Title: Proposal to add kZhuang to Unihan
Author: John Knightley
Date:15-03-2024

Zhuang characters have since CJK Extension F become an established and respected part of the Unicode community, and Unihan has accepted kZhuangNumeric therefore it is proposed that kZhuang also be added to Unihan. To date 2,471 Zhuang characters have been encoded having a GZ- source from the *Ancient Zhuang Character Dictionary* (古壮字字典), 1989, ISBN 7-5363-0614-8 and this can provide suitable guidance for a format for kZhuang. All readings in the dictionary are written using the official Zhuang orthography which uses the 26 letters of the English alphabet. This orthography is the same as that used for Standard Zhuang which is 1 of the 5 official languages of the People's Republic of China into which all national laws are translated and are included on paper bank notes. Rather than attempting to add kZhuang readings for over 10k encoded characters in one go, it is proposed that the readings be added in stages, starting with the 2,471 characters having GZ- sources and defined as follows:-

Property	**kZhuang**
Status	Provisional
Category	Readings
Delimiter	space
Syntax	[a-z]+*?
Description	The most customary Zhuang reading for this ideograph. Readings of words not part of the Standard Zhuang lexicon are suffixed by an asterisk. Among the sources used for the property data are the following: Ancient Zhuang Character Dictionary (古壮字字典), 1989, ISBN 7-5363-0614-8

In common with other Zhuang dictionaries the *Ancient Zhuang Character Dictionary* also notes if a word is not part of the Standard Zhuang lexicon, and it is proposed than kZhuang do likewise. In the *Ancient Zhuang Character Dictionary* readings of words not part of the Standard Zhuang lexicon have <方> written before them. Unihan should as indicate this in some way because the distinction is an integral part of Zhuang dictionaries. The use of an asterisk after the readings of words that are not a part of the standard Zhuang lexicon is a possible solution. A comparison of the different entries for two Zhuang characters meaning *sky* from Ext F, 𫶇 U+2D446 and 𭱇 U+2DC47 respectively, illustrates how this would work. The reading of the former is part of the Standard Zhuang lexicon but the latter is not (see figures 1 and 2).

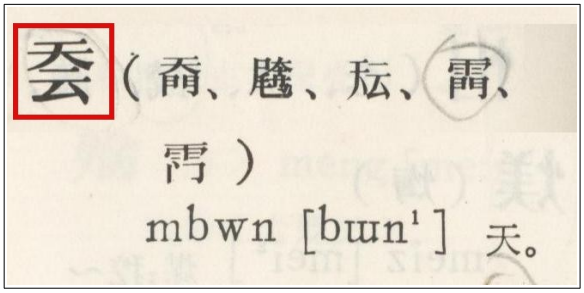


Figure 1: Ancient Zhuang Character Dictionary 'mbwn' page 322

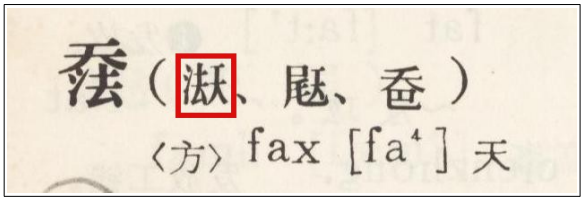


Figure 2: Ancient Zhuang Character Dictionary 'fax' page 156

𡈼 U+2D446 has source GZ-3222401, namely *Ancient Zhuang Character Dictionary* page 322 cloumn 2 entry 4 character 1. The reading is Standard Zhuang so the Unihan kZhuang entry would be:-

U+2D446 kZhuang mbwn

𡈼 U+2DC47 has source GZ-1561402, namely *Ancient Zhuang Character Dictionary* page 156 cloumn 1 entry 4 character 2. The reading is not Standard Zhuang so the Unihan kZhuang entry would be:-

U+2DC47 kZhuang fax*

This distinction is also noted in the standard reference *Sawloih Cuengh Gun (Zhuang-Chinese Dictionary)* which states that whilst both mean *sky* that *mbwn* is Standard Zhuang and *fax* is not (see figure 3).

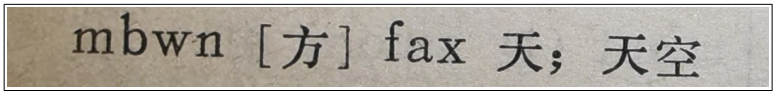


Figure 3: Sawloih Cuengh Gun(1984) 'mbwn' page 524

In the vast majority of cases the correct reading and whether or not it is part of the Standard Zhuang lexicon can easily be determined from the entry in *Ancient Zhuang Character Dictionary* but for 5 readings out of the 2,471 further investigation was required as shown below in Appendix 1.

For the initial list of 2,471 characters the value is based on the reading of GZ- source of the character, therefore each character has only one Zhuang reading even when the *Ancient Zhuang Character Dictionary* has several readings for the character. In the future some characters will require multiple Zhuang readings therefore the delimiter space is suggested.

Conclusion

The addition of kZhuang to Unihan at this time is logical and would help preserve useful information provided to IRG as a usual part of the encoding process. The 2,471 initial values proposed provide a good foundation on which to build. When more values are added the definition of kZhuang could be updated as required.

The author cordially invites those interested in helping in extending kZhuang to contact him via email knightleyjohnc@gmail.com or otherwise.

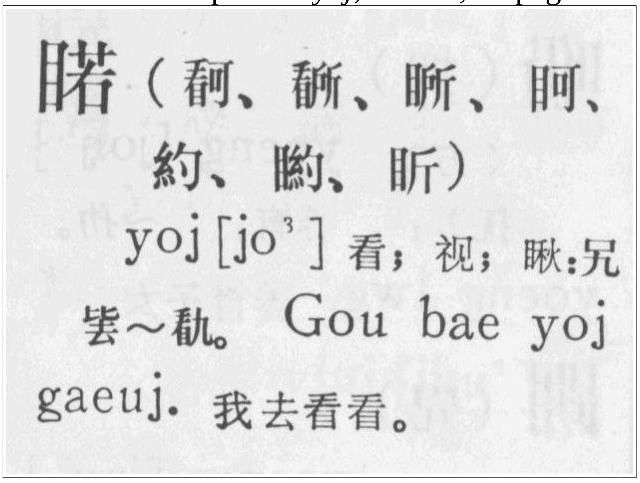
Acknowledgements:-

Thanks to Eiso Chan for discussions leading to this proposal, to Yi Bai for feedback on the data and to Ken Lunde for feedback on this document.

Appendix 1: Five readings that required further investigation

Readings 1-3:-

Whether or not a word is part of the Standard Zhuang lexicon is by no means naturally intuitive even for native speakers because the the Standard Zhuang lexicon includes words from different Zhuang languages. To determine whether or not a word is part of the Standard Zhuang lexicon is a matter of consulting appropriate reference works. Whilst the *Ancient Zhuang Character Dictionary* usually correctly indicates whether or not the readings are part of the Standard Zhuang lexicon, there are some exceptions. One such exception is *yoj*, to look, on page 510 (see figure 4).


睰（𭿋、𭿡、盺、𭾨、約、[illegible]、昕）

yoj [jo^{3}] 看；视；瞅：兄[illegible]～[illegible]。Gou bae yoj gaeuj. 我去看看。

Figure 4: Ancient Zhuang Character Dictionary ‘yoj’ page 510

That the word *yoj*, to look, is not Standard Zhuang is explicitly stated twice in *Sawloih Cuengh Gun*. Firstly, under the standard *yawj*, where *yoj* is listed as the 5th of 8 non standard alternatives (see figure 5) and second, under *yoj*, where the reader is directed to the standard *yawj* (see figure 6)

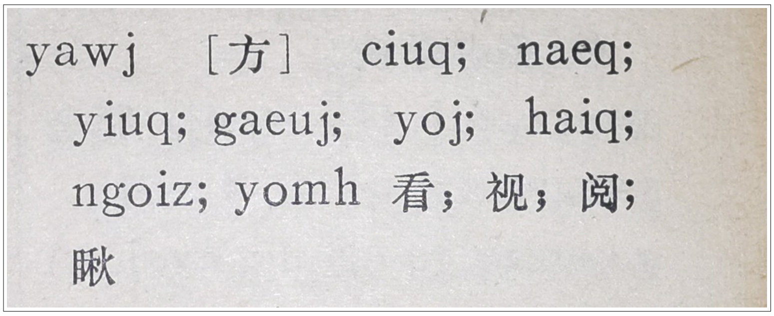

yawj [方] ciuq; naeq; yiuq; gaeuj; yoj; haiq; ngoiz; yomh 看；视；阅；瞅

Figure 5: Sawloih Cuengh Gun(1984) ‘yawj’ page 734

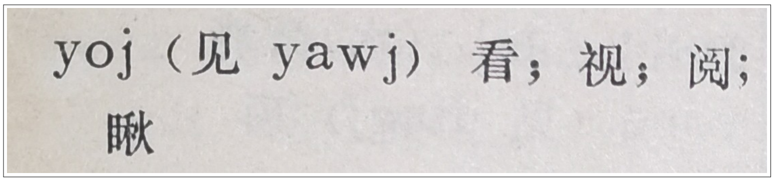

yoj（见 yawj）看；视；阅；瞅

Figure 6: Sawloih Cuengh Gun(1984) ‘yoj’ page 745

This affects the values of 3 characters: GZ-5101302 𭿋 U+2DFCB, GZ-5101303 𭿡 U+2DFE1, and GZ-5101305 𭾨 U+2DFA8 should be

U+2DFCB	kZhuang	yoj*
U+2DFE1	kZhuang	yoj*
U+2DFA8	kZhuang	yoj*

and not

U+2DFCB	kZhuang	yoj
U+2DFE1	kZhuang	yoj
U+2DFA8	kZhuang	yoj

Readings 4 and 5:-

Due to slight printing errors, the correct reading of two other characters whilst not immediately obvious can be deduced from context:-

4. The correct reading for 鐓 U+30f41 is *gu* not *guh*. Though the header in *Ancient Zhuang Character Dictionary* shows *guh*, the ipa and the two examples show *gu* is in fact the correct reading (see figure 7) and a comparison to *Sawloih Cuengh Gun* also confirms that *gu* not *guh* is correct.

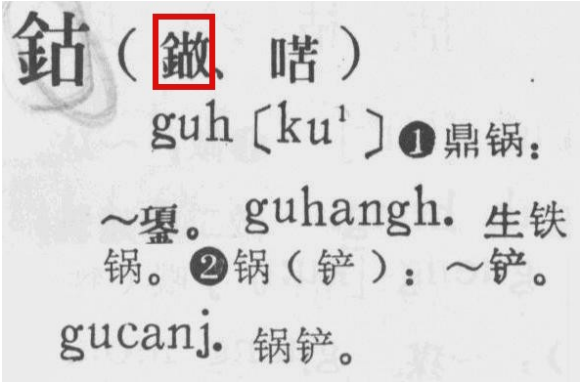


Figure 7: Ancient Zhuang Character Dictionary 'gu' page 205

5. Strictly speaking the reading for U+3193C of daem(lect) comes from not from GZ-0952501 itself but from the following entry which shows both the character and reading because GZ-0952501 itself only shows the character and has no reading. The evidence from the following entry was added via ORT (see https://hc.jsecs.org/irg/ws2017/app/index.php?find=GZ-0952501).

ISO/IEC JTC1/SC2/WG2/**IRGN2677EisoFeedback**
2023-03-15

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

Doc Type:	**Ideographic Research Group Document**
Title:	**Feedback on IRGN2677**
Source:	**Eiso Chan (陈永聪, Culture and Art Publishing House)**
Status:	**Individual Contribution**
Action:	**For consideration by IRG and UTC**
Date:	**2023-03-15**

Firstly, I need to say I totally support to accept the kZhuang property and all the data provided in IRGN2677. At the same time, I need to mention the relevant issues.

1) This property is to record the language, not only for the dictionary. The dictionary is the most important source now. That means we can add the property features based on other sources in future.

2) The current document only the readings of the inherent Zhuang words, but it is only a part of Zhuang words in Zhuang languages. So, we also need to consider how to handle the so-called 新借 and 老借 which are used for the common Hanzi and their variants. I suggest the following methods:

XXX for 新借; Xxx for 老借; xxx for the inherent ones

U+76AE	kZhuang	BIZ Beiz baez

For U+76AE 皮, biz is 新借, beiz is 老借, baez is the inherent one.

3) The authoritative source of 新借 is 《壮语新词术语汇编》 (*Vahcuengh Sinhswz Suzyij Veibenh*), which is often used for the person names, geographic names and the new words. I also found this system is related to 戏棚官话 (Theatre Mandarin) of Nanning Yongju Oprea (邕剧) and Cantonese Yueju Oprea (粤剧) and some kinds of "Mandarin" dialects in Guangxi Zhuang AR. The relationships among them are needed to clarify, but the readings are stable. There are more than 5,000 readings now.

4) There are not any authoritative sources to record 老借, which are used for some daily words and some person names. The scholars consider this system is related to Chinese Yue-dialects (even Chinese Pinghua-dialects), but it is not so easy to clarify after collecting more than 1,000 readings by myself.

5) For WS2021-03895 (⿰⻊党), it is treated as the variant of 蹚 in 《壮语新词术语汇编》, and the reading is dangh which has matched the reading provided by 《古壮字字典》. That means this character should be treated as the variant of common Hanzi, and the reading is 老借 not the inherent one.

6) The 新借 readings will be also submitted to UTC (and IRG?) in future.

(End of Document)

Response Feedback by Eiso Chan to *IRGN2677 Proposal to add kZhuang to Unihan* of 15-03-2024

(John Knightley: updated 06-04-2024)

The feedback and support for the introduction of kZhuang is appreciated. Here is a brief summary response, and a more detailed point by point response is available for reading below.

Whilst the readings in the initial proposal all come from the 《古壮字字典》which contains over 13k readings, it is and never was intended to be a complete list of Zhuang readings, other sources are required for a more complete list. The new 《古壮字大字典》 has 16k thousand readings and fills in many of the gaps in 《古壮字字典》. As for the the readings of characters in modern Chinese loans they are also suitable for Unihan and so should in the future be submitted to Unihan. What form the modern readings should take is yet to be decided.

The initial proposal covers only 2,471 characters from ExtF/G/H with a single reading, there are a further 8k readings of encoded characters in《古壮字大字典》and along with 《古壮字大字典》the number to add rises to over 10k, these figures do not include readings for at least 3k characters in these 2 dictionaries that have yet to be encoded. The plan is to prepare and submit the readings from these, supplementing from other sources if required. Assuming this can be completed in a timely manner then this should be done before submitting the Zhuang readings for modern Zhuang loans.

Whilst it is possible to list the Zhuang modern Chinese loan readings for characters, for other readings it is not possible with a sufficient degree of certainty to identify which readings of characters are those of ancient Chinese loans or which are not, therefore it would not be constructive to attempt to try to add this to Unihan. For some words like daj(打) the Zhuang modern Chinese loan reading is identical to the non-modern reading. Therefore a 3 part distinction between modern loans, ancient loans and indigenous words would not be possible. Since the use of upper and lower case to distinguish as suggested is unusual in that it does not appear to conform to existing Zhuang or Unihan conventions, therefore the existing lower case format should be kept.

Point by point response

1) This property is to record the language, not only for the dictionary. The dictionary is the most important source now. That means we can add the property features based on other sources in future.

It is entirely agreed that more sources should and will be added in the future. Adding a new source would itself entail updating the property, the *features* of kZhuang however should be those consistent with established practice and once introduced in general stable. Now therefore is a suitable time to discuss the features kZhuang should have. It should be noted the category is readings. The syntax proposed is [a-z]+*? (all lowercase letters with an optional asterisk at the end indicating a non-standard Zhuang reading).

2) The current document only the readings of the inherent Zhuang words, but it is only a part of Zhuang words in Zhuang languages. So, we also need to consider how to handle the so-called 新借 and 老借 which are used for the common Hanzi and their variants. I suggest the following methods:

XXX for 新借; Xxx for 老借; xxx for the inherent ones

U+76AE	kZhuang	BIZ Beiz baez

For U+76AE 皮, biz is 新借, beiz is 老借, baez is the inherent one.

The proposed format is already designed to accommodate readings that come from all Zhuang words be they indigenous (inherent) Zhuang words or loan words using lowercase only. Whilst the current list of 2,471 readings does not include any *modern loans* (新借) it does include some *ancient loans* (老借) or Zhuang-Chinese words of common origin such as

U+2DA21	kZhuang	saw

(The reading *saw* [book; character] of 𭓾 U+2DA21 has a common origin with the word 书 in Chinese.)

The 3 way distinction suggested above *BIZ Beiz baez* is not one used within Zhuang dictionaries, nor is it consistent any existing practice in Unihan and best not used. Not to mention that for some words there is debate among experts as to whether or not a word is an *ancient loan* (老借) so this part is unimplementable. Following the proposal format, and Unihan convention of alphabetical order we get the simpler and more manageable *baez beiz biz* rather than *BIZ Beiz baez*.

It is possible for a writing system to highlight some such differences. In the closely related Buyi modern loans are written differently from other Buyi words, however this is not the case for Zhuang were were they same orthography is used throughout. Modern loans in Zhuang are usually easily spotted because they are are Chinese words with a Zhuang accent. Hearing a sentence like, " Gou youq aen duzsuhguanj." (I'm in a library) a Chinese speaker of the Guiliu (桂柳) dialect who doesn't know Zhuang should probably recognize the word for library, duzsuhguanj (图书馆). *Ancient loans* are words of common origin with Chinese that have been completely assimilated into Zhuang and so for some words there is debate about whether or not they are ancient loans.

3) The authoritative source of 新借 is 《壮语新词术语汇编》 (*Vahcuengh Sinhswz Suzyij Veibenh*), which is often used for the person names, geographic names and the new words. I also found this system is related to 戏棚官话 (Theatre Mandarin) of Nanning Yongju Oprea (邕剧) and Cantonese Yueju Oprea (粤剧) and some kinds of "Mandarin" dialects in Guangxi Zhuang AR. The relationships among them are needed to clarify, but the readings are stable. There are more than 5,000 readings now.

Agreed that 《壮语新词术语词编》is *an* authoritative source that contains many modern loans with readings in it are based on the《新汉借词语音转写表》.《新汉借词语音转写表》is an appendix in both 《壮语新词术语词编》and 《壮汉词汇》The spelling of these modern loans is based on their Chinese pronunciation in South Western Mandarin and in particular what is commonly referred to as 桂柳话 so can indeed be calculated based a knowledge Chinese. The main body of《壮语新词术语词编》also includes some non-modern readings. To verify the list of Zhuang readings as such some processing would be required. Many of these readings are extremely rarely used in Zhuang character texts so priority should be given to preparing for addition the more common readings first. It may be that the 5k+ list mentioned in based on 《新汉借词语音转写表》in which case it may contain readings for characters not found in any word in the main body of the dictionary.

> 4) There are not any authoritative sources to record 老借, which are used for some daily words and some person names. The scholars consider this system is related to Chinese Yue-dialects (even Chinese Pinghua-dialects), but it is not so easy to clarify after collecting more than 1,000 readings by myself.

It is sometimes mistakenly thought that 《古壮字字典》and 《壮语新词术语词编》should together cover all the every day Zhuang readings but this is not the case. This is not a new discovery. I remember David Holm pointing this out to me over 20 years ago. Some common readings were included《古壮字字典》for various reasons. The readings for some common Chinese characters that are also Zhuang such as 三 sam, despite being included in example sentences in 《古壮字字典》does not have even have an entry. In other cases such as max (horse) several characters are shown but not the character for horse itself. This is a known issue and one reason that there has not sooner been a kZhuang proposal and why the initial proposal starts in Ext F. This problem is now considered soluble.

From the prospective of providing readings for Unihan it is not correct to say there are no authoritative sources for readings of those not 《古壮字字典》.《古壮字大字典》 would be an ideal second source for kZhuang. There are other sources are available. Care should be taken as some sources may have readings that are not reliable.

There is no authoritative list of *all* 老借, because there is debate over the status of some words. For a list of 老借 one reading per character is to be expected, unless the character has multiple Chinese readings. Some words may be pronounced differently in one location to elsewhere, so whilst a reading is correct for that location it is not representative, it is not a common reading. Sometimes words are written with tone sandhi that needs to be accounted for when writing the reading of syllables. There are a number of factors that can result in unreliable readings and these tend to inflate the number of readings of a long list. Comparison of a 老借 list and 1989 《古壮字字典》is indeed one way to find some obvious gaps. It would be useful also to do a comparison to 《古壮字大字典》.

> 5) For WS2021-03895 (⿰⻊党), it is treated as the variant of 蹚 in 《壮语新词术语汇编》, and the reading is dangh which has matched the reading provided by 《古壮字字典》. That means this character should be treated as the variant of common Hanzi, and the reading is 老借 not the inherent one.

An interesting question. If here you have found a 老借 reading 𠱓 党 in ws2021 it is further confirmation that your assumption in point 1 that there are no 老借 readings in the Ext F/G/H readings is not correct. In this case not all experts would accept as 老借. That the modern loan for 蹚 is dangh in 《新汉借词语音转写表》page 794 of《壮语新词术语词编》and 𠱓 党 dangh in《古壮字大字典》is a problem for which the multiple tones of 蹚 might offer a solution. One could also see if there is a possible a proto-Tai source for this word. From a Unihan prospective it would not be productive to try identify which readings are or are not ancient loans, or words of common origin.

6) The 新借 readings will be also submitted to UTC (and IRG?) in future.

Yes, as mention above the modern loan readings will also be submitted in the future to UTC.