L2/21-022

Status on the Update to the Unihan kCantonese Field

John H. Jenkins

14 January 2021

Summary

I've done an examination and comparison of Cantonese readings: the Unihan database for Unicode 13.0, list from the Hong Kong government's Office of the Government Chief Information Officer (OGCIO), and a list from the Linguistic Society of Hong Kong (LSHK). Between the three sources, there are a total of 30,168 characters with readings. Of these, 26,698 are sufficiently certain that they can be used. The remainder (3,470) need to be double-checked.

This current set covers all of HKSCS-2016 and may be considered a solid core for current-use Cantonese readings.

The goal of the process is to provide a set of data which is unlikely to require more than trivial corrections (as is the case with the `kMandarin` field now). Unlike `kMandarin`, however, it's anticipated that there will likely be block additions in the future as further research provides more readings.

The list of readings to accept and readings to confirm are attached as `kCantonese-Values.txt` and `kCantonese-Lookup.txt`, respectively.

Details

In L2/20-153, I recommended an overhaul to the `kCantonese` field in the Unihan database to improve its utility and accuracy. This was accepted by the UTC (see action items 164-A11 and 164-A12). We have since formed a group of Cantonese experts on Slack to discuss and work on this work.

The analysis involved separating the characters into classes based on the sources providing readings for them and on the number of readings available in the sources. In the case of the OGCIO and LSHK lists, multiple readings were eliminated, either by selecting the preferred reading (where unambiguous) or deletion (where ambiguous).

Characters with only one reading for any single source were accepted.

Characters from all three sources were accepted when in agreement.
Characters from the two of the three sources were accepted if the two sources agreed on a reading.

The remaining characters were flagged as needing confirmation.

The next step is to work with the Cantonese experts on Slack to confirm as many readings as possible for use in Unicode 14.0.