

Scripture 101

Gurmukhi and Unicode

🕒 5 min read · Updated 16 October 2026 · Verified against [6332e94](#) on 2026-09-25

The script of the Granth and its signs in Unicode: vowel signs, the sihari that prints before its consonant, the subjoined ya, the nasal signs, the dandas and ੲ— what the pipeline decodes.

<https://docs.gurbanisoul.com/scripture/gurmukhi-and-unicode/>

Explanation, not scripture — scholar review pending. These pages explain the scripture to engineers and students. Every quoted line is verbatim from the pinned database and cited by Ang; everything else is explanation, written by engineers and awaiting a Granthi's review. Where this page and the scripture differ, the scripture is right.

Every stored line is **Gurmukhi** — the script the Gurus standardised for Punjabi — encoded in the Unicode block U+0A00–U+0A7F. You do not have to read it to work here, but you do have to know how its signs behave in Unicode, because that is where the pipeline's only transforms live and where a careless string operation would alter scripture.

Consonants, vowel signs, and where they sit

A Gurmukhi syllable is a consonant with an optional **vowel sign** attached (

ਾ ਿ ਿ ਿ ੁ ੂ ੃ ੄ ੅ ੆ ੇ ੈ ੉ ੊ ੋ ੌ ੍

) and optional nasal or other marks. Logical Unicode order is *consonant, then signs*.

One sign breaks the visual rule: the **Sihari**

ਿ

(short i) is *printed before* the consonant it belongs to. A PDF extracts the printed order, so the pipeline moves each sihari after its consonant cluster (

ਿਕ੍ਪਾ

→

ਕ੍ਪਾ

) — a decoding step, registered as rule D5 in the **editorial ledger**. The skeleton search column strips all vowel signs, which is why a typo-tolerant search still finds a line.

Subjoined letters and the ya

The **Halant**

𑂔

joins a consonant to a subjoined one —

𑂔

,

𑂔

,

𑂔

and

𑂔

most often. The source font emitted several of these as private-use glyphs and, for the **subjoined ya** in Sahaskriti words, doubled the sequence (

𑂔𑂔

): 234 lines in the corpus (268 words on 78 Angs) carry a subjoined ya, always as that doubled pair. The printed Bir draws one ya tucked under the letter — Unicode’s yakash sign

𑂔

(U+0A75), which Sant Lipi draws the same way. The *traditional saroop* on the website and in the app draws each doubled pair as

𑂔

in rendered glyphs only (Ang 1400’s

ਜਲ੍ਹਨ

,

ਧਰ੍ਹਉ

), and anything you copy is mapped back to the stored characters. The stored text, the API, search and the app’s database are untouched — the **Saroop** is a display layer. The corpus contains no

𑂔

of its own, which is what makes that mapping exact; a test fails if a dataset ever changes either fact.

Nasal signs, nukta and editions

Tippi and bindi —

ੰ

and

ੰ

— both mark nasalisation; 17,755 lines carry a tippi and 1,817 a bindi in this edition. Editions differ in which they print for some words. The project keeps its edition’s choice

and, in the Answer Protocol, *notes* an edition difference rather than correcting it. The Nukta

marks borrowed sounds and is kept wherever printed.

Dandas, numerals and



The **double danda** ॥ closes a line or unit and carries the closing numerals (॥

॥
, ॥
२॥
); 60,176 of the 60,658 lines end with one, and the pipeline cuts the page stream into units on it. The numerals are Gurmukhi digits (0
—
ੳ
) ; the markers column keeps them as a list. ੳ — Ik Onkar, U+0A74 — opens compositions and the Granth itself: 568 lines carry it, and /api/health refuses a database with fewer than 560. The source font printed it as two glyphs; restoring the single character is decoding rule D3.

What the pipeline may do, and what it may not

Allowed (registered decoding)	Never
move a sihari after its consonant cluster	“correct” a spelling, however obvious
restore a font glyph to the sign it printed	normalise editions to each other
restore ੳ from two glyphs to one	change a tippi to a bindi or back
re-attach a vowel sign split off by a stray space	delete or reorder a word
strip print-job timestamps and stray Latin from the PDF stream	add or remove a danda or a numeral
apply one of the four reviewed editorial rules	apply an unregistered transform of any kind

Everything in the left column is an entry in the ledger; ledger_check.py fails a pull request in sggs-data whose code and register disagree. Everything in the right column is a fidelity concern for a scholar, never an edit.

For your code

- Treat every gurmukhi value as opaque: never lower-case it, never strip it, never normalise beyond NFC for *comparison* (the verifier does that on its own copy).
- Escape it before it reaches innerHTML — the front-end's central escape helper does; use it.
- Render it with the bundled Sant Lipi font and lang="pa"; the site's theme keeps Gurmukhi ink-coloured, never the accent colour.
- Search keys — text, translit, translit_norm, fl_g, skeleton — are *derived* columns. Query them; never rebuild them from the text at request time.

Check yourself

The sihari prints before its consonant. What does the pipeline store?

- ☐ The consonant first, then the sihari
- ☐ The printed order, sihari first
- ☐ Neither: it drops the sihari

The website's traditional-saroop toggle collapses the doubled subjoined ya. Where?

- ☐ In rendered glyphs only
- ☐ In the database
- ☐ In the API response

This edition prints a tippi where another prints a bindi. What does the project do?

- ☐ Keep its edition's sign and note the difference where known
- ☐ Change it to the more common form
- ☐ Store both

Read next: [Transliteration and the fold](#).