Background of Indic segmentation

(Input for UAX#29 technical report)

India has large linguistic diversity with 22 constitutionally recognized languages and 12 scripts. The mapping between languages and scripts is complex. Multiple languages may have common scripts, while a language can be written in multiple scripts. Each language and script is unique in nature and cannot be easily replicated, even if they share common characteristics. The orthographic changes may also occur in some languages and adoption of new orthography is a gradual process, thus posing additional challenges. However for segmentation purpose , all Indic Languages having phonetic correspondence at the alphabet level by and large follow a common definition of syllable across languages.

ABNF Indic Orthographic syllable definition:

V[m] I {CH}C [v][m] I CH

The definition is a combination of 3 rules:

Rule 1 : V[m]

Rule 2 : {CH}C[v][m]

Rule 3 : CH (This rule is applicable only at the end of the word)

V(upper case) is independent vowel

m is modifier(Anusvara/Visarga/Chandrabindu)

C is a consonant which may or may not include a single nukta

v (lower case) is any dependent vowel or vowel sign (mātrā)

H is halant / virama

| is a rule separator

[ ] - The enclosed items is optional under this bracket

{} - The enclosed item/items occurs zero or repeated multiple times

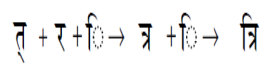
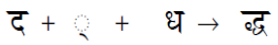
Testing of Devanagari data (Hindi, Marathi, Konkani, Nepali) using Unicode segmentation utility tool was carried out. The cases where utility tool does not give correct segmentation break points are tabulated and the correct desired out as per ABNF Orthographic Indic syllable definition is given below.

Devanagari :

Sl. No.	Rendering rules defined by Unicode	Words	Rendering as per Unicode segmentation utility tool (http://unicode.org/cldr/utility/breaks.jsp)	Segmentation as per Indic syllable definition
Devanagari Script				
Conjunct Formations				
1.	Example : ग् + ध → ग्ध क् + ष → क्ष क् + क → क्क	संदिग्ध	सं|दि|ग्|ध|	सं दि ग्ध Cm Cv CHC
		सुरक्षा	सु|र|क्|षा|	सु र क्षा Cv C CHCv
		टक्कर	ट|क्|क|र|	ट क्क र C CHC C
		सत्र	स|त्|र|	स त्र C CHC
		दक्षिण	द|क्|षि|ण|	द क्षि ण C CHCv C
		मिश्रणाने	मि|श्|र|णा|ने|	मि श्र णा ने Cv CHC Cv Cv
		मिनटापेक्षाही	मि|न|टा|पे|क्|षा|ही|	मि न टा पे क्षा ही Cv C Cv Cv CHCv Cv
		कुरुक्षेत्रांतल्या	कु|रु|क्|षे|त्|रां|त|ल्|या|	कु रु क्षे त्रां त ल्या Cv Cv CHCv CHCvm C CHCv

		क्षेत्रहरूमा	क्‌|षे|त्‌|र|ह|रू|मा|	क्षे त्र ह रू मा CHCv CHC C Cv Cv
Rules for rendering				
2.	If the dead consonant precedes a consonant, then it is replaced by the superscript nonspacing mark (U+0930 devanagari letter ra) positioned above or attached to the upper part of a base glyph form. Example : र्‌ + क → क + ◌	दुर्घटनाग्रस्त	दु|र्‌|घ|ट|ना|ग्‌|र|स्‌|त|	दु र्घ ट ना ग्र स्त Cv CHC C Cv CHC CHC
		मुहूर्त	मु|हू|र्‌|त|	मु हू र्त Cv Cv CHC
		शर्करायुक्त	श|र्‌|क|रा|यु|क्‌|त|	श र्क रा यु क्त C CHC Cv Cv CHC
		वर्सांतल्यान	व|र्‌|सां|त|ल्‌|या|न|	व र्सां त ल्या न C CHCvm C CHCv C
3.	For certain consonants, the mark (U+0930 devanagari letter ra positioned below or attached to the lower part of a base glyph form) may graphically combine with the consonant to form a conjunct ligature form. Example : फ् + र → फ + ◌्	अंतरराष्ट्रीय	अं|त|र|रा|ष्‌|ट्‌|री|य|	अं त र रा ष्ट्री य Vm C C Cv CHCHCv C
		राष्ट्रपति	रा|ष्‌|ट्‌|र|प|ति|	रा ष्ट्र प ति Cv CHCHC C Cv
		फ्रांस	फ्‌|रां|स|	फ्रां स CHCvm C
		ट्रैक्टर	ट्‌|रै|क्‌|ट|र|	ट्रै क्ट र

		सिट्रोनेलाचे टुक्रालाई	सि|ट्|रो|ने|ला|चे| टु|क्|रा|लाई|	CHCv CHC C सि ट्रो ने ला चे Cv CHCv Cv Cv Cv टु क्रा ला ई Cv CHCv Cv V
Ligature rules				
4.	If a dead consonant immediately precedes another dead consonant or a live consonant, then the first dead consonant may join the subsequent element to form a two-part conjunct ligature form Example : ट् + ठ → ट्ठ	इकट्ठा	इ|क|ट्|ठा|	इ क ट्ठा C C CHCv
5.	A conjunct ligature form can itself behave as a dead consonant and enter into further, more complex ligatures Example : स् + त्र → स्त्र	शास्त्र स्त्री	शा|स्|त्|र| स्|त्|री|	शा स्त्र Cv CHCHC स्त्री CHCHCv
6.	A conjunct ligature form can also produce a half-form Example : क्ष् + य → क्ष्य	लक्ष्य तीक्ष्ण	ल|क्|ष्|य| ती|क्|ष्|ण|	ल क्ष्य C CHCHC ती क्ष्ण Cv CHCHC

7.		त्रिवेदी कृत्रिम मात्रामा	त्रि|वे|दी| कृ|त्रि|म| मा|त्रा|मा|	त्रि वे दी CHCv Cv Cv कृ त्रि म Cv CHCv C मा त्रा मा Cv Cv Cv
8.	Example : 	सिद्धार्थनगर श्रद्धालुओं वृद्धिसँग	सि|द्धा|र्थ|न|ग|र| श्र|द्धा|लु|ओं| वृ|द्धि|सँ|ग|	सि द्धा र्थ न ग र Cv CHCv CHC C C C श्र द्धा लु ओं CHC CHCv Cv Vm वृ द्धि सँ ग Cv CHCv Cm C
9.	Modifying marks, apply to the orthographic syllable as a whole and should follow (in the memory representation) all other characters that constitute the syllable. 	अंत:ज्ञानी स्वत:ला	अं|त:|ज्ञा|नी| स्|व|त:|ला|	अं त: ज्ञा नी Vm Cm CHCv Cv स्व त: ला CHC Cm Cv

Recommendation:

The pitfalls observed through data analysis above reveal that the Indic ABNF definition 100 % correct and needs to be suitably integrated in UAX# 29.