L2/14-005

On the Telugu and Kannada Vowel Signs for O and OO

Shriramana Sharma, jamadagni-at-gmail-dot-com, India

2014-Jan-05

The Telugu and Kannada scripts exhibit two different ways of marking the dependent vowels O and OO (long O). This document documents this variation and can serve as a Unicode Technical Note on this matter. Text input, whether automated (OCR) or manual, needs to take this variation into consideration. First I describe the attested orthography, and then I discuss how Unicode can/should support the attested orthography.

Orthography

In both scripts, the variation for the short vowel O is between a one-part vowel sign to the top-right of the consonant and a two-part vowel sign, one to the top-right and one to the right. For the long vowel OO, both scripts add different length marks in different cases.

The following table summarizes the various forms:

	Telugu		Kannada	
	one-part	two-part	one-part[1]	two-part
ka	క		ಕ	
kŏ kō	కొ కో	కెు కెూ	ಕೊ ಕೋ[2]	ಕೊ ಕೋ[2]

General rule for the the short vowel ŏ:

1) The one-part vowel signs in both scripts are ◌ొ in Telugu and ◌ೊ in Kannada.
2) In Telugu, the two-part vowel sign is a composition of the vowel signs for short *ĕ* ◌ె and short *u* ◌ు. Thus for KA క we have:

[1] Attestation for the Kannada one-part signs is to be found in my separate document L2/14-004.

[2] Strictly speaking, the Kannada vowel signs for the long vowel ō are two-part and three-part respectively since they include the length mark ◌ೕ but I have ignored that for simplicity of terminology.

క + ◌ె + ◌ు → కొ

3) In Kannada, it is a composition of the vowel signs for short *ĕ* ◌ೆ and long *ū* ◌ೂ. Thus for KA ಕ we have:

ಕ + ◌ೆ + ◌ೂ → ಕೊ

General rule for the long vowel *ō*:

1) In Telugu, the one-part vowel sign for *ŏ* ◌ొ takes the length mark ◌ౕ to become ◌ో:

కొ + ◌ౕ → కో

2) The two-part vowel sign ◌ొ takes the sign for *ā* ◌ా as a length mark to become ◌ో:

కొ + ◌ా → కో

One may also analyse this as taking the sign for long *ū* ◌ూ instead of that of short *u* ◌ు for the second part:

క + ◌ె + ◌ూ → కో

3) In Kannada one simply adds the length mark ◌ೕ to the vowel sign for short *ŏ*, whether it is one part or two part. Thus we have:

ಕೊ + ◌ೕ → ಕೋ

ಕೊ + ◌ೕ → ಕೋ

Exceptions:

The main exception to the above rules is in the case of consonants which have the shape ుు as the right-most element and do not place the headstroke on the right-most ు.

In Telugu these are the consonants GHA ఘ, JHA ఝ, MA మ and YA య.

1) These rarely[3] use the single-part vowel signs for *ŏ* and *ō*.
2) Thus, in general, only the regular two-part form ◌ొ is found for the short vowel *ŏ* :

gha ఘ *ghŏ* (ఘొ) ఘొ

jha ఝ *jhŏ* (ఝొ) ఝొ

ma మ *mŏ* (మొ) మొ

[3] Campbell in p 10 of his 1849 Telugu Grammar (https://archive.org/details/grammarofteloogo00camprich) asserts that these *never* use the single-part form, but Arden in p 19 of his 1905 grammar (https://archive.org/details/aprogressivegra00ardegoog) shows the single-part form for GHA and JHA. Possibly other older sources such as manuscripts one may come across the same for MA and YA too.

ya య *yŏ* (యా) యొ

3) For the two-part form of the long vowel ō, the rule of *adding* the sign for *ā* ◌ా to the short *u* ◌ు to effectively make it long *ū* ◌ూ i.e. using ◌ెూ is discouraged[4]. Rather, ◌ా directly *replaces* the ◌ు to give ◌ెా for *ō* :

ghō	ఘ + ◌ె + ◌ా	→	ఘెా	(rarely ఘెూ or ఘో)
jhō	ఝ + ◌ె + ◌ా	→	ఝెా	(rarely ఝెూ or ఝో)
mō	మ + ◌ె + ◌ా	→	మెా	(rarely మెూ or మో)
yō	య + ◌ె + ◌ా	→	యెా	(rarely యెూ or యో)

Note that this still has a right-most element like long *ū* ◌ూ due to the consonant's ు .

Apart from these consonants, sometimes SA also[5] uses ◌ెా for *ō* :

sa స *sŏ* సొ సెు *sō* సో సెూ సెా [6]

HA హ which uses the length mark ◌ౕ and not ◌ా for indicating -ā (as it already has a ◌ా in its body) also sometimes replaces the ◌ు from ◌ొ by the length mark to give ◌ే for *ō* :

ha హ *hŏ* హొ హెు *hō* హో హెూ హే

In Kannada, exceptions to the general rule occur for JHA ಝ, MA ಮ and YA ಯ. (Note that this omits GHA ಘ from the Telugu list since in Kannada it does not have the ು right-side component.) Since Kannada consistently uses the length mark ◌ೕ to mark the long vowel ō, the description below relates only to the short vowel ŏ.

1) The two-part form of *ŏ* uses the vowel sign *ā* ◌ಾ instead of ū ◌ೂ for its second part. That is, the two-part form is a composition of *ĕ* ◌ೆ and *ā* ◌ಾ.
2) The combination of the right-most element ು of the consonant with this *ā* ◌ಾ may be cursively written identical to the *ū* ◌ೂ[7].

[4] Charles Brown in p 16 of his 1857 Telugu grammar (http://books.google.co.in/books?id=pnAIAAAAQAAJ) calls this a practice of "uneducated persons" but Grierson in his 1927 Linguistic Survey of India Vol IV p 585 (https://archive.org/details/rosettaproject_tel_ortho-2) shows *mō* and *yō* with ◌ెూ.

[5] Only Campbell mentions this.

[6] Note the slight difference from హె *hĕ* in the absence of the loop in the bottom left.

[7] Indeed, parallel to Telugu, the vowel sign *ū* ◌ೂ in Kannada is itself clearly only a (cursive) ligature of *u* ◌ು and *ā* ◌ಾ. And just as in Telugu, this two-part form makes it appear as if the consonant still retains a separate *ū* ◌ೂ as the second part as for the other consonants.

jhŏ ಝ + ◌ೆ + ◌ಾ → ಝೆಾ/ಝೊ

mŏ ಮ + ◌ೆ + ◌ಾ → ಮೆಾ/ಮೊ[8]

yŏ ಯ + ◌ೆ + ◌ಾ → ಯೆಾ/ಯೊ

Possibly in manuscripts and other older writings even more deviations from the general rule are seen. I have only attempted to summarize the more systematic among them.

Encoding model

Existing encoded characters

Despite the above multifarious (albeit systematic) variations, the existing encoding model for the vowel signs O and OO in Telugu and Kannada is quite simplistic. Only two characters are encoded for each script:

0C4A	◌ొ	TELUGU VOWEL SIGN O
0C4B	◌ో	TELUGU VOWEL SIGN OO
0CCA	◌ೊ	KANNADA VOWEL SIGN O
0CCB	◌ೋ	KANNADA VOWEL SIGN OO

Evidently, only the single-part form for Telugu and the two-part form for Kannada are encoded as these are the most prevalent in current use.

Canonical decomposition

Here, the long vowel for Kannada 0CCB ◌ೋ is given a canonical decomposition to its graphical components 0CCA ◌ೊ 0CD5 ◌ೕ i.e. the short vowel followed by the length mark. However, the long vowel in Telugu 0C4A ◌ో was, for whatever reason, *not* given the parallel decomposition to 0C4A ◌ొ 0C55 ◌ౕ even though it should have been[9].

The short vowel in Kannada 0CCA ◌ೊ is given a further decomposition to its own components: 0CC6 ◌ೆ 0CC2 ◌ೂ.

[8] Note the difference from *vŏ* ವೊ in the attachment of the ೂ shape.

[9] Canonical decomposition is also not provided for the other long Telugu vowels 0C40 ◌ీ II or 0C47 ◌ే EE even though these are also graphically formed from their respectively short vowels and the length mark. On the whole, 0C55 TELUGU LENGTH MARK ◌ౕ is not of use for known orthographies (Telugu, Sanskrit etc) in Telugu. For these reasons, it would be safe and advisable to prohibit this character from use in IDNs.

The forms that are not encoded

The two-part forms for the short and long vowels in Telugu: ◌ెు and ◌ెూ (◌ెా for GHA etc) are not encoded separately but can be composed as sequences of their components: i.e. 0C46 ◌ె 0C41 ◌ు and 0C46 ◌ె 0C42 ◌ూ (0C46 ◌ె 0C3E ◌ా for GHA etc).

However the single-part forms in Kannada ◌ೊ and ◌ೋ cannot be composed from existing characters and are hence proposed for encoding in my document L2/14-004.

Confusability issues with current practice

In Telugu, as mentioned, only the single-part forms are encoded. However, the consonants GHA ఘ, JHA ఝ, MA మ and YA య rarely use the single-part form. As a result, most fonts, including the default Telugu fonts on popular operating systems, always produce the two-part form even from the character which nominally represents the single-part form:

ghŏ	ఘ + ◌ొ	→	ఘొ	(= ఘ + ◌ెు)	*ghō*	ఘ + ◌ో	→	ఘో	(= ఘ + ◌ెా)
jhŏ	ఝ + ◌ొ	→	ఝొ	(= ఝ + ◌ెు)	*jhō*	ఝ + ◌ో	→	ఝో	(= ఝ + ◌ెా)
mŏ	మ + ◌ొ	→	మొ	(= మ + ◌ెు)	*mō*	మ + ◌ో	→	మో	(= మ + ◌ెా)
yŏ	య + ◌ొ	→	యొ	(= య + ◌ెు)	*yō*	య + ◌ో	→	యో	(= య + ◌ెా)

Obviously, the glyphs being produced are identical to those that would be produced by the sequences for the two-part forms.

I am given to understand that from a security viewpoint in relation to IDNs, it is not advisable to permit different encoded sequences that are not canonically equivalent to have the same final rendered shape. Given this, it is not clear if this confusability issue was taken into consideration when these system fonts were designed. Of course, IDNs are far from being widely accepted, but the potential issue remains, and is exacerbated by the fact that Unicode security mechanisms currently do not support multi-character confusables which the present situation entails. I hence feel obliged to point out this security issue.

The usage of the sequences 0C46 ◌ె 0C41 ◌ు etc for the two-part forms is quite legitimate and cannot be prohibited. Thus the fact that system fonts render 0C4A ◌ొ and 0C4B ◌ో identical to those sequences for the two-part forms is indeed a security risk.

Pedantically speaking, the recommended sequence/rendering correlation would be:

ghŏ	ఘ + ◌ొ → ఘొ	*ghō*	ఘ + ◌ో → ఘో
jhŏ	ఝ + ◌ొ → ఝొ	*jhō*	ఝ + ◌ో → ఝో
mŏ	మ + ◌ొ → మొ	*mō*	మ + ◌ో → మో
yŏ	య + ◌ొ → యొ	*yō*	య + ◌ో → యో

... though these shapes are less common[10]. This will force the use of the correct sequences 0C46 ◌ె 0C41 ◌ు (, 0C46 ◌ె 0C42 ◌ూ) and 0C46 ◌ె 0C3E ◌ా to get the two-part forms:

ghŏ	ఘ + ◌ె + ◌ు → ఘెు	*ghō*	ఘ + ◌ె + ◌ా → ఘెా
jhŏ	ఝ + ◌ె + ◌ు → ఝెు	*jhō*	ఝ + ◌ె + ◌ా → ఝెా
mŏ	మ + ◌ె + ◌ు → మెు	*mō*	మ + ◌ె + ◌ా → మెా
yŏ	య + ◌ె + ◌ు → యెు	*yō*	య + ◌ె + ◌ా → యెా

This applies to SA స and HA హ as well:

sŏ	స + ◌ొ → సొ	*sō*	స + ◌ో → సో
hŏ	హ + ◌ొ → హొ	*hō*	హ + ◌ో → హో
sŏ	స + ◌ె + ◌ు → సెు	*sō*	స + ◌ె + ◌ూ → సెూ
		sō	స + ◌ె + ◌ా → సెా
hŏ	హ + ◌ె + ◌ు → హెు	*hō*	హ + ◌ె + ◌ూ → హెూ
		hō	హ + ◌ె + ◌ా → హెా[11]

[10] Note that I only recommend this for system fonts in view of security issues. I am fully aware of heavy legacy use of 0C4A and 0C4B for the two-part signs of GHA etc. As such, I am not sure if popular fonts like Pothana would change their shaping for these technical/pedantic reasons. Hence the more practical and less intrusive alternative would be to add the sequences to a future multi-character version of the confusables database.

[11] In Telugu, since *ha* హ already has a feature similar to *ā* ◌ా, the long syllable *hā* is represented by adding the length mark ◌ౕ to get హా. Thus, fonts normally ligate 0C39 హ HA + 0C3E ◌ా AA to హా even though this would be graphically equivalent to 0C39 హ HA + 0C55 ◌ౕ LENGTH MARK. This confusability is not an issue since, as mentioned in footnote 2 on p 4, 0C55 ◌ౕ is unnecessary for normal Telugu orthography, causes other

In Kannada, the more common forms for *jhŏ*, *mŏ* and *yŏ* are ಝೊ, ಮೊ and ಯೊ. There is no problem with the consonants taking the regular vowel sign 0CCA ◌ೊ to give these forms:

jhŏ ಝ + ◌ೊ → ಝೊ

mŏ ಮ + ◌ೊ → ಮೊ

yŏ ಯ + ◌ೊ → ಯೊ

The one thing to be noted is the elision of one ು from the right side of the consonant along with the ligation. The less common forms *jhŏ* ಝೆಾ, *mŏ* ಮೆಾ, *yŏ* ಯೆಾ should be formed using 0CC6 ◌ೆ + 0CBE ◌ಾ:

jhŏ ಝ + ◌ೆ + ◌ಾ → ಝೆಾ

mŏ ಮ + ◌ೆ + ◌ಾ → ಮೆಾ

yŏ ಯ + ◌ೆ + ◌ಾ → ಯೆಾ

Of course, the long vowels in Kannada can always be formed using 0CD5 ◌ೕ.

I submit these recommendations to help a consistent representation of these written forms in Telugu and Kannada and also help reduce security risks related to confusability[12]. (See also the following appendix.)

Thanks

I wish to thank Srinidhi of Tumkur, Karnataka, for sending me the attestation samples for the single-part Kannada vowel signs O and OO thereby inciting me to write this document.

I also thank Elmar Kniprath of Hamburg, Germany for providing attestations from Arden, Grierson etc.

-o-o-o-

confusables too, and is hence better prohibited from IDNs. Thus it is also valid to use 0C39 హ HA + 0C46 ◌ె E + 0C3E ◌ా AA for హో. Note also that, if 0C55 ◌ౕ were to be used for హో from a purely graphical viewpoint, the resultant sequence హ + ◌ె + ◌ౕ would be ambiguous between the desired *hō* హో and *hē* హే (= హ + 0C47 ◌ే).

[12] In L2/10-416, the minutes of the 2010 Nov UTC meeting, under D.3.1 one finds mention of AIs for the Gov’t of India, the Gov’t of Andhra Pradesh and for the Editorial Committee to work to identify confusables for Telugu. It is unclear as to what progress has been made in this regard. This document, especially in its footnotes and including its appendix, identifies all the Telugu and Kannada confusables that I am aware of.

Appendix - Telugu and Kannada confusables involving VS-I/II

In Telugu, ◌ి is the vowel sign for short *i*. Long *ī* is represented by graphically adding the length mark ◌ౕ to give ◌ీ. Thus for *ka* క, *ki* is కి and *kī* is కీ. However, for consonants that have a right-side element ు but do not attach their short *i* sign ◌ి onto it, one finds[13] that instead of the length mark ◌ౕ being used, *ā* ◌ా may be added to the ు for the long *ī*. So:

gha	ఘ	*ghi*	ఘి	*ghī*	ఘీ	ఘిా
jha	ఝ	*jhi*	ఝి	*jhī*	ఝీ	ఝిా
ma	మ	*mi*	మి	*mī*	మీ	మిా
ṣa	ష	*ṣi*	షి	*ṣī*	షీ	షిా
sa	స	*si*	సి	*sī*	సీ	సిా

As usual, *ha* హ would take the length mark ◌ౕ attached to the consonant:

ha	హ	*hi*	హి	*hī*	హీ	హిౕ

In case of *ya* య, the regular form is not attested since the ligature of య with ◌ి gives యి (i.e. both the headstroke and vowel sign disappear) and this does not exhibit the required anchor for the length mark to attach at all!

ya	య	*yi*	యి	*yī*	యిా

In Kannada also[14], the consonants JHA, MA and YA show a similar behaviour, where the right-most ು of the consonant may ligate with the *ā* ◌ಾ to give ೂ:

jhī	ಝ + ◌ಿ + ◌ಾ → ಝಿಾ / ಝಿೂ	usu.	ಝ + ◌ೀ → ಝೀ
mī	ಮ + ◌ಿ + ◌ಾ → ಮಿಾ / ಮಿೂ	usu.	ಮ + ◌ೀ → ಮೀ
yī	ಯ + ◌ಿ + ◌ಾ → ಯಿಾ / ಯಿೂ	usu.	ಯ + ◌ೀ → ಯೀ

The confusability may be inferred as before from the above information.

[13] As per Campbell (link previously given), p 10. Note that graphically *pa* ప and *pha* ఫ also qualify, but they do not seem to be attested to exhibit this behaviour, probably since appending ◌ా to them makes them identical (in case of *pa* ప) or highly similar (in case of *pha* ఫ) to *ha* హ. (This is probably why *pā* పా etc make the right stroke intersect and rise above the vowel sign.) Also note that all these graphical criteria (such as having a right-side ు etc) mentioned in this entire document are deduced by me from analysing the behaviour of the consonants and I have not seen them mentioned in any texts.

[14] See Kittel (link previously given), p 18. Note that he does not show all forms for all consonants.