Defining Khmer clusters

Norbert Lindenberg
2020-10-26

Summary

There is no agreement on the definition of a valid Khmer cluster between the Unicode Standard, the documentation published by Khmer experts, the OpenType documentation, the various OpenType implementations, and Khmer AAT fonts. A unified definition is needed, taking into consideration the requirements of all languages written in the script as well as variations in usage over its history.

Problem

At least four different descriptions of the structure of a valid Khmer cluster encoded in Unicode have been published:

- in the Khmer section of the Unicode Standard itself,
- in the documentation of the OpenType Khmer shaping engine,
- by the Open Forum of Cambodia,
- in documentation of SIL International's Mondulkiri font package.

These four descriptions do not agree. The disagreements become most visible when OpenType implementations or AAT fonts insert dotted circles into what they consider invalid clusters. However, they also prevent smart keyboards from transforming user input into a standardized character sequence, can lead to search algorithms not finding character sequences that users would expect to match the search string, to sorting algorithms producing unexpected orderings of strings, and to systems designed to allow or block certain terms to make incorrect decisions.

Some of the resulting problems have been investigated and documented:

- Joshua Horton, Makara Sok, Marc Durdin, and Rasmey Ty (2017) investigated spoofing of Khmer words. In particular on Android at the time they found that many spelling variations of the same word were treated as equally valid (i.e., the rendering system did not insert dotted circles), and the default font in Android also rendered them the same way. Other implementations allowed fewer but still too many variations.
- I (2019) investigated validation in a number of corner cases where the first three descriptions of valid clusters differed, and found numerous differences in whether implementations inserted dotted circles into strings or treated them as valid by not doing so.

Towards a solution

A few steps have been made towards correcting the situation:

- The validation of Khmer clusters in HarfBuzz, the shaping engine used in Android and many other platforms, has been tightened to reject a number of the invalid spelling variations that Horton et al. found. The change is visible in the columns for Firefox 64 and 66 in my investigation.
- The default font on Android, Noto Sans Khmer, has been updated to make differences between spelling variations more visible (see last paragraph in the description of the HarfBuzz bug for the above change).
- Makara Sok (2020) provided a very thorough description of character usage in the languages written in the Khmer script today. As any definition of valid clusters has to enable writing in all these languages, this is extremely valuable information. At the Unicode conference, Makara also proposed a new cluster definition based on his work.

However, more information is needed:

- The spelling of the Khmer language was significantly simplified during the 20th century. We need better documentation of earlier usage, so that the cluster definition allows the transcription of earlier documents as well. One potential starting point might be Trent Walker (2018), which contains numerous transcriptions of older documents.
- The Khmer script has historically also been used to write the Thai language, in particular in temples, a usage that's sometimes called the Khom script. In this usage, additional characters were used that are not encoded in Unicode yet. Élie Roux at the Buddhist Digital Resource Center has started work an a proposal.