

भारतीय भाषाओं में पाठ प्रदर्शन, इनपुट, खोज तथा प्रोसेसिंग

Text Rendering, Input, Search and Processing in Indian Languages

करन मिश्र · Karan Miśra

Apple अंतरराष्ट्रीयकरण · Internationalization

३ सितंबर २०२४ · 3 September 2024

भारतीय भाषाओं और लिपियों की रचना के साथ साथ उन्हें यूनिकोड जैसे मानकों में एनकोड करने के तरीकों के कारण पाठ्य सामग्री के उपभोग और उत्पादन दोनों के लिए अत्यधिक उपयोगी और त्रुटिहीन अनुभव बनाने के संदर्भ में कई अहम बातें गौरतलब होती हैं।

इस दस्तावेज़ में निम्नलिखित विषयों पर सुझाव है :-

1. **पाठ प्रदर्शन**

वर्णों की अलग अलग सीक्वेंस एक जैसी न दिखें ताकि उन्हें धोखाधड़ी के लिए उपयोग न किया जा सके।

2. **पाठ इनपुट**

टाइप किए वर्ण और सीक्वेंस भाषा के लिए वैध हों।

3. **पाठ खोज**

पाठ्य सामग्री की खोज आसानी से की जा सके भले ही उसे इनपुट करने का तरीका खोज पाठ को इनपुट करने के तरीके से भिन्न हो।

4. **पाठ प्रोसेसिंग**

डेटा प्रोसेसिंग के दौरान अवैध वर्णों और सीक्वेंस को अस्वीकार या सामान्यीकृत किया जा सके।

नीचे दिए गए कई विचार उन भाषाओं के लिए विशिष्ट हैं जो ब्राह्मीजनित लिपियों का उपयोग करती हैं जैसे कि कन्नड़, गुरुमुखी, तमिल, देवनागरी आदि। हालाँकि कुछ भागों में उर्दू भाषा के लिए भी उदाहरण दिए गए हैं जो अरबी-फ़ारसी लिपि का उपयोग करती हैं। ब्राह्मीजनित लिपियों के लिए इस दस्तावेज़ में दिए उदाहरण हालाँकि देवनागरी लिपि का उपयोग करते हैं लेकिन दिए गए सुझाव व्यापक रूप से लागू होते हैं।

Due to the nature of Indian languages and scripts as well as the manner in which they have been encoded in standards such as Unicode, there are many important considerations in terms of creating a highly usable and error-free experience for both consuming and producing textual content.

This document proposes solutions for :-

1. **Text Rendering**

Ensuring different sequences of characters do not appear identically, to prevent their being used for nefarious purposes such as phishing attacks.

2. **Text Input**

Ensuring that characters and sequences conform to the set of valid set for the language being typed.

3. **Text Search**

Ensuring that users are able to easily search for text content even when it is input using different conventions from the target text.

4. **Text Processing**

Ensuring that invalid characters and sequences can be rejected or normalized during data processing.

Many of the considerations below are specific to languages that use Indic scripts, i.e. Brahmi-derived scripts such as Devanagari, Gurmukhi, Kannada, Tamil, etc., though certain examples are also provided for Urdu that uses the Perso-Arabic script. For Indic scripts, the examples used throughout this document use the Devanagari script. However, the concepts are broadly applicable.

पाठ प्रदर्शन

ब्राह्मीजनित लिपियों के इस्तेमाल में मुख्यतः दो समस्याएँ प्रकट होती हैं :-

1. जो स्वर अक्षर एवं चिह्न अखंड रूप से एनकोड किए गए हैं (उ० आ, ाँ आदि), उन्हें भाषा इस्तेमाल करने वाले अखंड नहीं समझते हैं (उ० आ → अ + ा, ाँ → ा + ँ आदि) और उन्हें इसी खंडित ढंग से टाइप करते हैं। लेकिन ऐसी सीक्वेंस को यूनिकोड में अवैध माना जाता है (यूनिकोड मानक §12.1 "Vowel Letters")।
2. स्वरों की मात्राओं और अन्य चिह्नों को अवैध तरीके से इनपुट किया जा सकते हैं। उदाहरणतः दोहराए गए इनपुट (केे, मेंं आदि) और गलत क्रम वाले इनपुट (कुटुंब, िक आदि)।

सुझाव

1. यूनिकोड में अखंड रूप से एनकोड किए गए स्वर अक्षरों के लिए केवल अखंड सीक्वेंस को ही वैध माना जाना चाहिए। खंडित सीक्वेंस को अवैध मान कर बिंदीदार गोले के साथ दिखाना चाहिए।

Text Rendering

For Indic scripts, there are two salient problems that emerge :-

1. For vowel letters and signs that are encoded atomically (e.g. आ, ाँ, etc.), many users naturally perceive these as consisting of a base and vowel sign (e.g. आ → अ + ा, ाँ → ा + ँ, etc.) and type them in this manner; however, such sequences are considered invalid in Unicode (Unicode Standard §12.1 "Vowel Letters").
2. Vowel signs and other modifier signs can be input in invalid combinations, including repeated inputs (e.g. केे, मेंं, etc.) as well as out-of-order input (e.g. कुटुंब, िक, etc.).

Recommendation

1. For vowel letters that are encoded atomically in Unicode, only the atomic sequence should be considered valid and decomposed sequences should be considered invalid and rendered with dotted circles.

अ + ा → अा

अा <U+0905, U+093E> को आ <U+0906> की तरह प्रदर्शित नहीं होना चाहिए बल्कि उसे बिंदीदार गोले के साथ प्रदर्शित होना चाहिए।

अा <U+0905, U+093E> should not render identically to आ <U+0906> and should render with dotted circles.

2. जहाँ एक अक्षर पर एकाधिक मात्राएँ या चिह्न चढ़ाने हों यूनिकोड मानक §5.13 में निर्धारित "अंदर से बाहर" वाले नियम का पालन करना चाहिए और अवैध सीक्वेंस को बिंदीदार गोले के साथ दिखाना चाहिए।

2. Sequences with multiple dependent marks or signs should follow the inside-out rule as laid out in the Unicode Standard §5.13, and invalid sequences should be shown with dotted circles.

अ + ँ + ँ → अँँ

अँँ <U+0905, U+0902, U+0945> को बिंदीदार गोलों के साथ दिखाना चाहिए और न कि अँ <U+0905, U+0945, U+0902> की तरह जो मराठी में वैध है।

अँँ <U+0905, U+0902, U+0945> should render with dotted circles and not like अँ <U+0905, U+0945, U+0902>, which is valid in Marathi.

3. दोहराई जाने वाली मात्राओं या चिह्नों वाली सीक्वेंस जिसमें किसी एक मात्रा या चिह्न को दूसरे के ऊपर चढ़ाना संभव न हो उन्हें बिंदीदार गोलों से दिखाना चाहिए।

3. Sequences with repeating marks or signs that cannot follow any stacking behaviour should be shown with dotted circles.

स + े + े → सेे
स + ा + ा → साा

चूँकि एकाधिक े <U+0947> की मात्रा को चढ़ाने की कोई विधि नहीं है इसलिए दोहराई मात्राओं को बिंदीदार गोलों के साथ दिखाना चाहिए। यह भी गौरतलब है कि सेे <U+0938, U+0947, U+0947> जैसी सीक्वेंस को हरगिज़ से <U+0938, U+0947> की तरह नहीं दिखाना चाहिए। किंतु यदि साा <U+0938, U+093E, U+093E> का उदाहरण लिया जाए तो इसमें दोहराई मात्राओं को बिंदीदार गोलों के साथ न दिखा कर उन्हें एक के बाद एक चढ़ाया जाना चाहिए।

There is no valid stacking behaviour for repeated occurrences of े <U+0947>. Thus, any repeated occurrences of े <U+0947> should be shown with dotted circles. Notably, सेे <U+0938, U+0947, U+0947> should not render identically to से <U+0938, U+0947>.

In contrast, for साा <U+0938, U+093E, U+093E>, the repeated vowel signs can be stacked and thus should not be rendered with dotted circles.

4. सं० 1-3 नियमों के दायरे के बाहर वाली सीक्वेंस को बिंदीदार गोलों के बिना दिखाना चाहिए ताकि इस्तेमाल में अभिव्यक्ति की व्यापकतम स्वतंत्रता प्राप्त हो। परिशिष्ट में “उपयोग के उदाहरण” देखें।

4. Sequences that do not fall into N^o1–3 should render without dotted circles to ensure that users retain the broadest possible freedom of expression. See “Usage Examples” in the appendix.

अ + े → ऐ
क + ु + ा → कुा
उ + ः → उँ

ऐ <U+0905, U+0947>, कुा <U+0915, U+0941, U+093E> तथा उँ <U+0909, U+093C> ऐसी सीक्वेंस हैं जो किसी और सीक्वेंस से मिलती जुलती हो कर भ्रम पैदा नहीं करती हैं और इन्हें बिंदीदार गोलों के बगैर दिखाना चाहिए।

ऐ <U+0905, U+0947>, कुा <U+0915, U+0941, U+093E> and उँ <U+0909, U+093C> are sequences that are not confusable with any other sequences and should render without dotted circles.

5. बिंदीदार गोलों को दिखाने के नियमों को मानकीकृत किया जाना चाहिए।

5. Sequences that cause dotted circle rendering to appear should be standardized.

पाठ इनपुट

“पाठ प्रदर्शन” भाग के अनुसार ब्राह्मीजनित लिपियों का उपयोग करने वाली भाषाओं के लिए जो कुछ भी लिपि स्तर पर अवैध है वह भाषा स्तर पर भी अवैध है। इसके अलावा भारतीय भाषाओं के लिए भाषा संबंधी विधिमान्य होने के लिए मुख्यतः दो समस्याएँ प्रकट होती हैं :-

1. लिपि में सारे वैध वर्णों और सीक्वेंस को भाषाएँ अकसर केवल आंशिक रूप से इस्तेमाल करती हैं। चूँकि कई लिपियों में दिखने में मिलते जुलते वर्ण और सीक्वेंस मौजूद होते हैं इससे पैदा हुए भ्रम के कारण या फिर कीबोर्ड लेआउट से अपरिचित होने के कारण अवैध वर्ण और सीक्वेंस टाइप हो सकते हैं। उ० “ने” की बजाए “नै” टाइप करना।
2. वैध वर्णों और सीक्वेंस के भीतर भी दिखने में मिलता जुलता होने के कारण या फिर कीबोर्ड लेआउट से अपरिचित होने के कारण भ्रम पैदा हो सकता है। उ० “ङ” की बजाए “ङ्” टाइप करना।

उदाहरण

देवनागरी लिपि

देवनागरी लिपि में कई भाषाओं की लिखावट होती है और इसमें कई भाषा आधारित परंपराएँ भी हैं। विशेष रूप से हिन्दी और मराठी को देखा जाए तो :-

- हिन्दी में ् नुक्ते और ँ चंद्रबिंदु का उपयोग व्यापक है।
- मराठी में “ळ” तथा “ऴ” का उपयोग व्यापक है।
- हिन्दी और मराठी में कुछ ऐसी मात्राएँ और चिह्न भी इस्तेमाल किए जाते हैं जो दिखने में एक जैसे लगते हैं लेकिन इस्तेमाल में फ़र्क़। उ० “टँग” को हिन्दी में ट ँ ग क्रम में और मराठी में ट ँ ँ ग क्रम में टाइप किया जाएगा और मायने तथा उच्चारण दोनों में भेद होगा।

Text Input

Per the “Text Rendering” section, for languages that use Indic scripts, anything that is invalid at the script level is also invalid at the language level. In addition to this, there are two salient problems that emerge for language-based validation for Indian languages :-

1. The valid characters and sequences for a given language are often a subset of valid characters and sequences for the script as a whole. However, language users often type characters and sequences outside of the valid set due to visual confusability or not being as familiar with the keyboard layout, e.g. ने *nē* being typed as नै *ne*.
2. One valid character or sequence may be confused with another due to visual similarity or not being as familiar with the keyboard layout, e.g. ङ *ṅa* being typed as ङ् *nga*.

Examples

DEVANAGARI SCRIPT

The Devanagari script is used to write many languages and there are many language-specific conventions. Looking at Hindi and Marathi specifically :-

- Hindi makes prolific use of ् *nuqta* and ँ *chandrabinḍu*.
- Marathi makes use of ळ *ḷa* and ळ *eyelash ra*.
- There are also sequences that are visually identical but are composed differently in Hindi and Marathi, e.g. टँग which should be typed as ट ँ ग in Hindi and as ट ँ ँ ग in Marathi; they would have different meanings and different pronunciations.

अरबी-फ़ारसी लिपि

कई भाषाओं को अरबी-फ़ारसी लिपि में लिखा जाता है जैसे कि अरबी, उर्दू, कश्मीरी, सिंधी इत्यादि। हर भाषा में लिपि को भिन्न रूप से इस्तेमाल किया जाता है।

- कुछ वर्णों के लिए उर्दू में अरबी व सिंधी के मुकाबले अलग कोडपॉइंट इस्तेमाल होता है। उ० ۛ "ये" के लिए उर्दू में U+06CC का उपयोग होता है बजाए U+0649 या U+064A।
- उर्दू में पूर्ण विराम के लिए U+06D4 चिह्न का इस्तेमाल होता है और अरबी में U+002E का।
- कश्मीरी में U+0620 ۛ "कश्मीरी ये" वर्ण का इस्तेमाल होता है जो सिर्फ कश्मीरी में ही पाया जाता है।

सुझाव

प्रत्येक भाषा के लिए कीबोर्ड लेआउट और इनपुट टूल यह सुनिश्चित कर सकते हैं कि भाषा उपयोगकर्ता इन समस्याओं से बचें और केवल उन वर्णों और सीक्वेंस का उपयोग करें जो उस भाषा के लिए वैध हों।

ऐसा अनुभव नीचे वर्णित "अक्षरवार कीबोर्ड लेआउट" के द्वारा रचित किया जा सकता है।

PERSO-ARABIC SCRIPT

The Perso-Arabic script is used to write many languages such as Arabic, Kashmiri, Sindhi, Urdu, etc. and each language uses the script differently.

- For the same conceptual letters, Urdu uses different code points compared to Arabic and Sindhi, e.g. for ۛ *yeh*, Urdu uses U+06CC and not U+0649 or U+064A.
- Urdu uses U+06D4 as its full stop character whereas Arabic uses U+002E.
- Kashmiri makes use of letters like U+0620 ۛ *kashmiri yeh* which are not used in any other language.

Recommendation

Keyboard layouts and input methods for each language can ensure that language users avoid these issues and only use characters and sequences that are valid for the particular language.

Specifically, this can be achieved via a "Contextual Keyboard Layout" as described below.

अक्षरवार कीबोर्ड लेआउट

IS16350:2016 में निर्धारित संवर्द्धित इनस्क्रिप्ट लेआउट प्रत्येक भाषा के लिए इनपुट को वैध वर्णों तक सीमित करते हैं। वर्तमान में यह एक अपरिवर्तनीय लेआउट है जिसमें प्रत्येक कुंजी पहले टाइप की गई कुंजियों को संदर्भ में नहीं लेती है और हर परिस्थिति में समान परिणाम आउटपुट करती है। उदाहरणतः निम्नचित्रित हिन्दी लेआउट के प्रथम प्लेन में कई मात्राओं की कुंजियाँ हैं जिन्हें टाइप करने पर ौ, ै, ा आदि जैसी सीक्वेंस आउटपुट होंगी।

Contextual Keyboard Layout

The Enhanced InScript layouts as specified in IS16350:2016 already limit the input to the characters that are valid for each language. Currently, the specification is a fixed layout; each key produces the same result ignoring previously-typed keys. For example, the default plane of the Hindi layout (shown below) has several dependent vowel sign keys, typing which will produce sequences like ौ, ै, ा, etc. :-



इसको ऐसे लेआउट में परिवर्तित किया जा सकता है जिसमें प्रत्येक कुंजी का आउटपुट संदर्भानुसार हो। उदाहरणतः े आउटपुट करने की बजाए वही कुंजी बिना किसी वर्ण के टाइप किए जाने वाली परिस्थिति में "ए" आउटपुट करेगी :-

This layout can be extended to produce a mechanism in which each key is contextual. For example, instead of producing े, the same key would produce ए when it is typed without any character preceding it :-



यदि कोई व्यंजन टाइप किया जाता है — जैसे “स” — तो वही कुंजी जो शुरू में “ए” आउटपुट करती थी अब े आउटपुट करेगी जिससे इन दो कुंजियों को टाइप करने से “से” आउटपुट होगा :-

Furthermore, if a consonant like स is typed, then typing the same key that produced ए initially will now produce े to form से :-



जब "से" टाइप हो चुका हो तो दोबारा उस स्वर कुंजी को टाइप करने से "सेे" की बजाए "सेए" आउटपुट होगा।

IS16350:2016 §4.3.1 में उल्लिखित सिद्धांतों का पालन करते हुए यह "अक्षरवार" कीबोर्ड लेआउट आज के संवर्द्धित इनस्क्रिप्ट लेआउट का उपयोग करने वाले लोगों के लिए संगतता बनाए रखता है क्योंकि प्रत्येक कुंजी की जगह अपरिवर्तित रहती है और केवल उन टाइपिंग सीक्वेंस को परिवर्तित किया गया है जो वर्तमान में अवैध आउटपुट उत्पन्न करती हों।

At this point, से has been typed and typing the same key again will not produce सेे but rather will produce सेए.

Notably, following the principle outlined in IS16350:2016 §4.3.1, this contextual keyboard layout maintains backwards compatibility for typists using Enhanced InScript today as the positions of each key remain stable and only those typing sequences that would currently produce invalid sequences are modified.

पाठ खोज

भारतीय भाषाओं में हस्तलिपि का लंबा इतिहास है और इसके परिणामस्वरूप एक ही भाषा के बोलने वालों के बीच भी लिखावट की परंपराओं में बहुत भिन्नता होती है। खोज करते समय उच्च मात्रा की स्मरण (रीकॉल) पाने के लिए इस भिन्नता को समझना आवश्यक है।

सुझाव

पाठ खोज में स्मरण की उच्च मात्रा को सुनिश्चित करने के लिए — यानी खोजबीन के लक्ष्य में सफल होने हेतु — वर्तनी में “नियमित भिन्नता” को इस्तेमाल करना चाहिए।

उदाहरण

नियमित भिन्नता कई भाषाओं में पाई जाती है। उदाहरणतः अंग्रेज़ी में “apple” खोजने पर “Apple” या “APPLE” वाले नतीजे मिलने की अपेक्षा होती है और जर्मन में “strasse” खोजने पर “straße” वाले नतीजे मिलने की। यह इसलिए है कि छोटे अक्षरों और बड़े अक्षरों के बीच नियमित संबंध होता है और ß व ss के बीच भी। इस तरह के नियम कई भारतीय भाषाओं में भी पाए जाते हैं।

हिन्दी में एक ऐसा नियम पंचमाक्षर और अनुस्वार वाली वर्तनी में दिखाई देता है जो खुद भाषा के नाम में भी नज़र आता है जिसे लिखने वाले की पसंद के अनुसार “हिंदी” (अनुस्वार) या “हिन्दी” (पंचमाक्षर) दो तरीकों से लिखा जा सकता है। इसके कुछ और उदाहरण नीचे दिए गए हैं :-

अनुस्वार - SIMPLIFIED

कवर्ग - KA

अंक

चवर्ग - CHA

मंच

टवर्ग - ṬA

झंडा

तवर्ग - TA

चंद्र

पवर्ग - PA

स्तंभ

Text Search

Indian languages have a long history of written use and as a result of this, there is a lot of natural variation in spelling conventions even among speakers of the same language. To ensure a high degree of recall when searching, it is essential to handle this variation.

Recommendation

Systematic variation should be handled for text search purposes to ensure a high degree of search recall.

Examples

Systematic variations exist in many languages, e.g. in English, searching for “apple” can be expected to match “Apple” or “APPLE”; in German, searching for “strasse” can be expected to match “straße”, etc. This is because there is a systematic mapping between lowercase and uppercase letters as well as other equivalences like ß = ss. Similar variation and equivalences also exist in Indian languages.

One such example in Hindi is अनुस्वार *anusvāra* (simplified) vs. पंचमाक्षर *pancamākṣara* (traditional) variation, which is exemplified in the name of the language itself, which can be written as हिंदी or हिन्दी depending on the spelling convention employed by the writer. Other examples are shown below :-

पंचमाक्षर - TRADITIONAL

अङ्क

मञ्च

झण्डा

चन्द्र

स्तम्भ

क्योंकि यह भिन्नता नियमित रूप की होती है इसे खोज/मिलान में इस्तेमाल किया जा सकता है। अधिक नियमों को भी लागू किया जा सकता है जिससे कि दिखने में मिलते जुलते वर्णों को खोजा जा सके। उदाहरणतः "ऐ" U+0910 की खोज करने से "ऐ" U+090E वाले परिणामों को शामिल किया जा सकता है, ँ U+06A9 को खोजने पर ँ U+0643 को आदि। इस प्रकार के नियम हर भाषा के लिए बनाए जा सकते हैं।

नीचे दिए गए iOS 18 के उदाहरण में टाइप किए गए पाठ "इंद्र" का "इन्द्र" से मिलान होता दिखाया गया है :-

Because this variation is systematic, it can be encoded in search/collation rules. Additional rules can also be encoded to fold similar characters, e.g. searching for ऐ U+0910 can also match ऐ U+090E, ँ U+06A9 can also match ँ U+0643, etc. due to their visual or conceptual similarities. These rules can be tailored per-language.

In the example below from iOS 18, the typed text इंद्र is able to match इन्द्र :-



पाठ प्रोसेसिंग

एनएलपी कार्यों में उच्च गुणवत्ता वाले परिणाम पाने के लिए साफ़ डेटासेट का उपयोग करना अत्यंत महत्वपूर्ण होता है जिसमें हर भाषा के लिए केवल वैध वर्णों और सीक्वेंस का उपयोग होता है।

सुझाव

प्रत्येक भाषा के लिए किसी भी एनएलपी कार्य के लिए प्रीप्रोसेसिंग चरण में वर्तनी को जाँचा जाना चाहिए ताकि यह सुनिश्चित किया जा सके कि केवल वैध वर्णों का उपयोग किया गया हो। इसके अलावा उपयोग के आधार पर वर्तनी में भिन्नताओं के लिए सामान्यीकरण का उपयोग भी किया जा सकता है।

वर्णों का विधिमान्यकरण

IS16350:2016 §3 प्रत्येक लिपि के लिए मान्य यूनिकोड कोडपॉइंट निर्धारित करता है। IS16350:2016 §4 प्रत्येक भाषा के लिए कीबोर्ड लेआउट को भी निर्धारित करता है जिसमें वैध कोडपॉइंट का आंशिक तौर पर उपयोग होता है। इन भाषा आधारित कोडपॉइंट का प्रत्येक भाषा के लिए वर्तनी जाँच टूल बनाने के लिए उपयोग किया जा सकता है ताकि प्रत्येक भाषा के लिए केवल वैध वर्ण ही उपयोग किए जाएँ। उदाहरणतः हिन्दी के लिए “सुरेंद्र” को अस्वीकार कर दिया जाना चाहिए क्योंकि इसमें े U+0946 मौजूद है जिसका हिन्दी में उपयोग वैध नहीं है।

सामान्यीकरण

ऊपर “पाठ खोज” के तहत प्रस्तावित तरकीबों के समान तकनीकों का उपयोग समानताएँ खोजने और पाठ को सामान्य रूप में ढालने के लिए किया जा सकता है। अधिक उन्नत सामान्यीकरण तकनीकों का भी उपयोग किया जा सकता है जो अंतरलिपीय सामान्यीकरण (उ० “Ganesh” → “गणेश”) या गैरसाधारण भिन्नताओं में सामान्यीकरण (उ० “गनेश” → “गणेश”) कर सकती हैं। इन तकनीकों का विवरण इस दस्तावेज़ के दायरे से बाहर है।

Text Processing

For NLP tasks, to ensure high-quality results, it is important to use a clean data set that uses only valid characters and sequences for each language.

Recommendation

For each language, the orthography should be validated as a preprocessing step for any NLP tasks to ensure that only valid characters are used. Additionally, normalization can be used for spelling variations depending on the use case.

CHARACTER SET VALIDATION

IS16350:2016 §3 specifies the valid Unicode code points for each script. IS16350:2016 §4 additionally specifies the keyboard layouts for each language, which uses a subset of the valid code points. These per-language code points can be used to build Orthography Checkers for each language to ensure that only the valid set of characters is used for each language. For example, सुरेंद्र should be rejected for Hindi NLP tasks, as it contains े U+0946 which is not used in Hindi.

NORMALIZATION

Similar mechanisms as those proposed under “Text Search” above can be used to normalize text and find equivalences. More advanced normalization techniques can also be employed which can normalize across scripts (e.g. “Ganesh” → “गणेश”) or across non-trivial variations (e.g. “गनेश” → “गणेश”). A description of these techniques is outside the scope of this document.

परिशिष्ट

उपयोग के उदाहरण

देवनागरी लिपि के प्रयोग में भिन्नता के कुछ उदाहरण नीचे दिए गए हैं।

1. "ओ", "ऐ" तथा "अु" हिन्दी में "ई", "ए" तथा "उ" के मान्य रूपांतर हैं जो ऐतिहासिक दस्तावेजों में और वर्तमान में अल्पसंख्या में उपयोग में हैं :-

Appendix

Usage Examples

Some examples are given below of variation in the use of Devanagari script.

1. Forms like ओ, ऐ and अु, which are valid variations of ई, ए and उ that are in use both in historical documents in Hindi as well as in minority use today :-

सूची

अल्लाह —के नामसे भगवानको पहचानना ३१, —यानी राम-नाम, खुदा, गॉड, २५, —राम व गॉड अेकार्थक ७, —वही जो गॉड व ओश्वर है १४	ओसा १४
अस्पताल ४१, —पर गाधीजीकी सम्मति ३०	अुपबास और प्रार्थना ३०
आतर्क शाति, प्रार्थना बिना नही मिलती १०	अुरुळीकाचन २३, ३७-४०, ५६
आत्मा ५१, —अमर है, शरीर नही २५	कटिस्नान ३७, ६९, —बाहरी मददके रूपमे ६
आयुर्वेद २६, ३६, ४२, —और कुदरती जिलाज २६, ३६	किशोरलाल मगरूवाला ५४
ऑपरेशन और रामनाम ४७	कुदरत —के नियम ४४-४५;
अिन्फेण्टाइल पैरेलिसिस ३३	—के कानूनोका पालन ४७,
ओश्वर —अेक शक्ति है २३, —और अुसका कायदा २६, —का दर्शन, अपने अन्दर ३४; —का सच्चा भवत ६८, —के अनेक नाम २५, २७, —के हजार नाम २७, ५३, —मनुष्य नही २२, —वैद्य भी है ३३	—के साथ मेल साधकर रहना ५७
	कुदरती जिलाज (अुपचार) २४-२५, २९-३०, ३५-३६, ३८-४२, ४६, —का जरूरी हिस्सा पथ्य परहेज ४१, —की गाधीजीकी कल्पना २४, —की हृद पाच महाभूतोका असल अुपयोग ३६, —के दो पहलू ४०, —गरीबोके लिअे ४१, —पर गाधीजीके विचार ३१, ४०-४१, —मे मसाले और पाक दगैराका स्थान ३६
	कुने ३१, ४३

2. "क्याा" जैसे रूप में मात्रा को कई बार दोहराकर स्वर की लंबी ध्वनि को व्यक्त जाता है। यह उपन्यासों, पत्रिकाओं और सोशल मीडिया में लोकप्रिय है :-

कई लेखक स्वर को लंबा खींचने के लिए अवग्रह — यानी "ऽ" — का उपयोग करते हैं खास तौर पर ऐसी मात्रा के लिए जिसे दोहराया नहीं जा सकता। उ० बड़ेऽऽऽ।

2. Forms like क्याा in which a vowel sound is exaggerated by repeating the vowel sign multiple times, which is popular in Hindi novels, magazines, as well as on social media :-

Many authors choose to use *S avagraha* for this purpose instead, especially when it follows a vowel that does not lend itself to visual reduplication, e.g. बड़ेऽऽऽ.

शीर्ष नवीनतम लोग तस्वीरें वीडियो



Rkrish @Rkrish227... · 17/10/22 ...

क्याा ????



Bad Company. @Ro... · 19/3/16 ...

क्याा 🖐️🖐️🖐️



Expi@e @rijal333 · 27/4/20 ...

कोरिया का मोटे सिकिस्टै भारच क्याा !!



SAURAV KUMAR @... · 14/4/18 ...

@Neeraj_chopra1 जी
आपने क्याा भाला फेंका हैं
वाह

3. "अ" जैसे रूप जिनमें ِ नुक्ते को व्यंजन की बजाए स्वर पर लगाया गया हो अरबी और फ़ारसी मूल के शब्दों को लिखने में आम इस्तेमाल में हैं जैसे कि यहाँ "अली" में दिखाया गया है :-
3. Forms like अ and other examples in which ِ *nuqta* is added to a vowel instead of a consonant, which is common in transcribing words of Arabic or Persian origin, as shown in अली :-



मूल / Source: <https://www.rekhta.org/editors/sayyad-jawwad-ali/ebooks?lang=hi>

4. देवनागरी लिपि में अंग्रेज़ी शब्दों के उच्चारण को इंगित करने वाली शब्दकोश प्रविष्टियों में "थ्र" और "व्र" जैसे रूपों के साथ-साथ U+02C8 तथा U+032E जैसे अन्य चिह्नों का उपयोग भी किया जाता है :-
4. Forms like थ्र and व्र as well as the use of Modifier Letter Vertical Line (U+02C8) and Combining Breve Below (U+032E) in dictionary entries indicating pronunciation of English words in Devanagari script :-

thrive | θraɪv | | थ्राइव् |

verb [I] (pt **thrived** or old-fashioned **throve** | θrəʊv | | थ्रोव् | ;pp **thrived**)

to grow or develop well

सफलतापूर्वक बढ़ना या विकसित होना

फलना-फूलना

identity | aɪ'dentəti | | आइडेंटिटि |

noun [C], [U] (pl. **identities**)

who or what a person or a thing is

किसी व्यक्ति या वस्तु की पहचान

► The region has its own cultural identity.

► The arrest was a case of mistaken identity (the wrong person was arrested).

मूल : ऑक्सफ़र्ड अंग्रेज़ी-अंग्रेज़ी-हिन्दी शब्दकोश

Source: Oxford English-English-Hindi Dictionary