Proposal for a Gujarati Script Root Zone Label Generation Ruleset (LGR)

LGR Version: 3.0

Date: 2018-07-10

Document version: 3.1

Authors: Neo-Brahmi Generation Panel [NBGP]

1 General Information/ Overview/ Abstract

The purpose of this document is to give an overview of the proposed Gujarati LGR in the XML format and the rationale behind the design decisions taken. It includes a discussion of relevant features of the script, the communities or languages using it, the process and methodology used and information on the contributors. The formal specification of the LGR can be found in the accompanying XML document:

Proposed-LGR-Gujarati-20180710.xml

Labels for testing can be found in the accompanying text document:

Gujarati LGR-Valid_Invalid Labels.txt

2 Script for which the LGR is proposed

ISO 15924 Code: Gujr

ISO 15924 Key N°: 320

ISO 15924 English Name: Gujarati

Latin transliteration of native script name: gujarâtî

Native name of the script: ગુજરાતી

Maximal Starting Repertoire (MSR) version: MSR-3

3 Background on the Script and the Principal Languages Using it[1]

Gujarati (ગુજરાતી) [also sometimes written as Gujerati, Gujarathi, Guzratee, Guujaratee, Gujrathi, and Gujerathi[2]] is an Indo-Aryan language native to the Indian state of Gujarat. It is part of the greater Indo-European language family. It is so named because Gujarati is the language of the Gujjars. Gujarati's origins can be traced back to Old Gujarati (circa 1100–1500 AD).

In India, it is the official language in the state of Gujarat, as well as an official language in the union territories of Daman and Diu and Dadra and Nagar Haveli. It is also a statutory provincial language in West Bengal State.

As per the 2011 census of India, 4.5% of the Indian population speaks Gujarati. There are about 65.5 million speakers of Gujarati worldwide, making it the 26th-most-spoken native language in the world. Gujarati is extensively spoken in large parts of Africa, Madagascar, UK and the USA as well as by emigrant communities around the world.

Of the approximately 65.5 million speakers of Gujarati in 1997, roughly 45.5 million resided in India, 150,000 in Uganda, 50,000 in Tanzania, 50,000 in Kenya and roughly 100,000 in Karachi, Pakistan. There is a certain number of the Mauritian population and a large number of Réunion Island people who are of Gujarati descent and some of these still speak Gujarati. A considerable Gujarati-speaking population exists in North America, most particularly in the New York City Metropolitan Area and in the Greater Toronto Area, which have over 100,000 speakers and over 75,000 speakers, respectively, but also throughout the major metropolitan areas of the United States and Canada

Besides being spoken by the Gujarati people, non-Gujarati residents of and migrants to the state of Gujarat also count as speakers, among them the Kutchis (as a literary language), the Parsis (adopted as a mother tongue), and Hindu Sindhi refugees from Pakistan[3].

[1] A considerable content in this section is from the Wiki articles on Gujarati Language and Gujarati Alphabet cf. Webography infra.

[2] Ethnologue (18th ed., 2015) also Mistry 2001, pp. 274 Mistry 2003, p. 115

[3] Devanāgarī has been mandated as the official script for writing Sindhi in India, although Perso-Arabic Sindhi is also used. Gujarati is used sparingly in some parts of Kutch.

3.1 The Evolution of the Script

Gujarati is a variant of Devanāgarī, the main difference being the absence of the shirorekha or the line above the character and also more rounded shapes. Since initially it was used for commercial ends, it has been referred to as śarāphi (banker's) or mahājani (trader's) script. The diagram below[4]shows the major stages in the evolution of Gujarati attesting its late divergence from Devanāgarī.

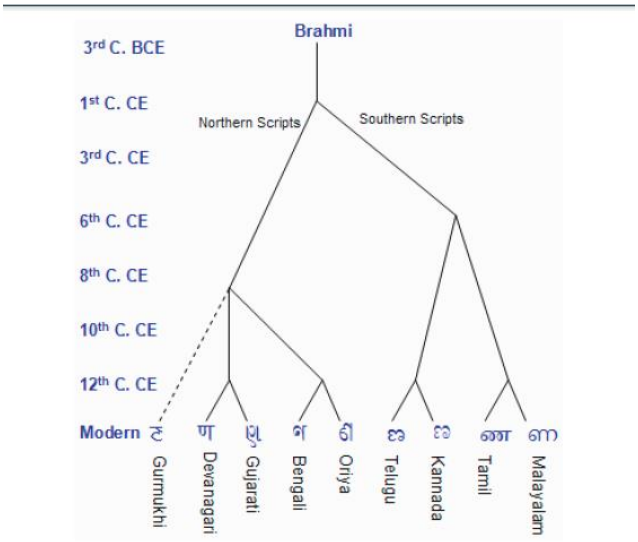

Figure 1: Pictorial depiction of Evolution of Gujarati

Gujarati is customarily divided into the following three historical stages[5]

- **Old Gujarati**
- **Middle Gujarati**
- **Modern Gujarati**

Old Gujarati (જૂનીગુજરાતી; also called ગુજરાતીભાખા Gujarati bhākhā or ગુર્જરઅપભ્રંશ Gurjar apabhraṃśa, 1100–1500 CE), the ancestor of modern Gujarati and Rajasthani,[2] was spoken by the Gurjars, who were residing and ruling in Gujarat, Punjab, Rajputana and central India. The language was used as literary language as early as the 12th century. Texts of this era display characteristic Gujarati features such as direct/oblique noun forms,

[4]Excerpted and adapted from Daniels and Bright, The World's Writing Systems. Oxford, Oxford University Press, 1996, p. 380

[5]This part is an emended version of the text on Gujarati Language from Wikipedia: https://en.wikipedia.org/wiki/Gujarati_language

postpositions, and auxiliary verbs. While generally known as Old Gujarati, some scholars prefer the name of Old Western Rajasthani, based on the argument that Gujarati and Rajasthani were not yet distinct. A sample of Old Gujarati is provided below from the Updeshmala, Manuscript in Jain Prakrit and Old Gujarati. The Old Gujarati prose commentary was written in 1487[6].

Figure 2: Upadeshmala

Middle Gujarati (AD 1500–1800)

According to Kausen[7] and Mistry[8], in this period Gujarati split from Rajasthani, and develop certain features which are the hall-marks of modern Gujarat such as the phonemes ɛ and ɔ, the auxiliary stem chh*, and the possessive morphological marker n*. A considerable amount of literature was created during this period.

Modern Gujarati (AD 1800-)

However, it is after 1800 that Gujarati came into its own and the language and script used today date from this period. The creation of metal types for printing Gujarati in 1815 saw a growth of Literature as well as Lexicography as is attested by the first printed book published: a Gujarati translation of Dabestan-e Mazaheb prepared and printed by the Parsi priest FardunjeeMarzban in 1815[9].

[6] https://en.wikipedia.org/wiki/File:Upadeshmala2.jpg
[7] Ernst Kausen, 2006. Die Klassifikation der indogermanischenSprachen
[8] Mistry 2003, pp. 115–116
[9] https://en.wikipedia.org/wiki/File:A_Page_from_the_Gujarati_translation_of_%27Dabist%C4%81n-i_Maz%C4%81hibm%27_prepared_and_printed_by_Fardunji_Marzban_(1815).jpg

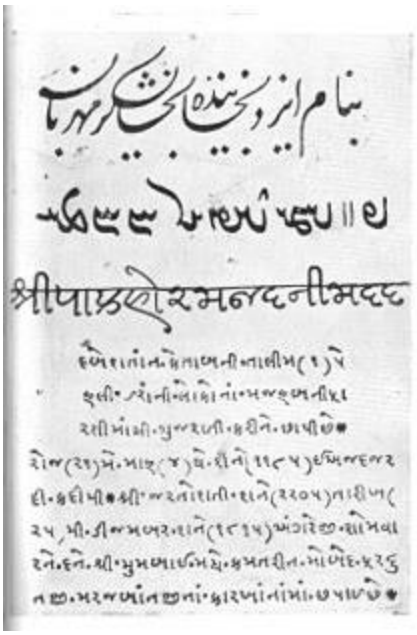

Figure 3: Dabestan-e Mazaheb

The advent of digital typography furthered the development of the language and Modern Gujarati has a rich literary and religious [Jaina] tradition.

3.2 Gujarati and its Dialects

3.2.1 "Standard Gujarati" and Dialects

The first researchers like Tisdall [1893][10]divided Gujarati into two dialects: a Hindu and a Parsi dialect. However, recent studies and analyses have shown that Gujarati admits a large number of dialects of which the major ones[11]are below:

- Standard Gujarati: primarily spoken in the Saurashtra region. This can be termed as something of a standardized variant of Gujarati across news, education and government
- Mumbai Gujarati, Nagari, Patnuli, Saurashtra Standard
- Gamadia: spoken primarily in Ahmedabad and the surrounding regions
- Ahmedabad Gamadia, Anawla, Brathela, Charotari, Eastern Broach Gujarati, Gramya, Patani, Patidari, Surati, Vadodari
- Parsi: spoken by the Zoroastrian Parsi minority[12]

[10]https://archive.org/details/simplifiedgramma00tisdiala

[11]Gujarati language at *Ethnologue* (16th ed., 2009)

- Khatiawari: spoken primarily in the Kathiawar region
- Bhawnagari, Gohilwadi, Holadi, Jhalawadi, Sorathi
- Kharwa, Kakari and Tarimuki also cited as additional varieties of Gujarati by Ethnologue.

The common feature of all these dialects is that they use the Gujarati script . The repertoire of Gujarati provided in Table 6: Code point repertoire below caters to all these dialects. The map below shows the administrative divisions of state of Gujarat in India since August 15, 2013.

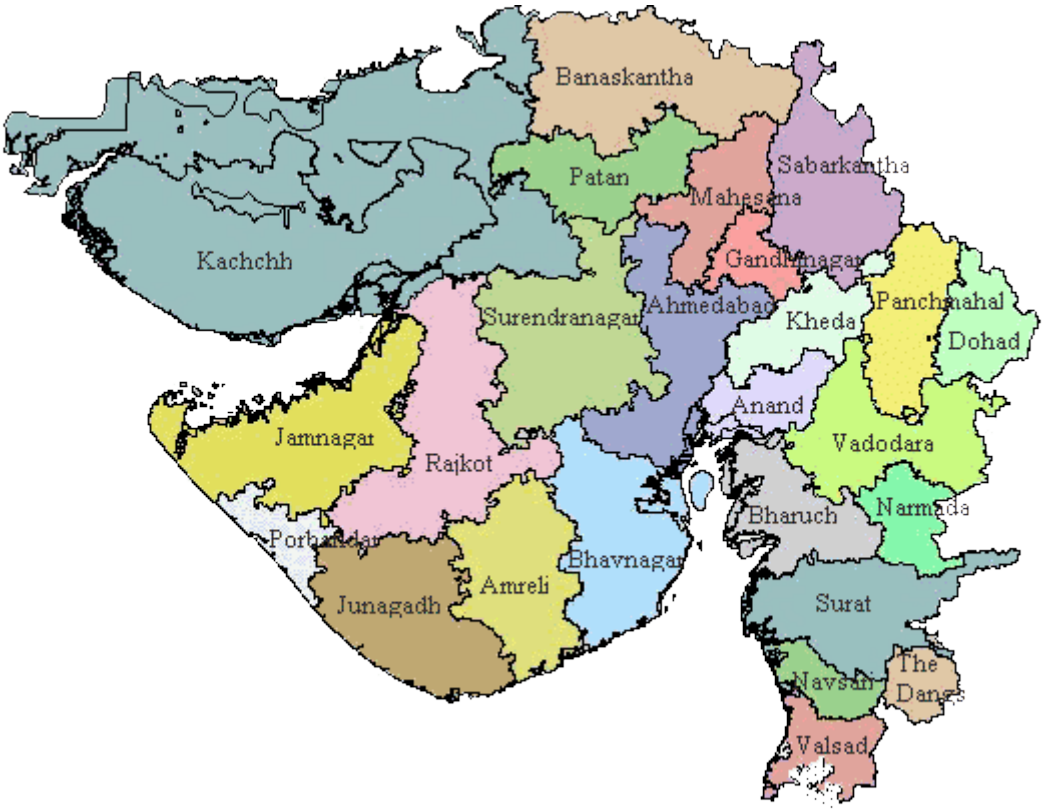

Figure 4: Administrative map of Gujarat

3.3 Language considered

Apart from the dialects listed above, 12 other languages use Gujarati for writing. A majority of these are EGIDS 5 and are developing. The only exception is Kukna, which is an Egids 4 language. After a study of the writing system of Kukna, it was found that it is in no way

[12]This dialect is endangered. The population of Parsis today is around 55,000 of which very few can read and write Gujarati. A majority can only read their liturgical texts in a Romanized form. Ethnologue, on being contacted in this regard, has agreed to class Parsi Gujarati as an endangered dialect, which needs to be preserved.

different from the standard Gujarati script and hence is not treated separately. Moreover, it has hardly any written system to speak of. Kachi Koli uses both Gujarati and Nashq to represent its written system. Present day Kacchi written in Gujarati is trying to evolve its own alphabet[13]. Sindhi was written in Gujarati, especially in the region of Kutch but in the present-day context, Sindhi in India is written mainly in Devanagari. The languages using Gujarati are as follow:

- Adiwasi Garasia
- Avestan
- Bhili
- Chodri
- Dungra Bhil
- Gamit
- Kachhi
- Kachi Koli
- Kukna
- Rajput Garasia
- Varli
- Vasavi

In developing this LGR, all known languages with a level between 1 and 4 on the EGIDS scale have been considered.

EGIDS Scale 1	**EGIDS Scale 2**	**EGIDS Scale 3**	**EGIDS Scale 4**
<None>	Gujarati	Kukna	<None>

Table 1: Main languages considered under Gujarati LGR

All efforts have been made to ensure that the writing system of the dialects and languages in the EGIDS scale are fully covered by the script inventory provided in the repertoire [cf. 5 infra]

[13] http://www.kutchimaadu.com/general/kutchi-language-gets-script/

3.4 The structure of written Gujarati

Gujarati is an alphasyllabary and the heart of the writing system is the Akshar. It is this unit, which is instinctively recognized by users of the script. To understand the notion of akshar, a brief overview of the writing system is provided in this Section and the akshar itself will be treated in depth in Section 3.4.

3.4.1 The Consonants

Gujarati consonants have an implicit schwa /ə/ included in them. As per traditional classification they are categorized according to their phonetic properties. There are 5 Varga groups (classes) and one non-Varga group. These Vargas are classified by the way they get pronounced i.e. Velar, Palatal, Retroflex, Dental and Bi-labial. Each Varga contains five homorganic consonants classified as per their properties. The first four consonants, which correspond to Stops, are classified on the basis of Voicing and Aspiration and the last is the corresponding nasal.

Varga	Unvoiced		Voiced		Nasal
	-Asp	+Asp	-Asp	+Asp	
Velar	ક	ખ	ગ	ઘ	ઙ
Palatal	ચ	છ	જ	ઝ	ઞ
Retroflex	ટ	ઠ	ડ	ઢ	ણ
Dental	ત	થ	દ	ધ	ન
Bi-labial	પ	ફ	બ	ભ	મ

Table 2: Varga classification of consonants

Non-Varga	ય	ર	લ	ળ	વ	શ	ષ	સ	હ

Table 3: Non-Varga consonants

3.4.2 The Implicit Vowel Killer: Halant[14]

All consonants contain an implicit vowel (schwa). A special sign is needed to denote that this implicit vowel is stripped off. This is known as the Halant ◌્ (U+0ACD). The Halant thus joins two consonants and creates conjuncts, which can be generally from 2 to 4 consonant combinations. In rare cases it can join up to 5 consonants. However, the notional maximum number of consonants joined to form an akshar is not given by rule, but is rather a constraint that has emerged in practice. It is just an observation drawn from the words that have been observed to date. Given the confluence of languages happening in the Internet age, the possibility that one may want a generic Top Level Domain [gTLD] which may have more than the observed maximum cannot be ruled out. Hence, in the LGR work, this limit will not be enforced.

3.4.2.1 Case of Vowel (V) preceded by Halant (H):

There could be cases involving multi-word domains where V may need to follow an H:

e.g. આમ્અચાર /aːm əchaːr/ (U+0A86 U+0AAE U+0ACD U+0A85 U+0A9A U+0ABE U+0AB0) (meaning: Mango pickle)

This is the case where two different words are joined together, and the former ends in an H and the latter begins with a V. By and large, writing the first word without an H is considered enough for full representation of the sound intended for the first word. Nevertheless, some parts of the linguistic community require the explicit presence of H for full representation of the sound intended.

This is a unique situation necessitated by the absence of any hyphen, space or the Zero Width Non-joiner character in the permissible set of characters in the Root zone repertoire. Otherwise in Gujarati spelling, V is never required to follow an H.

There is some concern that this may create a perceptual similarity among two labels (with and without H), confusing for the majority of the linguistic community.

However, having explicit halant in Gujarati text does survive, (even if increasingly as a rare and perhaps dying practice). Therefore, for practical reasons, the WLE rules (section 7 below) do not explicitly restrict this sequence.

[14]Unicode (cf. Unicode 3.0 and above) prefers the term Virama. In this report both the terms have been used to denote the character that suppresses the inherent vowel.

3.4.3 Vowels

Separate symbols exist for all Vowels, which are either pronounced independently at the beginning or attached to a consonant. To indicate the latter (other than the implicit one), a Vowel modifier (Matra) is attached to the consonant. Since the consonant has a built in schwa, there are equivalent Matras for all vowels excepting the અ (U+0A85). The correlation is shown below:

Vowel	**Corresponding vowel sign (Matra)**
અ U+0A85 (GUJARATI LETTER A)	
આ U+0A86 (GUJARATI LETTER AA)	◌ા U+0ABE (GUJARATI VOWEL SIGN AA)
ઇ U+0A87 (GUJARATI LETTER I)	◌િ U+0ABF (GUJARATI VOWEL SIGN I)
ઈ U+0A88 (GUJARATI LETTER II)	◌ી U+0AC0 (GUJARATI VOWEL SIGN II)
ઉ U+0A89 (GUJARATI LETTER U)	◌ુ U+0AC1 (GUJARATI VOWEL SIGN U)
ઊ U+0A8A (GUJARATI LETTER UU)	◌ૂ U+0AC2 (GUJARATI VOWEL SIGN UU)
ઋ U+0A8B (GUJARATI LETTER VOCALIC R)	◌ૃ U+0AC3 (GUJARATI VOWEL SIGN VOCALIC R)
એ U+0A8F (GUJARATI LETTER E)	◌ે U+0AC7 (GUJARATI VOWEL SIGN E)
ઐ U+0A90 (GUJARATI LETTER AI)	◌ૈ U+0AC8 (GUJARATI VOWEL SIGN AI)

ઓ U+0A93 (GUJARATI LETTER O)	ો U+0ACB (GUJARATI VOWEL SIGN O)
ઔ U+0A94 (GUJARATI LETTER AU)	ૌ U+0ACC (GUJARATI VOWEL SIGN AU)

Table 4: Vowels with corresponding Matras

In addition to show sounds borrowed from English, Gujarati admits two vowels and their corresponding Matras as in

Vowel	Corresponding vowel sign (Matra)
એ U+0A8D (GUJARATI VOWEL CANDRA E)	ે U+0AC5 (GUJARATI VOWEL SIGN CANDRA E)
ઑ U+0A91 (GUJARATI VOWEL CANDRA O)	ૉ U+0AC9 (GUJARATI VOWEL SIGN CANDRA O)

Table 5: "Borrowed" Vowels with corresponding Matras

as in એટલેન્ટિક /Atlantic/ U+0A8D U+0A9F U+0AB2 U+0AC5 U+0AA8 U+0ACD U+0A9F U+0ABF U+0A95

ઑર /or/ U+0A91 U+0AB0

3.4.4 The Anusvara (ં) (U+0A82)

In Gujarati, the Anusvara has a dual function. On the one hand, it acts as a homorganic nasal i.e. it replaces a conjunct group of a Nasal-Consonant+Halant+Consonant belonging to that particular varga. On the other hand, before a non-varga consonant the anusvara represents a nasal sound. Gujarati and its dialects prefer the anusvara to the corresponding half-nasal:

સન્ત	vs.	સંત	/sənt/ saint
U+0AB8 U+0AA8 U+0ACD U+0AA4		U+0AB8 U+0A82 U+0AA4	

3.4.5 Nasalization: Candrabindu (ઁ) (U+0A81)

The Candrabindu is rarely used in Gujarati and if at all, is used to represent content borrowed from Devanāgarī. It is therefore more in the nature of a transliterative character and traditional Gujarati grammars as well as Cardona[15] do not accept it in their inventory. As standard Gujarati does not use the Candrabindu, it is not included in the permissible code-point repertoire. A study of standard Gujarati dictionaries such as the Sarth Jodani Kosh[16], the Gujarati Lexicon[17] and the Brihad Gujarati Kosh[18] does not list the Candrabindu as a character acceptable in Gujarati.

3.4.6 Nukta[19] (઼) (U+0ABC)

Traditionally Gujarati does not admit the Nukta[20]. Gujarati grammarians in their inventory of the Gujarati alphabet do not admit this diacritic However the Nukta is used to represent content where Perso-Arabic characters have to be transliterated as in[21]:

> આગ़ાલિબ ની ૧૮મીગ़ઝલ નો શેર નંબર ૯ અને છેલ્લો શેર છે. ગ़ઝલ નો છેલ્લો શેર મક્તોકહેવાય છે. મક્તામાં શાયર નુ ઉપનામજે તખ़લ્લુસ કહેવાય છે તે સામેલહોય છે. હેફ! ઉસ ચારગિરેહ કપડેકીક़िસમત, ગ़ાલિબ જીસકીક़िસમતમેં હો, આશિક़કા ગરીબાં .
>
> */āġālibnī 18mī ġajhalnōśērnambar 9 anēchēllōśērchē. ġajhalnōchēllōśērmaktōkahēvāychē. maktāmāṁśāyar nu upnāmjētaḵẖalluskahēvāyhētēsāmēlhōychē. hēph! uscārgirēhkapḍēkīqismat, ġālibjīsakīqismatmēṁhō, āśiqkāgarībāṁ./*
>
> This is « Sher » [distich] number 9 of Ghalib's 18th Ghazal and it is the final one. The final « Sher » [distich] of a ghazal is called a « maktō». The pen name of the poet which is indicated therein is termed as the « takhallus. » « Alas. It is an iniquitous lot for a hand's breadth of cloth, Ghalib, To be allotted as the rent collar on a lover's robe ».[22]

---

[15]George Cardona. 1965. A Gujarati Reference Grammar. University of Pennsylvania Press

[16]Gujarat Vidyapith. 1967. Sarth Gujarati JodaniKosh. Amdavad

[17]http://www.gujaratilexicon.com/index.php/

[18]Sastri, KesavaramaKasirama. Brhad Gujarati kosa: Comprehensive Gujarati dictionary. Amdavad : University Granthnirman Board, Gujarat Rajya

[19]The possible sets of consonants have been derived from various sources viz. Prior research carried out by Centre for Development of Advanced Computing's [C-DAC] Graphics Intelligence based Script Technologies [GIST] Research Labs (https://cdac.in/index.aspx?id=mlc_gist_about)

[20]George Cardona. op. cit.

[21]http://roshan-safar.com/હેફ-ઉસ-ચારગિરેહ-કપડેકી-ક़/?lang=gu

[22]Translation referred from: http://www.caravanmagazine.in/poetry/four-ghazals-mirza-ghalib

Similarly, in Parsi Gujarati, the Nukta is used with ફ઼ and જ઼ in the name of a famous author of Munajats:મુલ્લાફ઼િરોજ઼બિન કૌસ / mullāfirōz bin kaus /MullaFiroz Bin Kaus

The Nukta can be adjoined to ક (GUJARATI LETTER KA -U+0A95),ખ (GUJARATI LETTER KHA -U+0A96),ગ (GUJARATI LETTER GA -U+0A97), જ (GUJARATI LETTER JA -U+0A9C), ફ (GUJARATI LETTER PHA -U+0AAB) to show that words having these consonants with a nukta are of Perso-Arabic origin and should be pronounced in the Perso-Arabic style.

3.4.7 Visarga (ઃ) and Avagraha (ऽ)

The Visarga is frequently used in Sanskrit and represents a sound very close to /h/. દુઃખ /duḥkha/ sorrow, unhappiness. It is used sparingly in Gujarati with a few words borrowed from Sanskrit.

The Avagraha (ऽ) creates an extra stress on the preceding vowel and is used in Sanskrit texts. It is rarely used in Gujarati. In the case of LGR, the Avagraha is not part of the repertoire as it is barred in the Maximal Starting Repertoire.

4 Overall Development Process and Methodology

Under the Neo-Brāhmi Generation Panel, there are many different scripts belonging to separate Unicode blocks. Each of these scripts has been assigned a separate LGR; however, Neo-Brāhmi GP ensured that the fundamental philosophy behind building those LGRs are all in accord with all other Brāhmi derived scripts. This is the Gujarati LGR, which caters to multiple languages written using Gujarati belonging to EGIDS scale 1 to 4.

4.1 Guiding Principles

The NBGP adopts following broad principles for selection of code-points in the code-point repertoire across the board for all the scripts within its ambit.

The main principle is that of Acknowledgement of Environmental Limitations. These comprise protocols or standards. All further principles are in fact subsumed under these limitations but have been spelt out separately for the sake of clarity.

4.1.1 External limits on scope:

The code point repertoire for root zone being a very special case, up the ladder in the protocol hierarchies, the canvas of available characters for selection as a part of the Root Zone code point repertoire is already constrained by various protocol layers beneath it. The following three main protocols/standards act as successive filters:

i. The Unicode Chart:

Out of all the characters that are needed by the given script, if the character in question is not encoded in Unicode, it cannot be incorporated in the code point repertoire. Such cases are quite rare, given the elaborate and exhaustive character inclusion efforts made by Unicode consortium. At present, the Unicode version compliant with the LGR is Unicode 7.0

ii. IDNA Protocol:

Unicode being the character encoding standard for providing the maximum possible representation of a given script/language, it has encoded as far as possible all the possible characters needed by the script. However the Domain name being a specialized case, it is governed by an additional protocol known as IDNA (Internationalized Domain Names in Applications). The IDNA protocol excludes some characters in the Unicode repertoire from being part of domain names.

IDNA Protocol also excludes invisible characters Zero Width Non-Joiner (U+200C) and Zero Width Joiner (U+200D), as they require a CONTEXTJ rule. These are required in certain cases where a typical visual shape of an akshar is desired.

iii. Maximal Starting Repertoire:

The Root-zone LGR being a repertoire of the characters which are going to be used for creation of the root zone TLDs, which in turn are an even more specialized case of domain names, the ROOT LGR procedure introduces additional exclusions on IDNA allowed set of characters.

Example: Gujarati Sign Avagraha (ऽ - U+0ABD) even if allowed by IDNA protocol, is not permitted in the Root Zone Repertoire as per the MSR.

To sum up, the restrictions start with admitting only such characters as are part of the code-block of the given script/language. This is further narrowed down by the IDNA Protocol and finally an additional filter in the form of Maximal Starting Repertoire restricts the character set associated with the given language even more.

4.1.2 Exclusion of Punctuation Marks:

The TLDs being identifiers, punctuation markers present in Brāhmi-based languages such as Danda (।) and double Danda (॥) will not be included.

4.1.3 No Symbols and Abbreviations:

Abbreviations, weights, currency and measures and other such iconic characters like BENGALI ISSHAR (৺ -U+09FA), GUJARATI ABBREVIATION SIGN (૰ -U+0AF0) etc. will not be included.

4.1.4 Exclusion of Rare and Obsolete Characters:

These are characters, which have been added to Unicode to accommodate rare forms especially like GUJARATI LETTER VOCALIC RR (ૠ) and GUJARATI LETTER VOCALIC LL (ૡ) as well as their matra forms GUJARATI VOWEL SIGN VOCALIC RR (ૃ) and GUJARATI VOWEL SIGN VOCALIC LL (ૣ). All such characters will not be included. This is in compliance with the Letter principle as laid down in the Root Zone LGR procedure.

5 Repertoire

Section 5.1 provides the section of the [MSR] applicable to the Gujarati script on which the Gujarati code-point repertoire is based.

Section 5.2 details the code-point repertoire that the Neo-Brahmi Generation Panel [NBGP] proposes to be included in the Devanagari LGR.

5.1 Gujarati section of Maximal Starting Repertoire [MSR] Version 3

Gujarati

Figure 5: Gujarati Code Page from [MSR]

Color convention[23]:

All characters that are included in the [MSR] - Yellow background

PVALID in IDNA2008 but excluded from the [MSR] - Pinkish background

Not PVALID in IDNA2008, or are ineligible for the root zone (digits, hyphen) - White background

[23]This document needs to be printed in color for this to be read correctly.

5.2 Code Point Repertoire

This section details the code-point repertoire[24] that the Neo-Brāhmi Generation Panel [NBGP] proposes to be included in the Gujarati LGR.

Sr. No.	Unicode Code Point	Glyph	Character Name	Indic Syllabic Category	Reference
1 .	0A82	ં	GUJARATI SIGN ANUSVARA	Anusvara (Bindu)	[Omniglot]
2.	0A83	ઃ	GUJARATI SIGN VISARGA	Visarga	[Omniglot]
3.	0A85	અ	GUJARATI LETTER A	Vowel	[Omniglot]
4.	0A86	આ	GUJARATI LETTER AA	Vowel	[Omniglot]
5.	0A87	ઇ	GUJARATI LETTER I	Vowel	[Omniglot]
6.	0A88	ઈ	GUJARATI LETTER II	Vowel	[Omniglot]
7.	0A89	ઉ	GUJARATI LETTER U	Vowel	[Omniglot]
8.	0A8A	ઊ	GUJARATI LETTER UU	Vowel	[Omniglot]
9.	0A8B	ઋ	GUJARATI LETTER VOCALIC R	Vowel	[Omniglot]
10.	0A8C	ઌ	GUJARATI LETTER VOCALIC L	Vowel	[Omniglot]

[24] https://www.omniglot.com/writing/gujarati.htm. Inaccuracies in the character set were pointed out to the site-owner and these have been corrected.

11.	0A8D	ઍ	GUJARATI VOWEL CANDRA E	Vowel	[Omniglot]
12.	0A8F	એ	GUJARATI LETTER E	Vowel	[Omniglot]
13.	0A90	ઐ	GUJARATI LETTER AI	Vowel	[Omniglot]
14.	0A91	ઑ	GUJARATI VOWEL CANDRA O	Vowel	[Omniglot]
15.	0A93	ઓ	GUJARATI LETTER O	Vowel	[Omniglot]
16.	0A94	ઔ	GUJARATI LETTER AU	Vowel	[Omniglot]
17.	0A95	ક	GUJARATI LETTER KA	Consonant	[Omniglot]
18.	0A96	ખ	GUJARATI LETTER KHA	Consonant	[Omniglot]
19.	0A97	ગ	GUJARATI LETTER GA	Consonant	[Omniglot]
20.	0A98	ઘ	GUJARATI LETTER GHA	Consonant	[Omniglot]
21.	0A99	ઙ	GUJARATI LETTER NGA	Consonant	[Omniglot]
22.	0A9A	ચ	GUJARATI LETTER CA	Consonant	[Omniglot]
23.	0A9B	છ	GUJARATI LETTER CHA	Consonant	[Omniglot]
24.	0A9C	જ	GUJARATI LETTER JA	Consonant	[Omniglot]
25.	0A9D	ઝ	GUJARATI LETTER JHA	Consonant	[Omniglot]

26.	0A9E	ઞ	GUJARATI LETTER NYA	Consonant	[Omniglot]
27.	0A9F	ટ	GUJARATI LETTER TTA	Consonant	[Omniglot]
28.	0AA0	ઠ	GUJARATI LETTER TTHA	Consonant	[Omniglot]
29.	0AA1	ડ	GUJARATI LETTER DDA	Consonant	[Omniglot]
30.	0AA2	ઢ	GUJARATI LETTER DDHA	Consonant	[Omniglot]
31.	0AA3	ણ	GUJARATI LETTER NNA	Consonant	[Omniglot]
32.	0AA4	ત	GUJARATI LETTER TA	Consonant	[Omniglot]
33.	0AA5	થ	GUJARATI LETTER THA	Consonant	[Omniglot]
34.	0AA6	દ	GUJARATI LETTER DA	Consonant	[Omniglot]
35.	0AA7	ધ	GUJARATI LETTER DHA	Consonant	[Omniglot]
36.	0AA8	ન	GUJARATI LETTER NA	Consonant	[Omniglot]
37.	0AAA	પ	GUJARATI LETTER PA	Consonant	[Omniglot]
38.	0AAB	ફ	GUJARATI LETTER PHA	Consonant	[Omniglot]
39.	0AAC	બ	GUJARATI LETTER BA	Consonant	[Omniglot]
40.	0AAD	ભ	GUJARATI LETTER BHA	Consonant	[Omniglot]

41.	0AAE	મ	GUJARATI LETTER MA	Consonant	[Omniglot]
42.	0AAF	ય	GUJARATI LETTER YA	Consonant	[Omniglot]
43.	0AB0	ર	GUJARATI LETTER RA	Consonant	[Omniglot]
44.	0AB2	લ	GUJARATI LETTER LA	Consonant	[Omniglot]
45.	0AB3	ળ	GUJARATI LETTER LLA	Consonant	[Omniglot]
46.	0AB5	વ	GUJARATI LETTER VA	Consonant	[Omniglot]
47.	0AB6	શ	GUJARATI LETTER SHA	Consonant	[Omniglot]
48.	0AB7	ષ	GUJARATI LETTER SSA	Consonant	[Omniglot]
49.	0AB8	સ	GUJARATI LETTER SA	Consonant	[Omniglot]
50.	0AB9	હ	GUJARATI LETTER HA	Consonant	[Omniglot]
51.	0ABC	◌઼[25]	GUJARATI SIGN NUKTA	Nukta	[Omniglot]
52.	0ABE	◌ા	GUJARATI VOWEL SIGN AA	Matra	[Omniglot]
53.	0ABF	◌િ	GUJARATI VOWEL SIGN I	Matra	[Omniglot]
54.	0AC0	◌ી	GUJARATI VOWEL SIGN II	Matra	[Omniglot]

[25]Cf. footnotes re. Nukta supra

55.	0AC1	ુ	GUJARATI VOWEL SIGN U	Matra	[Omniglot]
56.	0AC2	ૂ	GUJARATI VOWEL SIGN UU	Matra	[Omniglot]
57.	0AC3	ૃ	GUJARATI VOWEL SIGN VOCALIC R	Matra	[Omniglot]
58.	0AC4	ૄ	GUJARATI VOWEL SIGN VOCALIC RR	Matra	[Omniglot]
59.	0AC5	ૅ	GUJARATI VOWEL SIGN CANDRA E	Matra	[Omniglot]
60.	0AC7	ે	GUJARATI VOWEL SIGN E	Matra	[Omniglot]
61.	0AC8	ૈ	GUJARATI VOWEL SIGN AI	Matra	[Omniglot]
62.	0AC9	ૉ	GUJARATI VOWEL SIGN CANDRA O	Matra	[Omniglot]
63.	0ACB	ો	GUJARATI VOWEL SIGN O	Matra	[Omniglot]
64.	0ACC	ૌ	GUJARATI VOWEL SIGN AU	Matra	[Omniglot]
65.	0ACD	્	GUJARATI SIGN VIRAMA	Halant / Virama	[Omniglot]

Table 6: Code point repertoire

5.3 Code point not included:

Following code point has not been included in the repertoire.

Sr. No.	Unicode Code Point	Glyph	Character Name	Reason for exclusion
1.	U+0A81	ઁ	GUJARATI SIGN CANDRABINDU	Not used in standard Gujarati. See Section 3.4.5 for more information.

5.4 The Structural Formation of Gujarati:

All the languages written in Brāhmi derived scripts follow a particular way of formation of its words, known as "akshar". In the next section, detailed akshar formation rules as applicable to representation of languages written in Gujarati Script are provided.

In Section 7, the Whole Label Evaluation (WLE) rules are given which cover all the languages under the purview of the NBGP for Gujarati script.

5.5 Akshar formation rules for Gujarati:

This section details the Akshar formation rules as applicable to Gujarati. The first section lists the categories of the characters in the form of variables. In the rules, instead of their descriptive names, the variable names are used. The second section lists four operators along with their functions which are assumed while specifying the rules. The following two sections describe the two major categories of the Akshar formations first of which begins with the vowels and the second one with the consonants. These rules are based on an Indian Standard (IS 13194:1991) popularly known as "Indian Script Code for Information Interchange" [ISCII].

5.5.1 Variables involved

Dash → Hyphen -

Digit → Indo-Arabic digits [0-9]

C → Consonant

M → Matra

V → Vowel

B → Anusvara (Bindu)

X → Visarga

H → Halant / Virama

N → Nukta[26]

[26]Cf. Notes re. Nukta supra

5.5.2 Operators used:

Symbol	Function
|	Alternative
[ ]	Optional
*	Variable Repetition
( )	Sequence Group

Table 7: Symbol functions

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

5.5.3 The Vowel Sequence

A vowel sequence begins with a vowel. It may be optionally followed by an Anusvara (B), or a Visarga (X). The number of B or X which can follow a V in Gujarati are restricted to one[27].

The vowel sequence in Gujarati is therefore V [B | X]

Examples:

Sequence Description	Sequence	Example	Example Decomposition
Vowel	V	અ /a/ U+0A85	
Vowel + Anusvara	V[B]	અં /aṁ/ U+0A85 U+0A82	અ ં U+0A85 U+0A82
Vowel + Visarga	V[X]	અઃ /aḥ/ U+0A85 U+0A83	અ ઃ U+0A85 U+0A83

Table 8

[27]The possibility of a Visarga following an Anusvara is ruled out, since this combination is used only in Vedic and in Bengali script.

5.5.4 Consonant Sequence

A consonant sequence begins with a consonant. It may be optionally followed by a Nukta (N), Matra (M), Anusvara (B), Visarga (X) or a 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, M 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:

Sequence Description	Sequence	Example	Example Decomposition
Consonant	C	ક/ka/ U+0A95	
Consonant + Nukta	C[N]	ક़/क़a/ U+0A95 U+0ABC	ક़ U+0A95 U+0ABC

Table 9

2. A consonant optionally followed by dependent vowel sign/Matra [M] or Anusvara [B] or Visarga [X] or Halant [H]

C [M|B|X|H]

Examples:

Sequence Description	**Sequence**	**Example**	**Example Decomposition**
Consonant + Matra	C[M]	કિ/ki/ U+0A95 U+0ABF	કિ U+0A95 U+0ABF
Consonant + Anusvara	C[B]	કં/kaṁ/ U+0A95 U+0A82	કં U+0A95 U+0A82
Consonant + Visarga	C[X]	કઃ/kaḥ/ U+0A95 U+0A83	કઃ U+0A95 U+0A83

Consonant + Halant	C[H]	ક્/k/ (Pure Consonant) U+0A95 U+0ACD	ક્ U+0A95 U+0ACD

Table 10

2. A. A CM sequence can be optionally followed by B or X

(CM)[B|X]

Example:

Sequence Description	Sequence	Example	Example Decomposition
Consonant + Matra + Anusvara	CM[B]	કીં/kīṁ/ U+0A95 U+0AC0 U+0A82	કીં U+0A95 U+0AC0 U+0A82
Consonant + Matra + Visarga	CM[X]	કીઃ/kīḥ/ U+0A95 U+0AC0 U+0A83	કીઃ U+0A95 U+0AC0 U+0A83

Table 11

3. A sequence of consonants (up to 4) joined by a Halant *3(CH)C. These sequences are mainly found in loan words from English where the onset and coda of the syllable admit consonantal clusters..

Example:

Sequence Description	Sequence	Example	Example Decomposition
Consonant + Halant + Consonant + Halant + Consonant + Halant + Consonant	CHCHCHC	ૹ્રૅઋ/ rlds/ U+0AB9 U+0ACD U+0AB0 U+0ACD U+0AA1 U+0ACD U+0AB8	ર ્ લ ્ ડ ્ ઋ

Table 12

Subsets:

3. A. The combination may be followed by M, B or X

Example:

Sequence Description	Sequence	Example	Example Decomposition
Consonant + Halant + Consonant + Matra	CHC[M]	ક્કી/kkī/ U+0A95 U+0ACD U+0A95 U+0AC0	કકી U+0A95 U+0ACD U+0A95 U+0AC0
Consonant + Halant + Consonant + Anusvara	CHC[B]	કક્ં/kkaṁ/ U+0A95 U+0ACD U+0A95 U+0A82	કક્ં U+0A95 U+0ACD U+0A95 U+0A82
Consonant + Halant + Consonant + Visarga	CHC[X]	કક્ઃ/kkaḥ/ U+0A95 U+0ACD U+0A95 U+0A83	કક્ઃ U+0A95 U+0ACD U+0A95 U+0A83

Table 13

3. B. *3(CH)CM may be followed by a B or X

Example:

Sequence Description	Sequence	Example	Example Decomposition
Consonant + Halant + Consonant + Matra + Anusvara	CHCM[B]	કકીં/kkīṁ/ U+0A95 U+0ACD U+0A95 U+0AC0 U+0A82	ક ◌ુ ક ◌ી ◌ં U+0A95 U+0ACD U+0A95 U+0AC0 U+0A82
Consonant + Halant + Consonant + Matra + Visarga	CHCM[X]	કકીઃ/kkīḥ/ U+0A95 U+0ACD U+0A95 U+0AC0 U+0A83	ક ◌ુ ક ◌ી ◌ઃ U+0A95 U+0ACD U+0A95 U+0AC0 U+0A83

Table 14

Gujarati LGR is driven by these basic akshar rules. However, owing to Simplicity principle as laid down in the LGR Procedure, not all the rules described in this section have been considered in the final LGR rules provided in Whole Label Evaluation Rules (WLE). E.g. conjunct depth (maximum number of consonants joining each other to form a conjunct) of 4 is not enforced in the Gujarati LGR.

6 Variants

There are no characters/character sequences in Gujarati, which can be created by using the characters permitted as per the [MSR] and look exactly alike. Hence no variants are being proposed in Gujarati LGR. However, Gujarati has some cases of confusingly similar combinations which have been listed in Appendix A: In-script variant candidates.

7 Whole Label Evaluation Rules (WLE)

This section provides the WLEs that are required by all the languages mentioned in section 3.2 when written in Gujarati Script. The rules have been drafted in such a way that they can be easily translated into the LGR specification.

Below are the symbols used in the WLE rules, for each of the "Indic Syllabic Category" as mentioned in Table 6: Code point repertoire.

C → Consonant

M → Matra

V → Vowel

B → Anusvara (Bindu)

X → Visarga

H → Halant / Virama

N → Nukta

Below are the specific WLE rules:

1. N: must be preceded only by any of the specific set of Cs

   The specific Cs are:

   I. ક (U+0A95)
   II. ખ (U+0A96)
   III. ળ (U+0A97)
   IV. ક઼ (U+0A9C)
   V. ફ (U+0AAB)

2. H[28]: must be preceded by C or N
3. X: must be preceded by either of V, C, N or M
4. B: must be preceded by either of V, C, N or M
5. M: must be preceded either by C or N

[28]The limit of maximum 3 consonants joining to form a conjunct has not been enforced here as it may make the rules overly complex.

8 Contributors

- Dr. Raiomond Doctor
- Mr. Mahesh D. Kulkarni
- Mr. Arvind Bhandari
- Ms. Aparna A. Kulkarni
- Ms. Neha Gupta
- Mr. Akshat Joshi
- Mr. Nishit Jain

9 References

[MSR] Integration Panel, "Maximal Starting Repertoire — MSR-3 Overview and Rationale", 28 March 2018 https://www.icann.org/en/system/files/files/msr-3-overview-28mar18-en.pdf

[NBGP] Neo-Brāhmi Generation Panel, https://community.icann.org/display/croscomlgrprocedure/Neo-Brahmi+GP

[Omniglot], "Gujarati", https://www.omniglot.com/writing/gujarati.htm(Accessed on 6th Jan. 2018)

[EGIDS] Expanded Graded Intergenerational Disruption Scale, https://www.ethnologue.com/about/language-status (Accessed on 13th Nov. 2017)

[GIST] Graphics Intelligence based Script Technologies, https://cdac.in/index.aspx?id=gist (Accessed on 2nd Feb. 2018)

[C-DAC] Centre for Development of Advanced Computing, https://cdac.in (Accessed on 2nd Feb. 2018)

[ISCII] Indian Script Code for Information Interchange, https://cdac.in/index.aspx?id=mlc_gist_iscii(Accessed on 2nd Feb. 2018)

10 Materials and references used in the creation of this document

Following is a thematically sorted set of documents, books, articles and webographies consulted in the drafting of this report

10.1 ON WRITING SYSTEMS

1. Cardona, George; Jain, Dhanesh. 2003, The Indo-Aryan Languages, Philadelphia: Routledge.
2. Carol, Andrews. 1981, The Rosetta Stone ,London
3. Cohen, Marcel. 1953, L'Écriture. Paris.
4. Cohen, Marcel. 1958, La Grande Invention de l'Écriture et son Évolution, Paris.
5. Comrie, Bernard. 1987 ed., The World's Major Languages, London: Croom Helm; New York: Oxford University Press.
6. Diringer, David. 1962, Writing, London
7. Diringer, David. 1968. The Alphabet. A Key to the History of Mankind, 3rd ed. New York: Funk & Wagnalls.
8. Faulmann, Carl. 1990. Das Buch der Schrift. Frankfurt am Main: Eichborn
9. Haarmann, Harald. 1990. Die Universalgeschichte der Schrift. Frankfurt: Campus.
10. Kausen, Ernst, 2006. Die Klassifikation der indogermanischen Sprachen
11. Meillet, Antoine and Cohen, Marcel.1952. Les langues du monde. Collection linguistique, 16. Paris: Champion.
12. Taylor, Isaac. 1883. The alphabet: an account of the origin and development of letters. Vol. 1: Semitic alphabets; Vol. 2: Aryan alphabets. London: Kegan Paul.
13. Alan S. Kaye and Peter T. Daniels. 1997. (Eds).; Phonologies of Asia and Africa (Including the Caucasus). Volume 2. Winona Lake, Indiana. EISENBRAUNS
14. Cardona George. 1965. A Gujarati Reference Grammar. University of Pennsylvania Press
15. Cardona, George and Babu Suthar. 2003. Gujarati in Cardona, George; Jain, Danesh, The Indo-Aryan Languages, Philadelphia: Routledge.
16. Dave, Jagdish.1995. Colloquial Gujarati. Routledge.
17. Dave, T. N. 1995. Language of Gujarat, translated by Minaxi K. Patel Amadawad. University Book Production Board, Gujarat State.
18. Directorate of Languages. Gujarat State. 1979. An Introduction to Gujarati Language.. Rev. ed. 1991
19. Doctor, Raimond. 2004. A Grammar of Gujarati. LINCOM EUROPA. German Reprint 2007
20. Dwyer, Rachel. 1995. Gujarati : a complete course for beginners. Teach Yourself Books.
21. Gujarat Vidyapith. 1967. Sarth Gujarati Jodani Kosh. Amdavad.
22. Kothari, Jayant. 1983. Introduction to Language and Structure of Gujarati Language, Amadawad. University Book Production Board, Gujarat State.

23. Lambert H.M.1968. Introduction to the Devanāgarī Script for Students of Sanskrit, Hindi, Marathi, Gujarati and Bengali. Oxford University Press
24. Lambert H.M. 1971. Gujarati Language Course. Cambridge University press.
25. Mistry, P. 1996. "Gujarati Writing" in Daniels and Bright ed. The World's Writing Systems. OUP. pp 391-395
26. Modi, Bharati. 2011. Parsi Gujarati - Vanishing Dialect : Vanishing Culture. LINCOM EUROPA
27. Sastri, KesavaramaKasirama. Brhad Gujarati kosa: Comprehensive Gujarati dictionary. Amdavad : University Granthnirman Board, Gujarat Rajya
28. Vyas, Yogendra. 1977. Gujarati Bhashanu Vyakaran, Ahmedabad: Sahitya Mudralay.
29. William St. Clair Tisdall. 1961. A Simplified Grammar of the Gujarati Language: Together with a Short Reading Book and Vocabulary. F. Ungar Publishing Company.
30. ગુજરાત રાજ્ય ગાંધીનગર.1991. ભાષાનિયમક કચેરી ગુજરાતી ભાષા પરિચય (Gujarat Rajya Gandhinagar. 1991. BhashaNiyamKacheri, Gujarati BhashaParichay)
31. (Translation: Gujarat State Gandhinagar, 1991. Language Rules Office, Gujarati Language Introduction)

10.2 WEBOGRAPHY

1. Wikipedia, “Gujarati alphabet”, https://en.wikipedia.org/wiki/Gujarati_alphabet
2. Wikipedia, “Gujarati language”, https://en.wikipedia.org/wiki/Gujarati_language
3. Wikipedia, “Gujarati language”, https://en.wikipedia.org/wiki/Gujarati_(Unicode_block)
4. Scriptsource.com, http://scriptsource.org/cms/scripts/page.php?item_id=script_detail_use&key=Gujr
5. Ethologue, “Gujarati”, https://www.ethnologue.com/language/guj
6. Gujarati Lexicon, http://www.gujaratilexicon.com
7. ICANN, “Maximal Starting Repertoire 1”, https://www.icann.org/en/system/files/files/msr-overview-06jun14-en.pdf
8. ICANN, “Maximal Starting Repertoire 3”, https://www.icann.org/en/system/files/files/msr-3-overview-28mar18-en.pdf

11 Appendix A: In-script variant candidates

There are no characters/character sequences between Gujarati and other scripts under the NBGP ambit which can be created by using the characters permitted as per the [MSR] and look confusingly similar enough. Hence no cross-script variants are being proposed in Gujarati LGR.

There are certain combinations within the Gujarati script which may look confusingly similar but not enough to be termed as Variants. They are as follows.

Confusable 1	Confusable 2	Confusable 3
ચ U+0A9A	ચ U+0AAF	---
ફ્ચ U+0AAB U+0AAF	ફ્ચ U+0AAB U+0ACD U+0AAF	---
ફૂ U+0A95 U+0AC2	ફ U+0AAB	---
ઘ U+0AA6 U+0ACD U+0AB0	ઘ્ન U+0AA6 U+0ACD U+0AA8	ઘ્ન U+0AA6 U+0ACD U+0A97

Table 15: In-script confusables