Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type:	**Unicode Technical Committee Document**
Title:	**Proposal to update various readings values for CJK Ext. H in the Unihan Database**
Source:	**Eiso Chan (陈永聪, Culture and Art Publishing House)**
Status:	**Individual Contribution to UTC #171**
Action:	**For consideration by UTC**
Appendix:	**5**
Date:	**2022-04-02**

0. Background

IRG WS2017 has been accepted as CJK Ext. H, which will be included in Unicode Standard, 15.0.0 and ISO/IEC 10646:2020/Amd.1. Some pieces of the submitted evidence provided by the submitters show the different reading information. I think these are useful for us and the users, so I request to update some `kMandarin`, `kCantonese`, `kJapaneseKun`, `kHangul` and `kKorean` properties values based on the clear evidence on the IRG WS ORT. If any expert finds out anything wrong in this proposal, we can discuss and modify to the better ones.

I recorded almost all the reading values based on the clear enough evidence. If different evidence shows different readings, and we can't confirm easily, I will not include them in this document. For example, for WS2017-00788:GDM-00043;UK-10824 (⿰土最), the China-submitted evidence shows the Putonghua reading as zuì, and the UK-submitted evidence shows the one as zhuí, I can't confirm which one is better. If any expert shows more definitive evidence or the appropriate explanation, it will be OK to accept the readings.

1. Mandarin readings

There are 230 entries in AppA txt file.

Some UK-submitted characters are used for Sichuan or Chengdu dialects, the corresponding evidence shows the reading which looks like Chinese Pinyin, but it is the Romanization strings for the dialect in fact, so I don't include them in this part, such as WS2017-00264, 00764, 01395, 01401, 01499, 03868, 04045, 04139, 04224, 04496, 04710, 04744, 04751, 04853 and so on.

UK also submitted some characters used for Beijing dialect. As we know, the

Putonghua phonetic system is based on the Beijing one, but no all the Beijing dialect words are accepted in the modern Chinese and Putonghua. For example, WS2017-00559:UK-10427 (⿰口窜) reads as cuàn, which is not a Putonghua word, but it's common among the native speakers of Beijing dialect. In Beijing dialect, so many characters have literary reading and colloquial reading, but Putonghua doesn't include all the reading system. If the native speakers don't know one colloquial reading is related to which character, they will use a new character to record the word. When we need to confirm the Putonghua reading, we need to try to select the proper one. Therefore, I don't include them in this part as well.

2. Cantonese readings

There are 13 entries in AppB txt file.

I modify the Cantonese reading strings to match the Unihan conventions.

For WS2017-00185:GHC-0049.00;UTC-03155 (⿰先母), the submitted evidence don't show the Cantonese reading, but the evidence shows the real use in Guangdong Province. As what I wrote under Comment #3907, it's the variant of U+4E78 (𠜎), so we can confirm the Cantonese reading should be naa2.

3. Japanese Kunyomi readings

There are 2 entries in AppC txt file.

In this part, these two characters are all Japanese Kokuji, so there are only the Kunyomi readings.

4. Korean readings

There are 2 entries in AppD txt file.

This part is related to two properties, `kHangul` and `kKorean`. However, these two reading sources don't match the current description of `kHangul`. If we need to add these two reading values, I suggest adding one I tag for the IRG WS reading sources. So maybe the syntax should be updated as below.

```
[\x{1100}-\x{1112}][\x{1161}-\x{1175}][\x{11A8}-
\x{11C2}]?:[01E**I**NX]{1,3}
```

5. Candidates

AppE txt file is a candidates list to update Unihan database.

(End of Document)