Unicode Technical Note
ISCII to Unicode Conversion Issues for Gurmukhi

This document discusses issues that may be encountered when converting from ISCII to Unicode specifically for the Gurmukhi script (merely labelled Punjabi in ISCII-91).

Although it is possible to convert ISCII to Unicode and back again without loss of information, this will not result in readable Unicode Gurmukhi text. There are several changes to Gurmukhi in Unicode that require a change in the encoding and thus cannot be byte per byte round-tripped.

If the advice in this document is not heeded, any resulting conversion will not be legible to readers of the Gurmukhi script.

Bindi and Tippi

Bindi and Tippi are encoded using a single code point in ISCII (0xA2) and the underlying rendering engine selects the correct glyph. However, in Unicode they are given two separate code points.

Thus, 0xA2 should be converted to U+0A70 (Tippi) when:

- The preceding letter is a consonant (ignoring any Nuktas)
- The preceding letter is Vowel Sign I (◌ਿ - U+0A3F), Vowel Sign U (◌ੁ - U+0A41), Vowel Sign UU (◌ੂ - U+0A42),
- The preceding letter is Letter A (ਅ - U+0A05), Letter I (ਇ - U+0A07),

In all other cases, the sign should remain a Bindi (U+0A02).

When converting from Unicode to ISCII, both Bindi and Tippi should be converted to Bindi (0xA2).

Consonant Clusters

In general, consonant clusters are handled the same in Unicode and ISCII. However, Unicode differs in that it encodes geminate consonants using a separate Adhak sign. Thus, if the ISCII sequence:

C + Halant (0xE8) + C

Where both Cs are the same consonant, is converted to Unicode, it should be encoded:

Adhak (U+0A71) + C

For example:

ਰ + ◌੍ + ਰ → ◌ੱ + ਰ

In theory this same logic would apply if the first consonant was the unaspirated form of the second consonant, but this is not stipulated by the ISCII standard.

When converting from Unicode to ISCII, an Adhak followed by a consonant should be converted to the consonant, followed by Halant (0xE8), followed by the consonant again.

Gurmukhi Rra

Gurmukhi Rra (ੜ – U+0A5C) is treated as a Nukta character in ISCII but as a full character in Unicode. Thus the following conversions must be made:

ਡ਼ (0xBF + 0xE9) → ੜ (U+0A5C)
ਢ਼ (0xC0 + 0xE9) → ੜ੍ਹ (U+0A5C + U+0A4D + U+0A39)

In the latter of the two conversions, we see that in ISCII an aspirated version of ੜ is created using a Nukta on ḍha (as with Devanagari) but in Unicode it is represented with ੜ and a subjoined ਹ.