

Proposal to Add Two New Fields to the Unihan Database

John H. Jenkins

8 September 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: `kSoengmou2003Index`, and `kSoengmou2003Readings`. Descriptions of these two fields follows. Text files containing the data for these to proposed fields are attached to this document.

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.

kSoengmou2003Index

Property	kSoengmou2003Index
Status	Provisional
Category	Dictionary Indices
Introduced	TBD
Delimiter	space
Syntax	\d+\.\d{2}*?
Description	This represents the position(s) of the character in the <i>Soengmou San Zidin</i> (商務新字典, <i>New Commercial Press Character Dictionary</i>). The format is the page within the dictionary followed by the position on the page. An asterisk is used to mark characters printed in red instead of black, which indicates common use. 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 <i>Soengmou San Zidin</i> 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	kSoengmou2003Index	13.02*
U+771E	kSoengmou2003Index	460.12 72.05
U+9EBC	kSoengmou2003Index	823.05* 199.02 203.02

kSoengmou2003Readings

Property	kSoengmou2003Readings
Status	Provisional
Category	Readings
Introduced	TBD
Delimiter	space
Syntax	([a-z\x{300}-\x{302}\x{304}\x{308}\x{30C}]+,)*[a-z\x{300}-\x{302}\x{304}\x{308}\x{30C}]\x{7CB5}(([a-z\x{250}\x{25b}\x{254}\x{14b}\u{153}]+[\x{b9}\x{b2}\x{b3}\x{2074}-\x{2079}]),)*[a-z\x{250}\x{25b}\x{254}\x{14b}\u{153}]+[\x{b9}\x{b2}\x{b3}\x{2074}-\x{2079}]
Description	This represents the readings(s) of the character in the <i>Soengmou San Zidin</i> (商務新字典, <i>New Commercial Press Character Dictionary</i>). The full bibliographic information for this dictionary is found in the description of the kSoengmou2003Index field. The basic format is “Mandarin-reading(s)粵Cantonese-reading(s)”. The Mandarin readings are in <i>hànyǔ pīnyīn</i>. The Cantonese readings are in a custom romanization based on IPA. Individual readings are separated by commas. If a character has multiple entries, it means that the character has multiple definitions and the readings are grouped in order of those definitions.

Alternatively, it would be possible to normalize the Cantonese readings to *jyut6ping1*, which is what is used elsewhere in the UniHan database for Cantonese.

As an examples, we would have

U+4E95 kSoengmou2003Readings jǐng² 粵dzɿŋ², dzɿŋ²

U+884C kSoengmou2003Readings xíng⁴ 粵heŋ⁴ xíng, xìng⁶ 粵heŋ⁶ háng⁴ 粵hɔŋ⁴ háng⁴ 粵heŋ⁴

The Mandarin readings might be considered redundant given the presence of several other fields giving Mandarin readings.

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.