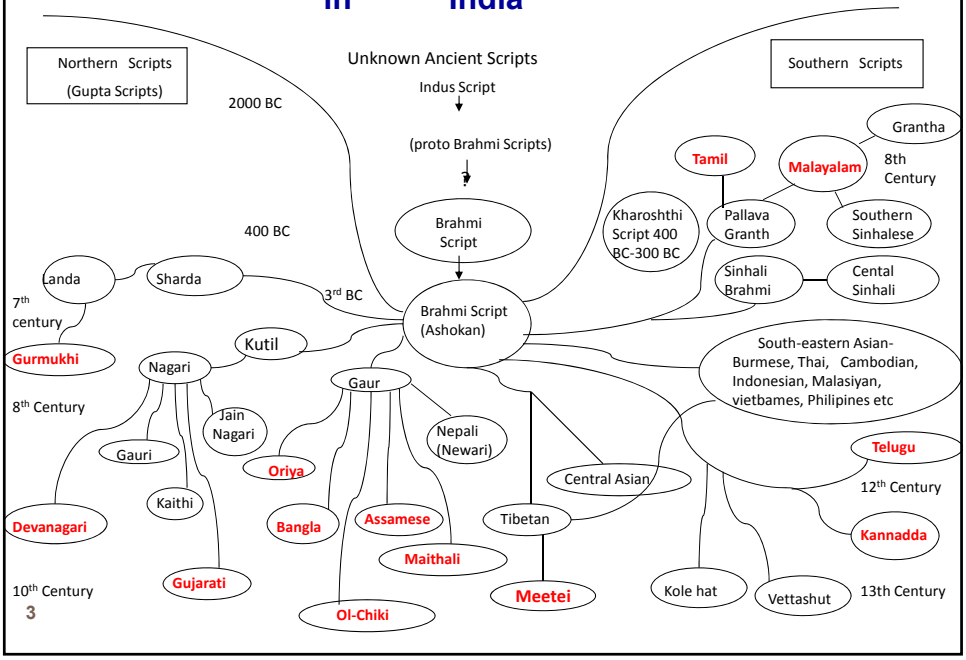
INDIC TEXT SEGMENTATION

Presented by :
Swaran Lata
Senior Director & HoD (TDIL Programme)
Department of Electronics and Information Technology (DeitY)
E-mail: slata@deity.gov.in

Diverse Multilinguality in India

My India is great.	English
मेरा भारत महान।	Hindi
আমার ভারত মহান।	bangla
ಮೇರಾ ಭಾರತ್ ಮಹನ್।	Kannada
మేరా భారత్ మహాన్।	Telugu
എന്റെ ഭാരതം വലിയത് ആണ്	Malayalam
माझा भारत मोठा आहे	Marathi
ମୋର ଭାରତ ଡୁହୁଡ୍ ଅଟେ	Oriya
ਮੇਰਾ ਭਾਰਤ ਮਹਾਨ ਹੈ	Punjabi
என்னுடைய இந்தியா மிகச்சிறப்பான இருக்கிறது	Tamil

Major Scripts and Corresponding Languages in India

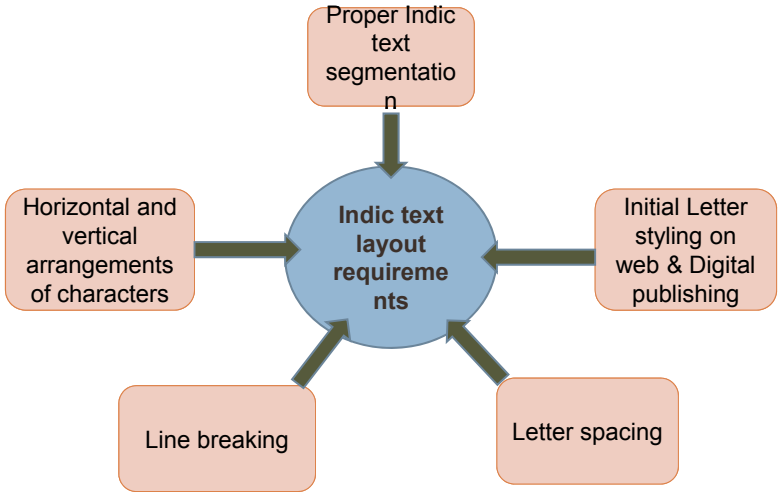


- Hindi Speaking region covers 40% of India.
- Any Localization effort Hindi is treated as test-bed.
- The efforts are iterated for other Indian languages using language specific requirements for Indic languages

Indian language complexities

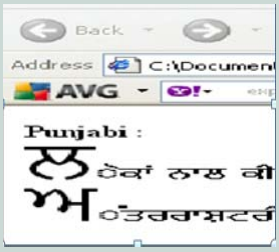
- India has large linguistic diversity with 22 constitutionally recognized languages and 12 scripts
- The mapping between languages and scripts is complex as multiple languages may have common scripts, and a language can be written in multiple scripts
- Each language and script is unique in nature and cannot be easily replicated , even if they share common characteristics

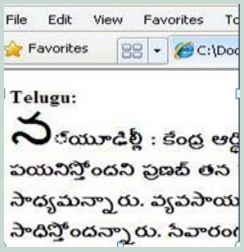
Indic Text layout requirements



Challenges in Indian languages

Use case Scenarios: Initial letter styling on Web publishing





Challenges in Indian languages

Use case Scenarios: Text input in a word processor

द्वारा	विद्यालय
द्वारका	विद्वान

Correct representation

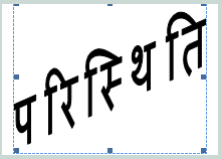
द्वारा	विद्यालय
द्वारका	विद्वान

Challenges in Indian languages

Use case Scenarios: Formatting and spacing on

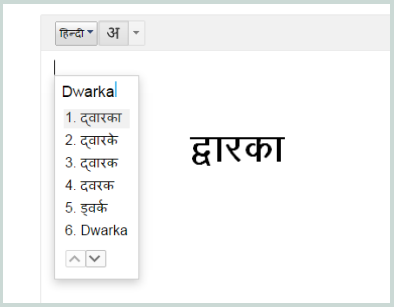
- Spacing

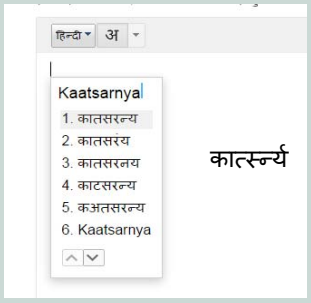
- Change shape



Challenges in Indian languages

Use case Scenarios: Phonetic Typing/Transliteration





Challenges in Indian languages

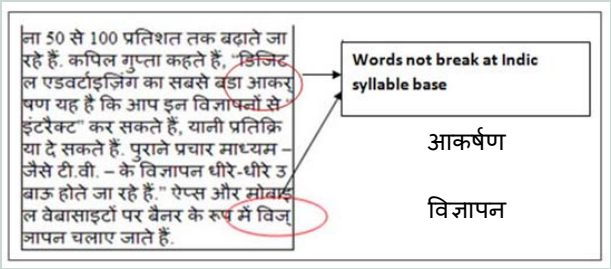
Use case Scenarios : Letter spacing on Web browsers

आर्थिक सुस्ती और इसकी वजह से मांग घटने के बावजूद देश की बड़ी कंपनियां इस साल अपने कर्मचारियों को बोनस दे रही हैं। इन कंपनियों का कहना है कि पिछले साल उन्होंने लाभ कमाने की अपनी क्षमता बढ़ाई थी, लिहाजा कर्मचारियों को उसका ईनाम दिया जा रहा है।

Challenges in Indian languages

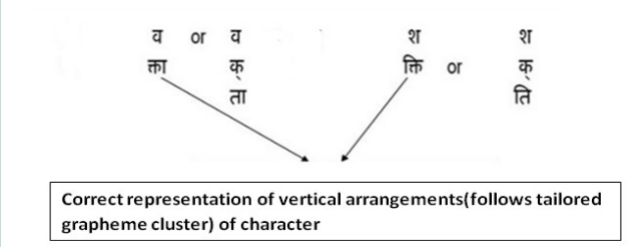
Use case Scenarios: Line breaking on applying

word wrap



Challenges in Indian languages

- Vertical arrangements of characters



Grapheme cluster boundaries defined in UAX#29

- ***legacy grapheme cluster*** :

  It is defined as a base followed by zero or more continuing characters.

- ***Extended grapheme cluster***

  It is the same as a legacy grapheme cluster, with the addition of some other characters.

- ***Tailored Grapheme cluster***

  Tailoring of Grapheme cluster to meet further requirements

Approach to be taken for Possible Solution

- Due to high complexities of Indian languages , it is required to tailored the grapheme cluster for Indian languages
- Indian languages Orthographic syllable should be based on tailored Grapheme Cluster as defined in UAX#29
- Rules for wrapping of Indian languages characters and identification of syllable boundaries needs to be evolved for tailoring of grapheme cluster so that segmentation in Indian languages seems logically.

Indic Orthographic syllable

- An Orthographic syllable includes Independent vowel or a base consonant and/or any combination of the following characters in the text stream:
  - Consonant/s and consonant + virama sequences
  - vowel signs
  - Modifiers

The above definition of Orthographic syllable is based on the tailored grapheme cluster discussed in section 3 of UAX#29 report.

Sample tailored Grapheme Cluster Boundaries for Indian languages

- Examples of Indic Orthographic syllable based on tailored grapheme cluster boundaries

क्या	0915 (क)DEVANAGARI LETTER KA 094D (्) DEVANAGARI SIGN VIRAMA 092F (य)DEVANAGARI LETTER SSA 093E (ा)DEVANAGARI SIGN AA	Devanagari kya
स्थि	0938 (स)DEVANAGARI LETTER SA 094D (्)DEVANAGARI SIGN VIRAMA 0925 (थ)DEVANAGARI LETTER THA 091C (ि)DEVANAGARI LETTER I	Devanagari sthi
स्तः	0938 (स) DEVANAGARI LETTER SA 0924 (त) DEVANAGARI LETTER TA 0903 (ः) DEVANAGARI Sign Visarga	Devana gari sth
त्क्ल	0924 (त) DEVANAGARI LETTER TA 094D (्) DEVANAGARI SIGN VIRAMA 0915 (क) DEVANAGARI LETTER KA 094D (्) DEVANAGARI SIGN VIRAMA 0932 (ल) DEVANAGARI LETTER LA	Devana gari tkl

Improving Indic text segmentation....

Formulation of ABNF based Indic Orthographic syllable definition for defining rules

- ABNF Valid Segmentation based Indic orthographic syllable definition is provided for correct and standardized representation of Indian languages text segmentation
- Augmented Backus–Naur Form (ABNF) is a meta-language based on Backus–Naur Form (BNF), but consisting of its own syntax and derivation rules. The motive principle for ABNF is to describe a formal system of a language to be used as a bidirectional communications protocol.

Indic Orthographic syllable definition

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

- The linguistic definition of Indic orthographic syllable has been mapped to ABNF(Augmented Backus–Naur Form) for the purpose of text segmentation, line breaking , drop letter, letter spacing in horizontal text and vertical text representation.

Indic Orthographic syllable definition

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 [**V_{vs} has been used as symbol in Unicode for dependent vowel of full vowel V e.g AA_{vs}**]
- H is Virama/ halant
- | is a rule separator
- [ ] - The enclosed items is optional under this bracket
- {} - The enclosed item/items occurs zero or repeated multiple times

Indic syllable boundary determination

No break rules for Indian languages

Rules	Do not break between
V[m]	Independent vowel and Modifier
{CH}C[v][m]	one or more consonant(N) + virama sequences and Consonant zero or more consonant(N) + virama sequences , Consonant and dependent vowel sign zero or more consonant(N) + virama sequences , Consonant and modifier zero or more consonant(N) + virama sequences, Consonant ,dependent vowel sign and modifier
CH	Consonant(N) with virama (applicable only for those Indian languages where pure consonant appears at the end of the word)

Note : Consonant may or may not include Nukta(N)

Categories values of Indic Orthographic syllable

The precise list of characters with their Unicode code points of all the categories i.e C, H, V etc defined in Indic syllable definition are enclosed as appendix 1 on the following link :

http://www.unicode.org/L2/L2016/16161-indic-text-seg.pdf

Boundary determination for line breaking

- In Indic writing system , it is preferred that line breaks at word boundaries ,if required following principle may be adhered : New line cannot begin with following symbols/Punctuation marks. Also these should be retain with the associated text :

Symbols	Character name	Unicode code-point
।	DEVANAGARI DANDA	U + 0964
॥	DEVANAGARI DOUBLE DANDA	U + 0965
)	RIGHT PARENTHESIS	U + 0029
+	PLUS SIGN	U + 002B
*	ASTERISK	U + 002A
-	HYPHENATIONPOINT-VISIBLE HYPHEN HYPHENATION-SOFT HYPHEN	U + 2027 U+ 00AD
/	SOLIDUS	U + 002F
,	COMMA	U + 002C
.	FULL STOP	U + 002E
:	COLON	U + 003A
;	SEMICOLON	U + 003B
=	EQUALS SIGN	U + 003D
>	GREATER-THAN SIGN	U + 003E
]	RIGHT SQUARE BRACKET	U + 005D
_	LOW LINE	U + 005F
|	VERTICAL LINE	U + 007C
}	RIGHT CURLY BRACKET	U + 007D
~	TILDE	U + 007E
%	PERCENT SIGN	U + 0025

Hyphenation at line boundary

- The definition of Indic orthographic syllable may be used to break the line and a hyphen should be at the breaking point so that word can be read intuitively.
- However the language specific morpho-phonemic rules and industry practices (from media, publishing and grammar books) could be used for hyphenation. U+ 00AD (soft hyphen) is used in some languages such as Tamil and Malayalam.

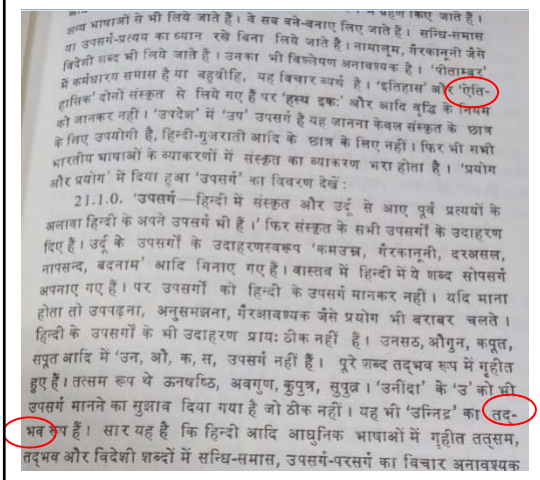
- The hyphenated words can be broken at the hyphenation point (U + 2027) e.g.:

  नर-नारी should be treated as:

  नर- on the first line and नारी on the next line

Hyphenation used in printed documents

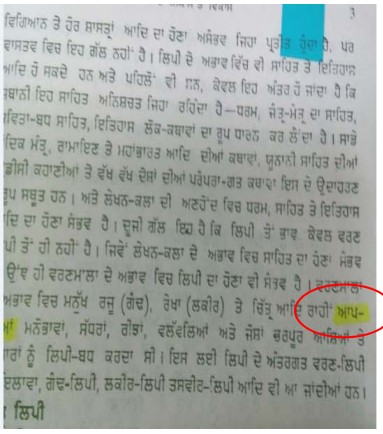
Hindi



अन्य भाषाओं से भी लिये जाते हैं। वे सब बने-बनाए लिए जाते हैं। या उपसर्ग-प्रत्यय का ध्यान रखे बिना लिये जाते हैं। सन्धि-समास विदेशी शब्द भी लिये जाते हैं। उनका भी विश्लेषण अनावश्यक है। 'पीताम्बर' में कर्मधारय समास है या बहुव्रीहि, यह विचार व्यर्थ है। 'इतिहास' और 'ऐति-हासिक' दोनों संस्कृत से लिये गए हैं पर 'ह्रस्य ह्रक:' और आदि वृद्धि के नियम को जानकर नहीं। 'उपदेश' में 'उप' उपसर्ग है यह जानना केवल संस्कृत के छात्र के लिए उपयोगी है, हिन्दी-गुजराती आदि के छात्र के लिए नहीं। फिर भी सभी भारतीय भाषाओं के व्याकरणों में संस्कृत का व्याकरण भरा होता है। 'प्रयोग और प्रयोग' में दिया हुआ 'उपसर्ग' का विवरण देखें :

21.1.0. 'उपसर्ग—हिन्दी में संस्कृत और उर्दू से आए पूर्व प्रत्ययों के अलावा हिन्दी के अपने उपसर्ग भी हैं।' फिर संस्कृत के सभी उपसर्गों के उदाहरण दिए हैं। उर्दू के उपसर्गों के उदाहरणस्वरूप 'कमउम्र, गैरकानूनी, दरअसल, नापसन्द, बदनाम' आदि गिनाए गए हैं। वास्तव में हिन्दी में ये शब्द सोपसर्ग अपनाए गए हैं। पर उपसर्गों को हिन्दी के उपसर्ग मानकर नहीं। यदि माना होता तो उपपढ़ना, अनुसमझना, गैरआवश्यक जैसे प्रयोग भी बराबर चलते। हिन्दी के उपसर्गों के भी उदाहरण प्रायः ठीक नहीं हैं। उनसठ, औगुन, कपूत, सपूत आदि में 'उन, औ, क, स, उपसर्ग नहीं हैं। पूरे शब्द तद्भव रूप में गृहीत हुए हैं। तत्सम रूप थे ऊनषष्ठि, अवगुण, कुपुत्र, सुपुत्र। 'उनीदा' के 'उ' को भी उपसर्ग मानने का सुझाव दिया गया है जो ठीक नहीं। यह भी 'उन्निद्र' का तद्-भव रूप हैं। सार यह है कि हिन्दी आदि आधुनिक भाषाओं में गृहीत तत्सम, तद्भव और विदेशी शब्दों में सन्धि-समास, उपसर्ग-परसर्ग का विचार अनावश्यक

Punjabi

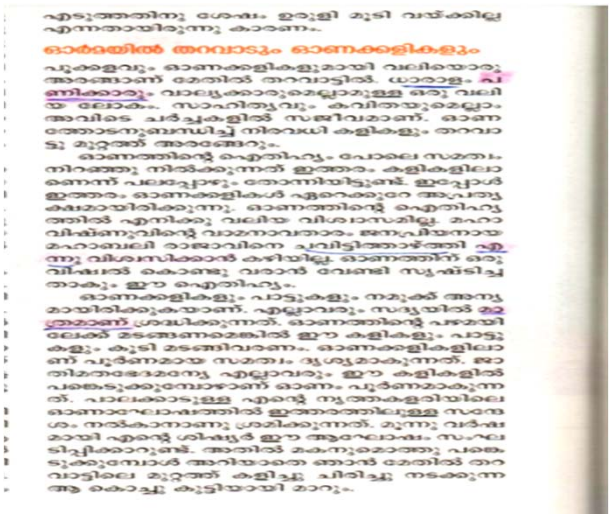


ਵਿਗਿਆਨ ਤੇ ਹੋਰ ਸ਼ਾਸਤ੍ਰਾਂ ਆਦਿ ਦਾ ਹੋਣਾ ਅਸੰਭਵ ਜਿਹਾ ਪ੍ਰਤੀਤ ਹੁੰਦਾ ਹੈ, ਪਰ ਵਾਸਤਵ ਵਿਚ ਇਹ ਗੱਲ ਨਹੀਂ ਹੈ। ਲਿਪੀ ਦੇ ਅਭਾਵ ਵਿੱਚ ਵੀ ਸਾਹਿਤ ਤੇ ਇਤਿਹਾਸ ਆਦਿ ਹੋ ਸਕਦੇ ਹਨ ਅਤੇ ਪਹਿਲੇ ਦੀ ਹਨ, ਕੇਵਲ ਇਹ ਅੰਤਰ ਹੈ ਜਾਂਦਾ ਹੈ ਕਿ ਭਾਸ਼ਾ ਦਿਹ ਸਾਹਿਤ ਅਨਿਸ਼ਚਤ ਜਿਹਾ ਰਹਿੰਦਾ ਹੈ—ਧਰਮ, ਮੰਤ੍ਰ-ਮੰਤ੍ਰ ਦਾ ਸਾਹਿਤ, ਇਤਿਹਾਸ-ਬਧ ਸਾਹਿਤ, ਇਤਿਹਾਸ, ਲੋਕ-ਕਥਾਵਾਂ ਦਾ ਰੂਪ ਧਾਰਨ ਕਰ ਲੈਂਦਾ ਹੈ। ਸਾਡੇ ਦਿਕ ਮੰਤ੍ਰ, ਰਾਮਾਇਣ ਤੇ ਮਹਾਂਭਾਰਤ ਆਦਿ ਦੀਆਂ ਕਥਾਵਾਂ, ਯੂਨਾਨੀ ਸਾਹਿਤ ਦੀਆਂ ਗ੍ਰੀਸੀ ਕਹਾਣੀਆਂ ਤੇ ਵੱਖ ਵੱਖ ਦੇਸ਼ਾਂ ਦੀਆਂ ਪਰੰਪਰਾ-ਗਤ ਕਥਾਵਾਂ ਇਸ ਦੇ ਉਦਾਹਰਣ ਰੂਪ ਸਬੂਤ ਹਨ। ਅਤੇ ਲੇਖਨ-ਕਲਾ ਦੀ ਅਣਹੋਂਦ ਵਿਚ ਧਰਮ, ਸਾਹਿਤ ਤੇ ਇਤਿਹਾਸ ਆਦਿ ਦਾ ਹੋਣਾ ਸੰਭਵ ਹੈ। ਦੂਜੀ ਗੱਲ ਇਹ ਹੈ ਕਿ ਲਿਪੀ ਤੋਂ ਭਾਵ ਕੇਵਲ ਵਰਣ ਮਾਲਾ ਤੋਂ ਹੀ ਨਹੀਂ ਹੈ। ਜਿਵੇਂ ਲੇਖਨ-ਕਲਾ ਦੇ ਅਭਾਵ ਵਿਚ ਸਾਹਿਤ ਦਾ ਹੋਣਾ ਸੰਭਵ ਹੈ ਉਵੇਂ ਹੀ ਵਰਣਮਾਲਾ ਦੇ ਅਭਾਵ ਵਿਚ ਲਿਪੀ ਦਾ ਹੋਣਾ ਵੀ ਸੰਭਵ ਹੈ। ਵਰਣਮਾਲਾ ਦੇ ਅਭਾਵ ਵਿਚ ਮਨੁੱਖ ਰਸੂ (ਗੰਢ), ਰੇਖਾ (ਲਕੀਰ) ਤੇ ਚਿੱਤ੍ਰ ਆਦਿ ਰਾਹੀਂ ਆਪ-ਸੀ ਮਨੋਭਾਵਾਂ, ਸੰਧਰਾਂ, ਰੀਤਾਂ, ਵਲੱਵਲਿਆਂ ਅਤੇ ਜੋਸ਼ਾਂ ਭਰਪੂਰ ਅਭਿਲਾਖਾਂ ਤੇ ਸਾਰਾਂ ਨੂੰ ਲਿਪੀ-ਬਧ ਕਰਦਾ ਸੀ। ਇਸ ਲਈ ਲਿਪੀ ਦੇ ਅੰਤਰਗਤ ਵਰਣ-ਲਿਪੀ ਇਲਾਵਾ, ਗੰਦ-ਲਿਪੀ, ਲਕੀਰ-ਲਿਪੀ ਤਸਵੀਰ-ਲਿਪੀ ਆਦਿ ਵੀ ਆ ਜਾਂਦੀਆਂ ਹਨ।

ਲਿਪੀ

Word-break at line boundary in south Indian language

Malayalam



എടുത്തതിനു ശേഷം ഉടുപ്പി മുടി വയ്ക്കില്ല എന്നതായിരുന്നു കാരണം.

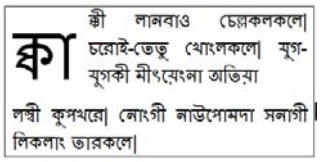
ഓർമയിൽ തറവാടും ഓണക്കളികളും

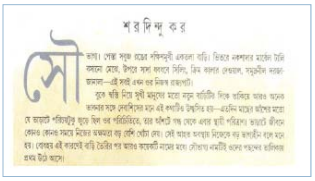
പൂക്കളും ഓണക്കളികളുമായി വലിയൊരു ആഘോഷമാണ് തറവാട്ടിൽ. ധാരാളം ബന്ധുക്കളും വായ്ക്കാരുമെല്ലാമുള്ള ഒരു വലിയ ലോകം. സാഹിത്യവും കവിതയുമെല്ലാം അവിടെ ചർച്ചകളിൽ സജീവമാണ്. ഓണത്തോടനുബന്ധിച്ച് നിരവധി കളികളും തറവാടു മുറ്റത്ത് അരങ്ങേറും.

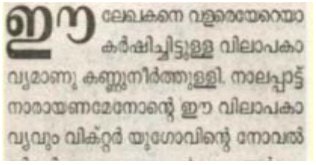
ഓണത്തിന്റെ ഐതിഹ്യം പോലെ സമത്വം നിറഞ്ഞു നിൽക്കുന്നത് ഇത്തരം കളികളിലാണെന്ന് പലരുമുണ്ട്. തോന്നിയിട്ടുണ്ട്. ഇപ്പോൾ ഇത്തരം ഓണക്കളികൾ ഏറെക്കുറെ അപ്രത്യക്ഷമായിരിക്കുന്നു. ഓണത്തിന്റെ ഐതിഹ്യത്തിൽ എന്നിരുന്നാലും വലിയ വിശ്വാസമില്ല. മഹാവിഷ്ണുവിന്റെ വാമനാവതാരം ജനപ്രിയനായ മഹാബലി രാജാവിനെ പാതാളത്തിലേക്കു ചവിട്ടിത്താഴ്ത്തി എന്നു വിശ്വസിക്കാൻ കഴിയില്ല. ഓണത്തിന് ഒരു വിശേഷ കൊണ്ടു വരാൻ വേണ്ടി സൃഷ്ടിച്ചതാകും ഈ ഐതിഹ്യം.

ഓണക്കളികളും പാട്ടുകളും നമുക്ക് അന്യമായിരിക്കുകയാണ്. എല്ലാവരും സമയത്തിനു പിന്നാലെ ശ്രമിക്കുന്നത്. ഓണത്തിന്റെ പഴയതിലേക്ക് മടങ്ങണമെങ്കിൽ ഈ കളികളും പാട്ടുകളും കൂടി മടങ്ങിവരണം. ഓണക്കളികളിലാണ് പൂർണമായ സമത്വം ദൃശ്യമാകുന്നത്. ജാതിമതഭേദമന്യേ എല്ലാവരും ഈ കളികളിൽ പങ്കെടുക്കുമ്പോഴാണ് ഓണം പൂർണമാകുന്നത്. പാലക്കാടുള്ള എന്റെ സുഹൃത്തുക്കളിലെ ഓണാഘോഷത്തിൽ ഇത്തരത്തിലുള്ള സമത്വം നടക്കാനാണു ശ്രമിക്കുന്നത്. മുന്നു വർഷം മാത്രമേ എന്റെ ശിഷ്യർ ഈ ആഘോഷം സംഘടിപ്പിക്കാറുള്ളൂ. അതിൽ മകനുമൊത്തു പങ്കെടുക്കുമ്പോൾ അറിയാതെ ഞാൻ തറവാട്ടിലെ മുറ്റത്ത് കളിച്ചു വീണ്ടും നടക്കുന്ന ആ കൊച്ചു കുട്ടിയായി മാറും.

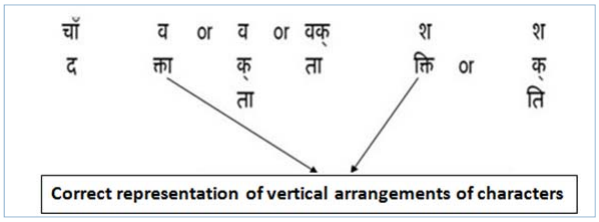
Indic text segmentation results based on Indic syllable definition







Indic text segmentation results based on Indic syllable definition



ক

ল

কা

তা

ধা রা বা হি ক উ প ন্যা স

পা ল্টা হা ও য়া

Proposal to incorporate Indian languages requirements in UAX#29

It is proposed to incorporate following Indian languages text segmentation requirements in UAX#29

- Additional information on Indic orthographic syllable boundaries based on tailored grapheme cluster define in UAX#29
- ABNF valid segmentation definition to define Indian languages orthographic syllable
- No break rules for determination of Indic syllable boundary
- Information for identification of boundaries of first letter styling, Guiding principles of line breaking at syllable level for Indian languages.
- Detailed report at L2/16-161

Thanks