Critical review of Rachana (L2/05-210) and arguments for encoding Malayalam Chillu

Author: Cibu C. J. email: cibu@yahoo.com phone: +1-847-630-2050

L2/05-334

Table of Contents

1 Rachana Document (L2/05-210) Review

Following sections point out that the arguments in Rachana document (L2/05-210) are in odds with facts.

1.1 Confusing code points as base characters

Rachana (L2/05-210) expresses a genuine concern that, by encoding, chillus are given base character status. That is not true. Code points are not base characters. This is a kind of explained in UTR#17 (Character Encoding Model).

Regarding collation, Rachana (L2/05-210) states:

> *"Only when two characters or sequences differ in [collation] value (or weight) at the primary level, is there a need to differentiate them at the encoding level."*

This is also not true:

- Even though, English lowercase and uppercase characters are encoded separately, they differ only in tertiary level in Default Unicode Collation Element Table (DUCET). Refer UTS#10 (Unicode Collation Algorithm).
- There are lot of codepoints without any primary weight. An indic example would be Visarga.

In a philosophical way, characters are for humans and codepoints are for computers. Human can understand the character from its word context which is not available/usable for a computer using codepoint. Collation Element Table is a way to connect these two.

1.2 Chillu-C1 + C2 is not equivalent to C1 + Virama + C2

Abstract

- Chillu-C1 + C2 is not equivalent to C1 + Virama + C2, in contrast to the claims in Rachana document (L2/05-210) section 3.
- Chillu issue is has existence independent of Samvruthokaram related issues.

Proof

It is proved thru following counter-example.

If Chillu-C1 + C2 is equivalant to C1 + Virama + C2, following two Malayalam words cannot have different Unicode encodings:

- /pin_nilaavum/പിൻനിലാവും
- /pinnilaavum/പിന്നിലാവും

Thus argument in the section 3 of Rachana (L2/05-210) is incorrect.

Why should they have different encodings?

The words /pin_nilaavum/ and /pinnilaavum/ are different in all 3 essential attributes of a word:

- Meaning. /pin_nilaavum/ means 'and shadow of moonlight'. /pinnilaavum/ means 'will be behind'
- Orthography. The first 'na' of /pin_nilaavum/ is chillu while we have the conjunct double of 'na' in /pinnilaavum/.
- Pronunciation. The second 'na' of /pinnilaavum/ is an alveolar and that of /pin_nilaavum/ is dental.

So these two words should have two different Unicode encodings. This argument is exactly same as why 'apple' and 'banana' should have two different Unicode encodings.

The difference should be in some non-joiner characters

See pages 389 to 391 in chapter 15 of Unicode 4.0.0

> "ZERO WIDTH NON-JOINER and ZERO WIDTH JOINER are format control
> characters. Like other such characters, they should be ignored by
> processes that analyze text content. For example, a spelling-checker

> or find/replace operation should filter them out. (See Section 2.11, Special Characters and Noncharacters, for a general discussion of format control characters.)"

(thanks to Mahesh Pai)

More examples

വൻയവനിക /van_yavanika/ meaning 'big curtain'
വന്യവനിക/vanyavanika/ meaning 'wild forest'

കൺവലയം /kaN_valayam/ meaning 'eye boundary'
കണ്വലയം/kaNvalayam/ meaning 'peace of Kanvan - the mythical character'.

തൻവിനയം /than_vinayam/ meaning 'his/her modesty'
തന്വിനയം /thanvinayam/ meaning 'policy of a woman'

മൻവിക്ഷോഭം /man_vikshObham/ meaning 'explosion of mind'
മന്വിക്ഷോഭം /manvikshObham/ meaning 'fury of a lady'

Above examples are just few from vast number of them possible with following generic patterns:

- character capable of forming chillu + semi-vowel
- ന + ന

Counter-challenge from Kenneth Whistler

If separate characters are encoded for Malayalam Chillus, so that the "challenge" distinction were to be encoded as:

"nn" is U+0D28, U+0D4D, U+0D28

"n_n" is U+0DXX, U+0D4D, U+0D28

implementers are then faced with determining what to do with the following sequence:

"???" is U+0D28, U+0D4D, U+200D, U+0D28

That sequence, of course, exists now, and would be a legitimate and possible sequence even if a Chillu-n is encoded. So how would a rendering engine render that sequence, and how would it be distinguished, by an end user or a text process such as a search engine, from the proposed U+0DXX, U+0D4D, U+0D28 sequence for "n_n"?

That counter-challenge needs a "solution" for the encoding of Chillu characters to make sense for Malayalam. For if there is no solution forthcoming, addition of Chillu characters would potentially be *increasing* the ambiguity potential for the Unicode representation of Malayalam text, rather than decreasing it.

Solution to Ken's challenge

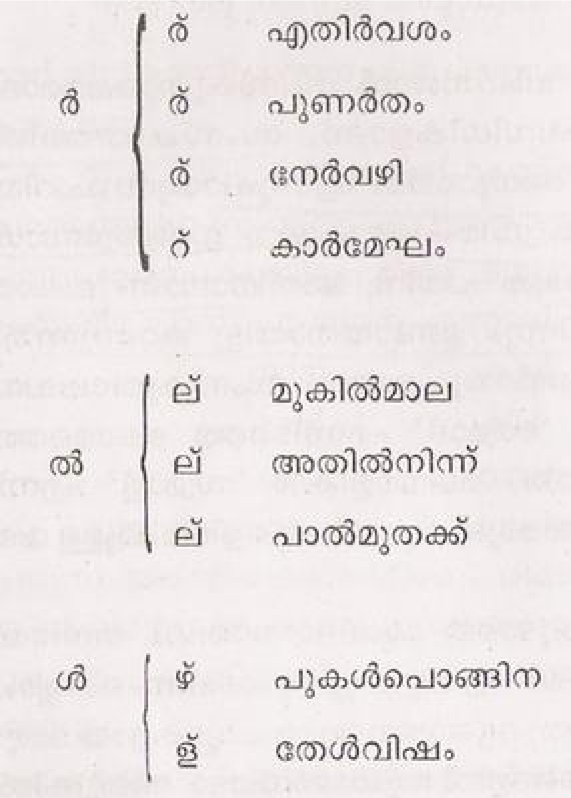
Half-form of NA (ന) is not chillu. It is described in detail in section 2.2.

Now, we can use the rules from PRI#37 (L2/04-279 - Functions of ZERO WIDTH JOINER in Indic Scripts) to show the behavior of the challenge sequence. It will form the conjunct double ന്ന /nna/ as per the second bullet in the section 7-proposal of PRI#37.

1.3 One chillu for two base characters

In section 4, Rachana Document (L2/05-210) establishes that chillu-ല is not chillu-ത. However, real issue is not with chillu-ല. It is with chillu-ര and chillu-ള; because they also are chillu-റ and chillu-ഴ, respectively.

This fact is clearly established in the foremost grammar book of Malayalam: Keralapaanineeyam by A. R. Rajaraja Varma. See the relevant scan from the section Peethika: 4. Varnnavikaarangal below:



ർ	ര്	എതിർവശം
	ര്	പുണർതം
	ര്	നേർവഴി
	റ്	കാർമേഘം
ൽ	ല്	മുകിൽമാല
	ല്	അതിൽനിന്ന്
	ല്	പാൽമുതക്ക്
ൾ	ഴ്	പുകൾപൊങ്ങിന
	ള്	തേൾവിഷം

Possible solutions and their implications

I am considering only chillu-ര/റ right now. These thoughts are applicable to chillu-ള/ഴ as well.

1. Encode this chillu as chillu-ര only. That is, only RA + VIRAMA + ZWJ will form the chillu. This would cause wrong collation ordering for words with chillu-റ. That is, /kaarr_mEgham/ from the above example will get wrong place in the collation order.
2. Both RA + VIRAMA + ZWJ and RRA + VIRAMA + ZWJ form chillu-ര/റ. This gives a uniqueness rule (refer section 5) warning: "if this scheme is allowed, a document (eg: a wiktionary.org document) written by multiple people using various inputting tools can quite possibly have different 'spellings' for a word, without reader or writer being aware of it. This can cause many problems including ineffective searches and inconsistent collation".

This is an example where codepoint and base characters of a language need to differ. A human can deduce base character from its word context which is not available/usable for a computer employing codepoints.

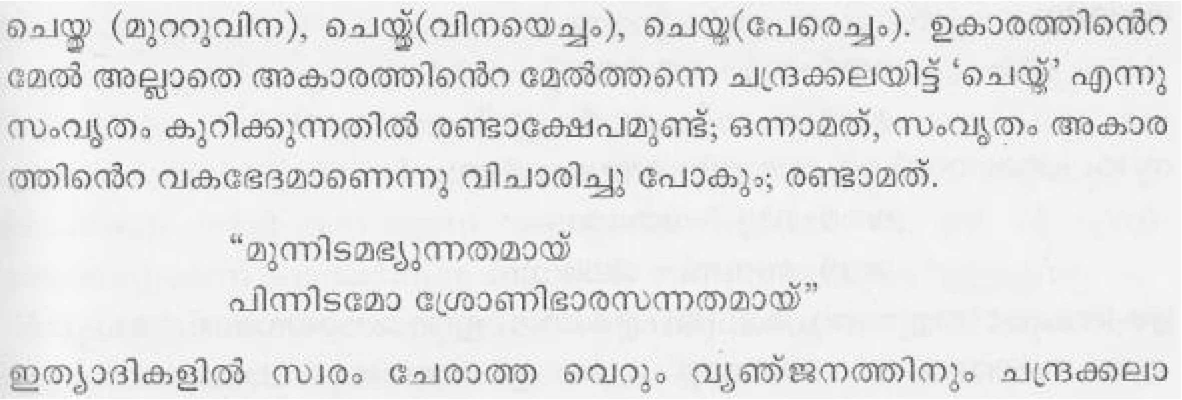
Thus collation correctness of chillu forming characters to their underlying letter identity is impossible in codepoint-space without the help of sophisticated text processing at higher levels. This would in turn mean, the collation correctness is not an argument until a new solution or perspective is proposed. Till then, both the choices of, encoding chillu with a control/format character and giving independant codepoint for it, have to be evaluated with respect to rest of the merits these options have.

1.4 Samruthokaram grapheme: history and current status

Summary
It is not what Rachana Document (L2/05-210) claims.

There are two different sets of graphemes used for samvruthokaram.
Usage 1: the sign of U + chandrakkala (visible virama)
Usage 2: chandrakkala (visible virama) alone.
These two practices co-exist today and had been like that for at least a century.

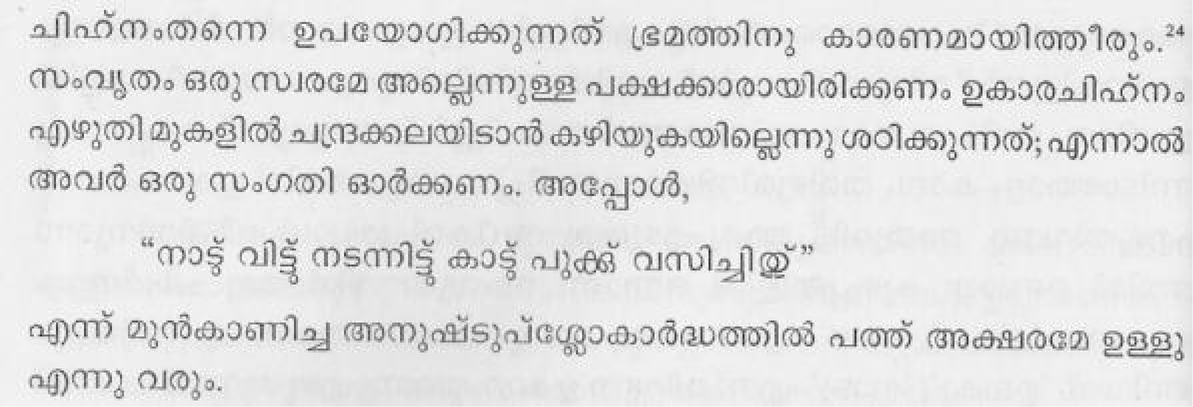
Details
In Keralapaanineeyam - the foremost grammar book of Malayalam - A. R. Rajaraja Varma criticizes those who argue for samvruthokaram to be written without the sign of U. This, in turn, tells us when Keralalpanineeyam was written (cir. 1896) the grapheme of samvruthokaram was an issue. See following scan from the book:



ചെയ്യു (മുററുവിന), ചെയ്യു്(വിനയെച്ചം), ചെയ്യ(പേരെച്ചം). ഉകാരത്തിന്റെ മേൽ അല്ലാതെ അകാരത്തിന്റെ മേൽത്തന്നെ ചന്ദ്രക്കലയിട്ട് 'ചെയ്യ്' എന്നു സംവൃതം കുറിക്കുന്നതിൽ രണ്ടാക്ഷേപമുണ്ട്; ഒന്നാമത്, സംവൃതം അകാരത്തിന്റെ വകഭേദമാണെന്നു വിചാരിച്ചു പോകും; രണ്ടാമത്.

"മുന്നിടമഭ്യുന്നതമായ്
പിന്നിടമോ ശ്രോണിഭാരസന്നതമായ്"

ഇത്യാദികളിൽ സ്വരം ചേരാത്ത വെറും വ്യഞ്ജനത്തിനും ചന്ദ്രക്കലാ

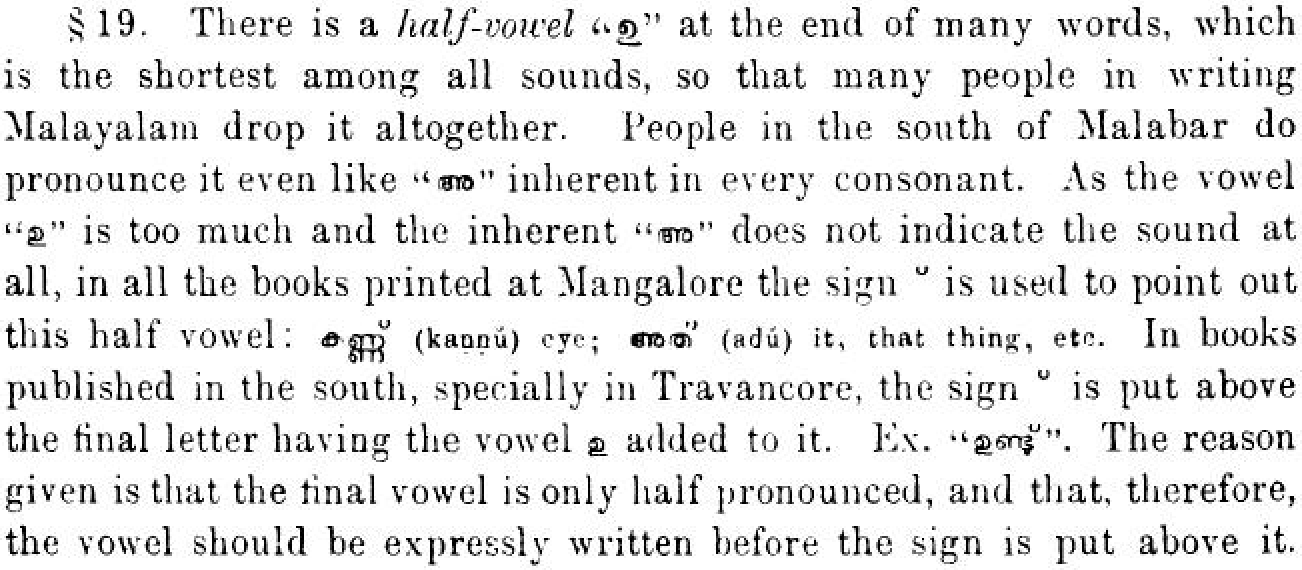


ചിഹ്നംതന്നെ ഉപയോഗിക്കുന്നത് ഭ്രമത്തിനു കാരണമായിത്തീരും.[24] സംവൃതം ഒരു സ്വരമേ അല്ലെന്നുള്ള പക്ഷക്കാരായിരിക്കണം ഉകാരചിഹ്നം എഴുതി മുകളിൽ ചന്ദ്രക്കലയിടാൻ കഴിയുകയില്ലെന്നു ശഠിക്കുന്നത്; എന്നാൽ അവർ ഒരു സംഗതി ഓർക്കണം. അപ്പോൾ,

“നാട്ടു് വിട്ടു് നടന്നിട്ടു് കാട്ടു് പുക്കു് വസിച്ചിതു്”

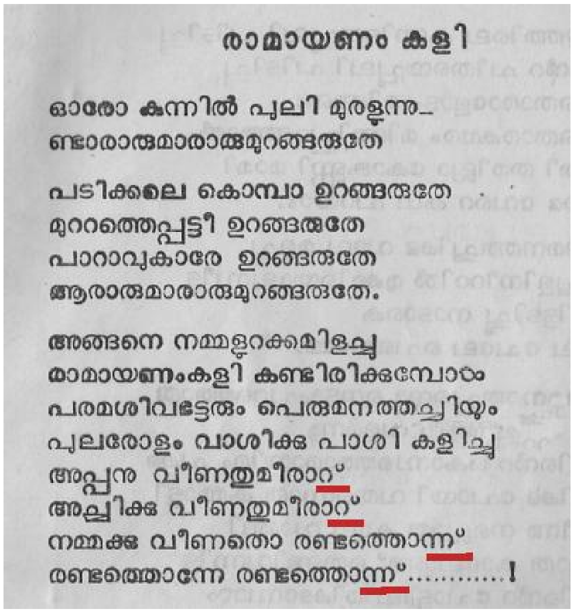
എന്ന് മുൻകാണിച്ച അനുഷ്ടുപ്ശ്ലോകാർദ്ധത്തിൽ പത്ത് അക്ഷരമേ ഉള്ളു എന്നു വരും.

Frohnmeyer writes about this in University of Madras book printed in 1913. (thanks to Eric Muller)



§ 19. There is a *half-vowel* “ഉ” at the end of many words, which is the shortest among all sounds, so that many people in writing Malayalam drop it altogether. People in the south of Malabar do pronounce it even like “അ” inherent in every consonant. As the vowel “ഉ” is too much and the inherent “അ” does not indicate the sound at all, in all the books printed at Mangalore the sign ˘ is used to point out this half vowel: കണ്ണ് (kaṇṇŭ) eye; അത് (adŭ) it, that thing, etc. In books published in the south, specially in Travancore, the sign ˘ is put above the final letter having the vowel ഉ added to it. Ex. “ഉണ്ടു്”. The reason given is that the final vowel is only half pronounced, and that, therefore, the vowel should be expressly written before the sign is put above it.

Even though, the following poem by N. N. Kakkad is printed (in 1988) in old orthography, the samvruthokaram is represented with chandrakkala (visible virama) alone.



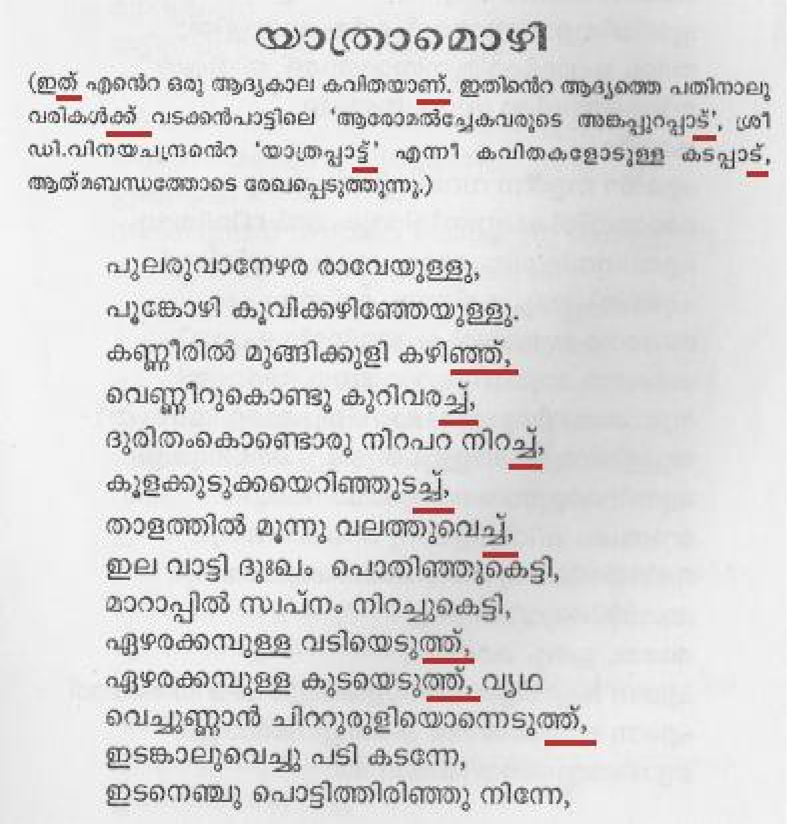
രാമായണം കളി

ഓരോ കുന്നിൽ പുലി മുരളുന്നു–
ണ്ടാരാരുമാരാരുമുറങ്ങരുതേ

പടിക്കലെ കൊമ്പാ ഉറങ്ങരുതേ
മുററത്തെപ്പട്ടീ ഉറങ്ങരുതേ
പാറാവുകാരേ ഉറങ്ങരുതേ
ആരാരുമാരാരുമുറങ്ങരുതേ.

അങ്ങനെ നമ്മളുറക്കമിളച്ചു
രാമായണംകളി കണ്ടിരിക്കുമ്പോഴും
പരമശിവഭട്ടരും പെരുമനത്തച്ചിയും
പുലരോളം വാശിക്കു പാശി കളിച്ചു
അപ്പനു വീണതുമീരാറ്
അച്ചിക്കു വീണതുമീരാറ്
നമ്മക്കു വീണതൊ രണ്ടത്തൊന്ന്
രണ്ടത്തൊന്നേ രണ്ടത്തൊന്ന്............!

When new orthography became common for printing in later years, the usage of chandrakkala alone for samvruthokaram became even more widespread. Vast majority of the printed material after ~1994 uses chandrakkala alone for samvruthokaram. As an example, see following poem by Balachandran Chullikkadu (written in 1970; printed in 1996 by DC books):

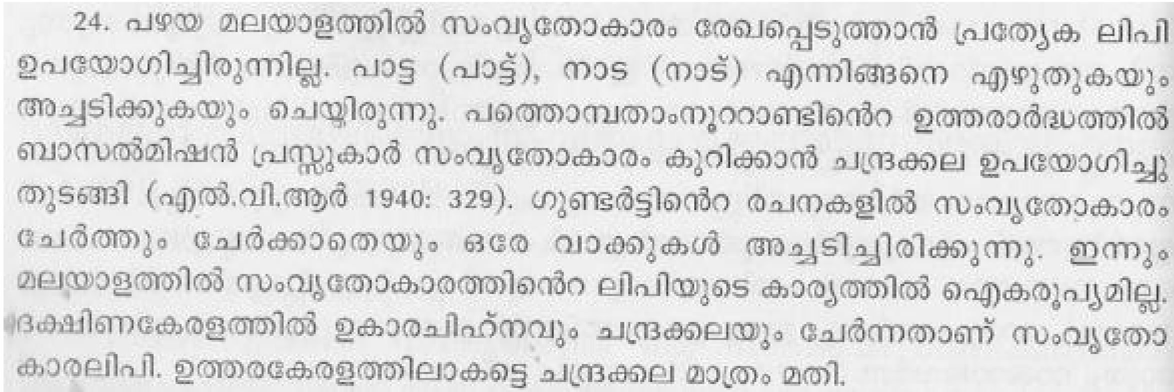


യാത്രാമൊഴി

(ഇത് എന്റെ ഒരു ആദ്യകാല കവിതയാണ്. ഇതിന്റെ ആദ്യത്തെ പതിനാലു വരികൾക്ക് വടക്കൻപാട്ടിലെ 'ആരോമൽച്ചേകവരുടെ അങ്കപ്പുറപ്പാട്', ശ്രീ ഡി.വിനയചന്ദ്രന്റെ 'യാത്രപ്പാട്ട്' എന്നീ കവിതകളോടുള്ള കടപ്പാട്, ആത്മബന്ധത്തോടെ രേഖപ്പെടുത്തുന്നു.)

പുലരുവാനേഴര രാവേയുള്ളു,
പൂങ്കോഴി കൂവിക്കഴിഞ്ഞേയുള്ളു.
കണ്ണീരിൽ മുങ്ങിക്കുളി കഴിഞ്ഞ്,
വെണ്ണീറുകൊണ്ടു കുറിവരച്ച്,
ദുരിതംകൊണ്ടൊരു നിറപറ നിറച്ച്,
കൂളക്കുടുക്കയെറിഞ്ഞുടച്ച്,
താളത്തിൽ മൂന്നു വലത്തുവെച്ച്,
ഇല വാട്ടി ദുഃഖം പൊതിഞ്ഞുകെട്ടി,
മാറാപ്പിൽ സ്വപ്നം നിറച്ചുകെട്ടി,
ഏഴരക്കമ്പുള്ള വടിയെടുത്ത്,
ഏഴരക്കമ്പുള്ള കുടയെടുത്ത്, വ്യഥ
വെച്ചുണ്ണാൻ ചിറ്റുരുളിയൊന്നെടുത്ത്,
ഇടങ്കാലുവെച്ചു പടി കടന്നേ,
ഇടനെഞ്ചു പൊട്ടിത്തിരിഞ്ഞു നിന്നേ,

Current status and brief history of grapheme(s) used for samvruthokaram is convincingly described by Dr. Scaria Zacharia in 1996 as a footnote description in Keralapanineeyam, centennial edition from DC books:



24. പഴയ മലയാളത്തിൽ സംവൃതോകാരം രേഖപ്പെടുത്താൻ പ്രത്യേക ലിപി ഉപയോഗിച്ചിരുന്നില്ല. പാട്ട (പാട്ട്), നാട (നാട്) എന്നിങ്ങനെ എഴുതുകയും അച്ചടിക്കുകയും ചെയ്തിരുന്നു. പത്തൊമ്പതാംനൂറ്റാണ്ടിന്റെ ഉത്തരാർദ്ധത്തിൽ ബാസൽമിഷൻ പ്രസ്സുകാർ സംവൃതോകാരം കുറിക്കാൻ ചന്ദ്രക്കല ഉപയോഗിച്ചു തുടങ്ങി (എൽ.വി.ആർ 1940: 329). ഗുണ്ടർട്ടിന്റെ രചനകളിൽ സംവൃതോകാരം ചേർത്തും ചേർക്കാതെയും ഒരേ വാക്കുകൾ അച്ചടിച്ചിരിക്കുന്നു. ഇന്നും മലയാളത്തിൽ സംവൃതോകാരത്തിന്റെ ലിപിയുടെ കാര്യത്തിൽ ഐകരൂപ്യമില്ല. ദക്ഷിണകേരളത്തിൽ ഉകാരചിഹ്നവും ചന്ദ്രക്കലയും ചേർന്നതാണ് സംവൃതോകാരലിപി. ഉത്തരകേരളത്തിലാകട്ടെ ചന്ദ്രക്കല മാത്രം മതി.

Translation: "*In the old Malayalam, there was no separate grapheme to indicate samvruthokaram. It used to be written and printed as /paaTTa/(patt~) and /naaTa/(naaT~) (refer transliteration scheme in section 6). In the second half of 19th centuary, Basel missionaries started to use chandrakkala (visible virama) to indicate samvruthokaram (L.V.R. 1940: 329). Same words have been printed with and* *without samvruthokaram in Gundert's works. There is no unity in Malayalam in the issue of grapheme for samvruthokaram. In southern Kerala, the sign of U and chandrakkala together is the grapheme for samvruthokaram. However, in northern Kerala, just chandrakkala (visible virama) alone is enough."*

1.5 Collation is a non-issue with encoding chillu

It is demonstrated thru a suggested scheme for Malayalam in AllKeys table of DUCET. The scheme is described thru examples:

(Please refer the transliteration scheme in section 6.)

```
n~                    = [1A2B.0020]
n_                    = [1A2B.0021]
nu~                   = [1A2B.0023]
na    =  n ~ -a       = [1A2B.0020], [1A08.0020]
naa   =  n ~ -aa      = [1A2B.0020], [1A0A.0020]
..
nau   =  n ~ -au      = [1A2B.0020], [1A17.0020]
n~a   =  n ~ a        = [1A2B.0020], [1A08.0021]
...
n~au  =  n ~ au       = [1A2B.0020], [1A17.0021]
nka   =  n ~ k ~ -a  = [1A2B.0020], [1A18.0020], [1A08.0020]
...
```

Together with anuswara

```
m_ka  = ng ~ k ~ -a  = [1A1C.0022], [1A18.0020], [1A08.0020]
...
m_ja  = ny ~ j ~ -a  = [1A21.0022], [1A1F.0020], [1A08.0020]
...
m_ta  = ny ~ j ~ -a  = [1A2B.0022], [1A27.0020], [1A08.0020]
...
m_ya  =  m ~ y ~ -a  = [1A30.0022], [1A31.0020], [1A08.0020]
...
```

Notes

- /n/ and /~/ together form a contraction.
- /na/ is represented by an expansion. The symbol of /a/ is a fictitious entity appear only in collation.
- /n_/, /n~/ and /nu~/ are same in primary level. But they are different in secondary level. One could very well differentiate /nu~/ from /n~/ in primary level also. This is just a usage example of an generic framework: splitting Consonant and Consonant+vowel-sign as /n ~ -x/

Thoughts

1. Chillus and virama forms are diacritics of base character. So diacritics themselves have to be level-1 ignorable; but should have some weight in level-2.
2. The AllKeys file containing the Default Unicode Collation Element Table (DUCET) does not currently handle Malayalam accurately. For example, ZWJ is by default ignorable, and NNA + VIRAMA + ZWJ, NNA + VIRAMA are treated as equal.
3. If Chillus are encoded, the following equivalence should NOT be used for tailoring:

   ```
   0D7F = 0D15 0D4D 200D
   ```

> . . .
> The behavior of `0D15 0D4D 200D` is different from chillu as explained in the solution to Ken's counter-challenge in section 1.2.

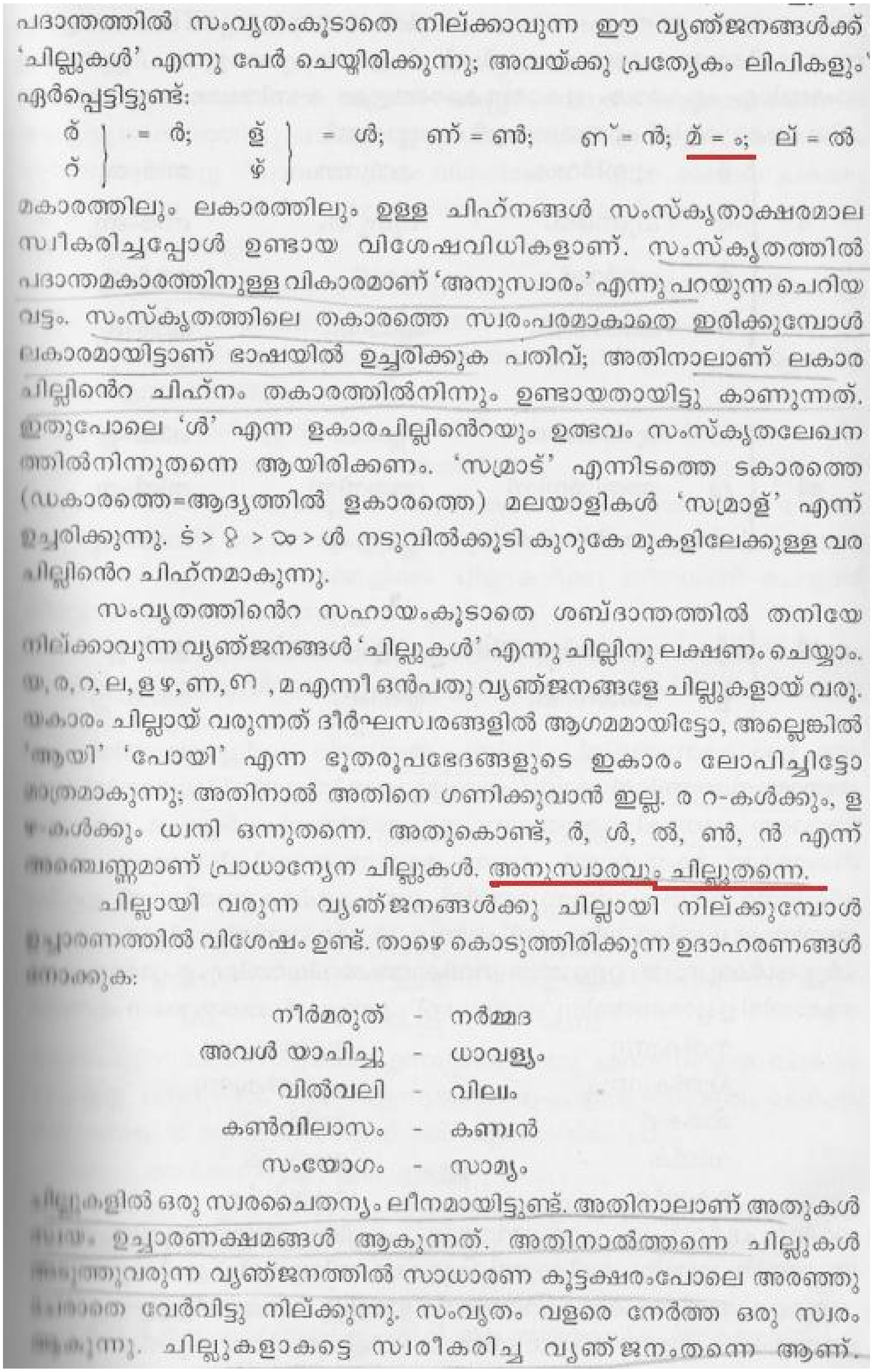
[Thanks to Åke Persson for educating me on UTS#10]

2 More arguments for encoding chillu

2.1 One chillu is already encoded years back

Knowingly or unknowingly Unicode has already encoded one chillu and that is Anuswara!

The fact that Anuswara is a chillu, is described in the foremost grammar book of Malayalam: Keralapaanineeyam by A. R. Rajaraja Varma. See the relevant scan from the section Peethika: 4.Varnnavikaarangal below:



പദാന്തത്തിൽ സംവൃതംകൂടാതെ നില്ക്കാവുന്ന ഈ വ്യഞ്ജനങ്ങൾക്ക് 'ചില്ലുകൾ' എന്നു പേർ ചെയ്തിരിക്കുന്നു; അവയ്ക്കു പ്രത്യേകം ലിപികളും' ഏർപ്പെട്ടിട്ടുണ്ട്:

ര് } = ർ; ള് } = ൾ; ണ് = ൺ; ന് = ൻ; മ് = ം; ല് = ൽ
റ് } ഴ് }

മകാരത്തിലും ലകാരത്തിലും ഉള്ള ചിഹ്നങ്ങൾ സംസ്കൃതാക്ഷരമാല സ്വീകരിച്ചപ്പോൾ ഉണ്ടായ വിശേഷവിധികളാണ്. സംസ്കൃതത്തിൽ പദാന്തമകാരത്തിനുള്ള വികാരമാണ് 'അനുസ്വാരം' എന്നു പറയുന്ന ചെറിയ വട്ടം. സംസ്കൃതത്തിലെ തകാരത്തെ സ്വരംപരമാകാതെ ഇരിക്കുമ്പോൾ ലകാരമായിട്ടാണ് ഭാഷയിൽ ഉച്ചരിക്കുക പതിവ്; അതിനാലാണ് ലകാര ചില്ലിൻെറ ചിഹ്നം തകാരത്തിൽനിന്നും ഉണ്ടായതായിട്ടു കാണുന്നത്. ഇതുപോലെ 'ൾ' എന്ന ളകാരചില്ലിൻെറയും ഉത്ഭവം സംസ്കൃതലേഖനത്തിൽനിന്നുതന്നെ ആയിരിക്കണം. 'സമ്രാട്' എന്നിടത്തെ ടകാരത്തെ (ഡകാരത്തെ=ആദ്യത്തിൽ ളകാരത്തെ) മലയാളികൾ 'സമ്രാള്' എന്ന് ഉച്ചരിക്കുന്നു. ട് > ള > ൾ > ൾ നടുവിൽക്കൂടി കുറുകേ മുകളിലേക്കുള്ള വര ചില്ലിൻെറ ചിഹ്നമാകുന്നു.

സംവൃതത്തിൻെറ സഹായംകൂടാതെ ശബ്ദാന്തത്തിൽ തനിയേ നില്ക്കാവുന്ന വ്യഞ്ജനങ്ങൾ 'ചില്ലുകൾ' എന്നു ചില്ലിനു ലക്ഷണം ചെയ്യാം. യ, ര, റ, ല, ള ഴ, ണ, ന , മ എന്നീ ഒൻപതു വ്യഞ്ജനങ്ങളേ ചില്ലുകളായ് വരൂ. യകാരം ചില്ലായ് വരുന്നത് ദീർഘസ്വരങ്ങളിൽ ആഗമമായിട്ടോ, അല്ലെങ്കിൽ 'ആയി' 'പോയി' എന്ന ഭൂതരൂപഭേദങ്ങളുടെ ഇകാരം ലോപിച്ചിട്ടോ മാത്രമാകുന്നു; അതിനാൽ അതിനെ ഗണിക്കുവാൻ ഇല്ല. ര റ-കൾക്കും, ള ഴ-കൾക്കും ധ്വനി ഒന്നുതന്നെ. അതുകൊണ്ട്, ർ, ൾ, ൽ, ൺ, ൻ എന്ന് അഞ്ചെണ്ണമാണ് പ്രാധാന്യേന ചില്ലുകൾ. അനുസ്വാരവും ചില്ലുതന്നെ.

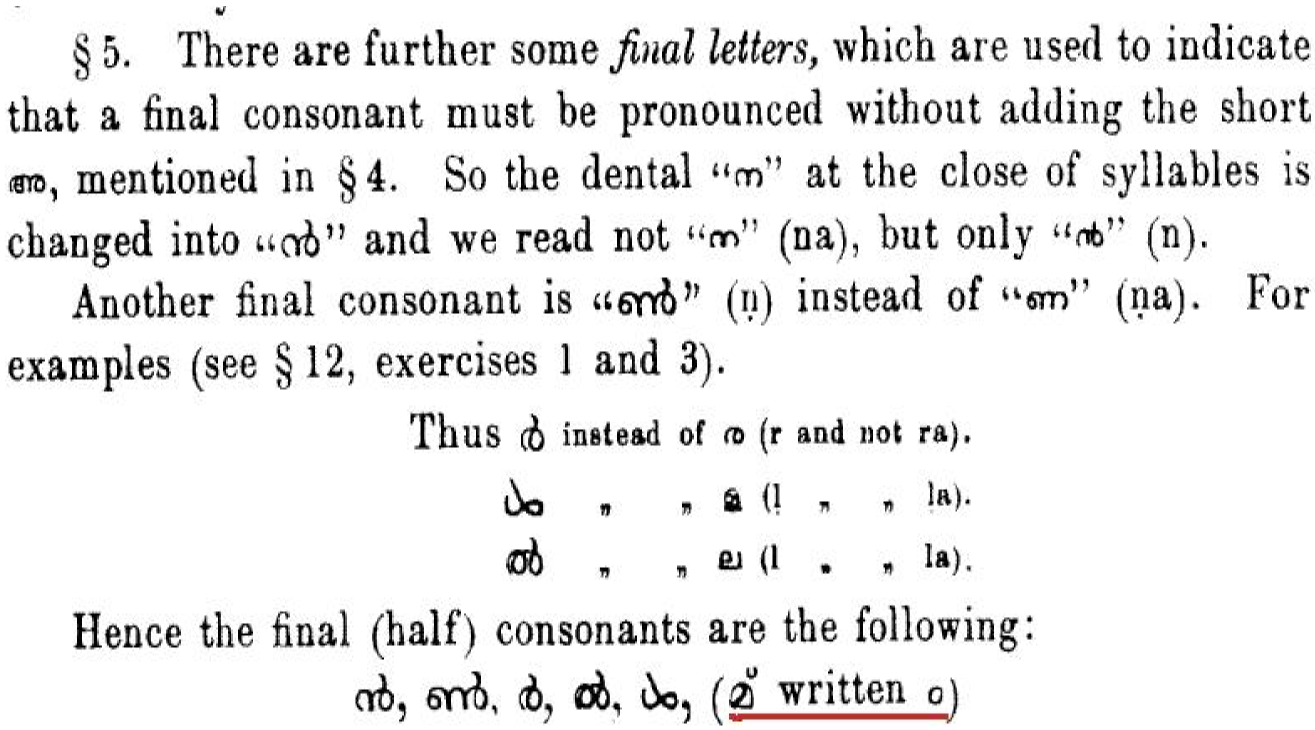
ചില്ലായി വരുന്ന വ്യഞ്ജനങ്ങൾക്കു ചില്ലായി നില്ക്കുമ്പോൾ ഉച്ചാരണത്തിൽ വിശേഷം ഉണ്ട്. താഴെ കൊടുത്തിരിക്കുന്ന ഉദാഹരണങ്ങൾ നോക്കുക:

നീർമരുത്	-	നർമ്മദ
അവൾ യാചിച്ചു	-	ധാവള്യം
വിൽവലി	-	വില്വം
കൺവിലാസം	-	കണ്വൻ
സംയോഗം	-	സാമ്യം

ചില്ലുകളിൽ ഒരു സ്വരചൈതന്യം ലീനമായിട്ടുണ്ട്. അതിനാലാണ് അതുകൾ സ്വയം ഉച്ചാരണക്ഷമങ്ങൾ ആകുന്നത്. അതിനാൽത്തന്നെ ചില്ലുകൾ അടുത്തുവരുന്ന വ്യഞ്ജനത്തിൽ സാധാരണ കൂട്ടക്ഷരംപോലെ അരഞ്ഞു ചേരാതെ വേർവിട്ടു നില്ക്കുന്നു. സംവൃതം വളരെ നേർത്ത ഒരു സ്വരം ആകുന്നു. ചില്ലുകളാകട്ടെ സ്വരീകരിച്ച വ്യഞ്ജനംതന്നെ ആണ്.

Translation of the underlined sentence: '*Anuswara also is a chillu.*'

Same opinion is echoed by L. J. Frohnmeyer (in 1913):



§ 5. There are further some *final letters,* which are used to indicate that a final consonant must be pronounced without adding the short അ, mentioned in § 4. So the dental “ന” at the close of syllables is changed into “ൻ” and we read not “ന” (na), but only “ൻ” (n).

Another final consonant is “ൺ” (ṇ) instead of “ണ” (ṇa). For examples (see § 12, exercises 1 and 3).

Thus ർ instead of ര (r and not ra).
ൾ „ „ ള (ḷ „ „ ḷa).
ൽ „ „ ല (l . „ la).

Hence the final (half) consonants are the following:
ൻ, ൺ, ർ, ൽ, ൾ, (മ് written ം)

(thanks to Eric Muller and Mahesh Pai)

The history of anuswara becoming chillu is something similar to that of how vowelless ത /tha/ became chillu-ല /la/. Initially Anuswara came to Malayalam along with rest of the Sanskrit package. Later Malayalam started to use it as chillu മ /ma/. Please note that, Malayalam anuswara should not be confused with the functions of anuswara of Devanagiri. They both are different.

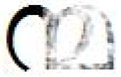
Current scenario can be compared to, assigning codepoints for some vowel signs and then on a later thought, encoding rest of the vowel signs as VIRAMA + ZWJ + Vowel.

So, if chillu of മ /ma/ can be encoded, then why not rest of the chillus?

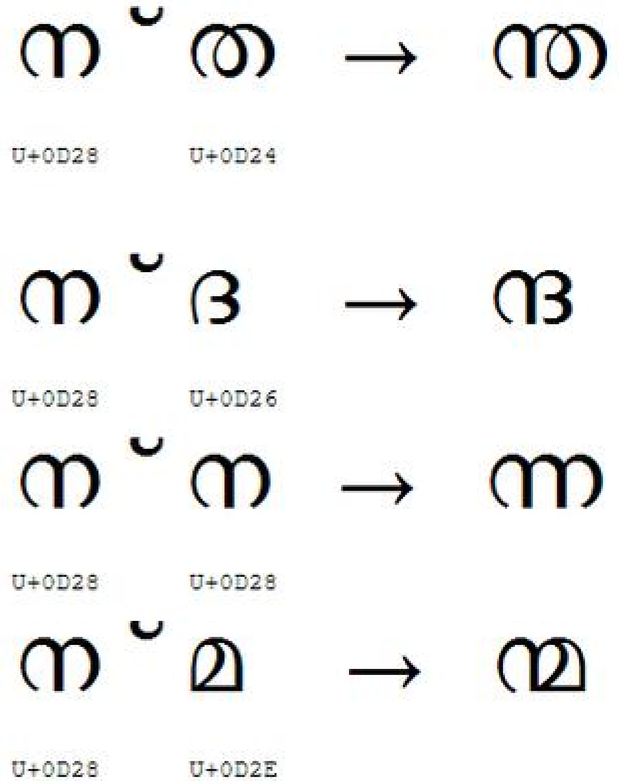
I understand that this argument alone is not enough for encoding rest of the chillus. But given that our options are limited, this makes the arguments to encode chillus more compelling.

2.2 Half forms are not chillus

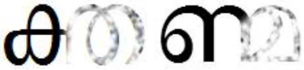
At least, 3 of the chillu forming consonants have half-forms different from chillus. As the first example, half-na is highlighted in this image:



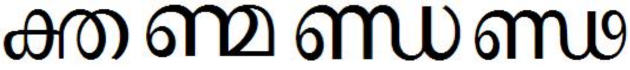
Notice that half-na is not chillu-na. Below, we can see the half-consonant in action forming conjuncts with various characters:



Similar to ന(NA), ക (KA) and ണ (NNA) also have following half-forms different from their Chillu forms:



Their conjunct formation examples are:



2.3 ZWJ shouldn't be overloaded for searching

Software Engineering Pragmatism

I have been in software industry for last 10 years. I know, as many of you do, this: If one entity does not have philosophical integrity, soon its minor functional anomalies will be interpreted differently or ignored by the developers and eventually it will be a perpetual bug in most of the software applications today or to come. You all can imagine, how much priority this Malayalam bug will get in each corporation's defect tracking system.

To tell an example, couple of months back we faced with an issue of Unicode Malayalam text getting truncated in Microsoft Outlook. After some google searches we stumbled upon this piece of information: http://www.landfield.com/usefor/1997/Aug/0142.html: *"...Unfortunately, the Unicode character 0x0d0a is used in the Malayalam set, so we couldn't really force 8 bit CR LF as the line terminator irrespective of the character set. Then again, how many people are posting messages in Malayalam, and how many would otherwise benefit from UCS-2 encoding?..."*

Undoubtedly that would the attitude development process will take towards this special case.

I can clearly understand UTC's fondness towards ZWJ solution - it is just a one line change in the standard and chillus are history or UTC. But for Malayalam community, life with Unicode is just begun. The potential bugs in the software tools and usage difficulties (which Kevin explained) will haunt us for decades. So I would ask all to think twice before committing to overload ZWJ with language specific functions.

I still wonder what Arabic community was thinking when they allowed this to happen for them. Taking Arabic as an example, soon ZWJ will get overloaded many such language specific functions and this simple ZWJ will turn into the most unscrupulous codepoint in entire standard.

2.4 Unanswered questions

1. Isn't using CGJ a dangerous thing? Because, a document (eg: a wiktionary.org document) written by multiple people using various inputting tools can quite possibly have different 'spellings' for a conjunct or word, without reader or writer being aware of it. This can cause many problems including ineffective searches and inconsistent collation.
2. My understanding about collation value of a codepoint is that it is directly tied both ways to search/sort functions. That is, searching and sorting is done using collation value and when collation values vary, search/sort can potentially give different results. Does collation value has any other purpose? If no, then by attaching search and sort meaning to ZWJ, aren't we actually adding a collation value to ZWJ? That is, ZWJ in turn becoming ZWJ + CGJ in case of chillus.
3. What was the reasoning behind giving vowel signs a different codepoint? Why they weren't encoded as, say, VIRAMA + AA = sign of AA
4. When do one say two words with different orthography and same meaning have two different spellings. Example: color & colour. Same way, can we say that the old and new orthography renderings of the same word, say 'ശബ്ദം' and 'ശബ്‌ദ' (/Sabdam/), qualify for two different spellings?
5. What is the assumption Unicode makes about the input methods? Does it assume the input method has word lookup feature or just a basic keyboard layout or inputting each Unicode codepoint by codepoint?

3 More Options

3.1 Why not encode the diacritic tail of Chillu?

I have shown in the section 1.2 that the issue of (chillu Vs chandrakkala) is not a derivative of the issue of (chandrakkala Vs samvruthokaaram). These issues have to be tackled separately. So I am taking (chillu Vs chandrakkala) alone right now.

In the primary example of section 1.2, the words differ precisely in /nn/ and /n_n/ location. We know, in the phoneme space they are different. That is, dental ന + dental ന and alveolar ന + dental ന. How is this distinction is indicated in orthography? Before explaining that, let us look at chandrakkala once more.

Chandrakkala when used for vowellessness, is acting as a language specific control character. It removes the default 'അ' from the consonant behind it. That is, it is acting as an attribute remover.

It is the property of any vowelless consonant to get 'help' from the consonant next to it as if that is a vowel and thus creating a conjunct. In /n_n/, this specific property is prevented. That is how alveolar n and dental n can stay close without undergoing conjunct forming transformations. How are we denoting the removal of this conjunct creation property? Are we using any specific symbols to indicate this? In fact, yes. It is a vertical tail across the letter.

Since chandrakkala exist as a separate orthographic entity detached from the letter, we can easy see its functionality. In contrast, conjunct-creation-preventer-symbol gets embedded in the orthography of a letter and that makes it difficult to recognize it.

So my conclusion (not a solution) is this: an Malayalam specific control symbol, different from chandrakkala, is present in a chillu letter. Its functionality is 1) remove the inherent അ vowel 2) then prevent the consonant from forming deep conjunct with the next letter. If we recognize, the function (1) alone, we will not solve the riddle of chillu.

This symbol could be encoded as level-1 ignorable in the collation table, as most of the diacritic marks are.

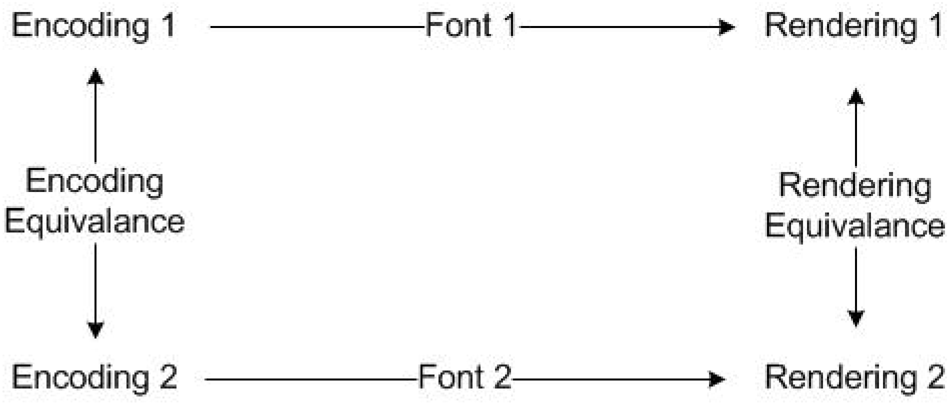
4 Final Bottom Line

Here are the two choices UTC will finaly need to choose from:

1. Encode Chillus as we did for ൔ /ma/ and handle the issue of equivalence to the base character at higher levels of text processing.
2. Encode Chillus other than ൔ /ma/ as Consonant + VIRAMA + ZWJ by overloading the format character ZWJ with Malayalam specific secondary or tertiary level collation weight. Also, both RA + VIRAMA + ZWJ and RRA + VIRAMA + ZWJ will represent exactly same chillu. Similarly, both LLA + VIRAMA + ZWJ and LLLA + VIRAMA + ZWJ represent same chillu. After all these, issue of correctness of the inputted text has to be handled at higher levels of text processing.

5 Uniqueness Rule

Consider this scenario:



Two encodings are equivalent if they differ only joiners. Encodings of 'ശബ്ദം' and 'ശബ്‌ദം' are equivalent because they are be different only by a ZWNJ. Meanwhile, encodings of 'അവൻ' and 'അവന്' are not because one has chillu letter, other hasn't.

Rendering equivalence is more of a subjective thing. We know 'ശബ്ദം' and 'ശബ്‌ദം' are equivalent renderings. 'അവൻ' and 'അവന്' are not equivalent renderings. 'അല്പം' and 'അൽപം' are sometimes considered equivalant, sometimes not. These pairs can not participate in this rule.

The fonts can vary from the new orthography font Nila from Bhasha Instituite to old orthography fonts like Anjali or Rachana. I don't think it is realistic to consider fonts which don't have lesser number of conjuncts than Nila.

Uniqueness Rule says:

If there is Encoding Equivalence then there should be Rendering Equivalence. (see details in section 5.1)
Also, if there is Rendering Equivalence then there should be Encoding Equivalence. (see details in section 5.2)

We can consider two versions of this rule. In the *lenient* version of this rule, at least one of the renderings should be *valid*. A rendering is *valid* when it is present in the dictionary or it is a word combination obeying grammar rules.

In the *aggressive* version of this rule, we consider all possible words, even those outside dictionary. This could be useful because these words can come from:

1. Colloquial phrases, often found in novels and stories.
2. Names of places, people etc.
3. Future words

5.1 Uniqueness Rule on Rendering

If a word is displayed correctly in one font then that word should be rendered correctly as the same word, in all fonts.

If Encodings 1 and 2 are exactly the same and Rendering 1 and 2 are different, then there is no guarantee that the text written by one will be readable to others.

If Encodings 1 and 2 are different by some joiners and Rendering 1 and 2 are different but valid, then there are following issues:

- search for rendering 1 will produce results with rendering 2
- search and replace would spoil the text
- Since typical language tools like spell checker or grammar checker is transparent to joiners (joiners are format controll characters consumed by shaping engine), we will need to make special tools which are joiners-aware for Malayalam.

If Encodings 1 and 2 are different by some joiners and one of the Renderings 1 and 2 is valid and other is invalid, then there is following issue:

- Since typical language tools like spell checker or grammar checker is transparent to joiners (joiners are format controll characters consumed by shaping engine), we will need to make special tools which are joiners-aware for Malayalam.

5.2 Uniqueness Rule on Encoding

Two different encodings should not render same, irrespective of the font or joiners used.

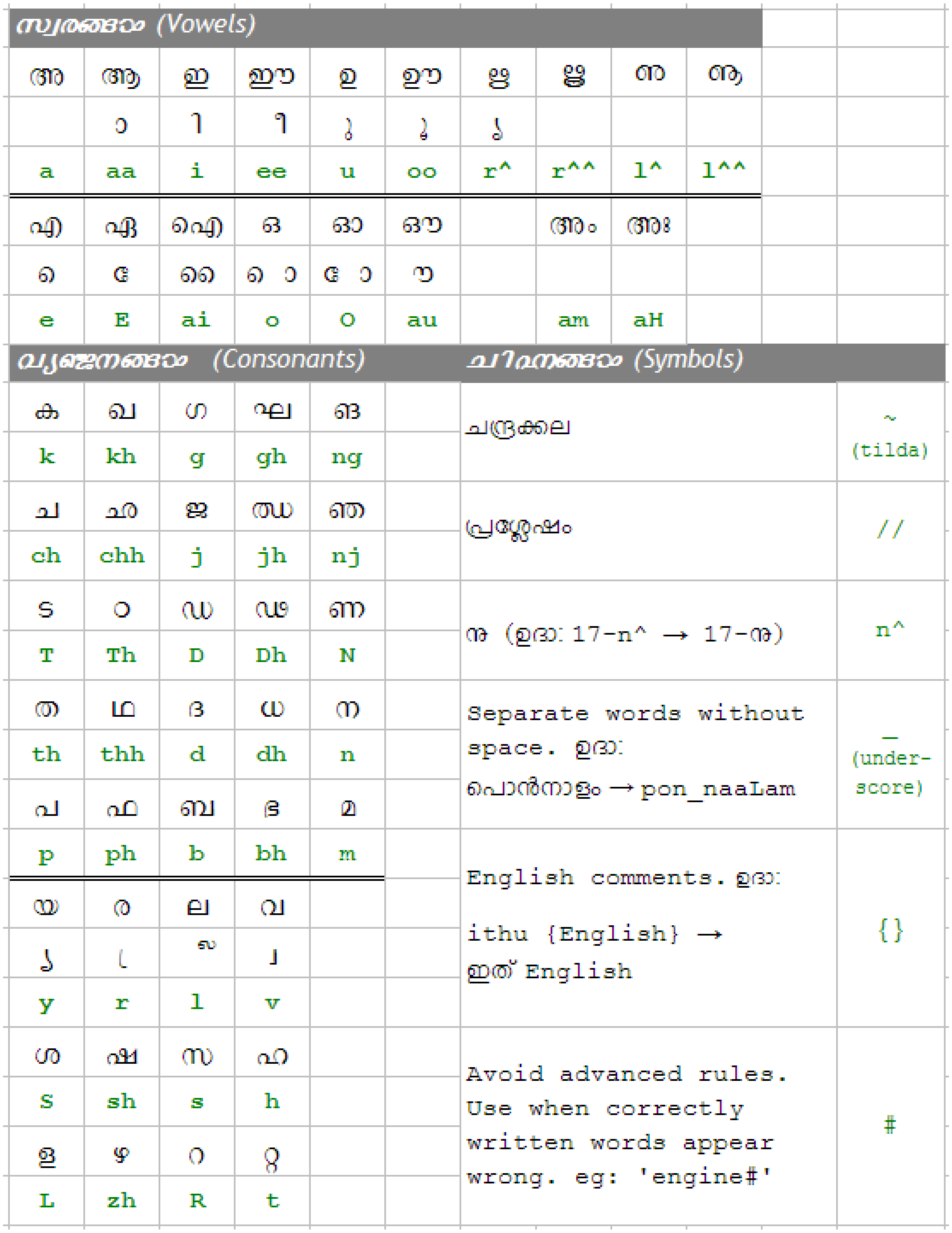
Two see why this rule is required, assume there is a conjunct formation rule for a subset of Chillu-C1 + C2 permutations and as per that rule, Chillu-NNA + DDHA (ൺ + ഢ) can form the conjunct (ണ്ഢ) in an old orthography font. Of course, NNA + VIRAMA + DDHA (ണ + ചന്ദ്രക്കല + ഢ) will also form the same conjunct.

There fore, a document (eg: a wiktionary.org document) written by multiple people using various inputting tools can quite possibly have both spellings for ണ്ഢ, without reader or writer being aware of it. This can cause many problems including ineffective searches and inconsistent sorted list of words.

Antoine Leca write this related text:
"Right now (for fifteen years really), we have a similar problem with Latin in Europe: our accentuated letters have two spellings, one which is the legacy one (using unique codepoints, for example U+00EE (î) which is the one everybody uses; and the other is the genuine Unicode encoding, the one we ought to use but nobody does in reality, using the base (English) letter and then another codepoint for the accent, i.e. U+0069, U+0302 (î)for my example above). You cannot normally see the difference, and if you do, it is just because of an imperfect Unicode support which does not render correctly the second form (things are getting better here, but still are not perfect). But if you are searching, the different spellings MAY be viewed as different, when of course it should not. Similarly, you could be allowed to enter both forms in a database field as "unique" key, when of course it should be prevented.

As this stuff is pretty evident to anyone in Europe developing in Unicode, this problems has been identified for years; and a "fix" has been developed, that is those two sequences are considered "canonically equivalent", so a "fully conforming" Unicode process should merge the two encodings for processes like searching or inserting. Please note that the majority of the tools used nowadays which deals with Unicode contents do not do that; only the tools specially prepared does it, and this comes with a noticeable performance impact."

6 Transliteration Scheme



സ്വരങ്ങൾ (Vowels)									
അ	ആ	ഇ	ഈ	ഉ	ഊ	ഋ	ൠ	ഌ	ൡ
	ാ	ി	ീ	ു	ൂ	ൃ			
a	aa	i	ee	u	oo	r^	r^^	l^	l^^
എ	ഏ	ഐ	ഒ	ഓ	ഔ		അം	അഃ	
െ	േ	ൈ	െ ാ	േ ാ	ൗ				
e	E	ai	o	O	au		am	aH	

വ്യഞ്ജനങ്ങൾ (Consonants)				
ക	ഖ	ഗ	ഘ	ങ
k	kh	g	gh	ng
ച	ഛ	ജ	ഝ	ഞ
ch	chh	j	jh	nj
ട	ഠ	ഡ	ഢ	ണ
T	Th	D	Dh	N
ത	ഥ	ദ	ധ	ന
th	thh	d	dh	n
പ	ഫ	ബ	ഭ	മ
p	ph	b	bh	m
യ	ര	ല	വ	
്യ	്ര	്ല	്വ	
y	r	l	v	
ശ	ഷ	സ	ഹ	
S	sh	s	h	
ള	ഴ	റ	റ്റ	
L	zh	R	t	

ചിഹ്നങ്ങൾ (Symbols)	
ചന്ദ്രക്കല	~ (tilda)
പ്രശ്ലേഷം	//
ൻ (ഉദാ: 17-n^ → 17-ൻ)	n^
Separate words without space. ഉദാ: പൊൻനാളം → pon_naaLam	_ (under-score)
English comments. ഉദാ: ithu {English} → ഇത് English	{}
Avoid advanced rules. Use when correctly written words appear wrong. eg: 'engine#'	#

-x is the symbol of vowel x

x_ is the chillu of x

m_ is the anuswara

~ is the virama