, the voter's list, dictionary, encyclopedia or the index of books, people expect the ordering of names and other words to be ordered in a unique way, so that it is easy to find out a name or a word from a list or a dictionary.

In Malayalam, different authors have used different ordering schemes for the lexicographic ordering of words and there has not been a clearly defined standard for the ordering. Thus, it is highly essential that a unique lexicographic ordering scheme be specified at the earliest.

The basic factors which were weighed to arrive at the lexicographic ordering are:

- Some studies have been made in this direction by eminent scholars and their inputs have been made use of.
- The ordering scheme is totally based on the script and the pronunciation of words will not have any role on the scheme.
- The scheme is independent of the coding scheme being used.

0 03 01

လယ် လယ် လယ် ...

လၢၣ် လၢၣ် လၢၣ် ..

ൺ ണ ണാ ണി ണീ ണ് ണു ണൂ ണൃ ണെ ണേ ണൊ ണോ ണൗ
 ണ്ക (ൺക).. ണ്ഖ (ൺഖ).. ണ്ഗ (ൺഗ) .. ണ്ഘ (ൺഘ) .. ണ്ഞ (ൺഞ).. ണ്ച (ൺച)

ത താ തി

ம ம மி

ଜି ୩୦ ଗି

ယ ယာ ယါ

நீ ந நா நி நீ ந் நு நு நெ நே நை நொ நோ நு
 ந்க (நீக்).. ந்வ (நீவ).. ந்ഗ (நீഗ).. ந்ഘ (நீഘ).. ந்ങ (நீങ).. ந்ച (நீച)..
 நீട... நீத (நீத, நீத)... நீப(நீப) நீയ(நீய, நீய)... நீந(நீந)... நீറ (நீറ, நீ)

બાબત ની

ഫ ഫോ ഫീ ...

ബ ബാ ബി ...

ଓ ଓ ଓ ...

o m ma mi mī dī mu mu me me me mo mo mo
m̐k (om).. m̐v (ov).. m̐r (or).. m̐l (ol).. m̐n (on).. m̐x (ox)....

യ യാ യി

ര രാ രി

രേ റാ റീ റ്റ് രൂ രു റെ റേ ഒരെ റൊ റോ റൗ
 റ്റക (രേക).. റ്റഖ (രേഖ).. റ്റഗ (രേഗ).. റ്റഘ (രേഘ).. റ്റങ (രേങ).. റ്റച (രേച)
 റ്റത (രേത)... റ്റപ (രേപ) രൂ (രേയ).. റ്റര (രേര).. റ്റ (രേന്).. റ്റല (രേല)

റ്ത (ർത)... റ്പ (ർപ) റ്യ (ർയ).. റർ (ർര).. റ്റ (റ്റ്).. റ്ല (ർല)

തേ ല ലാ ലി ലീ ല് ലു ലൂ ലെ ലേ ലൈ ലൊ ലോ ലൗ ലൗ
ല്ക (തേക).. ല്ഖ (തേഖ).. ല്ഗ (തേഗ).. ല്ഘ (തേഘ).. ല്ങ (തേങ).. ല്ച (തേച)

வ வா வி ...

ശ ശാ ശി ...

നഷ്ടം നഷ്ടം നഷ്ടം ...

ကံ ကံကံ ကံကံ ...

8 ഹെ ഹെ ഹെ

ശീ ഈ ഇം ഉീ ു് ലു ൂ ഒെ ൈെ ഒൊ ഒൊ ൊ ോ
 ്ക (ശീക).. ്വ (ശീവ).. ്ഗ (ശീഗ).. ്ഘ (ശീഘ).. ്ങ (ശീങ).. ്ച (ശീച)
 ഴ ഴാ ഴി ...

୫ ୬ ୭ ...

The following points may be noted in deciding the lexicographic ordering of Malayalam words:

75
1. The chillu characters are treated as the pure consonants of the corresponding consonants. (ൻ - ന; ഞ - ല; റ് - റ; ശ് - ള; ണ് - ണ)

2. Anuswaram (ണ) is treated as the pure consonant of മ.

3. Visargam (ള) is treated as the pure consonant of ഹ.

4. Space, punctuation marks takes precedence over (shall come before) any other characters. Pairs of words arranged in the sort order are given below.

അംബിക ടി	- അംബികകുമാരി
കവി വള്ളത്തോൾ	- കവിത
കേശവമേനോൻ	- കേശവമേനോൻ കെ.വി

5. Conjunct letter shall be treated as a single entity for deciding the lexicographic order.

Ex: അധോലോകം - അധ്യാപകൻ

Here the comparison is between ധോ and ധ്യാ (ധ്യാ).

5. Visargam (ള) is treated as the pure consonant of ഹ and is treated as a separate character.

Ex: അധഃപതനം - അധിപൻ

Here, the comparison is between ധ and ധി.

6. While comparing the conjunct letters for deciding the order, the character components of the conjunct letter shall be considered in the order of formation of the conjunct letter.

Ex: വക്ത്രം - വക്രം

Here, the comparison is between ക്രത and ക്ര. ക്രത is formed by characters ക്ക് ത് ണ in that order and ക്ര is formed by characters ക്ക് ണ in that order. Since ത takes precedence over ണ, ക്രത comes before ക്ര.

മന്ത്രശക്തി - മന്നത്തു

Here the comparison is between ന്ത്ര and ന്ന. ന്ത്ര is formed by characters ന് ത് ണ in that order and ന്ന is formed by characters ന് ന് in that order. Since ത takes precedence over ന്, ന്ത്ര comes before ന്ന.

7. The lexicographic order is decided exclusively based on the script. The pronunciation is not considered.

Ex: കിങ്സ്റ്റി - കിംഗ് (even though ണ is pronounced as ങ്)
അങ്കിൾ - അംകിസ്റ്റ്

74
8. Only the chandrakkala at the end of a word will be treated as സംവൃത ഉകാരം. Chandrakkala coming within a word (followed by other character(s) of the word) denotes a conjunct letter formed by the character(s) preceding and following the chandrakkala..

Ex: അക്കാദമി - അക്കാദി - അക്കാദ് (chandrakkala treated as സംവൃത ഉകാരം.)
 മായവദേവൻ - മായവദേവ് (chandrakkala treated as സംവൃത ഉകാരം.)
 ക്ളാസിൽ - ക്ളാസ് (chandrakkala treated as സംവൃത ഉകാരം.)
 അലക്സാണ്ടർ - അല്പം (ല്പ is a conjunct letter formed by ല് and പ)

7. Chillaksharas and Anuswaram will precede the corresponding consonants with chandrakkala. That is, ണ will come before ന്, ണ before ല്, ണ before റ്, ണ before ള്, ണ before ണ് and ണ before മ്.

Ex: അൻവർ - അന്വയം (അന്വയം)
 കൽക്കി - കല്ക്കി
 അവൾക്ക് - അവള്ക്ക്
 അക്രം - അക്രമ്

8. Chillu and Anuswaram will take precedence over the corresponding consonants. That is, ണ will come before ന്, ണ come before ല്, ണ before റ്, ണ before ള്, ണ before ണ് and ണ before മ്.

Ex: മാൻ - മാനഹാനി
 പാൽ - പാല
 വാൾ - വാളയാർ
 കാർ - കാറ
 ഫോൺ - ഫോണിച്ച്
 നിറം - നിറമാണ്

9. Chillu and anuswaram appearing within a word (between other characters) is treated combined with the following character, as a conjunct letter.

Ex: അനാദരം - അൻപെഴും (ൻപെ is treated as conjunct letter ന്പെ)
 പാൽ - പാല - പാൽക്കാരൻ - പാലം - പാലമരം - പാലാ
 (ൽക്കാ is treated as conjunct letter ല്ക്കാ, anuswaram is treated as മ്)
 കാറിന്റെ - കാർമേലം (ർമേ is treated as conjunct letter റ്മേ)
 ആളപായം - ആൾക്കൂട്ടം (ൾക്കൂ is treated as conjunct letter ല്ക്കൂ)
 കണക്ക് - കണമണി (ണമ is treated as conjunct letter ണമ)
 കമല - കംസൻ (ംസ is treated as conjunct letter മ്സ)

The examples shown below illustrate the guidelines for arranging the words in lexicographic order.

കുമാരൻ - കുറുൻ (ുര treated as the conjunct മ്ര)
 ചകോദരം - ചക്രപാണി
 വർധമാൻ - വർധമാനൻ
 (ൻ takes precedence over ന്, where ണ is the last character of a word.)
 രംഗം - രംഗമാല (ം takes precedence over മ, where ണ is the last character of a word)
 അനുക്രമണിക - അനുക്രമം (ണ takes precedence over ണ)

അനുനാസിക - അനുനാസം

അന്തരം - അൻവർ

(ൻവ treated as conjunct letter ന് വ)

അനുമാർ - അമ്പയം

(ന്നു is treated as conjunct letter ന്നു, which takes precedence over ന് വ)

അതിനുപുറമെ - അതിന്റെയും

(ന്നു treated as conjunct letter ന്നു)

ആൽ- ആല- ആല് - ആൽമരം

(അ treated as conjunct letter ല് മ)

അനന്തൻ - അനന്തനായണൻ

- അനന്തൻപിള്ള (ൻപ treated as conjunct letter ന് പ)

അജിത് - അജിത - അജിതേഷ്

അക്രീച്ചിയെ - അത്തർസിംഗ് - അത്തസാലിനി

(ക്രീ takes precedence to ത്ത, and രീ takes precedence to സാ)

അദമ്യൻ - അദംസ - അദാത്തുൽ

(മ് takes precedence to മ്സ and ട comes before ടാ)

പദ്മനാഭൻ - പദ്മം (ം is treated as pure consonant of മ)

6. OTHER RECOMMENDATIONS

6.1 Visual representation of characters

Malayalam characters can combine or change shape, depending on their context. A character's appearance may be affected by its ordering with respect to other characters, the font used to render the character, and the application and system environment.

Unicode, the coding scheme we have recommended with modifications, do not specify the visual representations of characters. It completely de-links the codes from the displayed fonts. The same text can be seen in different font styles by using a different font composition routine. A word can be displayed in a variety of styles depending on the conjunct repertoire used.

These standards define how the characters are interpreted, not how glyphs are rendered. The software or hardware rendering engine of a computer is responsible for the appearance of a character on the screen. The standards do not specify the shape, size or orientation of the characters on the screen or on the paper.

6.2 Malayalam Research Centre

Malayalam. Documents of the future will be generated and distributed through Internet and computer networks. We need to have a lot of contents in Malayalam to be generated and distributed through Internet and other electronic media, which will be useful to the common man. We also need to have the tools and technologies available in Malayalam, which will aid the generation and distribution of documents as well as other applications in Malayalam. The Committee recommends to the Government to set up a Malayalam Research Centre with the following broad objectives:

- To be a repository of Malayalam language tools
- To develop tools and technologies in Malayalam
- To exploit the emerging technologies for use in Malayalam computing
- Assist Government in technology dissemination

The Research Centre can either be a separate entity, or it can be instituted as a part of IT Department or its activities can be distributed among the various bodies like the Universities, R&D organizations, Department of Official languages and State Institute of Languages.

The Centre can take up activities like

- Development OS support and liaisoning with OS developers
- Development of database support and liaisoning with DBMS developers
- Develop resources like
 - Corpora and lexical resources
 - Malayalam fonts, to be distributed freely
 - Malayalam based AI tools
 - Electronic dictionary, Thesaurus, encyclopedia

- Malayalam language website
- Development of tools for content generation
 - Malayalam learning tools
 - Curriculum based educational software and games in Malayalam
 - Web-enabled and multimedia content generation
 - Development/ enhancement of Multimedia authoring tools in Malayalam
 - Malayalam searching and sorting engine
 - Spell-checker and grammar-checker
- Development of web-centric applications
 - Malayalam browser
 - Malayalam email
 - Malayalam search engine
 - Web publishing tools in Malayalam
- Malayalam Interface tools
 - OCR
 - Text to Speech synthesis system
 - Speech to text conversion system
 - Translation support systems, including Brail
 - Malayalam interfaces including voice

6.3 Official Malayalam Dictionary

The Committee recommends to the Government to publish an official Malayalam dictionary. The dictionary shall be prepared in such a way as to be used as a reference for the linguists, the software developers and the common man alike. The dictionary shall aid the standardising process of the script of Malayalam words and words will be arranged in the order mentioned in this recommendation. This will help in the software development activities like electronic dictionary, thesaurus, sorting engine, search engine, spell check etc.

Flexibility in the way of writing the scripts of Malayalam words will be one of the major stumbling blocks in the development of tools like search engine, spell check, thesaurus etc. Efforts are required to standardize the scripts of Malayalam words. Use of Style book (*Malayalam Achatiyum Ezhuthum - oru style pusthakam*) brought out by State Institute of Languages to be encouraged.

6.4 Government initiatives required for implementation of the standard

If the standard recommended in this document is to be widely used, Government may implement the following plan of action.

- Official declaration of the standard to be published
- A committee to be formed to evaluate any feedback and amend the standard, if required
- Approach BIS to amend the keyboard and character encoding standards in Malayalam
- Approach Unicode consortium through MIT, to modify Unicode specifications for Malayalam
- CDAC to be approached to make amendments in the applications developed by them

- Take initiative for the development of at least one Malayalam font, which will work in all major platforms, especially Windows, Unix, Linux and Mac systems and distribute it freely on Internet.
- All software developers in Malayalam to be intimated about the new standard.
- Make it mandatory for all the Malayalam software to conform to the standard within six months
- Any training in Malayalam software (DTP) to be recognized only if the package conforms to the standard.
- Government to procure only those Malayalam software packages, which conforms to the standard

Reference

1. Indian Standard - Indian Script Code for Information Interchange (ISCII), Bureau of Indian Standards, Delhi, 1991
2. Keyboard Manual, Centre for Development of Advanced Computing (CDAC), Pune, 1999
3. Malayalam Achatyumu Ezhuthum - oru style pusthakam, Kerala Bhasha Institute, 1999
4. Unicode Standard for Multilingual computing: <http://www.unicode.org>
5. Unicode Phase1-The Review:<http://www.lib.berkeley.edu/SSEAL/SouthAsia/Review.html>

Annexure 1

ISCII code table

<i>Position</i>		<i>Character</i>	<i>Name</i>	<i>Remarks</i>
<i>Hex</i>	<i>Decimal</i>			
A1	161			Reserved
A2	162	ഠ	Anuswara	
A3	163	ഡ	Visarga	
A4	164	അ	Vowel A	
A5	165	ആ	Vowel AA	
A6	166	ഇ	Vowel I	
A7	167	ഈ	Vowel II	
A8	168	ഉ	Vowel U	
A9	169	ഊ	Vowel UU	
AA	170	ഋ	Vowel R	
AB	171	എ	Vowel E	
AC	172	ഐ	Vowel EE	
AD	173	ഐ	Vowel AI	
AE	174			Reserved
AF	175	ഒ	Vowel O	
B0	176	ഓ	Vowel OO	
B1	177	ഔ	Vowel AU	
B2	178			Reserved
B3	179	ക	Consonant KA	
B4	180	ഖ	Consonant KHA	
B5	181	ഗ	Consonant GA	
B6	182	ഘ	Consonant GHA	
B7	183	ങ	Consonant NGA	
B8	184	ച	Consonant CHA	
B9	185	ഛ	Consonant CHHA	
BA	186	ജ	Consonant JA	
BB	187	ഝ	Consonant JHA	
BC	188	ഞ	Consonant NHA	
BD	189	ട	Consonant TA	
BE	190	ഠ	Consonant TTA	
BF	191	ഡ	Consonant hard DA	
C0	192	ഢ	Consonant hard DDA	
C1	193	ണ	Consonant hard NA	
C2	194	ത	Consonant THA	
C3	195	ഥ	Consonant TTHA	
C4	196	ദ	Consonant soft DA	
C5	197	ധ	Consonant soft DDA	
C6	198	ന	Consonant soft NA	
C7	199			Reserved

C8	200	പ	Consonant PA	
C9	201	ഫ	Consonant PHA	
CA	202	ബ	Consonant BA	
CB	203	ഭ	Consonant BHA	
CC	204	മ	Consonant MA	
CD	205	യ	Consonant YA	
CE	206			Reserved
CF	207	ര	Consonant RA	
D0	208	റ	Consonant RRA	
D1	209	ല	Consonant LA	
D2	210	ള	Consonant hard LA	
D3	211	ഴ	Consonant ZHA	
D4	212	വ	Consonant VA	
D5	213	ശ	Consonant soft SHA	
D6	214	ഷ	Consonant SHA	
D7	215	സ	Consonant SA	
D8	216	ഹ	Consonant HA	
D9	217	INV	Inverse key	See note 2
DA	218	ഓ	Vowel sign AA	
DB	219	ഐ	Vowel sign I	
DC	220	ഓ	Vowel sign II	
DD	221	ഓ	Vowel sign U	
DE	222	ഓ	Vowel sign UU	
DF	223	ഓ	Vowel sign RR	
E0	224	ഐ	Vowel sign E	
E1	225	ഐ	Vowel sign EE	
E2	226	ഐ	Vowel sign AI	
E3	227			Reserved
E4	228	ഐ	Vowel sign O	
E5	229	ഐ	Vowel sign OO	
E6	230	ഐ	Vowel sign AU	
E7	231			Reserved
E8	232	ച	Chandrakkala	
E9	233			Reserved
EA	234			Reserved
EF	239	ATR	Attribute code	See Note1
F0	240	EXT	Extension Code	See Note3
F1	241			Reserved
F2	242			Reserved
F3	243			Reserved
F4	244			Reserved
F5	245			Reserved
F6	246			Reserved
F7	247			Reserved
F8	248			Reserved
F9	249			Reserved

FA	250
FC	252
FD	253
FE	254

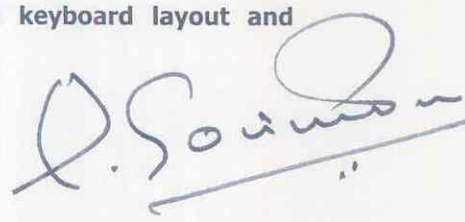
Reserved
Reserved
Reserved
Reserved

Note1 : Attribute code (ATR) defines a font attribute applicable to the following characters. The mechanism is meant for use in the medium where alternative font selection mechanism is not available. Details are given in Annexure E of the BIS document on ISCII .

Note 2: INVERSE key (code: D9) is used to modify visual representations of glyph sets from the normal display. Refer BIS document on ISCII and CDAC keyboard manual for details.

Note3: EXT character provides for an 8-bit code extension mechanism to encode more than 256 characters. Refer IS 13194:1991 Section 6.7 and Annexure G.

Members of the Committee for standardization of Malayalam keyboard layout and character encoding.

- | | | |
|--|----------|---|
| 1. Shri P.Govinda Pillai
Chairman, CDIT
Thiruvananthapuram | Chairman |  |
| 2. Smt Aruna Sundararajan
Secretary, Dept of IT
Thiruvananthapuram | Convener | |
| 3. Dr.S.P.Mudur
Associate Director, National Centre for Software Technology
Gulmohar Cross Road No 9
Juhu , Mumbai – 400 049 | Member | |
| 4. Dr.M.R.Thampan
Director, State Institute of Languages
Nalanda, Thiruvananthapuram | Member | |
| 5. Dr A.P.Andrews Kutty
Head of the Department, Department of Linguistics
University Of Kerala,
Thiruvananthapuram | Member | |
| 6. Shri Satish Babu
Chairman, Computer Society of India
Thiruvananthapuram Chapter | Member | |
| 7. Ms Kala Sriram
Team Coordinator, GIST
CDAC, Ganesh Khind
Pune – 411 007 | Member | |
| 8. Dr. Ezhumattoor Rajaraja Varma
Language expert
Personnel and Administrative Reforms
(Official Languages) Dept.
Govt of Kerala, Thiruvananthapuram | Member | |
| 9. Shri Sasi P.M
Mission Coordinator
IT Mission Group
Thiruvananthapuram | Member | |

Members co-opted to the Committee

1. Shri L. Natarajan
Special Officer (Official Languages) and Ex-Officio Secretary
Personnel and Administrative Reforms
(Official Languages) Dept.
Govt. Secretariat Annexe
Thiruvananthapuram
2. Dr. Prabhodhachandran Nayar
Former Head of the Dept (Linguistics)
University of Kerala
3. Shri R.Raveendra Kumar
Additional Director
ER&DCI, Thiruvananthapuram
4. Shri P.V. Unnikrishnan
Executive Director
Information Kerala Mission Group on IT
Thiruvananthapuram
5. Prof. K.S.Narayana Pillai
Language Expert