L2/03-068

UNICODE for Kannada
(U+0C80 to U+0CFF)

General Information & Description

Written by:

C V Srinatha Sastry

Issued by:

Director

Directorate of Information Technology

Government of Karnataka

Multi Storied Buildings, Vidhana Veedhi

BANGALORE 560 001

INDIA

UNICODE for Kannada

Introduction

The Kannada script is a South Indian script. It is used to write Kannada language of Karnataka State in India. This is also used in many parts of Tamil Nadu, Kerala, Andhra Pradesh and Maharashtra States of India. In addition, the Kannada script is also used to write Tulu, Konkani and Kodava languages. Kannada along with other Indian language scripts shares a large number of structural features. The Kannada block of Unicode Standard **(0C80 to 0CFF)** is based on ISCII-1988 (Indian Standard Code for Information Interchange). The Unicode Standard (version 3) encodes Kannada characters in the same relative positions as those coded in the ISCII-1988 standard.

The Writing system that employs Kannada script constitutes a cross between syllabic writing systems and phonemic writing systems (alphabets). The effective unit of writing Kannada is the orthographic syllable consisting of a consonant (Vyanjana) and vowel (Vowel) (CV) core and optionally, one or more preceding consonants, with a canonical structure of ((C)C)CV. The orthographic syllable need not correspond exactly with a phonological syllable, especially when a consonant cluster is involved, but the writing system is built on phonological principles and tends to correspond quite closely to pronunciation.

The orthographic syllable is built up of alphabetic pieces, the actual letters of Kannada script. These consist of distinct character types: Consonant letters and Independent vowels, the corresponding dependent vowel signs. In a text sequence, these characters are stored in logical phonetic order.

Rendering Kannada Characters

Kannada characters can combine or change shape depending on their context. A character's appearance is affected by its ordering with respect to other characters and the application or system environment. This variation can cause the appearance of Kannada characters to be different from nominal glyphs.

Vowels (*Swaras*)

Independent vowel letters

The independent vowels called *Swaras* in Kannada are letters that stand on their own. The writing system treats independent vowels as orthographic CV syllables in which the consonant is null. The Unicode character encoding for Kannada uses a distinct set of naming conventions for some mid vowels of the fourteen vowels in Kannada. Of these fourteen vowels, twelve vowels have been divided into six sets, each set consisting of a *Hrasva Swara* (short vowel) followed by a corresponding *Deerga Swara* (long vowel*)*. These are two types of *Swaras* depending on the time used to pronounce them.

Hrasva Swara is a freely existing independent vowel which can be pronounced in a single *matra* time (matra kala) whereas a ***Deergha Swara*** is the vowel which can be pronounced in two *matra*.time. The six sets of the swaras are :

ಅ, ಆ (0C85 , 0C86)

ಇ, ಈ (0C87 , 0C88)

ಉ, ಊ (0C89 , 0C8A)

ಋ, ೠ (0C8B , 0CE0)

ಎ, ಏ (0C8E , 0C8F)

ಒ, ಓ (0C92 , 0C93)

Of these, the vowel **ೠ**(0CE0) is not in present use.

The two Deergha swaras ಐ(0C90) and ಔ(0C94) have no Hrasva swara counterparts.

Further, the so-called swaras with code values 0C8C and 0CE1 are not used in Kannada and are not required for Kannada.

Dependent vowel signs (*Matras*)

The dependent vowel signs serve as the common manner of writing non-inherent vowels and are generally referred to as *Swara Chinhas* in Kannada or *Matras* in Samskrit. The dependent vowel signs do not appear stand-alone; rather, they are visibly depicted in combination with a base-letter form (generally a consonant). A single consonant or a consonant cluster may have a dependent vowel sign applied to it to indicate the vowel quality of the syllable, when it is different from the inherent vowel. Explicit appearance of a dependent vowel sign in a syllable overrides the inherent vowel ಅ(0C85) of a single consonant letter.

There are several variations with which the dependent vowels are applied to the base letterforms. Most of them appear as non-spacing dependent vowel signs when applied to base letterforms; above or to the right side of a consonant letter or a consonant cluster. The following are the exceptions and variations for the above rule:

A. The two dependent vowel signs ೃ(0CC3) & ೄ(0CC4) appear one level below and to the right of the consonant or the consonant cluster, separated by a small white space.

B. Each of the five dependent vowel signs ೆ (0CC0), ೇ(0CC7), ೈ(0CC8), ೊ(0CCA) & ೋ (0CCB) are depicted by two or three glyph components (two part or three part vowel signs) with one component appearing with a space to the right of the consonant or the consonant cluster.

   i) In the case of three of the above-mentioned two/three-part dependent vowels ೆ (0CC0), ೇ (0CC7) and ೋ(0CCB), the non-spacing component(s) of each of them is(are) the same as the vowel sign(s) of the corresponding preceding short vowels. The spacing component for each of these dependent vowels is the same length mark ೕ (0CD5) given in Unicode version 3. The logic for this is that these dependent vowels are nothing but the long forms (independent and phonetically distinct) of the preceding short vowels.

   ii) The first component of the dependent vowel ೈ(0CC8) mentioned above is the same as the dependent vowel ೆ(0CC6) and the second component is same as ೖ(0CD6). These are defined independently in Unicode version 3. The second part appears slightly below and to the right of the consonant or the consonant clusters.

C. In view of this, it is important to note that the two glyphs (the length mark ೕ and the second component of ೈ i.e. ೖ) represented with the codes at 0CD5 and 0CD6 in Unicode version 3 have no independent existence and do not play any part as independent codes in the collation algorithm.

D. Unlike Devanagari, the Kannada script does not have any character with a left-side dependent vowel sign.

E. A one-to-one correspondence exists between independent vowels and dependent vowel signs.

Consonant letters (*Vyanjanas*)

Each of the 36 consonant letters in Kannada (enumerated with codes 0C95 to 0CB9 and 0CDE) represents a single consonantal sound but also has the peculiarity of having an inherent vowel, generally the short vowel ಅ **(**/a/ 0C85).

Thus the Kannada letter at 0C95 represents not just ಕ್ (K) but ಕ (KA) with the inherent vowel ಅ(0C85). In the presence of a dependent vowel, however, this inherent vowel associated with

a consonant letter is overridden by the dependent vowel. The consonants ಱ(0CB1) and ೞ(0CDE) sound similar to ರ(0CB0) and ಳ(0CB3) respectively. These two appear in ancient Kannada texts but are not in present use. With this, consonants in modern Kannada are 34 in number (without ಱ and ೞ). These are classified as *Vargeeya Vyanjanas* (0C95 to 0CAE) and *Avargeeya Vyanjanas* (0CAF, 0CB0, 0CB2, 0CB3 and 0CB5 to 0CB9).

Vargeeya Vyanjanas: The five sets of ***Vargeeya Vyanjanas*** are (arranged row wise in the acceptable sorting order):

ಕ **0C95**	ಖ **0C96**	ಗ **0C97**	ಘ **0C98**	ಙ **0C99**
ಚ **0C9A**	ಛ **0C9B**	ಜ **0C9C**	ಝ **0C9D**	ಞ **0C9E**
ಟ **0C9F**	ಠ **0CA0**	ಡ **0CA1**	ಢ **0CA2**	ಣ **0CA3**
ತ **0CA4**	ಥ **0CA5**	ದ **0CA6**	ಧ **0CA7**	ನ **0CA8**
ಪ **0CAA**	ಫ **0CAB**	ಬ **0CAC**	ಭ **0CAD**	ಮ **0CAE**

Avargeeya Vyanjanas: The nine Avargeeya Vyanjanas (enumerated in the acceptable sorting order) are:

ಯ 0CAF	ರ 0CB0	ಲ 0CB2	ವ 0CB5	ಶ 0CB6	ಷ 0CB7	ಸ 0CB8	ಹ 0CB9	ಳ 0CB3

Halant

Like Devanagari, Kannada script also employs a sign known as *halant* or vowel omission sign. A halant sign (್, 0CCD) nominally serves to cancel (or kill) the inherent vowel of the consonant to which it is applied.

The *halant* functions as a combining character. When a consonant has lost its inherent vowel by the application of *halant*, it is known as a dead consonant. The dead consonants are the presentation forms used to depict the consonants without an inherent vowel. Their rendered forms in Kannada resemble the full consonant with the vertical stem replaced by the *halant* sign, which marks a character core. The stem glyph (್ at 0CBB) is graphically and historically related to the sign denoting the inherent /a/ (ಅ) vowel (0C85). In contrast, a live consonant is a consonant that retains its inherent vowel or is written with an explicit dependent vowel sign. The dead consonant is defined as a sequence consisting of a consonant letter followed by a *halant.* The default rendering for a dead consonant is to position the *halant* as a combining mark bound to the consonant letter form.

Avagraha (*s*)

A spacing mark *s*, called avagraha sign is used while rendering Samskrit tests. This is located at OCBD.

Encoding order

The traditional Kannada alphabetic encoding order for consonants follows articulatory phonetic principles, starting with velar consonants and moving forward to bilabial consonants, followed by liquids and then fricatives. ISCII (Indian Script Code for Information Interchange) & the Unicode standard both observe this traditional order.

Consonant conjuncts (Samyuktaksharas)

Like any other Indian script, Kannada is also noted for a large number of consonant conjunct forms that serve as orthographic abbreviations (ligatures) of two or more adjacent forms. This abbreviation takes place only in the context of a consonant cluster. An orthographic consonant cluster is defined as a sequence of characters that represent one or more dead consonants (denoted by C_d) followed by a normal live consonant (denoted by C_l).

Corresponding to each Kannada consonant, there exists a separate and unique glyph, which is specially used to represent the corresponding consonant in a consonant cluster. Most of these conjunct consonant glyphs resemble their original consonant forms (many without the implicit vowel sign, wherever applicable).

In Kannada, there is only one type of conjunct formation (consonant cluster) and it is depicted as follows:

A. The first consonant of the consonant cluster is rendered with the implicit or a different dependant vowel appearing as the terminal element of the consonant cluster.
B. The remaining consonants (consonants in between the first consonant and the terminal vowel element) appear in conjunct consonant glyph forms in the phonetic order. They are generally depicted directly below or sometimes below but to the right of the first consonant.

Thus, the systematically designed Kannada script font contains the conjunct glyph components, but they are not encoded as Unicode characters, because they are the resultant of ligation of distinct letters. Kannada script rendering software must be able to map appropriate combinations of characters in context to the appropriate conjunct glyphs in fonts.

Invisible consonant INV

There is a need to have a consonant, which provides an invisible base for the display of dependant vowels without any consonant base. This can be the Unicode Standard Zero Width Non-Joiner at 200C. This can also be used to provide proper collation of the words containing dead consonants.

Explicit *Halant*

Normally, a halant character serves to create dead consonants, which, in turn, combine with subsequent consonants in order to form conjuncts. This behaviour usually results in a *halant* sign not being depicted visually. Occasionally, however, this default behaviour is not desired when a dead consonant is need to be excluded from conjunct formation with the next consonant, in which case the *halant* sign is visibly rendered.

In order to accomplish this, the Unicode Standard character 200C (Zero Width Non-Joiner) is introduced immediately after the encoded dead consonant that is to be excluded from conjunct formation.

For example, the use of Zero Width Non-Joiner prevents the default formation of the conjunct form ರ್ಗ, resulting in ರ್‌ಗ.

The Kannada script adopts the convention of depicting the character (in this case the halant sign) as appropriate for the consonant to which it is attached.

In summary, each Kannada consonant may be encoded such that it denotes a live consonant, a dead consonant or a conjunct consonant glyph.

Memory Representations and Rendering Order

Notation

In the next set of rules, the following notation applies:

C_n Nominal glyph form of a consonant **C** as it appears in the code charts.

C_l A live consonant, depicted identically to **C_n.**

C_d Glyph depicting the dead consonant form of a consonant **C.**

C_h Glyph depicting the conjuct consonant glyph form of a consonant **C,** which appears as the second, third,….part of a conjucnct consonant.

L_n Nominal glyph form of a conjunct ligature consisting of two or more component consonants. A conjunct ligature composed of two consonants **X** and **Y** is also denoted by **$X.Y_n$.**

RA_{sub} A non-spacing combining mark glyph form positioned below the base glyph form.

V_{vs} Glyph depicting the dependent vowel sign form of a vowel **V**.

$HALANT_n$ The nominal glyph form non-spacing combining mark depicting 0CCD Kannada sign Halant.

A halant character is not always depicted; when it is depicted, it adopts this non-spacing mark form.

Memory representation and rendering order

The order for storage of plain text in Kannada always follows the phonetic order, that is, a CV syllable with a dependant vowel is always encoded as a consonant letter C followed by a vowel sign V in the memory representation. This order is employed by the ISCII standard and corresponds with phonetic and keying order of textual data. Unlike Devanagari and some other Indian Scripts, all the dependent vowels in Kannada are depicted to the right of their consonant letters. Hence there is no need to reorder the elements in mapping from the logical (character) store to the presentation (glyph) rendering and vice versa.

Rule R1 : Whenever a consonant is followed by a vowel, then the corresponding vowel sign attaches to the consonant suitably.

Character order	Glyph order
$KA_n + U$	® $KA_n + U_{vs}$
ಕ + ಊ	® ಕೂ

Further, Kannada script does not allow half-consonants, ligatures and half ligature forms.
The following provides more formal and complete rules for minimal rendering of Kannada as part of a plain text sequence. It describes the mapping between Unicode characters and the glyphs in a Kannada font. It also describes the combining and ordering of those glyphs.

The rules provide minimal requirements for legibly rendering Kannada text. As with any script, a more complex procedure can add rendering characteristics, depending on the font and application.

Dead Consonant Rule

The following rule logically precedes the application of any other rule to form a dead consonant. Once formed, a dead consonant may be subject to other rules described next.

Rule R2: When a consonant C_n precedes a **HALANT$_n$**, it is considered to be a dead consonant C_d. A consonant C_n that does not precede **HALANT$_n$** is considered to be live consonant C_l.

KA_n + **HALANT$_n$** ® KA_d

ಕ + ್ ® ಕ್

Consonant cluster (conjunct) rendering

As already explained in section 8, the conjunct formation (consonant cluster) with two or more consonants and a terminal vowel is as follows:

A. The first consonant of the consonant cluster is rendered with the terminal vowel.

B. The remaining consonants (in between the first consonant and the terminal vowel) are rendered in conjunt consonant glyph forms in the phonetic order.

Rule R3: Example1:

KA_d + KA_n ® KA_h

ಕ್ + ಕ ® ಕ್ಕ (conjunct consonant glyph of ಕ is ್ಕ)

Example 2:

SA_d + TH_d +RA_d + I_{vs} ® SAI_{vs} + TH_h + RA_h

ಸ್ + ತ್ + ರ್ + ೀ ® ಸ್ತ್ರೀ (್ತ and ್ರ are the conjunct consonant glyphs of ತ and ರ)

Consonant Clusters with two different display forms : Consonant RA Rules

Whenever a consonant cluster of two or more consonants is formed with the Kannada consonant letter RA ((ರ, 0CB0) as the first component of the consonant cluster, the component of this letter RA is depicted with two different presentation forms: one as the initial and the other as the final display element of the consonant cluster as detailed below.

Consonant clusters with RA as the first consonant : general method of rendering

Rule R4 : As explained before, the character ರ is rendered with the terminal vowel (implicit or dependent) and the in-between consonants are rendered below and/or to the right of ರ, in conjunct consonant glyph forms (ರ್ಕ, ರ್ಗ etc.).

Example 1:

RA_d + KA_l ® RA_l + KA_h

ರ್ + ಕ ® ರ್ಕ

Example 2:

RA_d + MA_l + U_{vs} ® RA_n + MA_h + U_{vs}

ರ್ + ಮ + ು ® ರ್ಮು

Example 3:

RA_d + TA_d + YA_n ® RA_n + TA_h + YA_h

ರ್ + ತ್ + ಯ ® ರ್ತ್ಯ

Consonant clusters with RA as the first consonant: Alternate method of rendering

Rule R5: In the alternate representation method also, the above procedure is followed assuming ರ is absent (which means that the conjunct formation starts from the second consonant) to obtain the consonant cluster (conjunct). This is followed by another distinct glyph ರ್ for ರ್ and this new glyph is depicted to the extreme right of the conjunct formed above. As per this representation, the conjuncts ರ್ಕ, ರ್ಮು and ರ್ತ್ಯ rendered in examples 1, 2 and 3 above are rendered as ರ್ಕ, ರ್ಮು and ರ್ತ್ಯ. The corresponding rule is as follows:

Example 1:

$RA_d + KA_l$ ® KA_l + Arkavottu

ರ್ + ಕ ® ರ್ಕ

Example 2:

$RA_d + MA_l + U_{vs}$ ® $MA_n + U_{vs}$ + Arkavottu

ರ್ + ಮ + ು ® ರ್ಮು

Example 3:

$RA_d + TA_d + YA_n$ ® $TA_n + YA_h$ + Arkavottu

ರ್ + ತ್ + ಯ ® ರ್ತ್ಯ

Exception for the alternate method

Rule R6: The exception for the rule R4 is that, whenever a conjunct is formed with both the first and second consonants as ರ (RA) (ie. a consonant conjunct using ರ with ರ itself), the rule R5 will not hold good. Instead, the general method of consonant conjunct formation is used (Rule R4). This means the conjunct consonant glyph ◌ of ರ is rendered.

$RA_d + RA_l + O$ ® $RA_n + RA_h + O_{vs}$

ರ್ + ರ + ೊ ® ರ್ರೊ

Nukta- Modifier Mark Rules

In addition to the vowel signs, one more type of combining mark may be applied to a component of an orthographic syllable or the syllable as a whole. The *NUKTA* sign, which modifies a consonant form, is placed immediately after the consonant (after the terminating vowel in case of a dependent vowel appearing after the consonant) in the memory representation and is attached to that consonant in rendering. If the consonant represents a dead consonant, then the *nukta* should precede halant in the memory representation. The *nukta* is represented by a double-dot mark placed at the location 0CBC. Two such modified consonants used in Kannada are ಜ಼(Pronounced as ZA) and ಫ಼(Pronounced as FA).

Diacritics

Diacritics are the principle class of non-spacing combining characters used with the Indian scripts. Diacritic is defined very broadly to include accents as well as other non-spacing marks. Kannada has a number of combining marks that could be considered diacritic. A set of five combining marks *Udattha* (◌ಾ above the character), *Anudattha* (◌ಿ below the character), *Guru* (◌ೀ above the character), *Laghu* (◌ು above the character) and *Deergha Swaritha* (◌ೂ above the character) located at 0CD1, 0CD2, 0CD3, 0CD4 and 0CF9 respectively. These are used in the transcription of Sanskrit texts (where ever needed) and for Kannada grammatical notations.

Digits

As in many Indian languages, Kannada also has a distinct set of appropriate digits. These are being used widely in ordinary texts, Government and public places. These are enumerated with code numbers 0CE6 to 0CEF.

Note: The charts show the allocation of Unicode for Kannada characters, special charcters and signs as accepted by the Ministry of Information Technology, Government of India.

Part-2
Issues related to sorting Kannada text

The sorting sequence for Kannada in Unicode is as per the collation chart enclosed with this document. However, the following are some important issues, which have to be addressed separately for proper sorting of data in Kannada.

ISCII – 91 provides direct sorting through its codes. It is the natural sorting method just based on code values. However, ISCII will not address some language specific issues for sorting the data resulting in non-conventional sorting in some specific cases. The scholars in Kannada have specified the sorting standards in Kannada. These standards are being followed in all dictionaries and other documents in Kannada. With this in view, the following special cases have been identified.

Sorting of *Nukta* characters

The modifying mark or *Nukta* located at 0CBC and included in the collation table is enough to take care of the sorting issues of characters ಫ಼ (modified ಪ) and ಱ(modified ಧ).

It also takes care of any other consonant, which may be modified using *Nukta.*

Sorting of words with dead consonants

- **Sorting of words terminating with dead consonants**
  Sorting in this case also violates the sorting rules of Kannada. The Unicode sorting places the word terminating with the dead consonant at the end of the list. The following list compares the sorting of a sample data using Unicode table and the acceptable sorting for this case.

Sorted data as per Unicode	Acceptable sorting
ರಾಕ	ರಾಕ್
ರಾಕ್	ರಾಕ
ರಾಗ	ರಾಗ್
ರಾಗೋ	ರಾಗ
ರಾಗ್	ರಾಗೋ

- **Dead consonants within words**
  Proper sorting of data with such words can be achieved by using the invisible zero width consonant just after the dead consonant.

To circumvent unacceptable situations mentioned in sections 2.2 and 2.3 above, the Unicode Standard character 200C (Zero Width Non-Joiner) can be used appropriately in the pre-processor and collation algorithms.

Sorting of Conjuncts having two different display forms

Two such conjuncts are rendered in Kannada at present.

- **Conjuncts with ರ (0CB0) as the first consonant**

  This has been explained at an earlier section as **Consonant Ra rules.**
  Words containing both the display forms of the same consonant cluster with ರ (0CB0) as the first consonant of the cluster has to be sorted as follows. Even though the display rendering are different, both are identical in all respects. It is therefore natural that they should appear at consecutive positions. Even though a separate glyph and a corresponding glyph code are present in the display/storage codes, such an arrangement in Unicode will not render for proper sorting.

  The only alternative is to represent both the display forms by the same set of codes with a distinguishing code (0CF5) within the string for the second display form.

In Unicode form, the distinguishing code value within the string of the consonant cluster for the second display form is to be considered as ignorable for the purpose of sorting (Ref. Implementation Guidelines, Section 5.17 of Unicode Standard Version 3 document). This can be achieved through a preprocessing software, with specific functions to generate proper glyph codes, storage codes, and the Unicode at different levels. Such a situation-specific code representation guarantees proper sorting of data containing consonant clusters with two different display forms. This condition has to be incorporated at the appropriate place in the sorting algorithm.

2.4 Sorting of Diacritic characters

Diacritic characters formed using symbols located at 0CD1, 0CD2, 0CD3 0CD4 and 0CF9 to render accents to consonants, are considered to be equivalent to the corresponding consonants for sorting purposes and hence the above procedure can be adopted in such cases also.

2.5 Conclusion

The sorting issues mentioned above may have multiple solutions. Similar issues might have been solved by different methods in respect of other Indian languages. Hence, it is desirable to evolve uniform procedures for issues common to all the Indian languages. However, solutions for sorting problems mentioned here with respect Kannada have been obtained by considering all the consonants from 0C95 to 0CB9 and the consonant 0CDE when they appear independently in a data field as pure consonants (i.e. as two part coded [Ex: 0C95 ≡ (0C95, 0CBB)]). The sorting of a data field is achieved by the indexing method. All these can be elaborated to give the actual algorithms and flow charts, if need be.

3. Acknowledgements

Acknowledgements are due from Directorate of Information Technology, Govt. of Karnataka, to the following persons who have taken the responsibility in arriving at the Unicode standard and prepared this document.

- Mr. C V Srinatha Sastry, Assistant Director, National Aerospace Laboratories, Bangalore 560 017, General Secretary, Kannada Ganaka Parishath, Bangalore 560 019 and Member, Technical Advisory Committee on Satndardisation and Usage of Kannada on Computers, Government of Karnataka.
- Prof. G Venkatasubbiah, Former President, Kannada Sahithya Parishath and Former Professor, Vijaya College, Bangalore.
- Mr. G N Narasimha Murthy, Secretary, Kannada Ganaka Parishath, Bangalore.
- Dr.Panditharadhya, Professor of Kannada, Institute of Kannada Studies, University of Mysore, Mysore.
- Prof. M H Krishnaiah, Former Professor, Bangalore University
- Dr. U B Pavanaja, CEO, VishvaKannada Softec., Bangalore
- Prof. Narahalli Balasubrahmanya, Professor, Bangalore University.

KANNADA UNICODE SET : 0C80-0CFF

The need for the suggested additions of character shapes at the code points 0CBA, 0CBB, 0CBC, 0CBD, 0CD1, 0CD2, 0CD3, 0CD4, 0CF5 and at 0CF9:

1. Kannada invisible letter at the code point 0CBA.

There is a need to have a consonant which provides an invisible base for the display of dependent vowels without a consonant base. This is especially needed while sorting data fields in Kannada which contain words with dead consonants. When we use the present Unicode set, the sorting places the words terminating with dead consonants at the end of the list. To circumvent such unacceptable situations, the Kannada invisible character at the code point 0CBA will be used appropriately in the pre-processor and collation algorithms.

2. Kannada Vowel Sign (ೕ) at the code point 0CBB.

In Unicode, each of the consonant letters (code points 0C95 to 0CB9 and 0CDE) is represented as a a single consonantal sound with the short vowel ಅ at 0C85 as inherent (i.e. these are depicted as orthographic CV syllables as the combination of the base consonant with the dependent vowel sign ೕ (code point 0CBB) of the vowel ಅ (code point 0C85). In the case of conjunct formation of a consonant with vowels other than this vowel ಅ, this inherent vowel sign is overridden by the corresponding dependent vowel. Because of the inherent presence of the vowel ಅ with a consonant, sorting and other language processing issues specific to Kannada language can be successfully addressed only by having an explicit vowel sign for the vowel ಅ at a code point in the Unicode table. Hence the need for the character shape at the code point 0CBB.

3. Kannada Nukta sign at the code point 0CBC

In addition to the vowel signs, one more type of combining mark may be applied to a component of an orthographic syllable or the syllable as a whole. The ***nukta*** sign, which modifies a consonant form, is placed immediately after the consonant (after the terminating vowel in case of a dependent vowel appearing after the consonant) in the memory representation and is attached to that consonant in rendering. If the consonant represents a dead consonant, then the ***nukta*** should precede halant in the memory representation. The ***nukta*** is represented by a double-dot mark placed at the location 0CBC. Two such modified consonants used in Kannada are ಫ಼ pronounced as ZA (obtained by modifying the consonant character ಜ at the code point 0C9C using the Nukta character) and ಫ಼ pronounced as FA (obtained by modifying the consonant character ಫ at the code point 0CAB using the Nukta character). This modifying mark or Nukta character at the code point 0CBC is necessary to take care of the sorting issues of characters ಫ಼, ಫ಼ and such other consonants which occupy independent specific positions in the sorting sequence.

4. Kannada sign Reph at the code point 0CF5.

Whenever a conjunct of consonant clusters is formed with the consonant character ರ (0CB0) as the first element of the conjugate cluster, there exist the following two different display forms of the conjunct.

Method 1. The character ರ is rendered with the terminal vowel (implicit or dependent) and the in-between consonants are rendered below and/or to the right of ರ in conjunct consonant glyph forms (ರ್ , ರ್ etc.).

Method 2. In this alternate representation method also, the above procedure is followed assuming ರ is absent (which means that the conjunct formation starts from the second consonant) to obtain the conjunct (consonant cluster). This is followed by another distinct glyph ್ for the base consonant ರ್ and this new glyph is depicted to the extreme right of the conjunct formed above. As per this representation, the conjuncts ರ್ಕ and ರ್ಗ are rendered as ಕ್ಕ and ಗ್ಗ. This procedure is applied for all the conjuncts with the consonant ರ as the first element, but with an exception. The exception for this rule is that, whenever a conjunct is formed using ರ with ರ itself, the first method of rendering the conjunct is followed. This means, only ರ್ರ is allowed (method 1) and not ರ್ರ (method 2).

Words containing both the display forms of the same consonant cluster with ರ (0CB0) as the first consonant of the cluster has to be sorted as follows. Even though the display rendering are different, both are identical in all respects. It is therefore natural that they should appear at consecutive positions. The present arrangement in Unicode will not render proper sorting.

The only alternative is to represent both the display forms by the same set of codes with a distinguishing code (0CF5 for the Reph symbol ್) within the string for the second display form. In Unicode form, the distinguishing code value within the string of the consonant cluster for the second display form is considered for the purpose of sorting (Ref. Implementation Guidelines, Section 5.17 of Unicode Standard Version 3 document). This can be achieved through preprocessing software, with specific functions to generate proper glyph codes, storage codes and the Unicode at different levels. Such a situation-specific code representation guarantees proper sorting of data containing consonant clusters with two different display forms. Hence the need for the character shape at the code point 0CF5.

Sorting and other Kannada language specific issues have already been addressed in the 8-bit ASCII environment by considering the basic consonants and the explicit vowel sign for the first vowel ಅ, the Nukta sign, the Reph sign etc. This has lead to a new coding system called Kannada Script Code for Language Processing (KSCLP) which is officially adopted by the Government of Karnataka. With this, sorting of large data base in Kannada and the development of search algorithms have been succesfully achieved. Similar procedure will have to adopted in the 16-bit UNICODE environment also. This proves the necessity of the charcter shapes at the code points 0CBA, 0CBB, 0CBC and 0CF5.

Ref: The Kannada-Kannada dictionary (8 volumes) published by the Kannada Sahithya Parishat which has been adopted as the official standard by the Government of Karnataka, Universities, Kannada teaching community and all Kannada speaking community.

5. The avagraha character at the code point 0CBD and the diacritic marks at the code points 0CD1, 0CD2, 0CD3, 0CD4 and 0CF9.

Diacritics are the principle class of non-spacing combining characters used with the Indian scripts. Diacritic is defined very broadly to include Vedic and musical accents, grammatical notations as well as other non-spacing marks. Kannada has a number of combining marks that could be considered diacritic. The four combining marks at the code points 0CBD (Kannada **Avagraha sign**), 0CD1 (Kannada diacritic sign **Udattha**), 0CD2 (Kannada diacritic sign **Anudattha**) and 0CF9 (Kannada diacritic sign **Deergha Swaritha**) are extensively in Musical texts and in Kannada transcription of Samskrit texts. The grammatical notations at at the code points 0CD3 (Kannada grammatical sign **Guru** used to identify the vowels, consonants and consonant conjuncts which require more than one matra

time to pronounce) and 0CD4 (Kannada diacritic sign **Laghu** used to identify the vowels, consonants and consonant conjuncts which can be pronounced in a single *matra* time). These are being widely used in Kannada grammer text books, regularly taught in the schools and colleges in Karnataka for grammatical and liguistic studies.

The above proposed additions with respect to Kannada UNICODE was finalised only after wide range of discussions with the members of Technical Adivisory Committee on Standardisation and Usage of Kannada in Computers, Government of Karnataka and many renouned Kannada Scholars like Prof. G Venkatasubbiah, Former President, Kannada Sahithya Parishath and Former Professor, Vijaya College, Bangalore, Prof M H Krishnaiah, Former Professor, Bangalore University, Prof. Narahalli Balasubrahmanya, Professor, Bangalore University, Prof. A R Mithra, Former Professor, Bangalore University and Prof. K V Narayana, Registar, Kannada University, Hampi.

Issued by:
Directer
Directorate of Information Technology
Government of Karnataka
Bangalore

Prepared by:
C V Srinatha Sastry
(Scientist, NAL and General Secretary, Kannada Ganaka Parishat, Bangalore)
Member, Technical Advisory Committee on Standardisation and Usage of Kannada in Computers, Government of Karnataka.

APPENDIX A

0C82 Unicode for Kannada 0CEF

Based on ISCII 1988

VARIOUS SIGNS					
0C82	ಂ	KANNADA SIGN ANUSVARA	0CB6	ಶ	KANNADA LETTER SHA
0C83	ಃ	KANNADA SIGN VISARGA	0CB7	ಷ	KANNADA LETTER SSA
INDEPENDENT VOWELS			0CB8	ಸ	KANNADA LETTER SA
0C85	ಅ	KANNADA LETTER A	0CB9	ಹ	KANNDA LETTER HA
0C86	ಆ	KANNADA LETTER AA	0CBA		KANNADA INVISIBLE LETTER
0C87	ಇ	KANNADA LETTER I	**DEPENDENT VOWEL SIGNS**		
0C88	ಈ	KANNADA LETTER II	0CBB	ಾ	KANNDA VOWEL SIGN A
0C89	ಉ	KANNADA LETTER U	0CBC	಼	KANNADA SIGN NUKTA
0C8A	ಊ	KANNADA LETTER UU	0CBD	ಽ	KANNADA SIGN AVAGRAHA
0C8B	ಋ	KANNADA LETTER VOCALIC R	OCBE	ಾ	KANNADA VOWEL SIGN AA
0C8C	ಌ	KANNADA LETTER VOCALIC L (Not in use)	0CBF	ಿ	KANNADA VOWEL SIGN I
0C8D		RESERVED	0CC0	ೀ	KANNADA VOWEL SIGN II
0C8E	ಎ	KANNADA LETTER E	0CC1	ು	KANNADA VOWEL SIGN U
0C8F	ಏ	KANNADA LETTER EE	0CC2	ೂ	KANNADA VOWEL SIGN UU
0C90	ಐ	KANNDA LETTER AI	0CC3	ೃ	KANNDA VOWEL VOCALIC R
0C91		**Reserved**	0CC4	ೄ	KANNADA VOWEL VOCALIC RR
0C92	ಒ	KANNADA LETTER O	0CC5		**RESERVED**
0C93	ಓ	KANNDA LETTER OO	0CC6	ೆ	KANNADA VOWEL SIGN E
0C94	ಔ	KANNDA LETTER AU	0CC7	ೇ	KANNADA VOWEL SIGN EE
CONSONANTS			0CC8	ೈ	KANNADA VOWEL SIGN AI
0C95	ಕ	KANNADA LETTER KA	0CC9		**RESERVED**
0C96	ಖ	KANNDA LETTER KHA	0CCA	ೊ	KANNADA VOWEL SIGN O
0C97	ಗ	KANNDA LETTER GA	0CCB	ೋ	KANNADA VOWEL SIGN OO
0C98	ಘ	KANNADA LETTER GHA	0CCC	ೌ	KANNADA VOWEL SIGN AU
0C99	ಙ	KANNDA LETTER NGA	**VARIOUS SIGNS**		
0C9A	ಚ	KANNADA LETTER CA	0CCD	್	KANNADA SIGN HALANT
0C9B	ಛ	KANNADA LETTER LETTER CHA	0CCE		**RESERVED**
0C9C	ಜ	KANNADA LETTER JA	0CCF		**RESERVED**
0C9D	ಝ	KANNADA LETTER JHA	0CD0		**RESERVED**
0C9E	ಞ	KANNADA LETTER NYA	0CD1	◌	KANNADA DIACRITIC SIGN UDATTA (Above any character)
0C9F	ಟ	KANNADA LETTER TTA	0CD2	◌	KANNADA DIACRITIC SIGN ANUDATTA (Below any character)
0CA0	ಠ	KANNADA LETTER TTHA	0CD3	◌	KANNADA DIACRITIC SIGN GURU-GRAVE (Above any character)
0CA1	ಡ	KANNADA LETTER DDA	0CD4	◌	KANNADA DIACRITIC SIGN LAGHU-ACUTTE (Above any character)
0CA2	ಢ	KANNADA LETTER DDHA	0CD5	ೕ	KANNADA LENGTH MARK
0CA3	ಣ	KANNADA LETTER NNA	0CD6	ೖ	KANNADA AI LENGTH MARK
0CA4	ತ	KANNADA LETTER TA	ADDITIONAL CONSONANTS		
0CA5	ಥ	KANNADA LETTER THA	0CDE	ೞ	KANNADA LETTER LLLA
0CA6	ದ	KANNADA LETTER DA	GENERIC ADDITIONS		
0CA7	ಧ	KANNADA LETTER DHA	0CEO	ೠ	KANNADA LETTER VOCALIC RR
0CA8	ನ	KANNADA LETTER NA	0CE1	ೡ	KANNADA LETTER VOCALIC LL (Not in use)
0CA9		**RESERVED**	**DIGITS**		
0CAA	ಪ	KANNADA LETTER PA	0CE6	೦	KANNADA DIGIT ZERO
0CAB	ಫ	KANNADA LETTER PHA	0CE7	೧	KANNADA DIGIT ONE
0CAC	ಬ	KANNADA LETTER BA	0CE8	೨	KANNADA DIGIT TWO
0CAD	ಭ	KANNADA LETTER BHA	0CE9	೩	KANNADA DIGIT THREE
0CAE	ಮ	KANNADA LETTER MA	0CEA	೪	KANNADA DIGIT FOUR
0CAF	ಯ	KANNADA LETTER YA	0CEB	೫	KANNADA DIGIT FIVE
0CB0	ರ	KANNADA LETTER RA	0CEC	೬	KANNADA DIGIT SIX
0CB1	ಱ	KANNADA LETTER RRA	0CED	೭	KANNADA DIGIT SEVEN
0CB2	ಲ	KANNADA LETTER LA	0CEE	೮	KANNADA DIGIT EIGHT
0CB3	ಳ	KANNADA LETTER LLA	OCEF	೯	KANNADA NINE
0CB4	**RESERVED**		0CF5	೯	KANNADA SIGN REPH
OCB5	ವ	KANNDA LETTER VA	0CF9	◌	KANNADA DIACRITIC SIGN DEERGHA SWARITHA (Above any character)

APPENDIX B – Chart 1								
	0C8	**0C9**	**0CA**	**0CB**	**0CC**	**0CD**	**0CE**	**0CF**
0		ಐ	ಠ	ರ	ೀ		ೠ	
1			ಡ	ಱ	ು	ೕ	ೡ	
2	ಂ	ಒ	ಢ	ಲ	ೂ	ೖ		
3	ಃ	ಓ	ಣ	ಳ	ೃ	೗		
4		ಔ	ತ		ೄ	೘		
5	ಅ	ಕ	ಥ	ವ		೙		೯
6	ಆ	ಖ	ದ	ಶ	ೆ	೚	೦	
7	ಇ	ಗ	ಧ	ಷ	ೇ		೧	
8	ಈ	ಘ	ನ	ಸ	ೈ		೨	
9	ಉ	ಙ		ಹ			೩	॥
A	ಊ	ಚ	ಪ	಺	ೊ		೪	
B	ಋ	ಛ	ಫ	಻	ೋ		೫	
C	ಌ	ಜ	ಬ	಼	ೌ		೬	
D		ಝ	ಭ	ಽ	್		೭	
E	ಎ	ಞ	ಮ	ಾ		೮	೮	
F	ಏ	ಟ	ಯ	ಿ			೯	

Suggested Unicode for Kannada

APPENDIX B – Chart 2				
Column 1	Column 2	Column 3	Column 4	Column 5
0C82 ಂ	0CBD ಽ	0C96 ಖ	0CA6 ದ	0CB9 ಹ
0C83 ಃ	OCBB ಾ	0C97 ಗ	0CA7 ಧ	0CB3 ಳ
0C85 ಅ	0CBE ಾ	0C98 ಘ	0CA8 ನ	0CDE ಱ
0C86 ಆ	0CBF ಿ	0C99 ಙ	0CAA ಪ	0CBC . .
0C87 ಇ	0CC0 ೀ	0C9A ಚ	0CAB ಫ	
0C88 ಈ	0CC1 ು	0C9B ಛ	0CAC ಬ	
0C89 ಉ	0CC2 ೂ	0C9C ಜ	0CAD ಭ	
0C8A ಊ	0CC3 ೃ	0C9D ಝ	0CAE ಮ	
0C8B ಋ	0CC4 ೄ	0C9E ಞ	0CAF ಯ	
0CE0 ೠ	0CC6 ೆ	0C9F ಟ	0CB0 ರ	
0C8E ಎ	0CC7 ೇ	0CA0 ಠ	0CB1 ಱ	
0C8F ಏ	0CC8 ೈ	0CA1 ಡ	0CB2 ಲ	
0C90 ಐ	0CCA ೊ	0CA2 ಢ	0CB5 ವ	
0C92 ಒ	0CCB ೋ	0CA3 ಣ	0CB6 ಶ	
0C93 ಓ	0CCC ೌ	0CA4 ತ	0CB7 ಷ	
0C94 ಔ	0C95 ಕ	0CA5 ಥ	0CB8 ಸ	

Sorting sequence of Kannada Unicode characters
The sequence is column wise, top to bottom