Proposal for a Gurmukhi Script Root Zone Label Generation Ruleset [LGR]

*LGR Version:* 3.0

*Date:* 2018-01-15

*Document version:*1.1

*Authors:* Neo-Brahmi Generation Panel [NBGP]

# General Information/ Overview/ Abstract

This document lays down the Label Generation Ruleset for Gurmukhi script. Three main components of the Gurmukhi Script LGR i.e. Code point repertoire, Variants and Whole Label Evaluation Rules have been described in detail here. All these components have been incorporated in a machine-readable format in the accompanying XML file named "Proposed-LGR-Gur-20171011.xml".

# Script for which the LGR is proposed

ISO 15924 Code: Guru

ISO 15924 Key N°: 310

ISO 15924 English Name: Gurmukhi

Latin transliteration of native script name: gurmukhī

Native name of the script: ਗੁਰਮੁਖੀ

Maximal Starting Repertoire [MSR] version: 2

# Background on Script and Principal Languages Using It

It is commonly accepted that Gurmukhi is a member of the BRAHMI family. Brahmi is an Aryan script which was developed by the Aryans and adapted to local needs. According to an opinion , the Brahmi script was introduced between the 8th and the 6th century BC. It does not concern us here whether the script was foreign or local, but it has now been established, on the basis of its name, that the Aryans did have a system of writing which must have been borrowed freely from local script.

The Iranians ruled the Punjab in 3rd and 4th centuries BC. They brought with them Aramaic script, which helped in the growth of Kharosthi largely used in the Punjab , Gandhar and Sindh between 300 BC and 3rd century AD. But even then Brahmi which in its development in the Punjab had undergone several changes , was commonly used along with Kharosthi. There are coins of the Bactrian kings and inscriptions of the Kushan rulers having both scripts on them. Brahmi was, of course, more popular on account of its simple curves alternated with straight strokes. Hence , in due course, it replaced Kharoshti and became the single script with composite features brought about by various local and neighborly influences. With the growth of literary and cultural activity during the Gupta period (4th-5th century AD) the Brahmi script improved further and became more expansive and common.

Immediately later, it developed, especially in northern India, fine curves and embellished flourishes with a small headline over each letter and became rather ornamental. This stage of Indian script was called Kutil (means curved) which had the widest use in Northern India. With the rise of regional languages taking the place of Sanskrit and Prakrit, regional scripts grew in number. Ardhanagari (west), Sharda (Kashmir) and Nagri (beyond Delhi) came into use and later both Sharda and Devanagari started their inroads into the land of the five rivers. This is evident from the coins of the Gaznavies and Goris minted at Lahore and Delhi. It is also known that the common (non-Brahman and non-official) people used a number of scripts for their temporal and commercial requirements. Of these scripts, Lande and Takre characters were most prevalent.

## The Evolution of the Script

Like most of the North Indian writing systems, the Gurmukhi script is a descendant of the Brahmi script. The Proto-Gurmukhi letters evolved through the Gupta script from 4th to 8th century, followed by the Sharda script from 8th century onwards and finally adapted their archaic form in the Devasesha stage of the later Sharda script, dated between the 10th and 14th centuries. The traditional accounts such as the references found in Janamsakhi literature say that the Gurmukhi script was invented by the second Sikh Guru, Guru Angad Dev. However it would be correct to say that the script was standardized rather than invented. E.P. Newton writes that at least 21 Gurmukhi characters are found in ancient manuscripts : 6 from 10th century, 12 from 3rd century B.C. Apparently, the first Sikh Guru, Guru Nanak Dev also used the Gurmukhi script for his writings. Scholars have tried to establish relationship of Gurmukhi script with different scripts, such as Devanagari (G.H. Ojha), Ardhanagari (G.B. Singh), Siddam/Siddamatirka (Pritam Singh), Sharda (Diringer) and Brahmi (generally). Some ascribe it to Lande and some others to Takre, a branch of Sharda used in Chamba area . The fact is that it is derived from or at least allied to all these and others mentioned above in their historical perspective. Regionally and contemporarily compared, Gurmukhi characters have direct similarities with Gujarati , Lande , Nagri, Sharda ,and Takre : they are either exactly the same or essentially alike. Internally, Aira (ਅ), Haha (ਹ), Chaccha (ਚ), Dadda (ਦ), Naana (ਣ) , Lalla (ਲ) letters of Gurmukhi had undergone some minor orthographical changes before 1610 A.D. A major change occurred in (ਙ) and (ਞ) letters. Babba (ਬ) letter was invented later.

Further changes came in the forms of Aira (ਅ), Haha (ਹ) and Lalla (ਲ) letters in the first half of the nineteenth century. Another reform carried out is the separation of lexical units of the sentence which previously formed one jumbled unit; lately punctuation marks borrowed from English have been incorporated besides the full stop. In place of full stop, dandi has been used which existed traditionally.

There are two major theories on how the Proto-Gurmukhi script emerged in the fifteenth century. G.B. Singh (1950) while quoting Al-Biruni’s Al-Hind (1030 AD), says that the script evolved from Ardhanagri. Al-Biruni writes that the Ardhanagri script was used in Bathinda, including Sindh and western parts of the Punjab in the tenth century. For some time Bathinda remained the capital of the kingdom of Bhatti Rajputs, who ruled North India before the Muslims occupied the country. Resulting from its connection with the Bhattis , the Ardhanagri script was also called Bhatachhari. According to Al-Biruni Ardhanagri was a mixture of Nagri ,used in Ujjain and Malwa, and Sidha Matrika is a variant of the Sharda script used in Kashmir. Pritam Singh (1992) also traced the origins of Gurmukhi to the Siddha Matirka.

Tarlochan Singh Bedi (1999) has suggested that the Gurmukhi script was developed during the tenth-fourteenth centuries from the Devasesha stage of the Sharda script. His argument suggests that regional differences started to appear in the Sharda script used in Punjab, partly Himachal Pradesh and Kashmir from the tenth century. The regional form of Sharda, used in Punjab, started to appear in the form of Gurmukhi in the 14th century. Bedi called this stage Pritham Gurmukhi or Proto-Gurmukhi.

The Sikh Gurus adopted the proto-Gurmukhi script to write the Guru Granth Sahib, the primary religious scripture of the Sikhs. Other contemporary scripts used in Punjab were Takri and the Landa alphabets. Takri was a script that developed through the Devasesha stage of the Sharda script and is found mainly in the Hill States such as Chamba, where it is called Chambyali and in Jammu, where it is called Dogri. The local Takri variants attained the status of official scripts in some of the Punjab, Hills States and were used for both administrative and literary purposes. When Himachal Pradesh was established ,the local Takri was replaced by Devanagari.

Meanwhile, the mercantile scripts of Punjab known as Landa were normally not used for literary purposes. In Punjab, there were at least ten different variants of Landa. Lande alphabets were used for household and trade purposes.

The letters no doubt existed before the time of Guru’s period. But Sikh Gurus not only modified and re-arranged certain letters but also shaped them into a script. They gave new shape and new order to the alphabet and made it precise and accurate. They fixed one letter for each of Punjabi phonemes, use of vowels-symbols was made obligatory, the letters meant for conjuncts were not adopted and only those letters were retained which depicted sounds of the then spoken language .There was some re-arrangement of the letters also such as sassa( ਸ ) and haha( ਹ ) were shifted to the first line and ura( ੳ ) was given the first place in the new alphabet.

**Gurmukhi Etymology**

The word Gurmukhi is the compound form of Guru and Mukh. It is commonly translated as from the mouth of Guru. However the term is used for the Punjabi script and has somewhat different connotations. The opinion given by traditional scholars is that as the Sikh holy writings, before they were scribed, were uttered by the Gurus, they came to be known as Gurmukhi (the Utterance of the Guru) and consequently, the script that was used for the scribing the utterance was also given the same name. However the prevalent view among Punjabi linguists is that as in the early stages, the Gurmukhi letters were primarily used by Gurmukhs (one who is living a life, as instructed by the Guru) or the Sikhs devoted to the Guru, the script came to be associated with them. The script associated with Gurmukhs came to be known as Gurmukhi.

Now Gurmukhi is the name of the script used in writing primarily for Punjabi language and secondarily was used for Sindhi language, but not now.

The development of Gurmukhi script is indicated by the following diagram ;

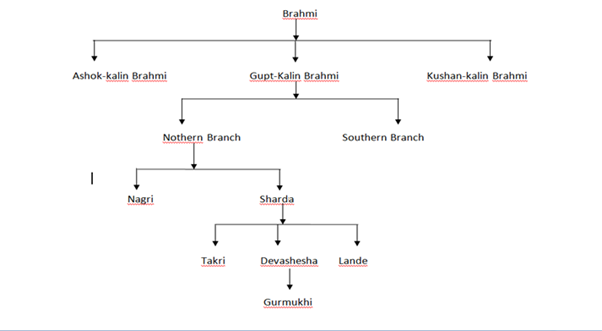


Table 1: Evolution of Gurmukhi



Figure 1: Pictorial depiction of Proto-Gurmukhi (13th century)

## Languages considered

Punjabi (EGIDS 2) is the only language which is currently using the Gurmukhi script.

## The structure of written Gurmukhi

Punjabi is written using the Gurmukhi script. It is an alphasyllabary with the akshar as its core. All scripts derived from Brahmi are Abugidas, i.e. syllabary driven systems. The main features of Abugidas are as under:

* The consonant has an implicit /ə/ vowel which is also known as the schwa.
* The inherent vowel can be modified by the addition of other vowels or muted by a diacritic termed as a Virama or Halant.
* Vowels can be handled as full vowels with a vocalic value .
* When two or more consonants join together they form ligatures. In Gurmukhi script, ligatures are formed only in condition of /h, r and v/ consonants

The writing system of Gurmukhi could be summed up as composed of the following:

### The Consonants

In Gurmukhi, all consonants have an implicit vowel sign schwa /ə/ within them. But these consonants are also used to represent consonant sounds where /ə/ vowel is not incorporated with them without any modification. Actually Gurmukhi script is of semi-syllabic nature where Punjabi consonant letters by itself may stand for the consonant sound as well as for the consonant plus following /ə/ vowel.

Punjabi is a tone language like Chinese but unlike Chinese, the tones are not represented by any letters or symbols in the Gurmukhi script. The same word in Punjabi can have different meaning depending on the pitch of voice or tone used in pronouncing it.

As per traditional classification, consonants are categorized according to their phonetic properties. There are 7 varga groups (classes) and one non-varga group. First Group contains five letters including three vowel carriers and two consonants. In this Group, both consonants are the fricatives. The first is dental and second is glottal. Next five vargas (groups), correspond to stops+nasals, contain five consonants classified as per their properties. The first four consonants are classified on the basis of Voicing and Aspiration and the last is the corresponding nasal. The fourth consonant in these five vargas that has been traditionally classified as voiced aspirated consonant is in fact used to denote tone. Punjabi does not contain voiced aspirated consonants. The pronunciation of these five voiced aspirated consonants is turned into tone. Actually Punjabi is a tonal language. When any of these letters comes in initial position then we are supposed to pronounce it as a unvoiced unaspirated consonant of that varga with a low tone, and in middle position we are supposed to pronounce it as a voiced unaspirated consonants of that varga with high and low tone, depending on the length of the preceding or the following vowel, and at the end of a word these letters are pronounced as voiced unaspirated consonants of that varga with high tone. So these letters could be pronounced only in two tones either it would be a high tone or a low tone. Next five consonants have different phonetic properties (in terms of place plus manner of articulation). So they do not correspond to a single varga. They are categorized as non-varga group. The last group has six letters. All the letters in this group have a bindi (dot) placed in their foot. So they are categorized as pairin bindi (having dot in the foot) letters.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Varga** | **Fricatives** | | | | |
|  | Vowel carrier | Vowel carrier | Vowel carrier | Dental | Glottal |
| **Mul Varga** | ੳ  U+0A73 | ਅ  U+0A05 | ੲ  U+0A72 | ਸ  U+0A38 | ਹ  U+0A39 |

Table 2: Mul varga

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Varga** | **Unvoiced** | | **Voiced** | | **Nasal** |
|  | -Asp | +Asp | -Asp | +Asp\* |  |
| **Velar** | ਕ  U+0A15 | ਖ  U+0A16 | ਗ  U+0A17 | ਘ  U+0A18 | ਙ  U+0A19 |
| **Palatal** | ਚ  U+0A1A | ਛ  U+0A1B | ਜ  U+0A1C | ਝ  U+0A1D | ਞ  U+0A1E |
| **Retroflex** | ਟ  U+0A1F | ਠ  U+0A20 | ਡ  U+0A21 | ਢ  U+0A22 | ਣ  U+0A23 |
| **Dental** | ਤ  U+0A24 | ਥ  U+0A25 | ਦ  U+0A26 | ਧ  U+0A27 | ਨ  U+0A28 |
| **Bi-labial** | ਪ  U+0A2A | ਫ  U+0A2B | ਬ  U+0A2C | ਭ  U+0A2D | ਮ  U+0A2E |

Table 3: Varga classification of consonants

\*Traditionally these letters indicate voiced aspirates but in Punjabi they are used to represent low+high tones

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Non Varga** | ਯ  U+0A2F | ਰ  U+0A30 | ਲ  U+0A32 | ਵ  U+0A35 | ੜ  U+0A5C |

Table 4: Non-Varga consonants

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Pairin Bindi Varga** | ਸ਼  U+0A36 | ਖ਼  U+0A59 | ਗ਼  U+0A5A | ਜ਼  U+0A5B | ਫ਼  U+0A5E | ਲ਼  U+0A33 |

Table 5: Pairin bindi consonants

### The Implicit Vowel Killer: Halant

In Gurmukhi and Devanagari, consonants have an implicit schwa /ə/ included in them. In Hindi, a special sign called halant "੍" (U+094D) is needed to indicate that this implicit vowel is stripped off, so to create conjuncts, halant is used with the consonants in Devanagari. Unlike Devanagari, Gurmukhi consonants are also used to represent consonant sounds where / ə / is not included in them.

So in Punjabi, halant is not used with the consonant that represents only consonant sound not consonant plus vowel sound. Therefore in Punjabi halant is not used to create conjunct except with the letters haha ਹ (U+0A39), rara ਰ (U+0A30)

and vava ਵ (U+0A35) in conjuncts. So when /h, r and v/ phonemes occur as the second member of consonant clusters, the halant joins these consonants in the foot of their preceding consonants and creates conjuncts. So in the consonant clusters, haha, rara and vava letters change their shape as pairin haha (), pairin rara () and pairin vava ().

### Vowels

Punjabi has ten vowels /ਅ(ə), ਆ(a), ਇ(I), ਈ(i), ਉ(U), ਊ(u), ਏ(e), ਐ(ɛ), ਓ(o) and ਔ(ͻ)/. The vowels are represented by nine matras (vowel signs) + three matra vahaks (vowel carriers). Out of these vowels, three /ਅ(ə), ਇ(I), ਉ(U)/ are short vowels and seven (ਆ(a), ਈ(i), ਊ(u), ਏ(e), ਐ(ɛ), ਓ(o) and ਔ(ͻ)/ are long vowels. Separate symbols exist for all vowels, when they occur at the initial position of syllables. To indicate a vowel sound after a consonant other than the implicit one that is / ə /, a vowel sign (Matra) is attached to the consonant. Since the consonant has a built in schwa, there are equivalent matras for all vowels except the ਅ. So Punjabi has ten vowels but it has signs for only nine of them.

The correlation is shown as under:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ਅ | ਆ | ਇ | ਈ | ਉ | ਊ | ਏ | ਐ | ਓ | ਔ |
| Mukta  (Akhars without any vowel sign) | ਾ | ਿ | ੀ | ੁ | ੂ | ੇ | ੈ | ੋ | ੌ |

Table 6: Vowels with corresponding Matras

### Supra-Segmental signs; bindi, tippi and addak

Gurmukhi script has three supra-segmental signs; bindi, tippi and addak, which are used with Gurmukhi vowel signs to represent the supra-segmental phonemes nasality, gemination and stress. These supra-segmental signs are called lagakhars in Punjabi. Every vowel in Punjabi has a nasalized version. Bindi and Tippi are allographic variants of the nasal. So in Gurmukhi, both bindi and tippi signs are used to nasalize the vowels. Addak is used to represent gemination and stress. The description of the usage of these signs is as below

#### The Bindi (ਂ-U+0A02)

The bindi (ਂ) represents a homo-organic nasal. Bindi is used with all long vowels/ਆ, ਈ, ਊ, ਏ, ਐ, ਓ, ਔ/ and the short vowel ਉ like in words - ਆਂਚਲ, ਜਾਈਂ, ਏਂਜਲ, ਐਂਗਲ, ਓਂਕਾਰ ,ਔਂਕੜ, ਉਂਗਲ, ਊਂਘ and with the matras of long vowels/ ਾ, ੀ, ੇ, ੈ, ੋ, ੌ / expect the matra ( ੂ) like in words – ਹਾਂ, ਟੀਂ, ਪੇਂਟ, ਦੈਂਤ, ਤੋਂ, ਜੌਂ .

#### The Tippi (ੰ -U+0A70)

Tippi (**ੰ**)is used to nasalize short vowels /ə/ and /I/ at all places and /U and u*/* after a consonant. So tippi comes with the matras of /ə/ and /I/ i.e. mukta (without any vowel sign) and ਿ with vowel carriers as ਅੰ and ਇੰ like in words ਅੰਗ and ਇੰਡੀਆ and with consonants as ਸੰ and ਸਿੰ like in words ਸੰਦ and ਸਿੰਘ. Matras of /U and u/i.e. (ੁ, ੂ ) after a consonant can be followed by tippi like in words- ਖੁੰਬ, ਗੂੰਦ. In addition to this, tippi is also used in gemination for nasal consonants ਙ, ਞ, ਨ and ਮ. The rules for placement of bindi and tippi are as under:

1. The initial forms of *u, uu* vowels i.e. ਉ and ਊ can be followed by bindi as ਉਂ, ਊਂ. In addition to this the forms of *u, uu* vowels after any other vowel i.e. ਉ and ਊ can be followed by bindi like in words ਆਉਂਦਾ and ਜਾਊਂ.
2. Matras of *U, u* (ੁ, ੂ ) after a consonant can be followed by tippi - ਖੁੰਬ, ਗੂੰਦ;
3. All other short vowels / matras (mukta, ਿ) can be followed by tippi like in words - ਅੰਗ, ਇੰਡੀਆ, ਸੰਦ, ਚਿੰਤਾ;
4. All other long vowels/mātrās (ਆ, ਈ, ਏ, ਐ, ਓ, ਔ/ ਾ , ੀ, ੇ, ੈ, ੋ, ੌ ) can be followed by bindi like in words – ਆਂਦਰ, ਸਾਈਂ, ਜਾਏਂ, ਐਂਠ, ਸਿਓਂਕ, ਔਂਤਰਾ/ਹਾਂ, ਟੀਂ, ਪੇਂਟ, ਦੈਂਤ, ਤੋਂ, ਜੌਂ.
   * + 1. The Addak (ੱ -U+0A71)

Addak is used to mark the gemination of the following consonant. In Punjabi, addak usually comes with mukta, aunkar (ੁ) and sihari (ਿ), the vowel signs of /ə, u and i/ short vowels and geminates the consonant which follows it. Actually gemination of consonants occurs only when their preceding vowels are short vowels. For example in ਟੱਪਾ, ਗਿੱਲਾ and ਮੁੱਕਾ, the geminated /ਪ/, /ਲ/ and /ਕ/ consonants have /ə, I and U/ short vowels as their preceding vowels which are represented by mukta(without any vowel sign), sihari (ਿ) and aunkar (ੁ)vowel signs. In addition to this, addak is also used to write English source words having English vowel /æ/. For example, set, jet and web are written in Gurmukhi as ਸੈੱਟ, ਜੈੱਟ and ਵੈੱਬ.

**We now look at some of the exceptions.**

Addak does not precede ਹ, ਙ, ਞ, ਣ, ੜ, ਖ਼, ਗ਼ and ਲ਼ letters. In these letters, ਙ and ਞ are nasal consonants so these are stressed or doubled by the nasal sign tippi. The rest of these letters cannot be pronounced with stress or elongation. So addak is not used before any of the above mentioned letters. Addak is also not used with the last letter of the word, as there is no letter after it which has to be geminated. Addak is used with geminated consonants and the sign is placed on the preceding syllable. Addak cannot be used at the beginning of a word.

#### Nukta (਼ - U+0A3C)

Termed as *pairin bindi* in Punjabi, nukta is used with the following consonants: ਸ, ਖ, ਗ, ਜ, ਫ and ਲ to represent the phonemes of words of Sanskrit and Perso-Arabic sources. ਸ਼ is used to represent the phoneme of Sanskrit source words. ਲ਼ is used to represent Punjabi’s retroflexal /ḷ/ phoneme and ਖ਼ ਗ਼ ਜ਼ ਫ਼ are used to represent Perso-Arabic sources words.

When pairin bindi is adjoined to ਸ, ਖ, ਗ, ਜ, ਫ and ਲ letters, these are written as below

ਸ਼(U+0A38+U+0A3C**),** ਖ਼(U+0A16+U+0A3C), ਗ਼(U+0A17+U+0A3C), ਜ਼(U+0A1C+U+0A3C), ਫ਼ (U+0A2B+ U+0A3C), ਲ਼ (U+0A32+ U+0A3C)

These letters are called pairin bindi letters. All the letters are combinations of Consonant+Nukta. But in Gurmukhi, these letters can also be written as a single unit like ਸ਼ (U+0A36) ਖ਼ (U+0A59) ਗ਼ (U+0A5A), ਜ਼ (U+0A5B), ਫ਼ (U+0A5E) and ਲ਼ (U+0A33). Thus

ਸ਼ (U+0A36)= ਸ਼(U+0A38+U+0A3C**)**

ਖ਼ (U+0A59)= ਖ਼(U+0A16+U+0A3C)

ਗ਼ (U+0A5A)= ਗ਼(U+0A17+U+0A3C)

ਜ਼ (U+0A5B)= ਜ਼(U+0A1C+U+0A3C)

ਫ਼ (U+0A5E)= ਫ਼(U+0A2B+ U+0A3C)

ਲ਼ (U+0A33)= ਲ਼(U+0A32+ U+0A3C)

#### Visarga (ਃ U+0A03)

The visarga is used in Sanskrit. It is rarely found in old Punjabi writings as “Sri Guru Granth Sahib” or “Mahan Kosh” where it act like a Sanskrit visarga where a voiceless 'h' sound is pronounced after the vowel. But its use is not common now.

# Overall Development Process and Methodology

Under the Neo-Brahmi Generation Panel, there are many different scripts belonging to separate Unicode blocks. Each of these scripts will be assigned a separate LGR; however Neo-Brahmi GP will ensure that the fundamental philosophy behind building those LGRs are all in sync with all other Brahmi derived scripts. This is the Gurmukhi LGR, which caters to Punjabi language written using Gurmukhi script.

## Guiding Principles

### Acknowledgement to Environment Limitations:

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. 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.

*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 introduces exclusion of some characters out of Unicode repertoire from being part of the domain names.

Example: Gurmukhi letters ਸ਼ (U+0A36), ਖ਼ (U+0A59), ਗ਼ (U+0A5A), ਜ਼ (U+0A5B), ਫ਼ (U+0A5E), ਲ਼ (U+0A33) are not allowed to be a part of domain name. Their decomposed forms, i.e. Gurmukhi letters ਸ (U+0A38), ਖ (U+0A16), ਗ (U+0A17), ਜ (U+0A1C), ਫ (U+0A2B), ਲ (U+0A32) followed by Gurmukhi Sign Nukta (pairin bindi) “਼” (U+0A3C) can be used instead.

*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 even more specialized case of domain names, the ROOT LGR procedure introduces additional exclusions on IDNA allowed set of characters.

To sum up, the restrictions start off 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.

### No Punctuation Marks:

The TLDs being identifiers, punctuation marks present in Brahmi based languages such as Dandi “।” and double Dandi "॥" will not be included.

### No Symbols and Abbreviations:

Gurmukhi sign Addak Bindi ਁ (U+ 0A01) will not be included as it is not used in modern Punjabi.

### No Rare and Obsolete Characters:

There are characters which have been added to Unicode to accommodate forms of Medieval writings such as of “Sri Guru Granth Sahib” like Gurmukhi sign Yakash “ ੵ” (U+ 0A75), and Visarga ਃ (U+ 0A03). All such characters will not be included. This is in consonance with the letter principle as laid down in the Root Zone LGR procedure.

### No Stress Markers of Medieval Punjabi:

Medieval Punjabi stress markers or tone marker sign e.g. Gurmukhi sign Uddat “ ੑ”, (U+ 0A51)will not be included. This is also in consonance with the Letter principle as laid down in the Root Zone LGR procedure.

### No Vowel Carriers which cannot come without a matra:

The vowel carriers Ura, ੳ (U+0A73) and Iri, ੲ (U+0A72), cannot occur without a matra in a word. They are used as vowel carriers and thus always need to be followed by some matra. In case they occur separately with a matra then they will create conflict with one of the vowel and this is not allowed in Unicode too. Thus ੳ (U+0A73) + ੁ (U+0A41), which looks same as ਉ (U+ 0A09), will create confusion and hence will not allowed. These characters can occur as single character words, but in TLD, single character labels are not allowed, so these letters will not be added.

# Repertoire

## Code Points

| **Sr. No.** | **Unicode Code Point** | **Glyph** | **Character Name** | **Unicode General Category (gc)** | **Indic Syllabic Category** | **Reference** |
| --- | --- | --- | --- | --- | --- | --- |
| 1. | 0A05 | ਅ | GURMUKHI LETTER A = Aira | Lo | Vowel/Vowel Carrier | [INSCRIPT] |
| 2. | 0A06 | ਆ | GURMUKHI LETTER AA | Lo | Vowel | [INSCRIPT] |
| 3. | 0A07 | ਇ | GURMUKHI LETTER I | Lo | Vowel | [INSCRIPT] |
| 4. | 0A08 | ਈ | GURMUKHI LETTER II | Lo | Vowel | [INSCRIPT] |
| 5. | 0A09 | ਉ | GURMUKHI LETTER U | Lo | Vowel | [INSCRIPT] |
| 6. | 0A0A | ਊ | GURMUKHI LETTER UU | Lo | Vowel | [INSCRIPT] |
| 7. | 0A0F | ਏ | GURMUKHI LETTER EE | Lo | Vowel | [INSCRIPT] |
| 8. | 0A10 | ਐ | GURMUKHI LETTER AI | Lo | Vowel | [INSCRIPT] |
| 9. | 0A13 | ਓ | GURMUKHI LETTER OO | Lo | Vowel | [INSCRIPT] |
| 10. | 0A14 | ਔ | GURMUKHI LETTER AU | Lo | Vowel | [INSCRIPT] |
| 11. | 0A38 | ਸ | GURMUKHI LETTER SA | Lo | Consonant | [INSCRIPT] |
| 12. | 0A39 | ਹ | GURMUKHI LETTER HA | Lo | Consonant | [INSCRIPT] |
| 13. | 0A15 | ਕ | GURMUKHI LETTER KA | Lo | Consonant | [INSCRIPT] |
| 14. | 0A16 | ਖ | GURMUKHI LETTER KHA | Lo | Consonant | [INSCRIPT] |
| 15. | 0A17 | ਗ | GURMUKHI LETTER GA | Lo | Consonant | [INSCRIPT] |
| 16. | 0A18 | ਘ | GURMUKHI LETTER GHA | Lo | Consonant | [INSCRIPT] |
| 17. | 0A19 | ਙ | GURMUKHI LETTER NGA | Lo | Consonant | [INSCRIPT] |
| 18. | 0A1A | ਚ | GURMUKHI LETTER CA | Lo | Consonant | [INSCRIPT] |
| 19. | 0A1B | ਛ | GURMUKHI LETTER CHA | Lo | Consonant | [INSCRIPT] |
| 20. | 0A1C | ਜ | GURMUKHI LETTER JA | Lo | Consonant | [INSCRIPT] |
| 21. | 0A1D | ਝ | GURMUKHI LETTER JHA | Lo | Consonant | [INSCRIPT] |
| 22. | 0A1E | ਞ | GURMUKHI LETTER NYA | Lo | Consonant | [INSCRIPT] |
| 23. | 0A1F | ਟ | GURMUKHI LETTER TTA | Lo | Consonant | [INSCRIPT] |
| 24. | 0A20 | ਠ | GURMUKHI LETTER TTHA | Lo | Consonant | [INSCRIPT] |
| 25. | 0A21 | ਡ | GURMUKHI LETTER DDA | Lo | Consonant | [INSCRIPT] |
| 26. | 0A22 | ਢ | GURMUKHI LETTER DDHA | Lo | Consonant | [INSCRIPT] |
| 27. | 0A23 | ਣ | GURMUKHI LETTER NNA | Lo | Consonant | [INSCRIPT] |
| 28. | 0A24 | ਤ | GURMUKHI LETTER TA | Lo | Consonant | [INSCRIPT] |
| 29. | 0A25 | ਥ | GURMUKHI LETTER THA | Lo | Consonant | [INSCRIPT] |
| 30. | 0A26 | ਦ | GURMUKHI LETTER DA | Lo | Consonant | [INSCRIPT] |
| 31. | 0A27 | ਧ | GURMUKHI LETTER DHA | Lo | Consonant | [INSCRIPT] |
| 32. | 0A28 | ਨ | GURMUKHI LETTER NA | Lo | Consonant | [INSCRIPT] |
| 33. | 0A2A | ਪ | GURMUKHI LETTER PA | Lo | Consonant | [INSCRIPT] |
| 34. | 0A2B | ਫ | GURMUKHI LETTER PHA | Lo | Consonant | [INSCRIPT] |
| 35. | 0A2C | ਬ | GURMUKHI LETTER BA | Lo | Consonant | [INSCRIPT] |
| 36. | 0A2D | ਭ | GURMUKHI LETTER BHA | Lo | Consonant | [INSCRIPT] |
| 37. | 0A2E | ਮ | GURMUKHI LETTER MA | Lo | Consonant | [INSCRIPT] |
| 38. | 0A2F | ਯ | GURMUKHI LETTER YA | Lo | Consonant | [INSCRIPT] |
| 39. | 0A30 | ਰ | GURMUKHI LETTER RA | Lo | Consonant | [INSCRIPT] |
| 40. | 0A32 | ਲ | GURMUKHI LETTER LA | Lo | Consonant | [INSCRIPT] |
| 41. | 0A35 | ਵ | GURMUKHI LETTER VA | Lo | Consonant | [INSCRIPT] |
| 42. | 0A5C | ੜ | GURMUKHI LETTER RRA | Lo | Consonant | [INSCRIPT] |
| 43. | 0A3E | ਾ | GURMUKHI VOWEL SIGN KANNA | Mc | Matra | [INSCRIPT] |
| 44. | 0A3F | ਿ | GURMUKHI VOWEL SIGN SIHARI | Mc | Nukta | [INSCRIPT] |
| 45. | 0A40 | ੀ | GURMUKHI VOWEL SIGN BIHARI | Mc | Matra | [INSCRIPT] |
| 46. | 0A41 | ੁ | GURMUKHI VOWEL SIGN AUNKAR | Mn | Matra | [INSCRIPT] |
| 47. | 0A42 | ੂ | GURMUKHI VOWEL SIGN DULAINKAR | Mn | Matra | [INSCRIPT] |
| 48. | 0A47 | ੇ | GURMUKHI VOWEL SIGN LANVAN | Mn | Matra | [INSCRIPT] |
| 49. | 0A48 | ੈ | GURMUKHI VOWEL SIGN DULANVAN | Mn | Matra | [INSCRIPT] |
| 50. | 0A4B | ੋ | GURMUKHI VOWEL SIGN HORA | Mn | Matra | [INSCRIPT] |
| 51. | 0A4C | ੌ | GURMUKHI VOWEL SIGN KANAURA | Mn | Matra | [INSCRIPT] |
| 52. | 0A02 | ਂ | GURMUKHI SIGN BINDI | Mn | Matra | [INSCRIPT] |
| 53. | 0A70 | ੰ | GURMUKHI SIGN TIPPI | Mn | Matra | [INSCRIPT] |
| 54. | 0A71 | ੱ | GURMUKHI SIGN ADDAK | Mn | Matra | [INSCRIPT] |
| 55. | 0A4D | ੍ | GURMUKHI SIGN VIRAMA | Mc | Virama/ Halant | [INSCRIPT] |
| 56. | 0A3C | ਼ | GURMUKHI SIGN NUKTA  = PAIRIN BINDI | Mn | Nukta | [INSCRIPT] |

Table 7: Code point repertoire

## Repertoire Excluded

Code Points in MSR-2 but excluded because they are either not in common use or used for special purpose only (Used as vowel carrier).

| **Sr. No.** | **Unicode Code Point** | **Glyph** | **Character Name** | **Note** |
| --- | --- | --- | --- | --- |
| 1. | 0A03 | ਃ | GURMUKHI SIGN VISARGA | Limited or declining use |
| 2. | 0A51 | $ੑ | GURMUKHI SIGN UDAAT | Limited or declining use |
| 3. | 0A72 | ੲ | GURMUKHI IRI | Does not occur alone |
| 4. | 0A73 | ੳ | GURMUKHI URA | Does not occur alone |
| 5. | 0A75 | $ੵ | GURMUKHI SIGN YAKASH | Limited or declining use |

## Syllable formation rules for Gurmukhi:

The syllable is a basic unit of speech studied on both the phonetic and phonological levels of analysis. It is very easy for a native language speaker to count the number of syllables in a sequence however orthographic syllable required for text processing need not correspond exactly with a phonological syllable. This section details the syllable formation rules as applicable to Gurmukhi. The definition represents a vowel, consonant, or a conjunct.

**Variables involved:**

C → Consonant, which may or may not include a single nukta

M → Matra

V → Vowel

B → Bindi

D → Tippi

H → Halant / Virama

A → Addak

**Operators used:**

|  |  |
| --- | --- |
| Symbol | Function |
| | | Alternative |
| [ ] | Optional |
| {} | Zero or One occurrence |
| ( ) | Sequence Group |

The definition is a combination of 2 rules :

Rule 1 : V[A|B|D]

Rule 2 : {CH}C[M][A|B|D]

Rule 1 : V[A|B|D]

|  |  |  |
| --- | --- | --- |
| Sl. No. | Examples | Definition |
| 1 | ਅ, ਆ, ਇ | V (Vowel) is a syllable |
| 2 | ਇੰ, ਉਂ, ਇੱ | V+ (A/B/D) is a syllable |

Rule 2 : {C(H|A)}C[M][B|D]

|  |  |  |
| --- | --- | --- |
| Sl. No. | Examples | Definition |
| 1 | ਕ, ਙ, ਧ | Consonant is a syllable where it has inherent ‘ə’ vowel |
| 2 | ਸ੍ਵ, ਲੱਗ | Zero or one Consonant + Virama/Addak sequence followed by consonant is a syllable |
| 3 | ਸ੍ਵੈ | Zero or one Consonant+ Virama/Halant sequence followed by a consonant followed by a matra or vowel sign is a syllable |
| 4 | ਸ੍ਵੰ | Zero or one Consonant+ Virama sequence followed by a consonant followed by bindi/tippi is a syllable |
| 5 | ਗ੍ਰਾਂ | Zero or one Consonant+ Virama sequence followed by a consonant followed by a matra and bindi/tippi is a syllable |

Examples of combination of the rules :

1. *ਕਰੰਸੀ* - C + CD + CM has following syllables :

|  |  |
| --- | --- |
| *ਕ* | C |
| *ਰੰ* | CD |
| *ਸੀ* | CM |

2. *ਪਰਿੰਦਾ* - CV + CMD + CM has following syllables :

|  |  |
| --- | --- |
| *ਪ* | CV |
| *ਰਿੰ* | CMD |
| *ਦਾ* | CV |

3. *ਅੰਦਰ* - VD + CV+ C

|  |  |
| --- | --- |
| ਅੰ | Vm |
| ਦਰ | CvC |

# Variants

There are no characters/character sequences in Gurmukhi which can be created by using the characters permitted as per the [MSR] and look exactly alike. However, Gurmukhi has ample cases of confusingly similar variants in both Gurmukhi and Devanagari scripts. We have categorized these confusingly similar variants in three groups.

**Group 1:** Visually similar Gurmukhi characters

**Group 2:** Visually similar Gurmukhi character combinations, due to presence dots and other characters

**Group 3:** Visually similar characters in other scripts such as Devanagari. Gurmukhi and Devanagari scripts are closely related to each other and there are many characters in both scripts, which can be confused with the characters in other script.

**Group 1:**

|  |  |
| --- | --- |
| ਟ (0A1F) | ਦ (0A26) |
| ਬ (0A2C) | ਥ (0A25) |
| ਚ (0A1A) | ਰ (0A30) |
| ਢ (0A22) | ਦ (0A26) |
| ਢ (0A22) | ਫ (0A2B) |
| ਤ (0A24) | ਭ (0A2D) |
| ੇ (0A47) | ੋ (0A4B) |

No cases belonging to Group 1 are proposed as variants, as there is another panel (String similarity assessment panel) entrusted to deal with such cases.

**Group 2:**

|  |  |
| --- | --- |
| Code Point Sequence | Code Point |
| ਖ਼ (0A16 + 0A3C) | ਖ (0A16) |
| ਗ਼ (0A17 + 0A3C) | ਗ (0A17) |
| ਫ਼ (0A2B + 0A3C) | ਫ (0A2B) |
| ਓਂ (0A13 + 0A02) | ਓ (0A13) |
| ਈਂ (0A08 + 0A02) | ਈ (0A08) |
| ਐਂ (0A10 + 0A02) | ਐ (0A10) |
| ਔਂ (0A14 + 0A02) | ਔ (0A14) |
| ਗੰ (0A17 + 0A70) | ਰੀ (0A30 + 0A40) |
| ਨੁ (0A28 + 0A41) | ਠ (0A20) |

No cases belonging to Group 2 are proposed as variants, as there is another panel (String similarity assessment panel) entrusted to deal with such cases.

**Group 3**

| # | Gurmukhi Code Points | Devanagari Code Points |
| --- | --- | --- |
| 1 | ਂ (0A02) | ं (0902) |
| 2 | ਙ (0A19) | इ (0907) |
| 3 | ਤ (0A24) | उ (0909) |
| 4 | ਗ (0A17) | ग (0917) |
| 5 | ਬ (0A2C) | घ (0918) |
| 6 | ਟ (0A1F) | ट (091F) |
| 7 | ਠ (0A20) | ठ (0920) |
| 8 | ਫ (0A2B) | ढ (0922) |
| 9 | ਜ (0A1C) | त (0924) |
| 10 | ਧ (0A27) | प (092A) |
| 11 | ਮ (0A2E) | भ (092D) |
| 12 | ਸ (0A38) | म (092E) |
| 13 | ਧ (0A27) | य (092F) |
| 14 | ਕ (0A15) | र (0930) |
| 15 | ਕ (0A15) | व (0935) |
| 16 | ਵ (0A35) | ह (0939) |
| 17 | ੂ (OA42) | ॗ (0957) |
| 18 | ੁ(OA41) | ॖ(0956) |
| 19 | ਿ (0A3F) | ि (093F) |
| 20 | ੀ (0A40) | ी (0940) |
| 21 | ੇ (0A47) | े (0947) |
| 22 | ੈ (0A48) | ै (0948) |
| 23 | ਇ (0A07) | प्टि (092A 094D 091F 093F) |
| 24 | ਈ (0A08) | प्टी (092A 094D 091F 0940) |
| 25 | ਏ (0A0F) | प्टे (092A 094D 091F 0947) |
| 26 | ਜ (0A1C) | त्त (0924 094D 0924) |
| 27 | ੱ (0A71) | ॅ(0945) |
| 28 | ੲ (0A72) | प्ट (092A 094D 091F) |

Group 3 is defined as variants

[TBD –The following temp table show the set of code points that need to be discussed by the NBGP]

| # | Gurmukhi Code Points | Devanagari Code Points | Current Views | NBGP decision |
| --- | --- | --- | --- | --- |
| 1 | ढ (0922) | ਢ (0A22) | (AJ) Red, (GL) Yellow | TBD |
| 2 | ड (0921) | ਛ (0A1B) | (AJ) Red, (GL) Green | TBD |
| 3 | ढ (0922) | ਛ (0A1B) | (AJ) Red, (GL) Green | TBD |
| 4 | ट (091F) | ਦ (0A26) | (AJ) Red, (GL) Yellow | TBD |

# Whole Label Evaluation Rules (WLE)

This section provides the whole label evaluation rules for text written in Gurmukhi 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 the Table 7: Code point repertoire. In addition, we have created a few more symbols related to matras and vowels for the explanation of the rules.

C → Consonant

M → Matra

V → Vowel

B → Bindi

D → Tippi

H → Halant / Virama

N → Nukta

M1 → { ਿ(U+0A3F), ੁ(U+0A41) } (Short matras)

M2 → M - M1 (Long matras)

V1 → { ਅ (U+0A05), ਇ (U+0A0F), ਉ (U+0A09)} (Short Vowels)

V2 → V - V1 (Long Vowel)

C1 → C – {ਹ(U+0A39),ਙ(U+0A19), ਞ(U+0A1E), ਣ(U+0A23), ੜ(U+0A5C)}

Below are the specific WLE rules:

## N: must be preceded only by a specific set of Cs

The specific Cs are:

1. ਖ (U+0A16)
2. ਗ (U+0A17)
3. ਸ (U+0A38)
4. ਜ (U+0A1C)
5. ਫ (U+0A2B)
6. ਲ (U+0A32)

## H: must be preceded by C or CN and followed by following Cs only

1. ਰ (U+0A30)
2. ਵ (U+0A35)
3. ਹ (U+0A39)

## M: must be preceded by C or CN

## B: must be preceded by specific V or M

The specific Vs are:

1. V2
2. ਉ (U+0A09)

The specific Ms are:

1. M2 – { ੂ (U+0A42)}

## D: must be preceded by, C, N or a specified set of V or M

The specific Vs are:

1. V1– { ਉ (U+0A09)}

The specific Ms are:

1. M1
2. { ੂ (U+0A42)}

## A: must be preceded by C, N or specific V or M and followed by C1

The specific Vs are:

1. V1
2. ਐ (U+0A10)

The specific Ms are:

1. M1
2. ੈ (U+0A48)

# Contributors

|  |  |
| --- | --- |
| Name | Address |
| Dr. Gurpreet Singh Lehal | Professor, Department of Computer Science, Punjabi University, Patiala |
| Dr. Harvinder Pal Kaur | Assistant Professor, Research Centre for Punjabi Language Technology, Punjabi University, Patiala |
| Dr. Boota Singh Brar | Professor, Punjabi University Regional Centre, Bathinda |
| Dr. Paramjit Singh Sidhu | Retd. Professor, School of Punjabi Studies, Guru Nanak Dev University, Amritsar |

# References

[MSR] Integration Panel, "Maximal Starting Repertoire — MSR-2 Overview and Rationale", 14 April 2015<https://www.icann.org/en/system/files/files/msr-2-overview-14apr15-en.pdf>

[NBGP] Neo-Brahmi Generation Panel

[gTLD] generic Top Level Domain

[1] <http://pt.learnpunjabi.org/av.aspx?l=9> (Accessed on 10th Nov. 2017)

[2] <http://pt.learnpunjabi.org/av.aspx?l=10> (Accessed on 10th Nov. 2017)

[3] <http://elearnpunjabi.com> (Accessed on 10th Nov. 2017)

[4] A start in Punjabi, <http://pt.learnpunjabi.org/assets/A%20START%20IN%20PUNJABI_web.pdf> (Accessed on 10th Nov. 2017)

[5] A reference Grammar of Punjabi, http://pt.learnpunjabi.org/assets/A%20Reference%20Grammar\_Final.pdf (Accessed on 10th Nov. 2017)

[6] Mangat Rai Bhardwaj, “Colloquial Panjabi: A Complete Language Course”

[7] Omniglot,<https://www.omniglot.com/writing/punjabi.htm> (Accessed on 10th Nov.. 2017)

[8] Unicode 10.0.0," South and Central Asia-I, Page 475-479 ",<http://www.unicode.org/versions/Unicode10.0.0/ch12.pdf> (Accessed on 13th Nov. 2017)

[9] Alberuni, 2000 (1910) Alhind, (Ed. Kyamu Din Ahmad and Trn. Gurcharn Singh Arshi), New Delhi : National Book Trust, India.

[10] Bedi, Tarlochan Singh, 1999, Gurmukhi Lipi da Janam te Vikas, Patiala: Punjabi University.

[11] Bhardwaj, Mangat Rai, 2016, Panjabi: A Comprehensive Grammar, Routledge.

[12] Bhatia, T. K., 1993, Panjabi: A Cognitive-Descriptive Grammar, Routledge.

[13] Brar, Boota Singh, 2016, Panjabi Viakarn:Sidhant ate Vihar, Ludhiana:Chetna Parkashan.

[14] Dringer, David, 1948, The Alphabet : A Key to the History of Mankind, London: Hutchinson Scientific and Technical Publication.

[15] Gill, H.S. and H.A. Gleason(Jr.), 1969, A Reference Grammar of Panjabi, Patiala: Punjabi University.

[16] Manmohan, 2009, Gurmukhi di Jugat, Ludhiana: Chetna Parkashan.

[17] Newton, E. P., 1961, Panjabi Grammar, Patiala: Language Department Punjab.

[18] Ojha, Gauri Shankar Hira Chand, 1962, Bharti Prachin Lipi Mala (Ed. Jagdish Chand & Others), Patiala : Bhasha Vibhag Punjab.

[19] G.B. Singh, 1950, Gurmukhi Lipi Da Janam te Vikas, Chandigarh: Punjab University.

[20] Pritam Singh, 1958, Gurmukhi Lipi Di Utpati te Vikas, Ludhiana: Lahor Book Shop.

# 