Indian language requirements for String identity matching on web

Introduction

The document formats and protocols that are based on character data are mainly prepared for the web. These protocols and formats can be accessed as resources that contain the various text files that cover syntactic content and natural language content in some structural markup language. In order to process these types of data , various string based operations such as searching, indexing, sorting, regular expressions etc are required. These documents inspect the text variations of different types and preferences of the user for string processing on the web. W3C has developed two documents Character Models: String Matching and searching that act as building blocks related two these problems on the web and defining rules for string manipulation i.e. string matching and searching on the web. These documents also focus on the different types of text variations in which same orthographic text uses different character sequences and encodings. The rules and facts of Indian languages defined in these documents act as a reference for the authors, developers etc. for consistent string manipulation on the web.

Goal of the document

The document discussed about the following two types of the character variations that seen in the Indian language text for five languages :

- Orthograhic representations(Alternate Spellings) that are not taken care by normalization
- Encoding variations i.e same written form with different character sequences and handled by normalization.

These variations may be ocur due to different user preferences and different keyboard representation. It is important that these types of variations need to be included in current W3C documents for the reference and consistent Indian languages string manipulations on the web. The document does not cover the inappropriate character sequences that are not made by normalization.

Orthographic variations

The Unicode Standard gives different alternatives to define text but requires that both texts should be treated identical. In order to improve efficiency, it is recommended that an application will normalize text before performing string manipulation operations such as search, comparisons on the web. In Indian languages, users can use different orthographic representations of the same text with different encoding as lack of language knowledge and no matter what the rules exist. This will cause unexpected results while searching and matching of string by the users as both string uses different code points. The spelling variations lead to inappropriate searching results. Some of the Indian languages rules/facts that lead to different written forms are shown below:

Hindi

Rules: Chandrabindu- Anunasik(◌ँ)

Rule 1: Anunasik is used only in conjunction with some vowels and matras i.e. [अ, आ, उ, ऊ] /ə, a, ʊ, u/.

Eg. हँस,चाँद,पूँछ,माँ, आँख,गाँव,उँड़ेल, बाँस, सँभाले, धँसकर, गाँव, मुँह, धुँधले, धुआँ, काँप, मँहगाई, उँगली, काँच, बूँदें, कुआँ, भाँति, जाऊँगा, ढूँढने, ऊँचे, पूँछ, काँच, झाँकते, अँधेर, माँ, फूँकना, भाँति, अँगूठा, बाँधकर, पहुँच, आँगन, कँप-कँपी, ठूँस, गूँथ, ऊँचाई, टाँग, पाँच, साँस, दाँत, झाँका, मुँहजोर। etc.

Rule 2: Anunasik is replaced with bindu if a vowel/matra extends above the line of the letters (superscript).

Eg. गोंद,कोंपल etc.

Rules: Bindu-Anuswar (◌ं)

Rule 1:Use of anuswar in place of the fifth letter i.e ङ, ञ, ण, न, म of the class as mentioned below:

Class 1: (क, ख, गand घ) we have; ङ

Class 2: (च, छ, जand झ) we have; ञ

Class 3: (ट, ठ, डand ढ) we have; ण

Class 4: (त, थ, दand ध) we have; न

Class 5: (प, फ, बand भ) we have; म

The spelling variations occur as user can use different orthographic representations of the same text such as:

ञ in place of Anuswar(◌ं) eg. मंजूषा --- मञ्जूषा etc.

णin place of Anuswar(◌ं) eg.ठंडा & ठण्डा, मंडी & मण्डी, झण्डा & झंडा , घंटी & घण्टी etc.

नin place of Anuswar(◌ं)eg. हिंदी & हिन्दी, धंधा & धन्धा, गन्दा & गंदा , दंत & दन्त, ग्रंथ & ग्रन्थ, बंदर & बन्दर, etc.

मin place of Anuswar(◌ं)eg. चम्पक & चंपक ,अम्वर & अंवर, कम्पन & कंपन भूकंप & भूकम्प, संभावना & सम्भावना, चंबल & चम्बल etc.

Different users can use different spellings of the same text, as both spellings are currently in use and thus do not meet the user requirements and give improper results while searching and matching.

Rule 2: If there is only nasal sound without short vowel then half nasal sound is written orthographically.

Eg. प्रसन्न, अन्न, सम्मेलन etc. Rule 3: /ŋ, ɲ, ɳ, n, m/ IPA sign is used for marking Anuswar.

If nasal consonant comes with another consonant before homorganic stops, then both are written orthographically in place of Anuswar eg.

अन्य, चिन्मय, उन्मुख etc.

Position of Anusvara at the end of the word -

Many times when there is a vowel-less म (म्) at the end of the word, it is also written as Anuswar. शिवम्-- शिवं (lord shiva), अहम्- अहं (ego), स्वयम् -स्वयं (self) etc.

Note: In some words, the use of Anunasik and Anusvara make a difference in the meaning and pronunciation of the word. eg.

हंस /həns/(Swan) & हँस /hə̃s/(Laugh), कंस /kəns/ (a name) & कँस /kə̃s/(tight), बंटा /bəɳʈɑː/ (a name) & बँटा /bə̃ʈɑː/ (divided) etc.

Bengali

Bengali, is one of the notorious languages with regard to spelling variation. The different spellings of word having same meaning are accepted in the Bengali language, it should be treated as different word although have same meaning. It has 5000+ words which record spelling variations. Typically, spelling variation ranges from 2 to 8 words. Majority of words have 2 variations; some have 3, 4, 8 and more variations. At least there is one word that records 16 spelling variations. Nearly 80% words show two spelling, 7% words show three variations, 7% words show four spellings, and 6% words show more than four variations. Some of the examples are shown below:

হল–হলো–হোল–হোলো,
রাণি–রাণী–রানি–রানী
চাপরাশ–চাপরাশি–চাপরাশী–চাপরাস–চাপরাসি–চাপরাসী
অঙ্ক–অংক,
নিচ–নীচ,
ভাল–ভালো,
মত–মতো,
গরু–গোরু,
চিন–চীন,
হিরা–হীরা
বাংলা–বাঙলা–বাঙ্গলা,
কলকাতা–কলিকাতা–কোলকাতা
অঙ্ক–অংক

Malayalam

Malayalam language also has different spellings for same words while typing. Only few words have spelling variations. Some of the spelling variations in Malayalam are shown below:

EX-1. അധ്യാപകൻ, അദ്ധ്യാപകൻ

EX-2. ദുഃഖം, ദുഖം

EX-3. തർപണം, തർപ്പണം

EX-4. കൽപന, കല്‍പന

EX-5. അർഥം, അർത്ഥം

EX-6. നന്ദൻകോട്, നന്തൻകോട്

EX-7. അദ്ദേഹം, അദ്ധേഹം

EX-8. അധ്വാനം, അദ്ധ്വാനം

EX-9. അത്ഭുതം, അൽഭുതം

EX-10. മാധ്യമം, മാദ്ധ്യമം

EX-11. വിദ്യുച്ഛക്തി, വിദ്യുഛക്തി, വിദ്യുത്ശക്തി

EX-12. സൽകീർത്തി, സത്കീർത്തി

EX-13. പഞ്ജരം, പഞ്ചരം

EX-14. പഞ്ചസ്സാര, പഞ്ചസാര

EX-15. കട്ടിള, കട്ടില

EX-16. പാരമ്പര്യം, പാരംപര്യം

EX-17. അധ്യാപകൻ, അദ്ധ്യാപകൻ

EX-18. അധ്യാപിക, അദ്ധ്യാപിക

EX-19. ക്ലിപ്തം, ക്നുപ്തം

EX-20. ചർമ്മം, ചർമം

Odia

The nasal consonants such as 'ଙ, ଞ, ଣ, ନ, ମ' can be the first letters in a conjunct such as 'ଙ୍କ, ଙ୍ଖ, ଙ୍ଗ, ଙ୍ଘ, ଞ୍ଚ, ଞ୍ଛ, ଞ୍ଜ, ଞ୍ଝ, ଣ୍ଟ, ଣ୍ଠ, ଣ୍ଡ, ଣ୍ଢ, ନ୍ତ, ନ୍ଥ, ନ୍ଦ, ନ୍ଧ, ମ୍ପ, ମ୍ଫ, ମ୍ବ, ମ୍ଭ' which can be written with Anuswar "ଂ (0B02)' instead of the nasals.Eq.

ନିଳକଣ୍ଠ - ନିଳକଂଠ

ପିମ୍ପୁଡି -ପିଂପୁଡି

ବନ୍ଧୁକ - ବଂଧୁକ

ଥଣ୍ଡା - ଥଂଡା

Gemination or doubling may take place in the case of dental and palatal consonants. So ତ or ତ୍ତ and ଜ or ଜ୍ଜ are found.

(Both are accepted)

ନିର୍ଜ୍ଜନ - ନିର୍ଜନ ('ନିର୍ଜ୍ଜନ' Not acceptable)

ପରିବର୍ତ୍ତନ - ପରିବର୍ତନ

The letter ଡ଼ is used only in the non-initial positions andଡ only in the initial position. Most educated people are writing ଡ଼ in the initial position and it is a mistake due to ignorance.

ଓଡ଼ିଶା - ଓଡିଶା ('ଓଡିଶା' Not acceptable)

Marathi

Writing परसवर्ण is optionally allowed only forतत्सम words and not for non-तत्सम words. Also, for meaning change as inवेदांत which means- 'in the Veda' as against, वेदान्त which means- end of the Veda or a school of philosophy.

Eq. पण्डित/पंडित

अन्तर्गत/ अंतर्गत

As per the Marathi Grammar rules, तत्सम words(Sanskrit words accepted by Marathi as they are) that end with इकार or उकार are always written with दीर्घ इकार or उकार Ex. कवि is written as कवी

Secondly all original Marathi wordsthat end with इकार or उकार are also always written with दीर्घ इकार or उकार Ex. गाडी, विडी, साडी, मऊ, खाऊ, माती

Gujarati

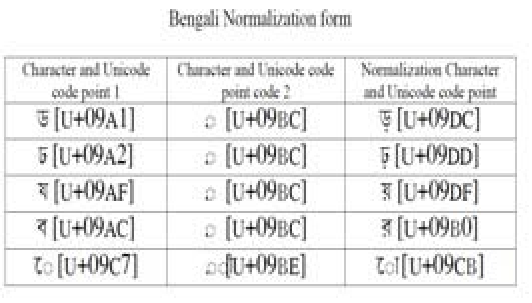
મંદિર/મન્દિર
ચંપલ/ચમ્પલ
શાંતિ/શાન્તિ
પ્રબંધક/પ્રબન્ધક
રંક/રઙક
પથ્થર/ પથ્ થર
પંડિત/પણ્ડિત *
* ण्ड
*The hand written form of the Gujarati nasal retroflex occuring in conjunction with a retroflex consonant, takes on the form similar to the Hindi nasal retroflex. However, that form is not generally found in typed texts.

1. Encoding variations

It is not always true that in the different keyboard layouts, keystrokes and input characters are same and one to one. It is depends on the type of the keyboard layout. Some keyboards can produce numerous characters from a single key press and some keyboards use different keystrokes to produce one abstract character. It is the limitations of Indian languages that too many characters need to be fit in one single keyboard. This leads to input more complex Indian languages input methods and which makeover keystrokes sequence in character sequences. In Indian languages, too many characters to fit on a keyboard and must rely on more complex input methods, which transform keystroke sequences into character sequences. It might occurs that different character sequences of the same text used by different users from the different keyboard creates the issues in string identity matching. Keyboard driver maps keybstrokes to normalized Unicode storge to overcome this problem.
In some cases the same chracter can be type using different key stokes as input which required normalization before saving.

Some examples of the encoding variations are shown below:

Language	Unicode Variation	Remarks
Hindi	Ex1. DEVANAGARI LETTER RRA •ऱ=0931 ऱ ≡ 0930 र 093C ◌़ Ex2. DEVANAGARI LETTER QA क़ =0958 क़≡ 0915 क 093C ◌़ Ex3. DEVANAGARI LETTER KHHAख़=0959 ख़ ≡ 0916 ख 093C ◌़ Ex4 DEVANAGARI LETTER GHHAग़=095A ग़ ≡ 0917 ग	

<table>
<tr><td></td><td>093C ़
Ex5. DEVANAGARI LETTER ZA ज़=095B ज़≡ 091C ज 093C ़
Ex6. DEVANAGARI LETTER DDDHAड़=095C ड़ ≡ 0921 ड 093C ़
Ex7. DEVANAGARI LETTER RHA ढ़=095D ढ़≡ 0922 ढ 093C ़
Ex8. DEVANAGARI LETTER FA फ़=095E फ़≡ 092B फ 093C ़
Ex9. DEVANAGARI LETTER YYA य़=095F य़≡ 092F य 093C ़
Ex10. DEVANAGARI LETTER NNNA ऩ=0929 ऩ≡ 0928 न 093C ़</td><td></td></tr>
<tr><td>Bengali</td><td>Ex 1. BENGALI VOWEL SIGN O ো=09CB
ো≡ 09C7 ে 09BE া Ex 2. BENGALI VOWEL SIGN AU ৌ=09CC
ৌ≡ 09C7 ে 09D7 ৗ

Bengali Normalization form
<table><tr><th>Character and Unicode code point 1</th><th>Character and Unicode code point code 2</th><th>Normalization Character and Unicode code point</th></tr><tr><td>ড [U+09A1]</td><td>় [U+09BC]</td><td>ড় [U+09DC]</td></tr><tr><td>ঢ [U+09A2]</td><td>় [U+09BC]</td><td>ঢ় [U+09DD]</td></tr><tr><td>য [U+09AF]</td><td>় [U+09BC]</td><td>য় [U+09DF]</td></tr><tr><td>ব [U+09AC]</td><td>় [U+09BC]</td><td>র [U+09B0]</td></tr><tr><td>ে [U+09C7]</td><td>া [U+09BE]</td><td>ো [U+09CB]</td></tr></table></td><td>Both encodings should be allowed for Vowel Sign O, leaving it to the search engines to normalize before searching. This should be the general policy for all bipartite representations of single characters, e.g. র, য়, জ় etc. Allow both unique and bipartite representations, but normalize before search. The alternative of allowing only single code representations of all these characters is also acceptable, but that would entail conversion of all existing texts containing bipartite representations of these characters.</td></tr>
<tr><td>Odia</td><td>Ex 1.OdIa VOWEL SIGN O ୋ=0B4B
ୋ ≡ 0B47 େ 0B3E ା

Ex 2. OdIa VOWEL SIGN AU ୌ =0B4C
ୌ≡ 0B47 େ 0B57ୗ

Ex 3. OdIa LETTER RRA = ddaଡ଼=0B5C
ଡ଼ ≡ 0B21 ଡ 0B3C ଼

Ex4.Odia LETTER RHA = ddhaଢ଼=0B5D
ଢ଼≡ 0B22 ଢ 0B3C ଼</td><td></td></tr>
<tr><td>Malayalam</td><td>Ex 1. MALAYALAM VOWEL SIGN O ൊ =0D4A
ൊ≡ 0D46 െ 0D3E ാ

Ex 2. MALAYALAM VOWEL SIGN OO ോ
=0D4Bോ= 0D47 േ 0D3E ാ

Ex3.MALAYALAM VOWEL SIGN AU • ൌ
=0D4Carchaic form of the /au/ dependent vowel →
ൌ =0D46 െ 0D57 ൗ

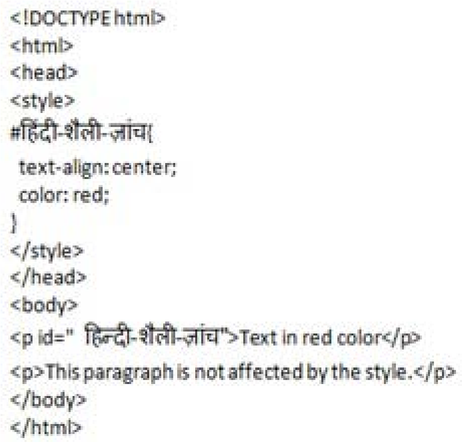
Ex 4. Chillu and Number
ർ = U+0D7C – Alphabet CHILLU RR</td><td>• Malayalam Chillu letters and Alphabets with vowel signs can be typed in different forms.
• Chillu has single atomic value and also in com-bined form. Both are in use.</td></tr>
</table>

	൪ = U+0D6A – Numeral representation 4 ൻ = U +0D7B - Alphabet CHILLU N ൯ = U +0D6F - Numeral representation 9 Ex 5. Chillu typing ൻ = U + 0D7B (single atomic value) ന്‍ = U + 0D28 ന U+ 0D4D ◌് with ZWJ ൺ = U + 0D7A (single atomic value) ണ്‍ = U + 0D23 ണ U+ 0D4D ◌് with ZWJ ർ = U + 0D7C (single atomic value) ര്‍ = U + 0D30 ര U + 0D4D ◌് with ZWJ ൽ = U + 0D7D (single atomic value) ല്‍ = U + 0D32 ല U + 0D4D ◌് with ZWJ ൾ = U + 0D7E (single atomic value) ള്‍ = U + 0D33 ള U + 0D4D ◌് with ZWJ	
Gujarati	Ex 1. GUJARATI RUPEE SIGN • preferred spelling is ૱=0AF1 ૱ =0AB0 ર 0AC2 $◌ૂ 0AF0 ▯	
Marathi	Ex 1. DEVANAGARI LETTER LLLA • ऴ=0934 ऴ≡ 0933 ळ 093C ◌़	

Use-Cases

1. Text variation in syntactic content under HTML/CSS & other application

The role of syntactic content in a document format and protocol is to represent the text that defines the structure of the document format and protocol. The different values used to define id, class name in markup languages and cascading style sheets are a part of syntactic content. In order to produce output as desired, we should ensure that the selectors and id or class name should be same as shown in the below example:



```
<!DOCTYPE html>
<html>
<head>
<style>
#हिंदी-शैली-जांच{
 text-align: center;
 color: red;
}
</style>
</head>
<body>
<p id=" हिन्दी-शैली-जांच">Text in red color</p>
<p>This paragraph is not affected by the style.</p>
</body>
</html>
```

In the above example, the id name defines in the HTML and CSS works on the same character sequences.

हिंदी = 0939+ 093F+0902+0926+0940

हिन्दी =0939+ 093F+0928+094D+0926+0940

ज़ांच = 095B+093E+0902+091A

ज़ांच =091C+093C+093E+0902+091A

Gaps will be there if id name uses different character sequences .This is particularly occurs and leads an issue if markup language and the CSS are being handled or maintained by different persons.

2. **Implementation of Internationalized domain name & Email addresses**

It does not require that user should have knowledge of text normalization it will be taken care by system before procced. There is another option in case of internationalized domain name which is called Variant table that can be used to normalize the visual similarty in case of IDN.

3. **Indian language search operations on the web**

User can search natural language content by using find command on the web. Different Users might use different character sequences of the same text by performing find command. There should be some common mapping and implementation in order to satisfy the users need.

Requirements for Indian languages

The above-defined W3C draft Standards specify the requirements while implementing string matching of syntactic content and search of natural language content by using matching rules. The following Indian language requirements need to introduce in the standards in order to perform proper string operations on the web:

- Different characters variations including orthographic and encoding variations in Indian languages need to address.
- In addition, it is recommended that Indian languages characters need to be post processed through normalization defined by Unicode for comparison and searching on the web. Character sequences can taken care by text Normalization but spelling variations is the natural fenomina of the language so we can not restrict them and implementaion is required for such cases.

Contributors

Sl. No.	Name	Organization
1.	Dr. Swaran Lata	NeGD
2.	Prof.Gautam Sengupta	University of Hyderabad
3.	Shri. NiladriShekhar Dash	Linguistic Research Unit Indian Statistical Institute, Kolkata
4.	Prof. Girish Nath Jha	JNU
5.	Pof. Panchanan Mohanthy	University of Hyderabad
6.	Prof. Shyamal Das Mandal	IIT kharagpur
7.	Dr. Rudranarayan Mohapatra	Utkal University
8.	Shri. Bhadran.V.K	Alibi Global
9.	Dr. Elizabeth Sherly	Director IIITM-K

10.	Shri. JITHESH.V.S.	IIITM-K
11.	Prof. Malhar Kulkarni	IIT Mumbai
12.	Dr. Mona Feroz Parakh	MS University
13.	Shri. Chinmay Madhu Ghaisas	Dept. Of Marathi, Goa University, Taleigao-Goa
14.	Members of IEEE Pre-Standardization “Language processing Text” Sub Committee	IEEE
15.	Shri Vijay Kumar	MeitY
16.	Shri. Bharat Gupta	MeitY
17.	Shri. Prashant Verma	WSI, MeitY
18.	Shri. Puneet Grover	CDAC Pune

References

1. Character Model for the World Wide Web: String Matching, 2021: https://www.w3.org/TR/charmod-norm/
2. W3C String Searching, 2020: https://w3c.github.io/string-search/
3. Unicode Consortium: https://home.unicode.org/
4. Normalization in HTML and CSS: https://www.w3.org/International/questions/qa-html-css-normalization/
5. Unicode Devanagari Code Chart,2020: https://unicode.org/charts/PDF/U0900.pdf
6. Unicode Normalization Forms, 2020: http://www.unicode.org/reports/tr15/
7. https://hindimiddleeast.com/difference-between-anuswar-and-anunasik/
8. https://www.cse.iitk.ac.in/users/cs671/2015/resources/pandey-14_akshara-to-sound-rules-for-hindi.pdf
9. http://information2media.blogspot.com/2016/12/blog-post_0.html