L2/07-279

Joiners (ZWJ/ZWNJ) with Semantic content for words in Indian subcontinent languages

N. Ganesan

This document gives examples of Unicode joiners, ZWJ and ZWNJ where the meanings of words differ substantially within Indian and Arabic scripts. The meanings of words depend totally upon whether ZWJ is used or not. Note particularly the existing Unicode sequences for India's languages like Marathi written in Devanagari script, Malayalam script. Since ZWJ carries semantic content in India's scripts, they need to be treated in collation of words, etc., in a given language's dictionary order. Several hundred examples exist in India's languages where the presence or absence of ZWJ is critical in semantics.

1.0 Canonical Equivalences for Atomic Chillus

Unicode rendering engines, fonts, … will continue to support the sequences for chillus in Malayalam, and eyelash repha in Devanagari indefinitely in the future. The situation of ZWJ in Malayalam and Devanagari scripts is quite the opposite of the deprecation done in Myanmar script,
http://www.unicode.org/notes/tn11/myanmar_uni-v2.pdf

In order to avoid destabilization and backward compatibility issues with existing and growing data in the web and implementations, atomic chillu letters, if encoded, can be given canonical equivalences with existing chillu sequences that use ZWJ.

CHILLU NN (U+0D7A) = <nna, virama, zwj>
CHILLU N (U+0D7B) = <na, virama, zwj>
CHILLU R (U+0D7C) = <ra, virama, zwj> [*]
CHILLU L (U+0D7D) = <la, virama, zwj>

CHILLU LL (U+0D7E) = <lla, virama, zwj>
CHILLU K (U+0D7F) = <ka, virama, zwj>

[*] U+0D7C is Chillu R first and foremost, Basing its name as Chillu RR on pronunciation is inappropriate for Malayalam script tradition. But ISCII and 100+ years tradition books on Malayalam script, plus Linux and Microsoft implementations has Chillu R as <RA, VIRAMA, ZWJ> (not as <RRA, VIRAMA, ZWJ>). Elementary schools in Kerala teach it as *chillu ir*, meaning Chillu R. Unicode name for U+0D7C has to respect that old tradition of naming the chillus. A. R. Rajarajavarma, in his Keralapaniniyam, an important grammar for Malayalam written in 1917, lists the 5 most important chillus with L, LL, R, N, NN. (References: University of Kerala transation of Keralapaniniyam into English, 1999. Also, see the last two pages of this pdf.)

In the case of Malayalam, there is a possibility of a subset of Chillus getting encoded atomically. Already, these chillus are available and will be used by developers, and users in the future more widely in software, fonts and in the web as sequences *<cons, virama, zwj>* with *cons = ll,l, r, n, nn*. As users find more archaic chillus (*m, s, k, y* & so on) the chillu sequences can be expanded to be inclusive. To avoid destabilization and backward compatibility problems in Unicode with vast and existing data in Malayalam, it is recommended that atomic chillus, if encoded, be given *canonical equivalences* with existing chillu sequences that employ ZWJ.

2.0 ZWNJ and ZWJ use/non-use determines Word Meanings

(a) ZWNJ in Farsi (Persian)

Farsi <Noon, Alef, Meem, Heh, Alef, Farsi Yeh>. Without a ZWNJ, it translates to "names"; with a ZWNJ between Heh and Alef, it means "a letter". Figure 1 illustrates this.

Figure 1.

	Code Points	Names (abbreviated)
نامهای	0646 + 0645 + 0627 + 0647 + 0645 + 06CC	NOON + ALEF + MEEM + HEH + ALEF + FARSI YEH
نامه‌ای	0646 + 0645 + 0627 + 0647 + 200C + 0645 + 06CC	NOON + ALEF + MEEM + HEH + ZWNJ + ALEF+ FARSI YEH

(b) ZWNJ in Tamil

Like Farsi, words in many of India's scripts depend for their meaning on whether ZWNJ is used or not. Let me take the example of Tamil. Western (Christian) and Muslim names and English loans are written with a ZWNJ, writing *'kṣ'* in a conjunct form without ZWNJ is not correct in Tamil script.

Without ZWNJ

பக்ஷி 'bird'

<0BAA 0B95 0BCD 0BB7 0BBF>

With ZWNJ

பக்ஷி 'Name of a Muslim person'

<0BAA 0B95 0BCD 200C 0BB7 0BBF>

English loans are written with ZWNJ:

செக்ஷன் (= 'section' from English)

<0B9A 0BC6 0B95 0BCD 200C 0BB7 0BA9 0BCD>

ZWNJ is needed for all English, etc., words written in Tamil script. Otherwise, incorrect forms of Tamil words appear.

(c) ZWJ in Malayalam

Cibu gave a constructed Malayalam example:

Without ZWJ:
വന്യവനിക vanya-vanika 'thick forest'
<0D35 0D28 0D4D 0D2F 0D35 0D28 0D3F 0D15>

With ZWJ
വൻയവനിക van-yavanika 'big curtain'
<0D35 0D28 0D4D 200D 0D2F 0D35 0D28 0D3F 0D15>

Note the word meaning very much depends upon the presence or absence of ZWJ, ZWNJ (for Farsi, Tamil, Malayalam, Marathi, Konkani, Newari or Nepali words).

(d) ZWJ in Marathi

Example 1:
आचार्यास "to the teacher" (without ZWJ)
आचाऱ्यास "to the cook" (with ZWJ)

Example 2:

दर्या "ocean" (without ZWJ) <0926 0930 094D 092F 093E>
दऱ्या "valleys" (with ZWJ) <0926 0930 094D 200D 092F 093E>

(e) ZWJ in Nepali

When the symbol र stands for the phoneme /r/ that is in the onset position of the second syllable as in CV.CGV structure, the /r/ is represented by symbol - , e.g. पऱ्यो *pa.ryo* 'it fell in', गऱ्यो *ga.ryo* 'he did it'

When the symbol र stands for the phoneme /r/ that is in the coda position of the the first syllable as in CVC.CV, the /r/ is represented by the symbol ◌् e.g. गर्न *gar.na* 'to do (inf.)', मर्न *mar.na* 'to die (inf.)'.

It is *not* correct in phonetics to use Dravidian RRA for these words. The contrast in word meanings in Devanagari is generated by ZWJ operating in a sequence, <RA, VIRAMA, ZWJ> & not RRA which is practically unknown in core Devanagari areas. RRA is just used to

transliterate the Dravidian words. Eyelash RA examples are quite different and have nothing to do with Dravidian/Tamil letter, RRA linguistically. Eyelash RA needs letter, RA in coded representation (e.g. दऱ्या "valleys" (with ZWJ)).

3.0 Semantic ZWNJ and ZWJ needed in more scripts (not just Sinhala)

http://www.alvestrand.no/pipermail/idna-update/2007-June/000622.html

Ken Whistler writes.," ZWJ seems only required in the Sinhala script.

That is not to say that ZWNJ and ZWJ aren't much more widely used in the Arabic script and in many Indian scripts for presentational purposes -- but the few instances above are the only ones we currently know about where important semantic distinctions require the presence of a ZWNJ or a ZWJ to be "spelled" correctly, from the point of view of an end user. "

As shown in this document, ZWJ is used in words in many languages of South Asia with semantic content. In addition to Sinhala, consider Marathi, Konkani, Nepali, Newari and Malayalam languages as well for semantic ZWJ. Like Farsi ZWNJ in Arabic script, Tamil, Malayalam, … need semantic ZWNJ also. Indian ZWJ (which includes Sinhala, Devanagari, Malayalam) can be treated in collation per dictionary order. In e-mail exchanges or in the web pages, ZWJ and ZWNJ joiners cannot be stripped from these words in India's languages just as Farsi words should not lose the ZWNJ to make sense of the meaning.

Ambiguities increase due to Atomic Chillus (Duplicate encoding)

A mail from Dr. Whistler to indic unicode list 2 years ago,
This is not adequately answered before Atomic Chillus are encoded.

From Ken Whistler k...@sybase.com Wed Aug 24 19:50:42 2005

[Begin Quote]
I agree that if you have distinct lexemes with distinct renderings (pronunciation is really irrelevant to the character encoding), you ought to have distinct character representations. But the problem is in the presupposition that a representation using ZWJ does not constitute a distinct character representation.

If there were a canonical equivalence involved, that presupposition would be correct. But in the absence of a canonical equivalence, the whole issue hinges on the interpretation of what "semantic" distinctions a ZWJ or ZWNJ should be expected to carry for a particular script.

The discussion of Chillus for Malayalam has continually come back to quoting the passage on p. 391 of TUS 4.0 that talks about ZWJ and ZWNJ as being ignored by processes that analyze text content. The problem is that that passage is in the middle of a long discussion about cursive joining in Arabic -- the original context for which ZWJ and ZWNJ were encoded -- where ZWJ and ZWNJ generally do not carry semantic distinctions. Even for the Arabic script, however, that is not entirely the case, because there are known situations for Persian where the presence or absence of a ZWNJ *can* carry significance in text. And furthermore, the paragraph immediately above the oft-cited text states:

"The ZWJ and ZWNJ also have specific interpretations in certain scripts as specified in this standard. ..."

That, of course, is an explicit reference to the fact that ZWJ and ZWNJ are used to make other distinctions in Indic scripts that are not merely controls over cursive connections between letters.

So while it is obvious that the standard as currently written is insufficiently precise about what distinctions ZWJ and ZWNJ *can* make in Indic scripts, for example, it is not the case that the intent of the standard is to preclude them from being able to make any distinctions for text processing. Making that determination more precise is exactly what Eric Muller and the rest of the UTC are wrestling over at this moment, so that it will be clear for the Indic scripts, in particular, what distinctions can or cannot be made with ZWJ and ZWNJ and how they differ in usage in Indic scripts from their use in the Arabic script. Only when such distinctions are clearly spelled out can general-purpose text processors such as internet search engines put in place the algorithms that will do the right thing for Arabic, but *also* do the right thing for Devanagari or Malayalam, for example.

Finally, as a kind of counter-challenge for Cibu, I need to point out the following.

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:

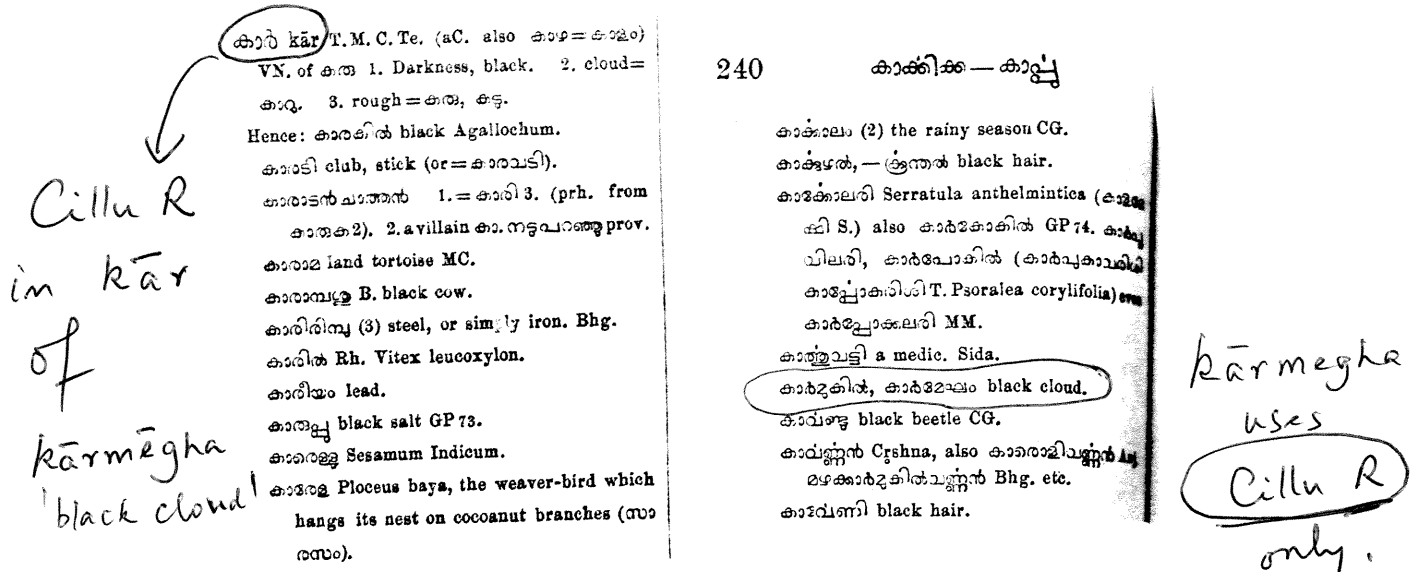
"ന്ന" is <U+0D28, U+0D4D, U+200D, U+0D28>

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.

Regards to all,
--Ken
[End Quote]

Figure 1. 'kaarmeegham' in H. Gundert's Malayalam-English dictionary.
(Note the transliteration r in kaar-megham. r = Unicode letter R)



കാർ kār T. M. C. Te. (aC. also കാഴ = കാളം) VN. of കരു 1. Darkness, black. 2. cloud = കാറു. 3. rough = കരു, കട്ട.
Hence: കാരകിൽ black Agallochum.
കാരടി club, stick (or = കാരവടി).
കാരാടൻ ചാത്തൻ 1. = കാരി 3. (prh. from കാരുകം 2). 2. a villain കാ. നട്ടപാഞ്ഞു prov.
കാരാമ land tortoise MC.
കാരാമ്പശു B. black cow.
കാരിരിമ്പു (3) steel, or simply iron. Bhg.
കാരിൽ Rh. Vitex leucoxylon.
കാരീയം lead.
കാരുപ്പു black salt GP 73.
കാരെള്ള Sesamum Indicum.
കാരേള Ploceus baya, the weaver-bird which hangs its nest on cocoanut branches (സാരസം).

240 കാക്കിക്ക — കാപ്പു

കാക്കാലം (2) the rainy season CG.
കാക്കുഴൽ, — ക്കുന്തൽ black hair.
കാക്കോലരി Serratula anthelmintica (കാമ്പു ക്കി S.) also കാർകോകിൽ GP 74. കാർപു ിലരി, കാർപോകിൽ (കാർപുകാരിക്കി കാപ്പോകാടി T. Psoralea corylifolia) കാർപ്പോക്കലരി MM.
കാത്തുവട്ടി a medic. Sida.
കാർമുകിൽ, കാർമാഘം black cloud.
കാവണ്ട black beetle CG.
കാവ്ണ്ണൻ Cṛshna, also കാരാമിവണ്ണൻ മഴക്കാർമുകിൽവണ്ണൻ Bhg. etc.
കാവുണി black hair.

Cillu R in kār of kārmēgha 'black cloud'

kārmegha uses Cillu R only.

Figure 2. Malayalam Cillaksharams
(K. P. Mohanan, "Malayalam writing", Peter Daniels & William Bright, The World's writing systems, p. 422, 1996)

TABLE 38.4: *Cillakṣarams*

ണ	ṇ >	ൺ	ചാണ്	cāṇ	'handbreadth'
ന	n >	ൻ	അവൻ	avan	'he'
ര	r >	ർ	മലർ	malar	'popped rice'
ല	l >	ൽ	പകൽ	pakal	'day'
ള	ḷ >	ൾ	അവൾ	avaḷ	'she'

Cillu R

(Note: no cillu RR in the table)

Figure 3. From page 4, Frohnmeyer, L. J., A progressive grammar of the Malayalam language . New Delhi : Asian Educational Services, 1979 (2nd edition)

§ 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).
ൽ „ „ ല (l „ „ la).
ൾ „ „ ള (l „ „ la).

അവൻ (avaṉ) he, that person; ഇവൻ (ivaṉ) this person; ഊൺ (ūṇ) meal; മൺ (maṇ) earth, soil; അവർ (avar) they, those persons; ഇവർ (ivar), these persons; മണൽ (maṇal) sand; കടൽ (kaḍal) sea; ആൾ (āḷ) man; അവൾ (aval) she; മരം (maram) tree; പണം (paṇam) money; മരണം (maraṇam) death.

(1) Figure 4. From page 92, Reinhold Grunendahl, South Indian Scripts in Sanskrit manuscripts and prints: Grantha Tamil - Malayalam - Telugu - Kannada - Nandinagari. (2001: Harrassowitz Verlag, Wiesbaden, Germany).

Prepausal Consonants

k	ക്ക് & ക്ക് & ഺ	r	ർ & ർ
ṇ	ൺ & ൺ	l	ൽ & ൽ
t	ൽ & ൽ & ൽ	ḷ	ൿ & ൿ
n	ൻ & ൻ	ḷ	ൟ
m	ൣ		

Cillu R