

BUREAU OF INDIAN STANDARDS
DRAFT FOR COMMENTS ONLY

(Not to be reproduced without the permission of BIS or used as a STANDARD)

**मसौदा भारतीय मानक
अक्षर
कोडित वर्ण सेट और रचना नियम**

Draft Indian Standard

AKSHAR
Coded Character Set and Composition Rules
Language: Hindi
ICS 35.040; 35.260

© BIS 2025

FOREWORD

(Formal clause will be added later)

This draft Indian Standard will be adopted by the Bureau of Indian Standards, after the draft is finalized by the Indian language technologies and products Sectional Committee LITD 20 and after the approval of Electronics and Information Technology Division Council.

The Ministry of Electronics and Information Technology (MeitY), Government of India has undertaken the task of developing digital standards for Indian languages to strengthen their integration with the internet and emerging technologies. The Department of Science and Technology (DST), Government of India, has also provided valuable support in this initiative.

In this endeavour, to define Coded Character Sets and Composition Rules for the scheduled Indian languages, the “AKSHAR Document” series has been envisaged. These documents will form the foundation for ensuring consistency, accuracy, and interoperability in digital processing of Indian scripts across platforms and applications.

“AKSHAR: Coded Character Set and Composition Rules – Language: Hindi”, is an outcome of extensive deliberations by Working Group setup by C-DAC under the guidance of MeitY on Script Grammar and Script-Language Gaps. Experts from academia, government organizations, industry, and language bodies actively contributed to the preparation of this draft through multiple deliberations. The list of experts is mentioned at Annex-A.

This standard provides:

- a) The character repertoire for Hindi,
- b) The Augmented Backus-Naur Formalism (ABNF) rules for validating permissible character combinations, and
- c) A reference implementation for testing and validation of rules.

The work represents an important milestone in harmonizing Indian language computing with international standards such as ISO/IEC 10646 and national standards like IS 16350 for keyboard layouts. By defining what constitutes valid and invalid characters and sequences, this document lays the foundation for preventing script variants and ensuring uniform digital representation of Hindi in software and hardware systems.

The composition of the Committee responsible for formulation of this standard is given in Annex B.

1. SCOPE

This document, titled “AKSHAR,” defines the Script Grammar for Hindi. The main component of this document is Unicode Code point repertoire, and Composition Rules described in the relevant sections.

The term script grammar refers to the behaviour pattern of the writing system of a given language. Languages that have written representations do not use a haphazard manner of storing the information within the system but use a coherent pattern which is like the linguistic grammar of a given language. Script Grammar is the term used to define:

- a) the writing system is a system that a language employs to encode its spoken form,
- b) the modifications brought to the writing system by a given language in terms of the addition of characters and deprecation of other characters along with a description of Unicode characters and their inclusion in the language,
- c) the syllabic structure of the writing system of the language,
- d) The rules governing the internal arrangement of characters within a syllable (Akshar)

This document defines the character sets of selected languages along with the rules for valid character combinations and is intended for a broad audience that includes both general users and subject matter experts such as linguists, font designers, language technologists and academicians.

It aims to present the information in a clear and accessible manner for those with limited technical background, while also providing the necessary detail and structure for professional and academic use. The document outlines the complete set of characters used in each language, including base characters and diacritics, and specifies how these characters may be combined according to linguistic, typographic, and encoding standards—such as those defined by UNICODE and recommended by the Central Hindi Directorate.

While everyday readers will benefit from the simplified explanations and illustrative examples, experts will find precise descriptions of combination logic, script-specific constraints, and normative versus descriptive usage patterns. The scope is limited to character encoding and composition rules and does not extend to higher-level grammatical or semantic structures/shapes.

2. STRUCTURE OF THE DOCUMENT

This section begins with an overview and background of the script, followed by the code point repertoire and a description of the Hindi syllable structure, known as Syllable/Akshar. These components aim to familiarize developers with the script and its fundamental units; which form the basis of all Brahmi-derived scripts. Section 6 introduces a formal framework for processing Syllable/Akshar also known as composition rules, which is essential for distinguishing valid syllables from invalid ones. Given that Unicode allows multiple representations for the same character, normalization also plays a critical role in ensuring consistency.

3. BACKGROUND

India is a linguist’s treasure trove, home to four major language families. According to the Census of India 2011, the country has 121 recognized languages and 270 mother tongues. Of these, 22 languages are listed in the Eighth Schedule of the Indian Constitution, known as the Scheduled Languages. These 22 languages belong to four distinct language families: Indo-Aryan, Dravidian, Tibeto-Burman, and Austroasiatic.

The writing systems of nearly all Indian languages—except Urdu and Santali—trace their origins back to the ancient Brahmi script. Urdu, which developed during the Mughal era, uses the Perso-Arabic script. Sindhi and Kashmiri also employ the Perso-Arabic script, although Devanagari is commonly used for Sindhi, and the historical Sharda script was once used for Kashmiri.

Manipuri uses the Meitei Mayek script which shows influences from Brahmi. Santali is primarily written in Devanagari today but also uses Ol Chiki, a modern script specifically designed for the language.

3.1 Script & language information:

ISO 15924 Script Code: Deva

Latin transliteration of native script name: Devanāgarī (ISO 15919)

Native name of the script: देवनागरी

Table 1: Information of Hindi Language
(Clause 3.1)

Sl. No.	Language	ISO 639-3	Family	Script
i)	Hindi	hin	Indo-Aryan	Devanagari

4. DEVANAGARI SCRIPT: AN OVERVIEW

Devanagari is the primary script used for several Indo-Aryan languages recognized as Scheduled Languages of India, including Hindi, Marathi, Maithili, Dogri, Konkani, Sanskrit, and Nepali. Beyond these, it is also used to write Boro(Bodo), a Tibeto-Burman language; Santali, an Austroasiatic language; and is occasionally employed for Sindhi and Kashmiri as spoken in India.

Devanagari extends beyond India's borders as well—it is used for Hindi in Fiji and for Nepali in Nepal. Additionally, it serves as a writing system for various previously unwritten languages. Closely associated with the classical languages Sanskrit and Prakrit, Devanagari plays a key role in India's linguistic heritage. The script is also widely Adopted by speakers of tribal languages in regions such as Arunachal Pradesh, Bihar, and the Andaman & Nicobar Islands, where it functions as a shared medium for literacy, communication, and cultural preservation.

4.1 Structure of Hindi represented in Devanagari Script:

All scripts derived from Brahmi, including Devanagari, are abugidas—writing systems that are syllable-based rather than purely alphabetic. The key features of Hindi as represented in the Devanagari script include:

- Each consonant inherently carries a vowel sound, typically the schwa (/ə/),
- This inherent vowel can be modified by adding vowel diacritics or suppressed using a special diacritic called the virama or halant,
- Vowels can appear independently as full vowel characters when they occur at the beginning of a syllable or stand alone,
- When two or more consonants occur together, they combine to form ligatures, which may retain recognizable elements of the original letters (e.g., क् + ल = क्ल) result in a completely new shape (e.g., क् + ष = क्ष).

4.2 Parivardhit Devanagari:

Parivardhit Devanagari is the extended Devanagari set for the representation of the other languages in Devanagari. Many languages share Devanagari Unicode block and Hindi uses a subset of it.

4.2.1 Extended Devanagari (Code charts as given in Unicode¹):

4.2.2 Devanagari Extended² - contains Vedic accents and listed characters that are irrelevant to Hindi.

4.2.3 Devanagari Extended A³ - contains characters from the 11th century Jain auspicious signs of Prakrit, which are also not relevant to modern Hindi but can be used when the text becomes multilingual

5. CODE POINT REPERTOIRE

The below table has code points from Unicode Code Block (0900-097F) for each varna (वर्ण) used in Devanagari. Each code point is being referred in “देवनागरी लिपि तथा हिंदी वर्तनी का मानकीकरण (Devanagari Lipi and Hindi Vartani ka Mankikaran) document”.

Table 2: Devanagari Code Point Repertoire for getting the information on Hindi
(Clause 5)

Sl. No.	Unicode Code Point (Decimal)	Unicode Code Point (Hexadecimal)	Glyph	Character Name	Category
i)	2305	0901	ँ	Devanagari SIGN CANDRABINDU	Vowel-Modifier
ii)	2306	0902	ं	Devanagari SIGN ANUSVARA	Vowel-Modifier
iii)	2307	0903	ः	Devanagari SIGN VISARGA	Vowel-Modifier
iv)	2309	0905	अ	Devanagari LETTER A	Vowel
v)	2310	0906	आ	Devanagari LETTER AA	Vowel
vi)	2311	0907	इ	Devanagari LETTER I	Vowel
vii)	2312	0908	ई	Devanagari LETTER II	Vowel
viii)	2313	0909	उ	Devanagari LETTER U	Vowel
ix)	2314	090A	ऊ	Devanagari LETTER UU	Vowel
x)	2315	090B	ऋ	Devanagari LETTER VOCALIC R	Vowel
xi)	2319	090F	ए	Devanagari LETTER E	Vowel

¹[Unicode 15.1 Character Code Charts](https://www.unicode.org/charts/PDF/UA8E0.pdf)

² <https://www.unicode.org/charts/PDF/UA8E0.pdf>

³ <https://www.unicode.org/charts/PDF/U11B00.pdf>

xii)	2320	0910	ऐ	Devanagari LETTER AI	Vowel
xiii)	2323	0913	ओ	Devanagari LETTER O	Vowel
xiv)	2324	0914	औ	Devanagari LETTER AU	Vowel
xv)	2325	0915	क	Devanagari LETTER KA	Consonant
xvi)	2326	0916	ख	Devanagari LETTER KHA	Consonant
xvii)	2327	0917	ग	Devanagari LETTER GA	Consonant
xviii)	2328	0918	घ	Devanagari LETTER GH A	Consonant
xix)	2329	0919	ङ	Devanagari LETTER NGA	Consonant
xx)	2330	091A	च	Devanagari LETTER CA	Consonant
xxi)	2331	091B	छ	Devanagari LETTER CHA	Consonant
xxii)	2332	091C	ज	Devanagari LETTER JA	Consonant
xxiii)	2333	091D	झ	Devanagari LETTER JHA	Consonant
xxiv)	2334	091E	ञ	Devanagari LETTER NYA	Consonant
xxv)	2335	091F	ट	Devanagari LETTER TTA	Consonant
xxvi)	2336	0920	ठ	Devanagari LETTER TTHA	Consonant
xxvii)	2337	0921	ड	Devanagari LETTER DDA	Consonant
xxviii)	2338	0922	ढ	Devanagari LETTER DDHA	Consonant
xxix)	2339	0923	ण	Devanagari LETTER NNA	Consonant
xxx)	2340	0924	त	Devanagari LETTER TA	Consonant
xxxi)	2341	0925	थ	Devanagari LETTER THA	Consonant
xxxii)	2342	0926	द	Devanagari LETTER DA	Consonant
xxxiii)	2343	0927	ध	Devanagari LETTER DHA	Consonant
xxxiv)	2344	0928	न	Devanagari LETTER NA	Consonant
xxxv)	2346	092A	प	Devanagari LETTER PA	Consonant

xxxvi)	2347	092B	फ	Devanagari LETTER PHA	Consonant
xxxvii)	2348	092C	ब	Devanagari LETTER BA	Consonant
xxxviii)	2349	092D	भ	Devanagari LETTER BHA	Consonant
xxxix)	2350	092E	म	Devanagari LETTER MA	Consonant
xl)	2351	092F	य	Devanagari LETTER YA	Consonant
xli)	2352	0930	र	Devanagari LETTER RA	Consonant
xlii)	2354	0932	ल	Devanagari LETTER LA	Consonant
xliii)	2357	0935	व	Devanagari LETTER VA	Consonant
xliv)	2358	0936	श	Devanagari LETTER SHA	Consonant
xliv)	2359	0937	ष	Devanagari LETTER SSA	Consonant
xlvi)	2360	0938	स	Devanagari LETTER SA	Consonant
xlvi)	2361	0939	ह	Devanagari LETTER HA	Consonant
xlvi)	2366	093E	ा	Devanagari VOWEL SIGN AA	Matra
xlvi)	2367	093F	ि	Devanagari VOWEL SIGN I	Matra
l)	2368	0940	ी	Devanagari VOWEL SIGN II	Matra
li)	2369	0941	ु	Devanagari VOWEL SIGN U	Matra
lii)	2370	0942	ू	Devanagari VOWEL SIGN UU	Matra
liii)	2371	0943	ृ	Devanagari VOWEL SIGN VOCALIC R	Matra
liv)	2375	0947	े	Devanagari VOWEL SIGN E	Matra
lv)	2376	0948	ै	Devanagari VOWEL SIGN AI	Matra
lvi)	2379	094B	ो	Devanagari VOWEL SIGN O	Matra
lvii)	2380	094C	ौ	Devanagari VOWEL SIGN AU	Matra
lviii)	2381	094D	्	Devanagari SIGN VIRAMA	Halant / Virama

lix)	2396	095C	ड़	Devanagari LETTER DDDHA	Consonant ⁴
lx)	2397	095D	ढ़	Devanagari LETTER RHA	Consonant ⁵
lxi)	2416	0970	॰	Devanagari ABBREVIATION SIGN	Abbreviation Marker

* The above table also contains characters that are overridden in UNICODE Normalization Form C (NFC)⁶

6. BORROWED CHARACTERS

Borrowed Characters (आगत वर्ण) have come into the Hindi language from other languages. These may be taken to represent Arabic, Farsi, English or any other foreign language and mentioned in देवनागरी लिपि तथा हिंदी वर्तनी का मानकीकरण (Devanagari Lipi and Hindi Vartani ka Manakikaran) document.

Table 3: Borrowed Character

(Clause 6)

Sl. No.	Unicode Code Point (Decimal)	Unicode Code Point (HexaDecimal)	Glyph	Character Name	Category
i)	2321	0911	औ	Devanagari LETTER CANDRA O	Vowel
ii)	2365	093D	ॱ	Devanagari SIGN AVAGRAHA	Avagraha
iii)	2377	0949	ँ	Devanagari VOWEL SIGN CANDRA O	Matra
iv)	2392	0958	क़	Devanagari LETTER QA	Consonant ⁷
v)	2393	0959	ख़	Devanagari LETTER KHHA	Consonant ⁸
vi)	2394	095A	ग़	Devanagari LETTER GHHA	Consonant ⁹
vii)	2395	095B	ज़	Devanagari LETTER ZA	Consonant ¹⁰
viii)	2398	095E	फ़	Devanagari LETTER FA	Consonant ¹¹

⁴Subject to Normalization NFC

⁵Subject to Normalization NFC

⁶<https://www.unicode.org/reports/tr15/>
<https://www.unicode.org/charts/normalization/>

⁷Subject to Normalization Form NFC

⁸ Subject to Normalization Form NFC

⁹ Subject to Normalization Form NFC

¹⁰ Subject to Normalization Form NFC



¹¹ Subject to Normalization Form NFC

7. ZWJ (U+200D) AND ZWNJ (U+200C)

These are code points that have been provided by the Unicode standard to instruct the rendering of a string where the script has the option between joining and non-joining characters. Without these control codes, the string may be rendered in an alternate form from what is intended.

For example, Examples of Composition Rules (ZWJ and ZWNJ)

Table 4: ZWJ/ZWNJ
(Clause 7)

Sl. No.	Unicode Code Point (Decimal)	Unicode Code Point (HexaDecimal)	Glyph	Character Name	Category
i)	8204	200C		ZERO WIDTH NON-JOINER (commonly abbreviated ZWNJ)	INV_MARK
ii)	8205	200D		ZERO WIDTH JOINER (commonly abbreviated ZWJ)	INV_MARK

8. अधोबिन्दु (ADHOBINDU) / NUKTA CASE OF (U+093C)

Unicode Normalization Forms standardize Unicode string representations to ensure that visually or semantically equivalent text can be reliably compared. For example, पेड़ can be written using either U+092A U+0947 U+095C (precomposed form), or U+092A U+0947 U+0921 U+093C (decomposed form) but still be treated as equivalent. Normalization forms like NFC and NFKC decompose and then recompose characters, except when blocked by exclusions. Indian scripts, including Devanagari, are subject to script-specific composition exclusions—certain characters like U+0958 (ऋ) are deliberately excluded from recomposition.

This is done to preserve canonical stability, maintain compatibility with legacy encodings (e.g., ISCII), and ensure interoperability. Once a character has a defined decomposition, Unicode forbids altering it, ensuring consistent normalization across versions and preventing data corruption in systems relying on decomposed forms. In simple language, Normalization takes place when a given character can be rendered in two or more ways. Since such multiple representations can affect searching, and sorting (and lead to spoofing and phishing), it is essential that these multiple representations be reduced to one single canonical form.

Unicode provides useful charts for the most common Normalization forms in all scripts. In the case of Hindi, normalisation is seen mainly in the case of additional consonants, as listed in Unicode for the characters below. Unicode 3.0 upwards permits two ways of representing the additional consonants of Hindi: Akhand form (i.e. single character with nukta/adhobindu inherent in it) vs. Consonant+adhobindu.

The Consonant+ adhobindu is preferred over the Akhand form during normalisation (NFC). To handle data in digital form, U+093C is added as a part of the permissible character set and treated as category “Adhobindu”, and it will only be permitted after seven consonants (C1) for Hindi (see section 6; specific rules).

Table 5: Nukta/Adhobindu

(Clause 8)

Sl. No.	Unicode Code Point (Decimal)	Unicode Code Point (Hexadecimal)	Glyph	Character Name	Category
i)	2364	093C	◌ं	Devanagari SIGN NUKTA • for extending the alphabet to new letters	Adhobindu

Table 6: Normalization Table

(Clause 8)

Sl. no.	Additional Consonants	Normalised Form
i)	क़ (U+0958)	क+◌ं (U+0915 093C)
ii)	ख़ (U+0959)	ख+◌ं (U+0916 093C)
iii)	ग़ (U+095A)	ग+◌ं (U+0917 093C)
iv)	ज़ (U+095B)	ज+◌ं (U+091C 093C)
v)	ड़ (U+095C)	ड+◌ं (U+0921 093C)
vi)	ढ़ (U+095D)	ढ+◌ं (U+0922 093C)
vii)	फ़ (U+095E)	फ+◌ं (U+092B 093C)

9. ADDITIONAL CHARACTERS:

In addition to the above, characters of Devanagari Script and related code blocks, the hindi text may also include Indo-Arabic Numerals, Devanagari Numerals, and various cultural, religious, financial and various domain specific symbols drawn from multiple Unicode code blocks. However, these characters are not part of the composition rules as outlined in Section 6. When encountered in text, such characters should be identified and excluded from text processing routines.

A list of commonly used symbols and punctuation marks is provided in the accompanying table. Implementer and stakeholders may extend this list by including additional characters as needed.

Table 7: Addition to Hindi Code Point Repertoire

(Clause 9)

Sl. No.	Unicode Code Point (Decimal)	Unicode Code Point (Hexadecimal)	Glyph	Character Name	Category
ii)	8377	20B9	₹	INDIAN RUPEE SIGN	Symbol
iii)	2384	0950	ॐ	DEVANAGARI OM	Symbol

iv)	4053	0FD5	卐	RIGHT-FACING SVASTI SIGN	Symbol
v)	48	0030	0	DIGIT ZERO	Digit
vi)	49	0031	1	DIGIT ONE	Digit
vii)	50	0032	2	DIGIT TWO	Digit
viii)	51	0033	3	DIGIT THREE	Digit
ix)	52	0034	4	DIGIT FOUR	Digit
x)	53	0035	5	DIGIT FIVE	Digit
xi)	54	0036	6	DIGIT SIX	Digit
xii)	55	0037	7	DIGIT SEVEN	Digit
xiii)	56	0038	8	DIGIT EIGHT	Digit
xiv)	57	0039	9	DIGIT NINE	Digit
xv)	2406	966	०	Devanagari DIGIT ZERO	Digit
xvi)	2407	967	१	Devanagari DIGIT ONE	Digit
xvii)	2408	968	२	Devanagari DIGIT TWO	Digit
xviii)	2409	969	३	Devanagari DIGIT THREE	Digit
xix)	2410	096A	४	Devanagari DIGIT FOUR	Digit
xx)	2411	096B	५	Devanagari DIGIT FIVE	Digit
xxi)	2412	096C	६	Devanagari DIGIT SIX	Digit
xxii)	2413	096D	७	Devanagari DIGIT SEVEN	Digit
xxiii)	2414	096E	८	Devanagari DIGIT EIGHT	Digit
xxiv)	2415	096F	९	Devanagari DIGIT NINE	Digit
xxv)	2404	0964		Devanagari DANDA	Punctuation
xxvi)	2405	0965	॥	Devanagari DOUBLE DANDA	Punctuation
xxvii)	46	002E	.	FULL STOP	Punctuation
xxviii)	63	003F	?	QUESTION MARK	Punctuation
xxix)	44	002C	,	COMMA	Punctuation
xxx)	33	0021	!	EXCLAMATION MARK	Punctuation
xxxi)	39	0027	'	APOSTROPHE	Punctuation
xxxii)	59	003B	;	SEMICOLON	Punctuation
xxxiii)	58	003A	:	COLON	Punctuation

xxxiv)	45	002D	-	HYPHEN-MINUS	Punctuation
xxxv)	47	002F	/	SLASH	Punctuation
xxxvi)	34	0022	"	QUOTATION MARK	Punctuation
xxxvii)	39	0027	'	APOSTROPHE SINGLE QUOTATION MARK	Punctuation
xxxviii)	40	0028	(LEFT PARENTHESIS	Punctuation
xxxix)	41	0029)	RIGHT PARENTHESIS	Punctuation
xl)	91	005B	[LEFT SQUARE BRACKET	Punctuation
xli)	93	005D]	RIGHT SQUARE BRACKET	Punctuation
xl ii)	123	007B	{	LEFT CURLY BRACKET	Punctuation
xl iii)	125	007D	}	RIGHT CURLY BRACKET	Punctuation

10. SAMYUKTA VYANJAN

संयुक्त व्यंजन (Samyukta Vyanjan) in Devanagari script refers to conjunct consonants/ligatures, which are combinations of two or more consonants joined together to form a single unit. These are used when two consonants come together without an intervening vowel sound.

संयुक्त व्यंजन as given in the document (देवनागरी लिपि तथा हिंदी वर्तनी का मानकीकरण)¹² are given below. The combination (CHC) would be handled by Composition Rules as given in section 6.

Table 8: Samyukta Vyanjan

(Clause 10)

Sl. No.	संयुक्त व्यंजन (Samyukta Vyanjan)	Combination	Unicode Value
i)	क्ष	क् ष	U+0915 U+094D U+0937
ii)	त्र	त् र	U+0924 U+094D U+0930
iii)	ज्ञ	ज ञ	U+091C U+094D U+091E
iv)	श्र	श् र	U+0936 U+094D U+0930

11. COMPOSITION RULES

1. All languages that use Brahmi-derived scripts follow a specific structure for word formation, known as Syllable/Akshar. This section outlines the rules for composition rules as they apply to Hindi.

¹²https://www.chdpublication.education.gov.in/ebook/Devanagari_Lipi_and_Hindi_Vartani_ka_Mankikaran/html5forpc.html?page=0 (Accessed on 4 July 2024)

2. It begins with a list of variables, each mapped to relevant character categories. Within the rules, these variables are referenced by their symbolic names rather than descriptive labels. The next part introduces the operators and their associated functions, which are assumed during rule application.
3. The following two sections present the core formation rules, divided into two main categories: one focusing on vowels and the other on consonants. These rules are based on the Indian Standard IS 13194:1991, commonly referred to as the "Indian Script Code for Information Interchange" (ISCII).
4. Any exceptions or constraints to these rules are provided in a separate section that follows. The syntax used to represent the composition rules is based on Augmented Backus-Naur Form (ABNF), as defined in RFC 5234¹³.

11.1 Variables involved

C → Consonant
 M → Matra
 V → Vowel
 D → Vowel-Modifier
 H → Halant / Virama
 N → Adhobindu
 Z → INV_MARK
 A → Avagraha
 S → Abbreviation_Marker

11.2 Operators used¹⁴

These are the Composition rules for Hindi.

In what follows, the Vowel Sequence and the Consonant Sequence pertinent to Devanagari, when used to write Hindi, are given.

Table 9: Symbol Table

Clause 11.2

Sl. No.	Symbol	Function
i)	/	Alternative /one of
ii)	[]	Optional

¹³<https://www.rfc-editor.org/rfc/rfc5234.html>

¹⁴ <https://www.rfc-editor.org/rfc/rfc5234.html>

iii)	*	<p>Variable Repetition</p> <p>The operator "*" preceding an element indicates repetition. The full form is: <a>*element</p> <p>where <a> and are optional decimal values, indicating at least <a> and at most occurrences of the element.</p> <p>Default values are 0 and infinity so that *<element> allows any number, including zero; *4 <element> allows element from 0 to 4</p>
iv)	()	Sequence Group
v)	=	Relation

11.3 Rules (used to form a syllable or Akshar)

- Full-Cons = C [N]
- Pure-Cons = Full-Cons H [Z]
- Cons-Syllable = *4 Pure-Cons
- Cons-Vowel-Syllable = [Cons-Syllable] Full-Cons [D/M[D]]
- Vowel-Syllable = V[D]

Syllable (Akshar) = (Vowel-Syllable ([S]/*A))/ (Cons-Vowel-Syllable([S]/*A))/ CH

CH - The Halant is allowed in word final positions to allow proper encoding of Tatsam words of Hindi as specified in para 3.8 Devanagari Lipi and Hindi Vartani ka Mankikaran document.

Examples: वाक्-(वाग्देवी), सत्-(सत्साहित्य), भगवन् (भगवद्भक्ति), साक्षात्-(साक्षात्कार), जगत्-(जगन्नाथ), तेजस्-(तेजस्वी)

11.4 Specific Rules

Below are the specific rules:

- N: must be preceded only by a member of C1

The set C1 consists of these consonants:

- क (U+0915)
- ख (U+0916)
- ग (U+0917)
- ज (U+091C)
- ड (U+0921)
- ढ (U+0922)
- फ (U+092B)

- H: must be preceded by Full-Cons
- M: must be preceded by Full-Cons
- D: must be preceded by either of V, Full-Cons or M
- D: To minimize visual ambiguity, D should not be placed immediately after the following characters:

- a) औ (U+0911)
- b) ौ (U+0949)

Examples of Composition Rules

ZWJ and ZWNJ:

Insofar as Hindi is concerned ZWJ/ZWNJ are used to render alternate rendering of ligatures. Thus the use of ZWJ/ZWNJ is restricted to the display level.

RULE 1: If a consonant+halant is followed by the ZWJ, the half-form of the consonant is formed in place of ligature.

Example:

शक्ति U+0936(श) U+0915(क) U+094D(्) U+0924(त) U+093F(ि)

शक्ति U+0936(श) U+0915(क) U+094D(्) U+200D(ZWJ) U+0924(त) U+093F(ि)

अक्षय U+0905(अ) U+0915(क) U+094D(्) U+0937(ष) U+092F(य)

अक्षय U+0905(अ) U+0915(क) U+094D(्) U+200D(ZWJ) U+0937(ष) U+092F(य)

This use of ZWJ serves a pedagogical purpose in that it allows the learner to study and master the half shapes of characters.

RULE 2: The use of ZWNJ in Hindi is restricted to representing a dead consonant within a string.

Thus to show the ग् within a word and retain the shape of the consonant followed by the halant; ZWNJ is used:

गङ्गा: U+0917(ग) U+0919(ङ) U+094D(्) U+0915(ग) U+092E(ा)

गङ्गा: U+0917(ग) U+0919(ङ) U+094D(्) U+200C(ZWNJ) U+0915(ग) U+092E(ा)

As can be seen from the above, ZWJ/ZWNJ are mainly used to render alternate forms.

Used as in the case of Hindi, to create alternate renderings, the insertion of these two signs can affect searching as well as during language processing tasks. Since storage is affected, these characters should be used only when necessary and mandatory. Text processing application developers should specifically pay attention to it and modify their applications to virtually consider it as not present, during processes like find, search, index etc.

12.1 Valid Sequences:

In what follows, the valid Vowel Sequence and the Consonant Sequence pertinent to Hindi are given.

12.1.1 Vowel Sequence

A vowel sequence begins with a vowel. It may be followed by one Vowel-Modifier, known as Yogavaha in Indian Grammatical Tradition. The Vowel-Modifier (or Yogavaha characters) for Hindi is three in number and are mutually exclusive. These include Anusvara, Chandrabindu and Visarga. That is, the number of D which can follow a V in Devanagari is restricted to one.

The vowel syllable is therefore V [D]

Examples:

Table 10: Vowel Sequence

(Clause 12.1.1)

Sl. No.	Sequence Description	Sequence	Example
i)	Vowel	V	अ U+0905
ii)	Vowel + Anusvara	V[D]	अं U+0905 U+0902
iii)	Vowel + Candrabindu	V[D]	अँ U+0905 U+0901
iv)	Vowel + Visarga	V[D]	अः U+0905 U+0903

A Vowel Syllable may be followed by Avagraha (A) or Abbreviation_Marker (S). Avagraha (A) may repeat after the Vowel Syllable.

(V[D]) [S]*A]

Table 11: (V[D])[S]/*A

(Clause 12.1.1)

Sl. No.	Sequence Description	Sequence	Example
i)	Vowel+ Avagraha	V[A]	अः U+0905 U+093D
ii)	Vowel + Abbreviation_Marker	V[S]	अ० U+0905 U+0970
iii)	Vowel + *Avagraha	V[*A]	अः U+0905 U+0901

12.1.2 Consonant Sequence

A consonant sequence begins with a consonant. It may be followed by a Nukta (N), Matra (M), Vowel-Modifier (D) or Halant (H). The number of instances of these characters occurring after a consonant is restricted to one. There is a possibility of further extension of the Consonant sequence after the N and H. Each of these has been discussed in the following sections:

1. A single consonant (C)

(The consonant shall be treated as coterminous with the Consonant along with the Nukta sign wherever such a case is pertinent.)

Examples:

Table 12: Consonant Sequence C[N]

(Clause 12.1.1)

Sl. No.	Sequence Description	Sequence	Example
i)	Consonant	C	क U+0915
ii)	Consonant + Nukta	C[N]	कृ U+0915 U+093C

2. A consonant may be followed by a dependent vowel sign/Matra [M], Vowel-Modifier[D] or Halant [H]

C [M/D]

Examples:

Table 13: Consonant Sequence C[M/D]

(Clause 12.1.1)

Sl. No.	Sequence Description	Sequence	Example
i)	Consonant + Matra	C[M]	कि U+0915 U+093F
ii)	Consonant + Anusvara	C[D]	कं U+0915 U+0902
iii)	Consonant + Candrabindu	C[D]	कँ U+0915 U+0901
iv)	Consonant + Visarga	C[D]	कः U+0915 U+0903

CH –The Halant is allowed in word final positions only with the Consonant. The consonant shall not be treated as coterminous with the Consonant along with Nukta. Halant after consonant sequence is also not permitted. Hence CNH, CHCH, CHCHCH, CHCHCHCH, CHCHCHCHCH shall not be valid.

Consonant + Halant	C[H]	क् U+0915 U+094D
--------------------	------	---------------------

3. A CM sequence can be followed by D

(CM)[D]

Example:

Table 14: Consonant Sequence CM[D]

(Clause 12.1.1)

Sl. No.	Sequence Description	Sequence	Example
i)	Consonant + Matra + Anusvara	CM[D]	कीं U+0915 U+0940 U+0902
ii)	Consonant + Matra + Candrabindu	CM[D]	काँ U+0915 U+093E U+0901
iii)	Consonant + Matra + Visarga	CM[D]	कीः U+0915 U+0940 U+0903

4. A sequence of consonants (up to 5) joined by Halant *4(CH[Z])C

This allows maximum of five consonants in one cluster as this is maximum limit for Sanskrit

Example:

Table 15: Consonant Sequence *4(CH[Z])C

(Clause 12.1.2)

Sl. No.	Sequence Description	Sequence	Example
i)	Consonant + Halant + Consonant	CHC	क्र U+0915 U+094D U+0930
ii)	Consonant + Halant + Consonant + Halant + Consonant	CHCHC	न्त्र U+0928 U+094D U+0924 U+094D U+0930
iii)	Consonant + Halant + Consonant + Halant + Consonant + Halant + Consonant	CHCHCHC	न्त्रय U+0928 U+094D U+0915 U+094D U+0930 U+094D U+092F

However, composition rules do not impose any restriction on the number of consonants that can be joined by a Halant; the restrictions are imposed by the language itself.

12.1.3 Subsets for Consonant sequence:

1. The combination may be followed by M or D

Example:

Table 16: Consonant Sequence *4(CH)C[M/D]

(Clause 12.1.3)

Sl. No.	Sequence Description	Sequence	Example
i.)	Consonant + Halant + Consonant + Matra	CHC[M]	क्की U+0915 U+094D U+0915 U+0940
ii.)	Consonant + Halant + Consonant + Anusvara	CHC[D]	क्कं U+0915 U+094D U+0915 U+0902
iii.)	Consonant + Halant + Consonant + Candrabindu	CHC[D]	क्कँ U+0915 U+094D U+0915 U+0901
iv.)	Consonant + Halant + Consonant + Visarga	CHC[D]	क्कः U+0915 U+094D U+0915 U+0903

2. *4(CH)CM may be followed by a D

Example

Table 17: Consonant Sequence *4(CH)CM[D]

(Clause 12.1.3)

Sl. No.	Sequence Description	Sequence	Example
i)	Consonant + Halant + Consonant + Matra + Anusvara	CHCM[D]	क्कीं U+0915 U+094D U+0915 U+0940 U+0902
ii)	Consonant + Halant + Consonant + Matra +Candrabindu	CHCM[D]	क्कीँ U+0915 U+094D U+0915 U+0940 U+0901
iii)	Consonant + Halant + Consonant + Matra +Visarga	CHCM[D]	क्कीः U+0915 U+094D U+0915 U+0940 U+0903

3. *4(CH)CM[D] may be followed by a [S]/*A

Example:

Table 18: Consonant Sequence (*4(CH)CM[D])[S]/*A

(Clause 12.1.3)

Sl. No.	Sequence Description	Sequence	Example
i)	Consonant + Matra + Avagraha	CMA	याऽ U+092F U+093E U+093D
ii)	Consonant + Matra + Candrabindu + Avagraha + Avagraha + Avagraha	CMDAAA	माँऽऽऽ U+092E U+093E U+0901 U+093D U+093D U+093D
iii)	Consonant + Matra + Abbreviation_Marker	CMS	रू० U+0930 U+0942 U+0970

12.2 Invalid Sequences

The composition rules define specific character combinations that are either disallowed or potentially problematic—particularly those that, due to rendering, may visually appear correct but are technically incorrect. In some cases, users may inadvertently use these combinations for various reasons. However, the ABNF-based validation will flag such sequences as invalid.

The following are the Invalid Sequences:

Table 19: Invalid Sequences

(Clause 12.2)

Sl. No.	Invalid Sequence	Example String	Character Sequence	Character Value in Decimal	Character Value in Hexa-Decimal
i)	VM	ओीशा	अ+ी+श+ा	2309, 2368, 2358, 2366	U+0905, U+0940, U+0936, U+093E
ii)	VH	अ्रत	अ+्र+र+त	2309, 2381, 2352, 2340	U+0905, U+094D, U+0930, U+0924
iii)	VZ	अंकुर	अ+ZWNJ+ं+क+ु +र	2309, 8205, 2306, 2325, 2369, 2352	U+0905, U+200D, U+0902, U+0915, U+0941, U+0930
iv)	CZ	करवट	क+ZWNJ+र+व+ट	2325, 8204, 2352, 2357, 2335	U+0915, U+200C, U+0930, U+0935, U+091F

v)	DZ	दुःख	द+ु+ः +ZWJ+ख	2342, 2369, 2307, 8205, 2326	U+0926, U+0941, U+0903, U+200D, U+0916
vi)	MZ	खून	ख+ू+ZWJ+न	2326, 2370, 8205, 2344	U+0916, U+0942, U+200D, U+0928
vii)	NZ	खुश	ख+़+ZWJ+ु+ष	2326, 2364, 8205, 2369, 2359	U+0916, U+093C, U+200D, U+0941, U+0937
viii)	DM	बूँद	ब+ँ+ू+द	2348, 2305, 2370, 2342	U+092C, U+0901, U+0942, U+0926
ix)	DH	मंद्	म+ं+्+द	2350, 2306, 2381, 2342	U+092E, U+0902, U+094D, U+0926
x)	MH	क्लि	क+ि+्+ल+ा	2325, 2367, 2381, 2354, 2366	U+0915, U+093F, U+094D, U+0932, U+093E
xi)	HD	चंदन	च+्+ं+द+न	2330, 2381, 2306, 2342, 2344	U+091A, U+094D, U+0902, U+0926, U+0928
xii)	HM	दूत	द+्+ू+त	2342, 2381, 2370, 2340	U+0926, U+094D, U+0942, U+0924
xiii)	HV	खाओ	ख+ा+्+ओ	2326, 2366, 2381, 2323	U+0916, U+093E, U+094D, U+0913
xiv)	VN	ऊंट	ऊ+ं+ं+ट	2314, 2364, 2306, 2335	U+090A, U+093C, U+0902, U+091F
xv)	MN	कफि	क+ा+फ+ि+ं	2325, 2366, 2347, 2367, 2364	U+0915, U+093E, U+092B, U+093F, U+093C
xvi)	DN	जंग	ज+ं+ं+ग	2332, 2306, 2364, 2327	U+091C, U+0902, U+093C, U+0917

xvii)	HN	प्रश	फ+्+र्+र+्+श	2347, 2381, 2364, 2352, 2381, 2358	U+092B, U+094D, U+093C, U+0930, U+094D, U+0936
xviii)	ZZ	गद्दी	ग+द+् +ZWJ +ZWNJ+द+ी	2327, 2342, 2381, 8205, 8204, 2342, 2368	U+0917, U+0926, U+094D, U+200D, U+200C, U+0926, U+0940
xix)	CIN	ज़ब्त	ज़+र्+ब+्+त	2395, 2364, 2348, 2381, 2340	U+095B, U+093C, U+092C, U+094D, U+0924
xx)	NN	फ़िर	फ+र्+र्+ि+र	2347, 2364, 2364, 2367, 2352	U+092B, U+093C, U+093C, U+093F, U+0930
xxi)	MM	याेग	य+ा+े+ग	2351, 2366, 2375, 2327	U+092F, U+093E, U+0947, U+0917
xxii)	HH	राज्य	र+ा+ज+्+्+य	2352, 2366, 2332, 2381, 2381, 2351	U+0930, U+093E, U+091C, U+094D, U+094D, U+092F
xxiii)	DD	अतः	अ+त+ं+ः	2309, 2340, 2306, 2307	U+0905, U+0924, U+0902, U+0903
xxiv)	HA	परम्	प+र+म+्+	2346, 2352, 2350, 2381, 2365	U+092A U+0930 U+092E U+094D U+093D
xxv)	HS	परम्	प+र+म+्+	2346, 2352, 2350, 2381, 2416	U+092A U+0930 U+092E U+094D U+0970

CONCLUSION

This document outlines the Script Grammar for Hindi, focusing primarily on the Unicode code point repertoire used in the language and the composition rules, as described in the relevant sections. It also identifies invalid character combinations that users may input—combinations that can result in misspellings or inconsistencies—and offers guidelines for detecting and flagging them during text processing.

Annex-A

Working Group–1 of MeitY on Script Grammar and Script-Language Gaps

1. Dr. Narayan Choudhary, CIIL Mysore (Chairman)
2. Dr. L. Sobha, AUKBC, Chennai - Co-Chair
3. Shri. Vijay Kumar, TDIL- MeitY
4. Shri Mahesh Kulkarni, Ex C-DAC
5. Dr. G S Lehal, IIIT Hyderabad
6. Dr. Malhar Kulkarni, IIT Bombay
7. Dr. Niladri Sekhar Dash, ISI Kolkata
8. Prof Sarbajit Singh, IIIT Manipur
9. Shri Abhijit Dutta, Ex IBM
10. Shri Cibu Johny, Google
11. Shri. Vivekananda Pani, RLT- Bengaluru-Industry
12. Shri Prashant Verma, DIBD
13. Ms. Shraddha Kalele, C-DAC
14. Shri Md. Shahzad Alam, C-DAC
15. Shri Atiur Rahman Khan, C-DAC
16. Shri Chandrakant Dhutadmal, C-DAC
17. Shri Vainateya Koratkar, C-DAC
18. Ms. Lenali Singh, C-DAC
19. Ms. Neha Gupta, C-DAC (Member and Convener)

Special Acknowledgements:

1. Smt. Kavita Bhatia, MeitY
2. Shri Akshat Joshi, Thinktrans
3. Dr. Anupam Mathur, CHD
4. Shri Baskaran Sankaran, Maadhyamik Tech
5. Shri Devansh Deolekar, BIS
6. Shri N Rajesha, CIIL
7. Shri Nutan Pandey, CHD
8. Shri Pranjal Nayak, Reverie Inc
9. Shri Purushottam Patil, Kendriya Hindi Sansthan
10. Ms. Swati Bhaskar, Reverie Inc

ANNEX-B

LITD 20 Committee Composition

(Formal clause will be added later)