

# Khojki Confusable Sequences

Peter Constable  
March 30, 2022

## Summary

Khojki script was accepted by UTC in 2010 ([125-C1](#)) and added to Unicode 7.0. Three additional characters were accepted in 2021 ([168-C25](#), [168-C26](#), [168-C27](#)) for addition to Unicode 15.0, and are currently in alpha review ([PRI #442](#)). Eduardo Marin Silva has submitted feedback ([L2/22-056](#)) in which he indicates that one of the new characters introduces confusability involving existing characters, and that this should be called out by adding an informative alias and an annotation to one of those characters, 11202.

Silva has proposed a change pertaining to one case of script-internal confusability, but it is one of several such cases in Khojki script. Calling out one instance of confusability while ignoring other similar cases in the same script would not be good as it suggests that the other cases present no significant risks, which is not the case. If a remedy is to be made for Khojki confusable sequences, a more comprehensive remedy should be made.

## Background Details

Khojki is a Brahmi-derived script with typical characteristics of that family of scripts. In particular, it has independent vowel letters and dependent vowel signs. As happens in many other Indic scripts, some of the independent vowel letters have components that are visually similar or identical to other letter or sign characters, and are confusable with combinations thereof.

For example, KHOJKI LETTER AA is confusable with a combination of KHOJKI LETTER A plus KHOJKI VOWEL SIGN AA:



Such cases in Indic scripts are not unlike cases of Latin pre-composed base/diacritic characters that are visually identical to base + combining mark sequences. Those Latin instances of confusability are mitigated by canonical decomposition mappings and normalization. When Indic scripts were first incorporated into Unicode, however, normalization was not applied to those scripts, given the complex nature of satellite marks in Indic scripts.

When Khojki was encoded in 2010, the precedents established for Indic scripts were maintained, and characters such as LETTER AA were encoded without any decomposition.






The Khojki proposal from 2010 ([L2/10-236](#)) had a section discussion confusability, but only called out cross-script confusable characters. UTC was not systematically addressing cases of script-internal

confusable sequences at that time, though it had begun addressing it in some cases. In particular, in Unicode 5.0 (2006), a table was added in [section 9.1](#) showing Devanagari vowel letters that are confusable with letter-mark sequences, and indicating that the sequences “should not be used”.

In more recent versions, other “Do Not Use” tables have been added to the block descriptions of various Indic scripts. In Unicode 14.0, chapter 12 has twelve “Do Not Use” tables; chapters 13 and 14 each have one, and chapter 15 has four. Such “Do Not Use” tables are an imperfect solution: the guidance not to use sequences is only a recommendation, and in practice sequences do get used in user data. Recently, the Script Ad Hoc has begun to apply a different approach for new Indic scripts, incorporating canonical decomposition mappings for such cases when feasible. But canonical decompositions cannot be added to already-encoded characters. Thus, for other encoded Indic scripts, including Khojki, inclusion of a “Do Not Use” table in the block description is the best mitigation that Unicode currently can offer.

## Recommendation

To mitigate cases of script-internal confusable sequences related to vowel letters in Khojki script, the following “Do Not Use” table should be incorporated into the Khojki block description<sup>1</sup> in a future version of the Standard.

For	Use	Do Not Use
	11201	11200 + 1122C
	11202	11240 + 1122C 11240 + 1122E
	11203	11206 + 1122C
	11205	11200 + 11231
	11207	11200 + 11233

These confusable character/sequence pairs could potentially also be added to the `confusables.txt` file of UTS #39 (e.g., similar “Do Not Use” pairs for Devanagari are included), although the `Identifier_Status` property in UTS #39 for Khojki characters is `Restricted`, and that file currently doesn’t include such sequences involving `Restricted` characters.

<sup>1</sup> [Section 15.7](#) in Unicode 14.0