L2/26-061

Discussion points for UNICODE
UTC#185 Meeting
29 October 2025

UNICODE is used for all 22 scheduled languages of India. Data is useful for Mission Bhashini under which various AI based language technologies are being developed by MeitY, GoI. The UNICODE standard is script based, however teachings in schools and implementations by industry are language based.

Background:
India's linguistic and script diversity is rooted in the concept of the **Akshar**, which represents the fundamental orthographic unit in Indic writing systems. Each Akshar typically corresponds to a Grapheme cluster composed of consonants, vowels, and modifiers and functions as the true linguistic building block of Indian languages.

While the **UNICODE** effectively encodes individual characters for all major Indic scripts, it does not explicitly represent the language specific **Akshar-level structure** or the rules that govern how characters combine to form syllabic units. This **character-based encoding** approach limits the faithful digital representation of Indic languages, resulting in:

- Inconsistent rendering and conjunct formation across platforms,
- Difficulties in text search, indexing, and sorting,
- Ambiguities in speech-text alignment and linguistic modeling, and
- Challenges in interoperability across AI, NLP, and transliteration systems.

Purpose of the document:
The purpose of this document is to propose the integration or formal recognition of "*AKSHAR Coded Character Set and Composition Rules Language: Hindi"* within the UNICODE Standard, either as a supplementary guideline or an annexure to the existing Indic script specifications.
The proposal seeks to:

- Introduce standardized rules for Hindi composition rules that complement UNICODE's existing character encoding.
- Provide language-specific framework , currently for Hindi as Akshar Document

Proposed Solution:

To bridge this gap, AKSHAR Coded Character Set and Composition Rules Language: Hindi is a draft standard by the Bureau of Indian Standards (BIS), supported by MeitY and Department of Science and Technology (DST), to define coded character sets and rules for Indian languages starting with Hindi. It establishes formal grammar for valid syllable formation / grapheme cluster rules, ensuring consistent digital text representation across systems. AKSHAR standardizes the handling of special script elements like Nukta, Virama, ZWJ, and ZWNJ, preventing ambiguity in rendering and processing.

By providing validation tools and test resources, it helps developers, font designers, and technology providers ensure interoperability. AKSHAR forms the foundation for digital use of languages using Indian scripts, enabling applications in search, storage, publishing, NLP, and AI/ML. The draft Hindi document has following components:

1. Charset for Hindi language
2. Composition rules as Augmented Backus Naur Formalism (ABNF) for Hindi
3. Sample test strings

Current status:

The Hindi Akshar Document, developed under the supervision of the Ministry of Electronics and Information Technology (MeitY), Department of Science & Technology (DST). It has been formally submitted to the Bureau of Indian Standards (BIS) for review and standardization and available at
https://www.services.bis.gov.in/tmp/WCLITD23526504_22082025_1.pdf (since archived)

The submission was made under the BIS Technical Committee LITD 20 – Language Technology and Localization, which oversees the development and adoption of standards related to Indian language computing and digital text representation.

Specific Discussion Points

1	A B N F formalism for Devanagari	Existing Unicode Annexure: UAX#29 (Unicode Text Segmentation Report), https://unicode.org/reports/tr29/ Syllable Definition/Grapheme Cluster Rules Useful for Cluster Formation, Rendering, Editing, Cursor / Caret - Movement, Hyphenation / Dropcaps, etc.

2	A B N F formalism for Hindi (Language Specific rules)	*AKSHAR Coded Character Set and Composition Rules Language: Hindi draft* (https://www.services.bis.gov.in/tmp/WCLITD23526504_22082025_1.pdf) **Why Hindi Charset:** Devanagari script is shared across languages such as Hindi, Bodo, Dogri, Konkani, Maithili, Marathi, Sanskrit, Santali, Sindhi, etc. Users get confused due to Devanagari Support in Fonts, Keyboards, etc. 0912 ऒ DEVANAGARI LETTER SHORT O • Kashmiri, Bihari languages • also used for transcribing Dravidian short o 0913 ओ DEVANAGARI LETTER O 094A ॊ DEVANAGARI VOWEL SIGN SHORT O • Kashmiri, Bihari languages • also used for transcribing Dravidian short o 094B ो DEVANAGARI VOWEL SIGN O For example : Users feel that 0912 is artistic rendition of 0913 For Hindi Charset, we will include only 0913 and 094B. Or update annotation as below: 0913 DEVANGARI LETTER O: *List out languages* 094B DEVANGARI VOWEL SIGN O: *List out languages* *Even for single script-language cases, there are archaic characters which are causing confusions **Why Hindi Rules:** A sequence which is valid in a language, may be invalid in another language. During the process of adding support to more languages, existing implementations may have got broken. Inconsistent implementation across platforms / OS may be due to lack of documentation of language specific behaviour. See reference links below for variations in such rules UNICODE, Font, ICANN, Harfbuzz, W3C, etc. While many of these cater to Devanagari, we have tuned the same for Hindi Character set and Hindi specific behavior including special signs viz. Avagraha and Abbreviation Sign. Examples of language specific rules Examples : Hindi - Santali क ा ़ = का़ का़ Hindi - Kashmiri च ़ = च़ च़ * Combinations in Red are invalid and Green are valid for respective languages. **Ask:** (1) “AKSHAR Document: Coded Character Set and Composition Rules Language: Hindi , can this document be distributed as UTR/UTN or both? (2) Where in UNICODE can we contribute / enrich such information : UTR / CLDR / Indic mailing list / https://www.unicode.org/ucd/ Chapter 12 – Unicode 16.0.0 / UTN (procedure / WG / responsible person) ? (3) UNICODE may guide which WG can we contribute to / who will deal with it (4) Can a focused / dedicated WG be formed so that we can expedite similar exercise for all 22 scheduled languages and encourage rollout of implementations

3	Valid / Invalid Combinations	Visually, the same but encoding different causes problems in Search, TTS output, Sorting. सहयोग = स ह य ो ग सहयोग = स ह य ा े ग योगा = य ो ग ा योगा = य ा े ग ा चाँंद, किंताब On Notepad in Windows the invalid (2nd string) is rendered with a dotted circle (U+025CC), while in most browsers it joins and both strings look alike. A list for Hindi is submitted for feedback in August/September 2025. https://drive.google.com/file/d/1Sk8mqiUitt0AI2vjK9KERxDjtYPJSTgP/view?usp=drive_link

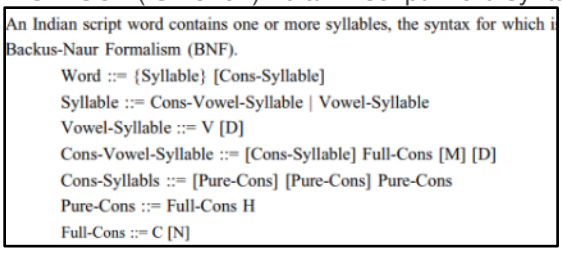
Category	For	Use	Do Not Use		Rationale
Indic_Vowel_Sign	◌ॅ	U+0949	◌ॆ◌ॅ	U+093E U+0945	Ambiguity
Indic_Vowel_Sign	◌ॅ	U+0949	◌ॅ◌ा	U+0945 U+093E	Ambiguity
Indic_Vowel_Sign	◌ॊ	U+094A	◌ॆ◌ॅ	U+093E U+0946	Ambiguity
Indic_Vowel_Sign	◌ॊ	U+094A	◌े◌ा	U+0946 U+093E	Ambiguity
Indic_Vowel_Sign	◌ो	U+094B	◌ॆ◌ा	U+093E U+0947	Ambiguity
Indic_Vowel_Sign	◌ो	U+094B	◌े◌ा	U+0947 U+093E	Ambiguity
Indic_Vowel_Sign	◌ौ	U+094C	◌ॆ◌ौ	U+093E U+0948	Ambiguity
Indic_Vowel_Sign	◌ौ	U+094C	◌ै◌ा	U+0948 U+093E	Ambiguity

Preferred Combinations out of Valid List:
Specific cases have also been submitted for consideration in “DoNotEmit” list

Category	For	Use	Do Not Use		Rationale
Devanagari_Eyelash_Ra	र्	U+0931 U+094D	ऱ	U+0930 U+094D U+200D	Compatibility

Reference Link:
http://www.unicode.org/reports/tr39
https://unicode.org/reports/tr29/Combination
https://www.unicode.org/Public/17.0.0/ucd/DoNotEmit.txt

Ask:
(1) The submitted list to consider (Status of the list) under DoNotEmit.txt or confusables.txt (Script specific). Any further clarification required to complete the current exercise?
(2) Can DoNotEmit be language-wise, so that language specific list can also be proposed
E.g. ◌ॊ -> ◌ो
[DEVANAGARI VOWEL SIGN SHORT O of Kashmiri if typed by user may be normalised to Hindi DEVANAGARI VOWEL SIGN O]
These strings can be added in ‘do-not-emit’ for language Hindi.
(3) How are suggestions verified ?

4	CLDR	The Hindi Character set from the Akshar document may be used in CLDR. For language specific Keyboard Layout: https://www.unicode.org/reports/tr35/tr35-keyboards.html https://www.unicode.org/resources/keyboards.html https://cldr.unicode.org/index/keyboard-workgroup Emoji: https://st.unicode.org/cldr-apps/v#/hi/People/ Language wise Charset (Currently some characters in link below do not belong to Hindi): https://www.unicode.org/cldr/cldr-aux/charts/28/collation/hi.html **Ask:** (1) How can we contribute to Keyboard WG? Or attend Keyboard WG meetings? (2) Is the repository of keyboard Public? If not, how can we access and give our feedback? (3) We have an understanding that, there will be rules for accepting keyboards, such as they have to output normalized text and avoid DoNotEmit sequences, otherwise they would not be allowed If Donotemit is script specific, how we can avoid language specific "DoNotEmit" sequences?
8	Inviting voluntary Participation of Unicode in C-DAC Working Groups	We invite experts for voluntary participation in the C-DAC working groups, so that exercise may be taken up for 22 scheduled languages. WG 1: Script Grammar and Script-Language gaps : Language experts + Academia + Implementing agencies WG 2: Hardware & Software (Keyboard, Fonts, Rendering Engine, Applications, Security) : Implementing agencies WG 3: Training & Capacity Building : Awareness and capacity building
9	References	**Central Hindi Directorate (CHD) :** https://www.chdpublication.education.gov.in/ebook/pdf/devanagarilipiandhindivartanikamankikaran.pdf **Other links with grapheme / cluster formation rules** BIS - ISCII (IS:13194)Indian Inscript Word syntax : rule 8.1  An Indian script word contains one or more syllables, the syntax for which i Backus-Naur Formalism (BNF). Word ::= {Syllable} [Cons-Syllable] Syllable ::= Cons-Vowel-Syllable | Vowel-Syllable Vowel-Syllable ::= V [D] Cons-Vowel-Syllable ::= [Cons-Syllable] Full-Cons [M] [D] Cons-Syllabls ::= [Pure-Cons] [Pure-Cons] Pure-Cons Pure-Cons ::= Full-Cons H Full-Cons ::= C [N] UNICODE Grapheme https://unicode.org/reports/tr29/ (A grapheme is a user-perceived unit of text) and https://www.unicode.org/L2/L2021/21112-deva-cluster-valid.pdf

Expected Outcomes

- **Standardized / consistent Representation:** Language specific Akshar-level encoding rules leading to uniform text behavior across systems.
- **Improved Search and Indexing:** Search engines and NLP pipelines can match, sort, and retrieve words based on Akshar units.
- **Enhanced Linguistic Processing:** Better segmentation and recognition in OCR, ASR, and TTS pipelines for AI engines.
- **Cross-Script Interoperability:** Easier transliteration, translation, and rendering across Indic languages.
- **Dataset Standardization:** Support for consistent corpus creation and linguistic annotation for NLP and AI.

Acknowledgements:

The **Akshar Document** has been developed through extensive consultation and collaborative efforts involving a diverse group of stakeholders including Language Experts, Language Bodies, CDAC, BIS Technical Committee, Academia, Cultural organizations, industry and technology experts.

Way forward:

Submission of the knowledgebase (*AKSHAR Document: Coded Character Set and Composition Rules Language: Hindi*) to the UNICODE Consortium for review and consideration. The intent is to facilitate the reference / integration within the relevant UNICODE Technical Annex (UAX) or associated documentation pertaining to Indic scripts.