L2/22-204R

Proposal to Add Two New Fields to the Unihan Database

John H. Jenkins

25 October 2022

Summary

The *Soengmou San Zidin*[1] is a popular mid-sized character-based dictionary published in Hong Kong. It has 11,968 main entries covering 10,700 distinct Unicode code points, providing multiple definitions and multiple readings in both Mandarin and Cantonese. (Some characters occur more than once because of unifiable variants and characters with multiple radicals.)

It is recommended that two new fields be added to the Unihan database: `kSMSZD2003Index`, and `kSMSZD2003Cantonese`. Descriptions of these two fields follows. Text files containing the data for these two proposed fields are attached to this document.

The field names use the initials of the romanization of the dictionary's title, following the precedent of `kSBGY`, `kTGHZ2013`, and `kXHC1983`.

The current `kCantonese` field provides only one reading. Earlier versions provided multiple readings ordered alphabetically. Adding these new fields will provide multiple Cantonese readings ordered by significance.

Question for the UTC

The experts in the Unihan ad hoc were unable to fully reach agreement on one issue:

The dictionary includes both Mandarin and Cantonese readings. Should the data be restricted to the Cantonese readings, or should the Mandarin readings be

---

[1] The Chinese name is 商務新字典 (*New Commercial Press Character Dictionary*), which in Mandarin would be *Shāngwù Xīn Zìdiǎn*. I use the Cantonese name over the Mandarin one because this is a dictionary published in Hong Kong and targeted at Cantonese speakers. Note that its name currently found in section 4.5 of UAX #38 is incorrect.

included? Concern was expressed that Mandarin readings would be redundant, given that the Unihan database already has numerous fields with Mandarin readings. On the other hand, including them provides correlation between the pronunciations in the two dialects. If we include Mandarin readings, the field would need to be renamed, and typical entries might look like:

```
U+4E95 kSMSZD2003Readings      jǐng粵zing2,zeng2
U+884C kSMSZD2003Readings      xíng粵hang4 xíng,xìng粵hang6 háng粵hong4 háng粵hang4
```

The current recommendation is not to include them.

kSMSZD2003Index

Property	kSMSZD2003Index
Status	Provisional
Category	Dictionary Indices
Introduced	TBD
Delimiter	space
Syntax	`\d{1,3}\.\d{2}`
Description	This represents the position(s) of the character in the *Soengmou San Zidin* (**商務新字典**, *New Commercial Press Character Dictionary)*. The format is the page within the dictionary followed by the position on the page. If multiple values are present, the first is the primary entry for the character. Other entries are simply cross-references to the primary entry and are in numeric order. The complete bibliographic information for the *Soengmou San Zidin* is: Wong Gongsang **黃港生**, ed. Shangwu Xin Zidian / Soengmou San Zidin **商務新字典** (New Commercial Press Character Dictionary). Hong Kong: **商務印書館(香港)有限公司** (Commercial Press [Hong Kong], Ltd.), 2003. ISBN 962-07-0140-2.

Examples:

```
U+4E95 kSMSZD2003Index  13.02
U+771E kSMSZD2003Index  460.12 72.05
U+9EBC kSMSZD2003Index  823.05 199.02 203.02
```

kSMSZD2003Cantonese

Property	kSMSZD2003Cantonese
Status	Provisional
Category	Readings
Introduced	TBD
Delimiter	space
Syntax	`([a–z]+[1–6])?[a–z]+[1–6]`
Description	This represents the Cantonese readings(s) of the character in the *Soengmou San Zidin* (商務新字典, *New Commercial Press Character Dictionary)*. The full bibliographic information for this dictionary is found in the description of the `kSMSZD2003Index` field. Readings are in *jyut6ping1*, the Cantonese romanization used elsewhere in the Unihan database. Note that some characters have readings which would ordinarily be considered invalid, such as polysyllabic readings. If a character has multiple entries, it means that the character has multiple definitions and the readings are grouped in order of those definitions.

Examples:

```
U+4E95 kSMSZD2003Cantonese    zing2 zeng2
U+884C kSMSZD2003Cantonese    hang4 hang6 hong4
```

Other Data

Other data are available, specifically radical-stroke data and variant data. The recommendation is not to use include these data in their own fields. Instead, these can be used to add information to other fields such as `kRSUnicode`, `kSemanticVariant`, and so on.