L2/20-153

Proposal to Repurpose the kCantonese Field in the Unihan Database

John H. Jenkins

24 June 2020

Summary

The data in the kCantonese field of the Unihan database currently has little practical use. It is recommended that it be repurposed as was the kMandarin field to provide a single reading targeted specifically for use by CLDR collation and transliteration.

Proposal

The kMandarin and kCantonese fields were originally added to the Unihan database to provide the same sort of data as one might expect from a dictionary: an exhaustive set of readings in the order of frequency or significance. Because of the practical difficulties in generating such data, the kCantonese field adopted an alphabetical order for its values and the kMandarin field was repurposed.

The kMandarin field now provides a single reading—occasionally two. The field contains most customary pinyin reading for a character. When there are two values, then the first is preferred for zh-Hans (CN) and the second is preferred for zh-Hant (TW). When there is only one value, it is appropriate for both. The kMandarin field is targeted specifically for use by CLDR collation and transliteration. As such, it is subject to considerations that help keep pinyin-based Han collation (and its tailorings) and transliteration reasonably stable. The values may not in all cases track the preferred use in some dictionaries.

We recommend that the kCantonese field be similarly repurposed. In this case, there is need for only one value, preferred for yue-Hant-HK generally. As with Mandarin, we can add fields tied to specific dictionaries to provide information on multiple readings.

Of the 23,111 characters with a kCantonese value, 4882 have multiple values. We recommend two sources to determine the preferred reading in these cases: http://www.iso10646hk.net/jp/document/jyut_yam_table.jsp (generated by the Hong Kong government), and the 商務新字典 (*Soeng^{1}mou^6 San^1 Zi^{6}din^2*), which is a common student's dictionary published in Hong Kong. (The latter was selected because I'm already involved in a project to extract Cantonese readings from it.) We note that the Hong Kong governments data is encumbered by a license which prevents our using it directly.

There are three advantages to the greater community of continuing to include the `kCantonese` in the Unihan database.

First, our license is less restrictive.

Second, the data is available in plain-text form rather than being embedded in a PDF.

Third, the data from the Unihan database can be incorporated more easily into the CLDR.

We should also follow the example set by `kMandarin` and change the field from Provisional to Informative. We should also institute a policy of having proposed changes vetted and formally approved by the UTC before being incorporated into the Unihan database.

I've done some tests on how long it takes to make the necessary changes to the Unihan database, and it can be completed well in time for Unicode 14.0.