L2/22-111

Title: Unihan Numeric Fields Problems and Proposed Updates

Author: Wang Yifan, Eiso Chan

Date: 2022-06-05

This document accounts for (1) existing problems in Unihan `k*Numeric` data that we observed, with our suggested fixes for them; and (2) proposed new additions to those fields that we find desirable.

All English number names referred in this document are in the short scale, unless otherwise stated.

Please see the attached text file for the summary of proposed updates in the UAX #38-like format.

1. Problems

1.1. Incompleteness

- U+5169 兩 has `kOtherNumeric` value 2, but its Simplified counterpart U+4E24 两 does not. We suggest that U+4E24 should also have `kOtherNumeric` value 2.
- The `kOtherNumeric` repertoire contains many composite numbers such as U+5EFF 廿 (20), U+5345 卅 (30) etc. but lacks U+7695 皕, which means "two hundred" as included in *Hanyu Da Zidian* and *Shuowen Jiezi*. We suggest that U+7695 should have `kOtherNumeric` value 200.

1.2. Inconsistency

- U+4E07 万 has `kPrimaryNumeric` value 10,000 while its Traditional variant U+842C 萬 has `kAccountingNumeric` value 10,000. This seems inconsistent where both U+4EBF 亿 and U+5104 億 has a `kPrimaryNumeric` value. We suggest that U+842C should instead have `kPrimaryNumeric` value 10,000.

1.3. Non-uniqueness

- U+5146 兆 currently have `kPrimaryNumeric` value 1,000,000,000,000 (= 10^{12}, one trillion), but this number name traditionally had multiple definitions as well: 10^6 or 10^{16}, of which the 10^6 reading still survives in some PRC and Vietnamese conventions. We suggest that U+5146 should have a numeric value 1,000,000 in any form in the database.

2. Additions

2.1. Larger number names

With the progress of digital computing, more larger units are recently coming into practical use, especially in Japan. we can easily find examples such as: 京, popularized by the namesake supercomputer (K computer) describing its 10 petaflops performance; 澗, often employed in non-technical explanation of the magnitude of 2^{128}, principally in the context of IPv6 address space (an example). Therefore, we see the usefulness of adding following `kPrimaryNumeric` entries for each character, based on the modern 10^4 scale system.

Glyph	Code	Value	Notes
京	U+4EAC	10,000,000,000,000,000	10^{16}; ten quadrillion
垓	U+5793	100,000,000,000,000,000,000	10^{20}; one hundred quintillion
秭	U+79ED	1,000,000,000,000,000,000,000,000	10^{24}; one septillion
秆	U+25771		
穰	U+7A70	10,000,000,000,000,000,000,000,000,000	10^{28}; ten octillion
穣	U+7A63		
溝	U+6E9D	100,000,000,000,000,000,000,000,000,000,000	10^{32}; one hundred nonillion
沟	U+6C9F		
澗	U+6F97	1,000,000,000,000,000,000,000,000,000,000,000,000	10^{36}; one undecillion
涧	U+6DA7		

General notes:

- While they are expected to be mostly used in Japan, Simplified/Traditional variants are also supplied.
- Although U+25771 秆 is thought to be a misanalyzed form of U+79ED 秭, it is better included as a regional variant in view of the preference in Japan (coded in JIS X 0213).
- Similar to the issue in Section 1.3, U+79ED 秭 means "billion" (short or long scale) in Vietnamese, which may need a special treatment to the same effect.

2.2. Vietnamese numbers

Vietnamese native numerals can be written in chữ Nôm using CJK Ideographs. We suggest assigning Unihan numeric values to the following 42 characters in any form (shaded cells are upcoming Extension H characters).

Glyph	Code	Quốc ngữ	Value
空	U+7A7A	không	0
沒	U+6C92	một/mốt	1
没	U+6CA1		

Glyph	Code	Quốc ngữ	Value
蔑	U+8511		
艾	U+20B20		
歿	U+31357		

Glyph	Code	Quốc ngữ	Value
𱥺	U+3197A		
台	U+53F0	hai	2
𠄩	U+20129		
𱎔	U+31394		
𠀧	U+20027	ba	3
奔	U+5954	bốn	4
本	U+672C		
𦊚	U+2629A		
𠄼	U+2013C	năm	5
𫡵	U+2B875		
𱎖	U+31396		
𠄻	U+2013B	lăm	
𬂽	U+2C0BD		
𠄶	U+20136	nhăm	
𦒹	U+264B9	sáu	6
罢	U+7F62	bảy	7
罷	U+7F77		
𦉱	U+26271		

Glyph	Code	Quốc ngữ	Value
𬙞	U+2C65E		
𠔭	U+2052D	tám	8
𫤯	U+2B92F		
㐱	U+3431	chín	9
𠃩	U+200E9		
𢒂	U+22482		
𫡦	U+2B866		
迈	U+8FC8	mười/mươi	10
邁	U+9081		
𠦩	U+209A9		
𨑮	U+2846E		
𨒒	U+28492		
𱑕	U+31455		
啉	U+5549	trăm	100
𤾓	U+24F93		
彦	U+5F66	ngàn/nghìn	1,000
𠦳	U+209B3		
𡶨	U+21DA8		

2.3. Zhuang numbers

Numbers in Zhuang language(s) can be written in Zhuang characters (Sawndip). We suggest assigning Unihan numeric values to the following 13 characters from *Sawndip Sawdenj* 古壮字字典 in any form.

Glyph	Code	Reading	Value
了	U+4E86	ndeu	1
吊	U+540A		
尞	U+5C1E		
𠮩	U+20BA9		

Glyph	Code	Reading	Value
𠲢	U+20CA2		
𬺴	U+2CEB4		
𰀌	U+3000C		
能	U+80FD	nwngh	

Glyph	Code	Reading	Value
𠦂	U+2B9C7		
双	U+53CC	song	2
松	U+677E		

Glyph	Code	Reading	Value
侮	U+4FC9	ngux	5
呸	U+3576	haj	

Acknowledgments

Our thanks to Tao Yang who provides us with a Zhuang Extension G font.

(End of document)