

Chandrakkala. Samvruthokaram. Chillaksharam.

From the perspective of Malayalam Collation

R. Chitrajakumar and N. Gangadharan
Rachana Akshara Vedi

L2/05-210

1

For clarity about the distribution of characters we maintain the following values and terminology in this document:

1. Samvruthokaram
The central mid-vowel written as ള് (ള) following the scheme of the Original script.
2. Chandrakkala
 1. as minus-vowel marker : the normal meaning of chandrakkala
 2. pseudo-samvruthokaram: which is the usage of chandrakkala as a substitute for the samvruthokaram in the Typewriter script.
3. Chillaksharam
Chillaksharams are the special forms of some consonants which can occur without a vowel at the word-end position, within combinations of words, or within combinations of a base form and suffixes.
4. Vowellessness
That state of a consonant when it exists without the inherent vowel, i.e., the pure form.
5. Original Script
The script which was established over several centuries of development and natural evolution. It is the predominant script used in writing and, until the introduction of computer typesetting in the 1990s, in printing.
6. Typewriter Script
The reform mainly intended to include entire Malayalam syllables into the QWERTY keyboard of the typewriter, chiefly for official use. (When the Typewriter script entered into printing a large number of scripts emerged as a correction of the illogical nature of this script,. The number of scripts multiplied when the same reforms were applied to DTP. Since all these scripts are derived from the Typewriter keyboard, we collectively call them Typewriter script. The Reformed script as characterized by Unicode, is only one of these scripts, which itself is derived from the Typewriter script. In effect, the Malayalam script implemented in these various packages are very different from each other when considering the formation of conjuncts, vowel markers, consonant markers, etc.)

2

The latest events in Malayalam Unicode regarding the Chillaksharam was the allocation of codepoints in the Malayalam block to them. This is a dangerous development, both for the Malayalam language and for the computer technology surrounding it.

We believe that the encoding of chillu forms will cause major problems in the implementation of Malayalam language software. We counter this proposal in 2 ways:

1. showing that the reasons for chillu encoding, i.e., differentiating pseudo-samvruthokaram and chillu is not as important as differentiating samvruthokaram (whether pseudo or not) from vowellessness.
2. showing that the pseudo-samvruthokaram is an artificial construct of the Typewriter script which causes problems in higher level applications: it is only the limitation of the various schemes

available on the computer so far that prevents Malayalees from using the samvruthokaram.

In Malayalam, it is the sort order which provides the rationale for distinctions in an encoding. This is because, it is the sort order which assigns a value (or weight). Only when two characters or sequences differ in value (or weight) at the primary level, is there a need to differentiate them at the encoding level. Actual pronunciation of the character does not imply a difference in sorting value.

3

Consider the following 3 renderings of a Malayalam sequence ന + ് + മ:

- ന്മ (nma)
- ന്മ (nma)
- ന്മ (nma)

These three represents the same combination, and has the same reading*. They did not take these shapes accidentally. For historic reasons, they were used selectively in various contexts such as in loan words, Dravidian word formations, etc. Today, these ancient distinctions are not self-evident at a plain text level.

In the first (ന്മ nma), the operative principle is that of the chillaksharam. In the second (ന്മ nma), we see the chandrakkala, which acts as a minus-vowel marker (which is applicable to all consonants). Finally in the third (ന്മ nma), the consonants ligate to form a conjunct.

Linguistically, all three are equivalent i.e., all three have equal value. They differ only in visual appearance. At the primary level, the sort keys generated for the 3 sequences are the same. If we do not treat them as such, it will cause inconsistencies in various applications in Malayalam computing.

The ZWJ/ZWNJ-system is the most apt for representing these differences, since by definition ZWNJ/ZWJ only encodes rendering differences.

4

Giving specific code points to the chillu divorces the chillu characters from their base (or underlying characters), which is not logical. As a matter of fact, some of the comments regarding the Chillu encoding proposal, specifically seeks to do this in order to satisfy some unscientific claims, such as the false requirement that ഌ (l) be equated to the vowellessness of ത (ta). In addition, rarely misrepresented sequences such as ത്സ (tsa), can easily be handled by Input Methods and spell checkers.

Divorcing the chillus from their underlying characters also causes more trouble, due to the extra rules that will be required in the already complicated implementations of spell checkers and grammar checkers, not to mention collation rules.

It is quite interesting to see some persons advocating that the ഌ (l) is the vowellessness of both ല and ത. We have no idea how these persons plan to sort ഌ (l) in this case, i.e, whether ഌ should be sorted before ല or ത. We also do not await a quick solution, since such a collation would be based on a false premise:

1. Some persons are in utter confusion regarding the actual value and identity of the chillu ഌ, ല്

*In word-final position, a consonant with chandrakkala can have a different reading in Typewriter Script from the chillu form of that consonant. This is related to the pseudo-samvruthokaram, which is explained later.

and ത്. The main reason for this confusion is the accidental resemblance of ള and ത.

2. In certain contexts and in certain combinations, the pronunciation of ത comes near to that of ല. This resemblance, has given impetus to the confusion mentioned above.
3. The fact is that, this occurs only in Sanskrit clusters, where ത occurs as the first member (e.g., തസ). It is important to note that such clusters occur in Sanskrit only.
4. The ള-chillaksharam (and all other chillaksharams) is purely a Malayalam feature. Therefore, this feature cannot occur in a Sanskrit cluster.
5. Certain Sanskrit ത clusters have a pronunciation that comes very near to that of ല. When facts points to this, one cannot consider that the Dravidian ള (chillaksharam) and the Sanskrit ത are the same.
6. Even though this similiarity of pronunciation could have been one of many possible factors in determining the graphical form of ള chillaksharam, that has no bearing in the consideration of whether ള is the chillaksharam of ത. These two graphemes originated from entirely different environments.
7. It is also important to note that the ല pronunciation is not restricted to ത only. In Sanskrit clusters, where ള is the first member, there too the pronunciation changes to ല. Even though ഉദ്ഘാടനം is pronounced as ഉൽഘാടനം, noone is proposing that ള is the chillu of ള.

5

There have been concerns that since ZWNJ/ZWJ are ignorable under the Unicode collation rules, a pseudo-samvruthokaram will appear in the results in a search for a chillu. This is a real problem, and it occurs only because of the dual interpretation of the written symbol ് as minus-vowel marker and as pseudo-samvrutokaram. The only scientifically correct solution in the existing scheme is the following:

1. accepting only ് (u) form as the samvruthokaram, following its predominance in the Original system,
2. giving the chandrakkala the sole function of minus-vowel marker when used with consonants, and
3. retaining the existing situation of chillaksharam using Joiners

Also, note that even if the encodings were accepted in order to allow the distinction between a chillu and the pseudo-samvruthokaram, it is still not complete: there is still no way to differentiate between the vowellessnes in ത് (t) and the pseudo-samvruthokaram in ത് (tu), leaving the majority of consonants yet to be considered.

It should be noted that there was no such problem in the Original system of Malayalam. It is solely the continuance of the Typewriter reforms in this age of computing that is causing this chaos.

The issue of semantic differences in the usage of chillaksharam, samvruthokaram and chandrakkala cannot be oversimplified in this manner. The problem of so-called semantic difference between അൻ and അന് occurs only in the Typewriter script. The reality is that there are 4 cases of distinctions to consider:

1. A vowelless consonant manifested as chillu (അൻ - avan)
2. A vowelless consonant written with chandrakkala mostly at word-ending position (രാജീവ് - rajiv)
3. The chandrakkala used to show vowelless consonant components in a cluster (ഉദ്ഘാടനം - udghāṭanam)
4. The chandrakkala used as a substitute for samvruthokaram in the Typewriter script i.e., pseudo-samvruthokaram (അന് - avanu)

The first three cases, shows a chandrakkala (and its prescence in the chillu) which is equivalent in value, i.e., as a minus-vowel marker. Also the semantic differences in the first three cases, is not apparent at plain text level.

The case 4 has to be considered separately from the other three. This chandrakkala represents the samvruthokaram which has very vast and deep functions in the grammatical process of Malayalam. This substitute for the samvruthokaram, sneaked into the language via the Typewriter script. This script was created considering only the "easiness" of implementing Malayalam rendering on the QWERTY keyboard, excluding grammatical implications.

This pseudo-samvruthokaram also paved the way for all the problems of encoding chillaksharam. The real problem is not in distinguishing chillaksharam from samvruthokaram, rather it is in the distinction between samvruthokaram and chandrakkala. Please refer our document on Samvruthokaram to understand the importance of samvruthokaram

There is a spelling and presentation distinction between C+് and C+ു്. So, it requires a distinction at an encoding level also, which already exists. What should be clarified is the interpretation of a sequence C+് which occurs at the word-ending position: whether it represents a pseudo-samvruthokaram or a vowelless consonant, or both. In the case of C+ു്, this is unambiguous- it is the samvruthokaram. For the case of C+്, it is not obvious.

The distinction between samvruthokaram and chandrakkala cannot be overstressed. It causes extreme problems in higher-level applications in Malayalam such as spellchecker and grammar checker.

For a very simple e.g, consider the words and their components below:

സത്കാര്യം (satkāryam), and
വാതുകാര്യം (vātukāryam)

സത്കാര്യം (satkāryam) => സത് (sat) + കാര്യം (kāryam)
വാതുകാര്യം (vātukāryam) => വാതു് (vātu) + കാര്യം (kāryam) (in Typewriter script, വാത് (vātu) + കാര്യം (kāryam))

സത്കാര്യം (satkāryam) means 'good event, thing or issue': it is a Sanskrit loan compound word widely used in Malayalam and governed purely by the Sanskrit word formation rules.

വാതുകാര്യം (vātukāryam) means 'the subject of debate': it is a compound word, in which വാതു് has the samvruthokaram at word-ending position and it behaves under Dravidian rules.

Spell check programs in Malayalam do not enumerate all compound words. Instead, it dynamically applies grammatical rules to a list of base words (i.e., സത് (sat - meaning good), വാതു് (vātu - meaning debate), കാര്യം (kāryam - meaning issue)) and checks if the candidate word (i.e., സത്കാര്യം (satkāryam) and വാതുകാര്യം (vātukāryam)) is formed according to these rules. If so, the word is spelt correctly, else, the word is spelt wrong.

If C+് is given the interpretation as pseudo-samvruthokaram, then the combination of സത് + കാര്യം becomes സതുകാര്യം (satukāryam) which is a meaningless word.

If C+് is given the interpretation as vowellessness, then the combination വാത് + കാര്യം (correctly spelt as വാതു് + കാര്യം) becomes വാത്കാര്യം (vātkāryam) which is also a meaningless word.

We could consider that an algorithm could check both possibilities. However, the problem with this approach is that both സതുകാര്യം (satukāryam) and വാത്കാര്യം (vātkāryam) will be shown as correct.

As can be seen, the issue is whether to formalize the psuedo-samvruthokaram, i.e., whether the C+് can be interpreted as a samvruthokaram or not. Since this causes problems in the implementation of higher

level applications in Malayalam, this interpretation cannot be accepted. If the interpretation cannot be granted, the major reason for the chillu encoding falls apart. So, chillu encoding is not meaningful. If this interpretation has to be granted, then the pseudo-samvruthokaram itself has to be distinguished from vowellessness. This is the real problem in Malayalam, and is still unresolved.

To our surprise, some persons have commented that vowellessness at the word-ending position is an exception in Malayalam, citing the example of രഘുനാഥ്. The truth is that vowellessness is not an exception, rather it is one of the features of Malayalam; there are thousands of words which require it.

A large number of base/root forms (ധാതു dhātu) end with vowelless consonants. In several important circumstances, these base/root forms have to be represented in their original forms without ambiguity. They are an important foundation in grammatical analysis, in dictionary compilation and in computer applications such as spell check, etc. These forms have to be used in their natural form, i.e., with chandrakkala, in order to achieve a proper implementation of the above mentioned areas. E.g. :

തുമ്മ് (tumm),
ഉറു (ūr),
തുവ് (tūv),
കുലു (kuluḥ),
ഇളക് (iḷak),
കോത് (kōt), etc

There is yet another case of Malayalam words, ending in vowelless യ (ȳ), for e.g.,
കായ് (kāy),
നായ് (nāy),
പായ് (pāy).

This usage is also prevalent in the ever growing list of thousands loan words from Sanskrit, English, etc. e.g.,

in English:

മൈക്രോസോഫ്റ്റ് (maikrōsōphṭṭ),
യൂണിക്കോഡ് (yūṇikkōḍ), etc

in Sanskrit/Hindi:

തത് (tat),
സദ് (sad),
ഋക് (rk),
ആശിസ് (āśis),
ആശിർവാദ് (āśirvād),
ഛലന്ത് (halant), etc

in Arabic:

നിക്കാഹ് (nikkah),
റുഹ് (ruh), etc

The interpretation of word-final chandrakkala as pseudo-samvruthokaram, thus leads to insurmountable problems in Malayalam language applications.

6

The claim that chillu should be encoded for differentiating ന്ത (nṭa) and ന്റ (nra) is unnecessary. This is because:

1. In both ന്ത (nṭa) and ന്റ (nra), the first member is ന് (n̄) (= n (na) + ൾ , i.e., the chillu form of the alveolar nasal). So there is nothing gained by encoding a chillu as far as ന is concerned.

2. In Tamil and Malayalam, the alveolar class comprises of \underline{t} (alveolar stop) and \underline{n} (alveolar nasal).
3. In Tamil, \underline{ta} is written with the same glyph as \circ (\underline{ra}). The nasal is written as \circ (\underline{na}).
4. Malayalam also accepted the same practice, i.e., we write $\circ\circ$ (\underline{nta}) just as in Tamil. $\circ\circ$ (\underline{nta}) is considered as a conjunct.
5. In loan words, $\circ\circ$ (\underline{nra}) is written as in the source language. Unlike $\circ\circ$ (\underline{nta}), it is not at all considered as a conjunct. It is pronounced as separate characters. However, pronunciation is not a basis for encoding differences, just as "read" can be pronounced in 2 different ways.
6. In the case of rendering, the ZWJ is sufficient, since this is a rendering difference: the semantic difference between $\circ\circ$ (\underline{nta}) and $\circ\circ$ (\underline{nra}) is based on a distinction which is not apparent at a plain text level (the distinction lies in whether the combination \circ (\underline{na}) + ZWJ + \circ (\underline{ra}), occurs in a loan word or not). The use of the ZWJ in $\circ\circ$ (\underline{nra}) is sufficient for it to appear one after the other (and not conjoined).
7. The subjoined \circ is a typographic convention: in the Typewriter script, it is also common to write $\circ\circ$ (\underline{nta}) as $\circ\circ$. (It is better to avoid confusion by writing \underline{nta} as $\circ\circ$ and \underline{nra} as $\circ\circ$).
8. In the case of sorting, at the primary level, both are equal, since the cluster consists of the same components (except for ZWJ).

7a

Sorting in Malayalam is just like in other languages, with one small difference: in Malayalam, the operational unit is the cluster.

When sorting, we take individual clusters, split them into their component parts, and sort them. The splitting and sorting operations are formalized according to the following axioms:

Axiom 1: The basic order follows the Aksharamala (or alphabet set)

Just like any other language, Malayalam sorting order follows the natural order established in the alphabet set. In Malayalam, the natural order begins with vowel \circ and then it continues in the ascending order as follows:

1. Vowel
2. Vowelless consonant
3. Vowelless consonant + vowel
4. Vowelless consonant + vowelless consonant
5. Vowelless consonant + vowelless consonant + vowel
6. ... and so on

So,

\circ (\underline{ta}) = \circ (\underline{ta}) + ZWJ + \circ (\underline{a})

$\circ\circ$ (\underline{ti}) = \circ (\underline{ta}) + ZWJ + \circ (\underline{i})

$\circ\circ$ (\underline{ti}) = \circ (\underline{ta}) + ZWJ + $\circ\circ$ (\underline{i})

...

$\circ\circ\circ$ (\underline{ttu}) = \circ (\underline{ta}) + ZWJ + \circ (\underline{ta}) + ZWJ + \circ (\underline{u})

...

തൃ (tṣa) = ത (ta) + ി̣ + സ (sa) + ി̣ + അ (a)

So, തൃ (tu) < തൃ (tṣa), because

തൃ (tu) = ത (ta) + ി̣ + തൃ (tu),
തൃ (tṣa) = ത (ta) + ി̣ + സ (sa),
and തൃ (u) < സ (sa)

Axiom 2: ത (ta) = ത (ta) + ി̣ + അ (a)

i.e., the normal form of the consonant in the Aksharamala has an inherent vowel അ.

Axiom 3: ത് (t) < ത (ta)

This is because, in ത there is an inherent vowel അ, and in ത്, the inherent vowel was removed by chandrakkala.

ത് (t) = ത (ta) + ി̣
ത (ta) = ത (ta) + ി̣ + അ (a)

Similarly, chillu forms also have a weight less than their corresponding normal forms. This is because chillu expresses the vowellessness of the normal form.

So, ന് (n) < ന (na).

Axiom 4: The samvruthokaram (whether pseudo or not) is the central mid vowel, sorted in between തൃ (ṭ) and തൃ (u).

തി (ti) < തൃ (tu) (ത് in the Typewriter script) < തൃ (tu).

Even those who argue for chillu codepoints, admits that the chillu is a vowelless consonant (ന് (n) = ന (na) + ി̣). They propose that ന (na) + ി̣ (pseudo-samvruthokaram = ഴ) equals ന (na) + ി̣ + ഴ (u)

7b

If the chillus are given codepoints in order to circumvent the problem of differentiating chillaksharam from the pseudo-samvruthokaram, the rules of sorting would then require extraordinary effort to implement and it would be very confusing to the user. For e.g., consider the following cases:

a) Placing the pseudo-samvruthokaram

According the distribution of the chandrakkala in the Typewriter script, the following sort order would have to be maintained:

- ത് (vowelless ത = t)
- ത (ta)
- താ (tā)
- തി (ti)
- തി (tī)
- ത് (pseudo-samvruthokaram = tu)

തു (tu)

തു (tū)

...

which does not fit the natural chandrakkala model for sorting Malayalam words. A collation designer would literally have to choose between two evils, that is a sort order like തി (tī) < തു (tsa) < തു (tu), which would confound a Malayalee, or തു (tsa) < ത (ta) which is equally unusable.

If the vowelless (chandrakkala) form is placed before ത (ta), then pseudo-samvruthokaram ത് (tū) and samvruthokaram ത് (tū) will also be placed before ത (ta).

Also, since ന് (n) = ന (na) + ്, words that end in pseudo-samvruthokaram will also appear before ന (na).

If pseudo-samvruthokaram ത് (tū) is placed between തി (tī) and തു (tu), then the vowelless (chandrakkala) form should also be placed between തി (tī) and തു (tu) and this will cause തി (tī) < തു (tsa) < തു (tu).

Therefore we cannot place the sequence C+് at any of these positions, since both would lead to extremely wrong sorting order. The only resolution of this problem, is to drop the interpretation of pseudo-samvruthokaram, and consider this sequence as manifesting only vowellessness. With this solution, chillu and vowelless consonant will be placed before its base character (ന് (n) < ന (na)) and (ത് (n) < ന (na)). This corresponds exactly with the Original sorting scheme of Malayalam.

Even if it were possible to provide a different collation weight for a chandrakkala at a word boundary by making suitable changes to UCA, the problem of distinguishing the pseudo-samvruthokaram from vowellessness still exists.

This is an important problem, since the samvruthokaram has a special place in Malayalam, and it is imperative to be able to distinguish it from vowellessness, as mentioned earlier.

b) Placing the divorced റ് chillu

According to some persons, the റ് chillu codepoint is justified since, according to them, there is a practice of considering the റ് chillu as the chillu of ത. However, this causes ambiguity about the position of റ് in the sort order: either റ് = ത് or റ് = ത്, or in the extreme case, both. This is clearly wrong, and is against the natural sorting order of Malayalam. According to the Original sorting scheme of Malayalam, റ് is the chillu of only ത and not of any other character, and so, റ് = ത്.

To resolve such cases, one would have to develop a comprehensive list of Malayalam words, along with their phonological and grammatical transformations, which is a Herculean task. Such a program or list would be far more complicated than the UCA.

8

Considering all the points given above, it is obvious that the said programme to encode chillus will be a failure, both technically and linguistically. If the pseudo-samvruthokaram is accepted in order to provide a rationale for encoding chillus, then significant and unscientific changes to the Malayalam sorting scheme would also have to be made. Any new sorting scheme will not be acceptable to the Malayalee: it will not be as intuitive and natural as the Original sorting scheme of the script of Malayalam.

Malayalam computing is at a nascent stage in our industry and culture. Such drastic actions at this stage

would scuttle efforts to bring Malayalam up to speed with other languages, chiefly in the area of applications like spelling and grammar checkers, and AI-based applications.

Therefore, the Chillu encoding proposal must be dropped in its entirety.

The statement that the pseudo-samvruthokaram is in vogue, and that the samvruthokaram is not used at all, is absolutely wrong. 90% of printed material in Malayalam is typeset with the samvruthokaram. The vast majority of writing today also uses samvruthokaram. Only since the introduction of computer typesetting in the 1990s, and even then only due to the limitations in DTP software, has the pseudo-samvruthokaram been in use.

It was after feeling the limitations of the Typewriter script in this regard, that several successful attempts were made to develop software for typesetting the Original script, including samvruthokaram. Development has also begun for several Unicode-compliant packages.

It is without recognizing the wide acceptance and celebration of the Original script among Malayalees, that the practice of using pseudo-samvruthokaram has continued.

Before the Typewriter reform, there was no such issue between chillu and samvruthokaram. The principle of the Original script - that chillu and vowelless consonant are the same, and that the samvruthokaram is a separate vowel- is the only way to resolve this problem.

Above all, it should be recognized that the Typewriter script was constructed solely for the purpose of rendering. Consideration of other forms of linguistic processing, such as sorting and spell-checking, were not a factor in its development, and as we have shown certain problems for such processes arise as a result.

On the other hand, Unicode is the foundation for building a very broad set of applications of which rendering is only one. We have to move towards developing several such applications. The embedding of Malayalam in Unicode must take this under deep consideration. Towards this purpose, we are fully ready to provide additional clarification on any of the issues described in this document. It is our hope that this will lead to the most logical encoding for the Malayalam script in the Unicode.

Date: August 5, 2005

About the Authors:

R. Chitrajakumar is the General Secretary of the Rachana Akshara Vedi (A Forum for Linguistic and Language studies of Indic languages, especially Malayalam), and Senior Editor of the Malayalam Lexicon, Dept of Malayalam Lexicon, University of Kerala. He is an accomplished lexicographer and linguist, and is an expert on the Malayalam script.

N. Gangadharan is Joint Secretary of the Rachana Akshara Vedi, and Senior Editor of the Malayalam Lexicon, Dept of Malayalam Lexicon, University of Kerala. He is a scholar of Sanskrit as well as Malayalam.