TO: UTC
FROM: Swaran Lata, MeitY
SUBJECT: Recommendations to UTC #152 , July 31-August 4 2017 on Text segmentation in Indian languages
DATE: 16/06/2017

1. Background

a) **Document:** L2/16-161 : Proposal to add ABNF based Orthographic Indic syllable definition(appended at Annexure 'A') and guidelines for proper text segmentation in Indian languages and guiding principles of line breaking in Unicode text segmentation (UAX#29) . This document covers:

i. Indic Orthographic syllable definition
ii. Additional sample examples of tailored grapheme clusters
iii. Indic syllable boundary determination
iv. Examples of Initial letter styling and Vertical text segmentation
v. Guiding principles of line breaking in Indian languages
vi. The precise list of characters with their Unicode code points of all the categories of the characters of 9 scripts which constitute ABNF definition (Refer Annexure 'A')

b) **Discussions on** L2/16-161: Presentation was made at UTC meeting #148 held on August 2-5, 2016 for highlighting the issues of text segmentation in Indian languages. It was suggested that the Unicode utility tool should be used to test data to identify the gaps in implementation of text segmentatiosn for Indian languages and to report specific problems w.r.t UAX #29.

2. Action undertaken as a followup of UTC#148

a) Using the utility tool of Unicode (http://unicode.org/cldr/utility/breaks.jsp), the sample data of Devanagari based languages[Hindi, Sanskrit (Excluding Vedic Extensions), Marathi, Konkani, Nepali, Bodo , Dogri , Maithili, Kashmiri, Sindhi] was used to test the existing UAX#29 rules viz - a - viz the proposed Orthographic Indic syllable definition (Refer Annexure 'A'). The test data mainly covers the combinations of conjunct characters, conjunct formations of consonants etc which sometimes don't get rendered properly.

b) The output of the tool has been compared with correct segmentation expected as per Orthographic Indic syllable definition. It was observed that Rule 1 and Rule 3are already implemented in the Unicode utility tool which is evident from the tested data presented at Annexure B. Following examples also illustrate this :

1. अंतराल
   Vm C Cv C

2. अयम्
   V C CH

c) The words covered in Annexure 'B' having CHC as a part are not being resolved properly by the Unicode Utility tool (Refer column 4) , which when segmented as per ABNF definition

gives the correct output (Refer column 5). This reveals improper implementation with reference to Rule 2 of ABNF.

3. Gaps in Utility tool implementation:

The following gaps in implementation of Rule 2 have been identified:

Improper break points of the following combinations:

a. consonant + virama sequences + consonant (CHC)
b. consonant + virama sequences + consonant + vowel sign (CHCv)
c. consonant + virama sequences + consonant + modifier (CHCm)
d. consonant + virama sequences + consonant + vowel sign + modifier (CHCvm)

The combinations at a, b are already covered in UAX#29 in the table 1a of section 3. The combinations at c & d need to examined and addressed.

4. Recommendations:

The data correlation was done for the above identified gaps. The errors due to non - implementation of Rule 2 in the Unicode Utility tool are marked in red colour. The corresponding output using ABNF Orthographic Indic syllable definition is shown in green colour. The corrective action w.r.t Rule 2 of ABNF implementation will require following additions in the UAX reports:

a) UAX#29 (UNICODE TEXT SEGMENTATION)

i. Proposed addition of Tailored Grapheme clusters :

The combinations at above defined 3(c) & 3(d) needs to be added as tailored grapheme cluster in the Table 1a.

Examples	Characters	Comments
र्गं (CHCm) Word : दुर्गंध	0930 र DEVANAGARI LETTER RA 094D ् DEVANAGARI SIGN VIRAMA 0917 ग DEVANAGARI LETTER GA 0902 ं DEVANAGARI SIGN ANUSVARA	Combining character sequences
फ्रां(CHCvm) Word : फ्रांस	092B फ DEVANAGARI LETTER PHA 094D ् DEVANAGARI SIGN VIRAMA 0930 र DEVANAGARI LETTER RA 093E ा DEVANAGARI VOWEL SIGN AA 0902 ं DEVANAGARI SIGN ANUSVARA	Combining character sequences

ii. Proposed addition of new section

The standard Indic syllable

- An V(Independent vowel) or a combination of V(Independent vowel) and m(modifier) is Indic syllable - V or Vm

  or
- A sequences of consonant(C) + virama(H) followed by consonant(C) – CHC

  or
- A sequences of consonant(C) + virama(H) followed by consonant(C) with dependent vowel/modifier or both - CHCv, CHCm, CHCvm

  or
- Consonant with virama (applicable only at the end of the word) - CH

Indian languages syllable boundary determination:

Using the above Standard Indic syllable definition, the Indic syllable break never occurs between the pairs of characters as shown in the table below:

Do not break between		Examples
V	m	V X m
C	v, m	C X v , C X m, C X v X m
C	H	a) C X H X C, C X H X C X H X C , C X H X C X H X C X H X C , C X H X C X H X C X H X C x H x C b) C X H (applicable only at the end of the word in Sanskrit , Nepali, Dogri)
C	H, v , m	C X H X C X v, C X H X C X m, C X H X C X v X m , C X H X C X H X C X v , C X H X C X H X C X m , C X H X C X H X C X v X m , C X H X C X H X C X H X C X v , C X H X C X H X C X H X C X H X C X m , C X H X C X H X C X H X C X v X m

Theoretically there is no practical limit of CHC, but the known limit is four in Sanskrit language (Excluding Vedic Extensions) and two in other languages. Example of maximum four CHC combinations is त्स्न्र्य is formed viz CHCHCHCHC.

b) UAX#14 (UNICODE LINE BREAKING ALGORITHM)

The definition of Standard Indic syllable (Refer UAX#29) may be used to break the line and a hyphen should be inserted at the breaking point so that word can be read intuitively however the language specific morpho-phonemic rules could be used for hyphenation as U+ 00AD (soft hyphen) is used in some Indian languages such as Tamil and Malayalam rather than inserting the hyphen character (U + 002D).

Annexure A

ABNF based Orthographic Indic syllable definition

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

This 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 etc)

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

The Unicode code points of above categories of the characters is given below :

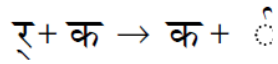
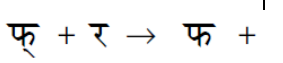
S. No.	Scripts	Consonant(C)	Vowel Independent V	Vowel dependent v	Modifier m	Halant (Virama) H
1.	Devanagari	A) DEVANAGARI LETTER KA...DEVANAGARI LETTER HA (0915..0939) B) DEVANAGARI LETTER QA....DEVANAGARI LETTER YYA. (0958..095F) C) DEVANAGARI LETTER MARWARI DDA...DEVANAGARI LETTER BBA (0978..097F)	A) DEVANAGARI LETTER SHORT A...DEVANAGARI LETTER AU (0904..0914) B) DEVANAGARI LETTER VOCALIC RR...DEVANAGARI LETTER VOCALIC LL (0960..0961) C) DEVANAGARI LETTER CANDRA A...DEVANAGARI LETTER UUE (0972..0977)	A) DEVANAGARI VOWEL SIGN OE (093A) B) DEVANAGARI VOWEL SIGN OOE (093B) C) DEVANAGARI VOWEL SIGN AA...DEVANAGARI VOWEL SIGN II (093E...0940) D) DEVANAGARI VOWEL SIGN U...DEVANAGARI VOWEL SIGN AI (0941..0948) E) DEVANAGARI VOWEL SIGN CANDRA O...DEVANAGARI VOWEL SIGN AU(0949..094C) F) DEVANAGARI VOWEL SIGN PRISHTHAMATRA E...DEVANAGARI VOWEL SIGN AW (094E...094F) G) DEVANAGARI VOWEL SIGN CANDRA LONG E...DEVANAGARI VOWEL SIGN UUE (0955..0957) H) DEVANAGARI VOWEL SIGN VOCALIC L...DEVANAGARI VOWEL SIGN VOCALIC LL (0962..0963)	A) DEVANAGARI SIGN INVERTED CANDRABINDU. ..DEVANAGARI SIGN ANUSVARA (0900...0902) B) DEVANAGARI SIGN VISARGA (0903) C) DEVANAGARI SIGN AVAGRAHA (093D) D) DEVANAGARI SIGN NUKTA (093C)	DEVANAGARI SIGN VIRAMA (094D)

Annexure B

Sample Test Results of 10 Indian languages(Devanagari script) based on Unicode Utility tool

1. Hindi

Sl. No.	Rendering rules defined by Unicode	Words	Rendering as per Unicode segmentation utility tool (http://unicode.org/cldr/utility/breaks.jsp)	Correct Segmentation as per Indic syllable definition
Conjunct Formations				
2.	Example : ग् + ध → ग्ध क् + ष → क्ष क् + क → क्क	संदिग्ध	सं|दि|ग्|ध| Cm Cv CH C	सं दि ग्ध Cm Cv CHC
		सुरक्षा	सु|र|क्|षा| Cv C CH Cv	सु र क्षा Cv C CHCv
		टक्कर	ट|क्|क|र| C CH C C	ट क्क र C CHC C
		सत्र	स|त्|र| C CH C	स त्र C CHC
		दक्षिण	द|क्|षि|ण| C CH Cv C	द क्षि ण C CHCv C
		मिश्रणाने	मि|श्|र|णा|ने| Cv CH C Cv Cv	मि श्र णा ने Cv CHC Cv Cv

	Rules for rendering			
3.	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 : 	दुर्घटनाग्रस्त मुहूर्त शर्करायुक्त	दुर्घटनाग्रस्त Cv CH C C Cv CH C CH C मुहूर्त Cv Cv CH C शर्करायुक्त C CH C Cv Cv CH C	दु र्घ ट ना ग्र स्त Cv CHC C Cv CHC CHC मु हू र्त Cv Cv CHC श र्क रा यु क्त C CHC Cv Cv CHC
4.	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 : 	अंतरराष्ट्रीय राष्ट्रपति फ्रांस ट्रैक्टर सिट्रोनेलाचे	अंतरराष्ट्रीय Vm C C Cv CH CH Cv C राष्ट्रपति Cv CH CH C C Cv फ्रांस CH Cvm C ट्रैक्टर CH Cv CH C C सिट्रोनेलाचे	अं त र रा ष्ट्री य Vm C C Cv CHCHCv C रा ष्ट्र प ति Cv CHCHC C Cv फ्रां स CHCvm C ट्रै क्ट र CHCv CHC C सि ट्रो ने ला चे

		टुक्रालाई	Cv CH Cv Cv Cv Cv टु|क्रा|लाई| Cv CH Cv Cv V	Cv CHCv Cv Cv Cv टु क्रा ला ई Cv CHCv Cv V
	5. Ligature rules			
6.	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|ट्ठा| C C CH Cv	इ क ट्ठा C C CHCv
7.	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| Cv CH CH C स्|त्|री| CH CH Cv	शा स्त्र Cv CHCHC स्त्री CHCHCv

8.	A conjunct ligature form can also produce a half-form Example : क्ष् + य → क्य	लक्ष्य तीक्ष्ण	लक्ष्य C CH CH C तीक्ष्ण Cv CH CH C	ल क्ष्य C CHCHC ती क्ष्ण Cv CHCHC
9.	त् + र + ि → त्र + ि → त्रि	त्रिवेदी कृत्रिम मात्रामा	त्रिवेदी CH Cv Cv Cv कृत्रिम Cv CH Cv C मात्रामा Cv CH Cv Cv	त्रि वे दी CHCv Cv Cv कृ त्रि म Cv CHCv C मा त्रा मा Cv CHCv Cv
10.	Example : द + ् + ध → द्ध	सिद्धार्थन गर श्रद्धालुओं वृद्धिसँग	सिद्धार्थनगर Cv CH Cv CH C C C श्रद्धालुओं CH C CH Cv Cv Vm वृद्धिसँग Cv CH Cv Cm C	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

11.	Modifying marks, apply to the orthographic syllable as a whole and should follow (in the memory representation) all other characters that constitute the syllable. क + ◌ा + ◌ँ → काँ	अंतःज्ञानी	अं|त:|ज्|ञा|नी| Vm Cm CH Cv Cv	अं त: ज्ञा नी Vm Cm CHCv Cv

2. Sanskrit(Excluding Vedic Extensions)

Words	Rendering as per Unicode segmentation utility tool (http://unicode.org/cldr/utility/breaks.jsp)	Correct Segmentation as per Indic syllable definition
गन्नदीधिम	ग|न्|न|दी|धि|म| C CH C Cvm Cv C	**ग न्न दी धि म** C CHC Cvm Cv C
प्रपंद्ये	प्|र|प|द्|ये| CH CH C Cm CH Cvm	**प्र पं द्ये** CHCHC Cm CHCvm
मनस्तापः	म|न|स्|ता|पः| C C CH Cv Cm	म न स्ता पः C C CHCv Cm
हविष्करोमि	ह|वि|ष्|क|रो|मि| C Cv CH C Cv Cv	ह वि ष्क रो मि C Cv CHC Cv Cv
अहर्पतिः	अ|ह|र्|प|ति:| V C CH C Cvm	अ ह र्प तिः V C CHC Cvm
गच्छति	ग|च्|छ|ति| C CH C Cv	ग च्छ ति C CHC Cv
अयम्	अ|य|म्| V C CH	अ य म् V C CH
शिवश्चोदति	शि|व|श्|चो|द|ति| Cv C CH Cv C Cv	शि व श्चो द ति Cv C CHCv C Cv
मनष्टालयति	म|न|ष्|टा|ल|य|ति|	म न ष्टा ल य ति

	C C CH Cv C C Cv	C C CHCv C C Cv
अश्वष्ठक्कस्य	अ|श्|वा|ष्|ठ|क्|क|स्|य| V CH C CH C CH C CH C	अ श्व ष्ठ क्क स्य V CHC CHC CHC CHC
दुष्पुत्रः	दु|ष्|पु|त्|रः| Cv CH Cv CH Cm	दु ष्पु त्रः Cv CHCv CHCm
द्विःपक्वम्	द्|वि|ः|प|क्|व|म्| CH Cvm C CH C CH	द्विः प क्व म् CHCvm C CHC CH
द्विष्कामः	द्|वि|ष्|का|म|ः| CH Cv CH Cv Cm	द्वि ष्का मः CHCv CHCv Cm
भर्तुर्भोगः	भ|र्|तु|र्|भो|ग|ः| C CH Cv CH Cv Cm	भ र्तु र्भो गः C CHCv CHCv Cm

3. Kashmiri

Words	Rendering as per Unicode segmentation utility tool (http://unicode.org/cldr/utility/breaks.jsp)	Correct Segmentation as per Indic syllable definition
शॅत्युल	शॅ|त्|युल| Cv CH Cv C	शॅ त्यु ल Cv CHCv C
महारॅन्य	म|हा|रॅ|न्|य| C Cv Cv CH C	म हा रॅ न्य C Cv Cv CHC
सॅक्युल	सॅ|क्|युल| Cv CH Cv C	सॅ क्यु ल Cv CHCv C
ल्यॉदुर	ल्|यॉ|दु|र| CH Cv Cv C	ल्यो दु र CHCv Cv C
फयॉक	फ्|यॉ|क| CH Cv C	फ्यो क CHCv C
मॉहन्युव	मॉ|ह|न्|युव| Cv C CH Cv C	मो ह न्यु व Cv C CHCv C
अन्यर	अ|न्|य|र|	अ न्य र

	V CH C C	V CHC C
ख्वजि	ख्|वा|जि| CH C Cv	ख्व जि CHC Cv
ख्युज	ख्|यु|ज| CH Cv C	ख्यु ज CHCv C
ख्योमुत	ख्|यो|मु|त| CH Cv Cv C	ख्यो मु त CHCv Cv C
खॅरिन्य	खॅ|रि|न्|य| Cv Cv CH C	खॅ रि न्य Cv Cv CHC
ग्यॅड	ग्|यॅ|ड| CH Cv C	ग्यॅ ड CHCv C
छन्यर	छ|न्|या|र| C CH C C	छ न्य र C CHC C
छिर्युव	छि|र्|यु|व| Cv CH Cv C	छि र्यु व Cv CHCv C
जॉल्य	जॉ|ल्|य| Cv CH C	जॉ ल्य CH CHC
छ्वम्बुन	छ्|वा|म्|बु|न| CH C CH Cv C	छ्व म्बु न CHC CHCv C
डॉगिन्य	डॉ|गि|न्|य| Cv Cv CH C	डॉ गि न्य Cv Cv CHC
टिक्युल	टि|क्|यु|ल| Cv CH Cv C	टि क्यु ल Cv CHCv C
तिक्याज़ि	ति|क्|या|ज़ि| Cv CH Cv Cv	ति क्या ज़ि Cv CHCv Cv
थ्यकुन	थ्|या|कु|न| CH C Cv C	थ्य कु न CHC Cv C
त्यम्बर	त्|या|म्|बा|र| 	त्य म्ब र

	CH C CH C C	CHC CHC C
थ्वसिलद	थ्‌व‌सि‌ल‌द‌ CH C Cv C C	थ्व सि ल द CHC Cv C C

4. Marathi

Words	Rendering as per Unicode segmentation utility tool (http://unicode.org/cldr/utility/breaks.jsp)	Correct Segmentation as per Indic syllable definition
उच्छ्वास	उ‌च्‌छ्‌वा‌स‌ V CH CH Cv C	उ च्छ्वा स V CHCHCv C
व्यक्तिमत्व	व्‌य‌क्‌ति‌म‌त्‌व‌ CH C CH Cv C CH C	व्य क्ति म त्व CHC CHCv C CHC
आपल्या	आ‌प‌ल्‌या‌ V C CH Cv	आ प ल्या V C CHCv
दातांच्यामध्ये	दा‌तां‌च्‌या‌म‌ध्‌ये‌ Cv Cvm CH Cv C CH Cv	दा तां च्या म ध्ये Cv Cvm CHCv C CHCv
टिपांच्या	टि‌पां‌च्‌या‌ Cv Cvm CH Cv	टि पां च्या Cv Cvm CHCv
ह्याच्यासाठी	ह्‌या‌च्‌या‌सा‌ठी‌ CH Cv CH Cv Cv Cv	ह्या च्या सा ठी CHCv CHCv Cv Cv
मिनटापेक्षाही	मि‌न‌टा‌पे‌क्‌षा‌ही‌ Cv C Cv Cv CH Cv Cv	मि न टा पे क्षा ही Cv C Cv Cv CHCv Cv
दुर्गंधी	दु‌र्‌गं‌धी‌ Cv CH Cm Cv	दु र्गं धी Cv CHCm Cv
शर्करायुक्त	श‌र्‌क‌रा‌यु‌क्‌त‌ C CH C Cv Cv CH C	श र्क रा यु क्त C CHC Cv Cv CHC
त्यामध्ये	त्‌या‌म‌ध्‌ये‌	त्या म ध्ये

	CH Cv C CH Cv	CHCv C CHCv
दुर्गंधीपासूनसुद्धा	दु|र्गं|धी|पा|सू|न|सु|द्|धा| Cv CH Cm Cv Cv CV C Cv CH Cv	दु र्गं धी पा सू न सु द्धा Cv CHCm Cv Cv CV C Cv CHCv
कोलेस्टेरॉलला	को|ले|स्|टे|रॉ|ल|ला| Cv Cv CH Cv Cv C Cv	को ले स्टे रॉ ल ला Cv Cv CHCv Cv C Cv
नित्यकर्मामध्ये	नि|त्|य|क|र्|मा|म|ध्|ये| Cv CH C C CH Cv C CH Cv	नि त्य क र्मा म ध्ये Cv CHC C CHCv C CHCv
मिश्रणाने	मि|श्|र|णा|ने| Cv CH C Cv Cv	मि श्र णा ने Cv CHC Cv Cv
प्रत्येकवेळी	प्|र|त्|ये|क|वे|ळी| CH C CH Cv C Cv Cv	प्र त्ये क वे ळी CHC CHCv C Cv Cv
ठेवण्यामध्ये	ठे|व|ण्|या|म|ध्|ये| Cv C CH Cv C CH Cv	ठे व ण्या म ध्ये Cv C CHCv C CHCv
मिश्रणाने	मि|श्|र|णा|ने| Cv CH C Cv Cv	मि श्र णा ने Cv CHC Cv Cv
मधुमेहासारख्या	म|धु|मे|हा|सा|र|ख्|या| C Cv Cv Cv Cv C CH Cv	म धु मे हा सा र ख्या C Cv Cv Cv Cv C CHCv
नित्यकर्मामध्ये	नि|त्|य|क|र्|मा|म|ध्|ये| Cv CH C C CH Cv CH Cv	नि त्य क र्मा म ध्ये Cv CHC C CHCv CHCv
आजारांपासूनसुद्धा	आ|जा|रां|पा|सू|न|सु|द्|धा| V Cv Cvm Cv Cv C Cv CH Cv	आ जा रां पा सू न सु द्धा V Cv Cvm Cv Cv C Cv CHCv
समाविष्ट	स|मा|वि|ष्|ट| C Cv Cv CH C	स मा वि ष्ट C Cv Cv CHC
भाज्यांमध्ये	भा|ज्|यां|म|ध्|ये| Cv CH Cvm C CH Cv	भा ज्यां म ध्ये Cv CHCvm C CHCv
स्वत:ला	स्|व|त:|ला| CH C Cm Cv	स्व त: ला CHC Cm Cv
उच्छ्वासाच्या	उ|च्|छ्|वा|सा|च्|या|	उ च्छ्वा सा च्या

	C CH CH Cv Cv CH Cv	C CHCHCv Cv CHCv
मिनटापेक्षाही	मि|न|टा|पे|क्|षा|ही| Cv C Cv Cv CH Cv Cv	मि न टा पे क्षा ही Cv C Cv Cv CHCv Cv

5. Konkani

Words	Rendering as per Unicode segmentation utility tool (http://unicode.org/cldr/utility/breaks.jsp)	Correct Segmentation as per Indic syllable definition
सुकिल्लीं	सु|कि|ल्|लीं| Cv Cv CH Cvm	सु कि ल्लीं Cv Cv CHCvm
कुडक्यांक	कु|ड|क्|यां|क| Cv C CH Cvm C	कु ड क्यां क Cv C CHCvm C
रोंवपाच्या	रों|व|पा|च्|या| Cvm C Cv CH Cv	रों व पा च्या Cvm C Cv CHCv
रोंप्याचो	रों|प्|या|चो| Cvm CH Cv Cv	रों प्या चो Cvm CHCv Cv
स्लिपां	स्|लि|पां| CH CV Cvm	स्लि पां CHCV Cvm
सिट्रोनेलाचे	सि|ट्|रो|ने|ला|चे| Cv CH Cv Cv Cv Cv	सि ट्रो ने ला चे Cv CHCv Cv Cv Cv
मिसळिल्ल्यान	मि|स|ळि|ल्|ल्|या|न| Cv C Cv CH CH Cv C	मि स ळि ल्ल्या न Cv C Cv CHCHCv C
रोंप्यांची	रों|प्|यां|ची| Cvm CH Cvm Cv	रों प्यां ची Cvm CHCvm Cv
खोडाच्या	खो|डा|च्|या| Cv Cv CH Cv	खो डा च्या Cv Cv CHCv
रोंप्यांक	रों|प्|यां|क| Cvm CH Cvm C	रों प्यां क Cvm CHCvm C

वर्सांतल्यान	व|र्|सां|त|ल्|या|न| C CH Cvm C CH Cv C	व र्सां त ल्या न C CHCvm C CHCv C
म्हण्टात	म्|ह|ण्|टा|त| CH C CH Cv C	म्ह ण्टा त CHC CHCv C
कृत्रीम	कृ|त्|री|म| Cv CH Cv C	कृ त्री म Cv CHCv C
सगल्यांत	स|ग|ल्|यां|त| C C CH Cvm C	स ग ल्यां त C C CHCvm C
रोंप्याच्या	रों|प्|या|च्|या| Cvm CH Cv CH Cv	रों प्या च्या Cvm CHCv CHCv
संस्थानान	सं|स्|था|ना|न| Cm CH Cv Cv C	सं स्था ना न Cm CHCv Cv C
केल्ल्यान	के|ल्|ल्|या|न| Cv CH CH Cv C	के ल्ल्या न Cv CHCHCv C
म्हळ्यार	म्|ह|ळ्|या|र| CH C CH Cv C	म्ह ळ्या र CHC CHCv C
सगल्यांत	स|ग|ल्|यां|त| C C CH Cvm C	स ग ल्यां त C C CHCvm C
नाशिल्ल्यान	ना|शि|ल्|ल्|या|न| Cv Cv CH CH Cv C	ना शि ल्ल्या न Cv Cv CHCHCv C
म्हटलां	म्|ह|ट|लां| CH C C Cvm	म्ह ट लां CHC C Cvm
शेतकर्यां नी	शे|त|क|र्|यां|च|नी| Cv C C CH Cvm Cv	शे त क र्यां नी Cv C C CHCvm Cv
कुडक्यांनी	कु|ड|क्|यां|नी| Cv C CH Cvm Cv	कु ड क्यां नी Cv C CHCvm Cv
जिल्ल्याच्या	जि|ल्|ल्|या|च्|या| Cv CH CH Cv CH Cv	जि ल्ल्या च्या Cv CHCHCv CHCv

ब्लॉकांत	ब्|लॉ|कां|त CH Cv Cvm C	ब्लॉ कां त CHCv Cvm C
कुरुक्षेत्रांतल्या	कु|रु|क्षे|त्रां|त|ल्|या Cv Cv CH Cv CH Cvm C CH Cv	कु रु क्षे त्रां त ल्या Cv Cv CHCv CHCvm C CHCv
भाज्ज्यांची	भा|ज्|ज्|यां|ची Cv CH CH Cvm Cv	भा ज्ज्यां ची Cv CHCHCvm Cv

6. Nepali

Words	Rendering as per Unicode segmentation utility tool (http://unicode.org/cldr/utility/breaks.jsp)	Correct Segmentation as per Indic syllable definition
काटिन्छ	का|टि|न्|छ Cv Cv CH C	का टि न्छ Cv Cv CHC
त्यसलाई	त्|य|स|लाई CH C C Cv V	त्य स ला ई CHC C Cv V
भनिन्छ	भ|नि|न्|छ C Cv CH C	भ नि न्छ C Cv CHC
स्लिप्स	स्|लि|प्|स CH Cv CH C	स्लि प्स CHCv CHC
टुक्रामा	टु|क्|रा|मा Cv CH Cv Cv	टु क्रा मा Cv CHCv Cv
बिरुवामध्ये	बि|रु|वा|म|ध्|ये Cv Cv Cv C CH Cv	बि रु वा म ध्ये Cv Cv Cv C CHCv
टुक्रालाई	टु|क्|रा|लाई Cv CH Cv Cv V	टु क्रा ला ई Cv CHCv Cv V
छिद्रमा	छि|द्|र|मा Cv CH C Cv	छि द्र मा Cv CHC Cv
सिट्रोनेलाका	सि|ट्|रो|ने|ला|का Cv CH Cv Cv Cv Cv	सि ट्रो ने ला का Cv CHCv Cv Cv Cv
त्यसो	त्|य|सो	त्य सो

	CH C Cv	CHC Cv
मात्रामा	मा|त्रा|मा Cv CH Cv Cv	मा त्रा मा Cv CHCv Cv
गरिनुपर्छ	ग|रि|नु|प|र्|छ C Cv Cv C CH C	ग रि नु प र्छ C Cv Cv C CHC
महिनाभन्दा	म|हि|ना|भ|न्|दा C Cv Cv C CH Cv	म हि ना भ न्दा C Cv Cv C CHCv
एकपल्ट	ए|क|प|ल्|ट V C C CH C	ए क प ल्ट V C C CHC
अतिरिक्त	अ|ति|रि|क्|त V Cv Cv CH C	अ ति रि क्त V Cv Cv CHC
सुक्नु	सु|क्|नु Cv CH Cv	सु क्नु Cv CHCv
सामान्यतः	सा|मा|न्|य|तः Cv Cv CH C Cm	सा मा न्य तः Cv Cv CHC Cm
सिम्वोपोगान	सि|म्|वो|पो|गा|न Cv CH Cv Cv Cv C	सि म्वो पो गा न Cv CHCv Cv Cv C
कृत्रिम	कृ|त्|रि|म Cv CH Cv C	कृ त्रि म Cv CHCv C
प्रकारको	प्|र|का|र|को CH C Cv C Cv	प्र का र को CHC Cv C Cv
मध्यस्थकारीले	म|ध्|य|स्|थ|का|री|ले C CH C CH C Cv Cv Cv	म ध्य स्थ का री ले C CHC CHC Cv Cv Cv
क्षेत्रहरूमा	क्|षे|त्|र|ह|रू|मा CH Cv CH C C Cv Cv	क्षे त्र ह रू मा CHCv CHC C Cv Cv
वृद्धिसँग	वृ|द्|धि|सँ|ग Cv CH Cv Cm C	वृ द्धि सँ ग Cv CHCv Cm C

7. Bodo

Words	Rendering as per Unicode segmentation utility tool (http://unicode.org/cldr/utility/breaks.jsp)	Correct Segmentation as per Indic syllable definition
बाराद्राय	बा|रा|द्|रा|य| Cv Cv CH Cv C	बा रा द्रा य Cv Cv CHCv C
खान्थियाव	खा|न्|थि|या|व| Cv CH Cv Cv C	खा न्थि या व Cv CHCv Cv C
सरासनस्रायै	स|रा|स|न|स्|रा|यै| C Cv C C CH Cv Cv	स रा स न स्रा यै C Cv C C CHCv Cv
रांखान्थियारि	रां|खा|न्|थि|या|रि| Cvm Cv CH Cv Cv Cv	रां खा न्थि या रि Cvm Cv CHCv Cv Cv
सिम्वपगान	सि|म्|व|प|गा|न| Cv CH C C Cv C	सि म्व प गा न Cv CHC C Cv C
ग्रामनी	ग्|रा|म|नी| CH Cv C Cv	ग्रा म नी CHCv C Cv
बयनिखिुइ	ब|य|नि|ख्|रि|इ| C C Cv CH Cv V	ब य नि ख्रि इ C C Cv CHCv V
हाग्रा	हा|ग्|रा| Cv CH Cv	हा ग्रा Cv CHCv
बोसोरब्रै	बो|सो|र|ब्|रै| Cv Cv C CH Cv	बो सो र ब्रै Cv Cv C CHCv
दिहुनग्राया	दि|हु|न|ग्|रा|या| Cv Cv C CH Cv Cv	दि हु न ग्रा या Cv Cv C CHCv Cv
प्रभुआ	प्|र|भु|आ| CH C Cv V	प्र भु आ CHC Cv V
बेनिफ्राय	बे|नि|फ्|रा|य| Cv Cv CH Cv C	बे नि फ्रा य Cv Cv CHCv C
दैथाइहरग्रा	दै|था|इ|ह|र|ग्|रा| Cv Cv C C C CH Cv	दै था इ ह र ग्रा Cv Cv C C C CHCv
झज्जरनि	झ|ज्|ज|र|नि|	झ ज्ज र नि

	C CH C C Cv	C CHC C Cv
फार्माव	फा|र्|मा|व| Cv CH Cv C	फा र्मा व Cv CHCv C
थाग्रा	था|ग्|रा| Cv CH Cv	था ग्रा Cv CHCv
हानिफ्राय	हा|नि|फ्|रा|य| Cv Cv CH Cv C	हा नि फ्रा य Cv Cv CHCv C
रिपोर्टनि	रि|पो|र्|ट|नि| Cv Cv CH C Cv	रि पो र्ट नि Cv Cv CHC Cv
बादिब्ला	बा|दि|ब्|ला| Cv Cv CH Cv	बा दि ब्ला Cv Cv CHCv
जौखोन्दो	जौ|खो|न्|दो| Cv Cv CH Cv	जौ खो न्दो Cv Cv CHCv
जौखोन्दोल	जौ|खो|न्|दो|ल| Cv Cv CH Cv C	जौ खो न्दो ल Cv Cv CHCv C

8. Dogri

Words	Rendering as per Unicode segmentation utility tool (http://unicode.org/cldr/utility/breaks.jsp)	Correct Segmentation as per Indic syllable definition
मनुक्खो	मा|नु|क्|खो| C Cv CH Cv	म नु क्खो C Cv CHCv
मित्तर	मि|त्|त|र| Cv CH C C	मि त्त र Cv CHC C
दिक्खी	दि|क्|खी| Cv CH Cv	दि क्खी Cv CHCv
टोस्टर्ज़	टो|स्|ट|र्|ज़| Cv CH C CH C	टो स्ट र्ज़ Cv CHC CHC

आह्न्नो	आ|ह्|न्|नो| V C CH Cv	आ ह् न्नो V C CHCv
वक्तव्य	व|क्|त|व्|य| C CH C CH C	व क्त व्य C CHC CHC
पत्तर	प|त्|त|र| C CH C C	प त्त र C CHC C
पुश्टी	पु|श्|टी| Cv CH Cv	पु श्टी Cv CHCv
प्रक्रिया	प्|र|क्|रि|या| CH C CH Cv Cv	प्र क्रि या CHC CHCv Cv
प्रविश्ट	प्|र|वि|श्|ट| CH C Cv CH C	प्र वि श्ट CHC Cv CHC
बक्खरा	ब|क्|ख|रा| C CH C Cv	ब क्ख रा C CHC Cv
सक्खना	स|क्|ख|ना| C CH C Cv	स क्ख ना C CHC Cv
खोह्ल्लना	खो|ह्|ल्|ल|ना| Cv C CH C Cv	खो ह् ल्ल ना Cv C CHC Cv

9. Maithili

Words	Rendering as per Unicode segmentation utility tool (http://unicode.org/cldr/utility/breaks.jsp)	Correct Segmentation as per Indic syllable definition
निर्दिष्ट	नि|र्|दि|ष्|ट| Cv CH Cv CH C	नि र्दि ष्ट Cv CHCv CHC
भ्रमणक	भ्|र|म|ण|क| CH C C C C	भ्र म ण क CHC C C C
कूटशब्दक	कू|ट|श|ब्|द|क| Cv C C CH C C	कू ट श ब्द क Cv C C CHC C

तालाबन्न	ता|ला|ब|न्न| Cv Cv C CH C	ता ला ब न्न Cv Cv C CHC
अस्वीकृत	अ|स्|वी|कृ|त V CH v Cv C	अ स्वी कृ त V CHv Cv C
संख्यासँ	सं|ख्|या|सँ Cm CH Cv Cm	सं ख्या सँ Cm CHCv Cm
श्रेणीक	श्|रे|णी|क CH Cv Cv C	श्रे णी क CHCv Cv C
प्रयोक्ता	प्|र|यो|क्|ता CH C Cv CH Cv	प्र यो क्ता CHC Cv CHCv
निर्गम	नि|र्|ग|म्म Cv CH C C	नि र्ग म Cv CHC C
फँटवर्क	फँ|ट|व|र्|क Cm C C CH C	फँ ट व र्क Cm C C CHC
अनुच्छेदसँ	अ|नु|च्|छे|द|सँ V Cv CH Cv C Cm	अ नु च्छे द सँ V Cv CHCv C Cm
दहिन्ना	द|हि|न्|ना C Cv CH Cv	द हि न्ना C Cv CHCv
स्वचालित	स्|व|चा|लि|त CH C Cv Cv C	स्व चा लि त CHC Cv Cv C
सँपत्तिक	सँ|प|त्|ति|क Cm C CH Cv C	सँ प त्ति क Cm C CHCv C
निर्गत	नि|र्|ग|त Cv CH C C	नि र्ग त Cv CHC C
प्रकार्यक	प्|र|का|र्|य|क CH C Cv CH C C	प्र का र्य क CHC Cv CHC C
संख्याक	सं|ख्|या|क 	सं ख्या क Cm CHCv C
स्थानधारक	स्|था|न|धा|र|क	स्था न धा र क

	CH Cv C Cv C C	CHCv C Cv C C
भविष्यकँ	भ|वि|ष्|य|कँ| C Cv CH C Cvm	भ वि ष्य कँ C Cv CHC Cvm
पँक्तिक	पँ|क्|ति|क| Cm CH Cv C	पँ क्ति क Cm CHCv C

10.Sindhi

Words	Rendering as per Unicode segmentation utility tool (http://unicode.org/cldr/utility/breaks.jsp)	Correct Segmentation as per Indic syllable definition
ईश्वरु	ई|श्|व|रु| V CH C Cv	ई श्व रु V CHC Cv
गर्मीपदु	ग|र्|मी|प|दु| C CH Cv C Cv	ग र्मी प दु C CHCv C Cv
ट्रेक्टरु	ट्|रे|क्|टा|रु| CH Cv CH C Cv	ट्रे क्ट रु CHCv CHC Cv
कलेक्टरु	क|ले|क्|टा|रु| C Cv CH C Cv	क ले क्ट रु C Cv CHC Cv
ग्यानु	ग्|या|नु| CH Cv Cv	ग्या नु CHCv Cv
शास्त्रु	शा|स्|त्|रु| Cv CH CH Cv	शा स्त्रु Cv CHCHCv
शिकस्त	शि|क|स्|त| Cv C CH C	शि क स्त Cv C CHC
इन्द्री	इ|न्|द्|री| V CH CH Cv	इ न्द्री V CHCHCv
श्रधालू	श्|र|धा|लू| CH C Cv Cv	श्र धा लू CHC Cv Cv
सकार्थो	स|का|र्|थो| C Cv CH Cv	स का र्थो C Cv CHCv

मुख़्तियारी	मु|ख़्|ति|या|री| Cv CH Cv Cv Cv	मु ख़्त या री Cv CHCv Cv Cv
बेवक़्ति	बे|व|क़्|ति| Cv C CH Cv	बे व क़्त Cv C CHCv
आमदरफ़्त	आ|म|द|र|फ़्|त V C C C CH C	आ म द र फ़्त V C C C CHC
आस्तिकु	आ|स्|ति|कु| V CH Cv Cv	आ स्ति कु V CHCv Cv
इन्किलाबी	इ|न्|कि|ला|बी| V CH Cv Cv Cv	इ न्कि ला बी V CHCv Cv Cv
इब्तिदा	इ|ब्|ति|दा| V CH Cv Cv	इ ब्ति दा V CHCv Cv
यज़ु	य|ज़्|ज़ु| C CH Cv	य ज़ु C CHCv
सन्बंधु	स|न्|बं|धु| C CH Cm Cv	स न्बं धु C CHCm Cv
त्रिलोकु	त्|रि|लो|कु| CH Cv Cv Cv	त्रि लो कु CHCv Cv Cv
इस्पंगुरु	इ|स्|पं|गु|रु| V CH Cm Cv Cv	इ स्पं गु रु V CHCm Cv Cv