ISO/IEC JTC1/SC2/WG2/IRG N2337

Title: Request to disunify U+2F83B and discuss the feasibility of IVD usage

Source: Ming Fan

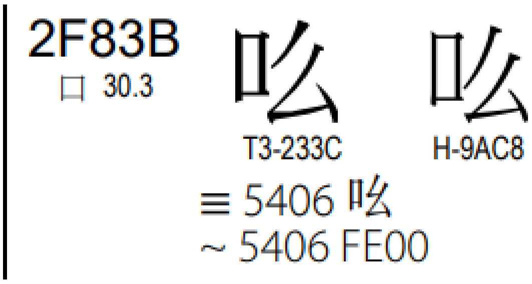
Status: Individual Contribution

Action: For consideration by ISO/IEC JTC1/SC2/WG2/IRG

Date: 2018.9.24

1. Background

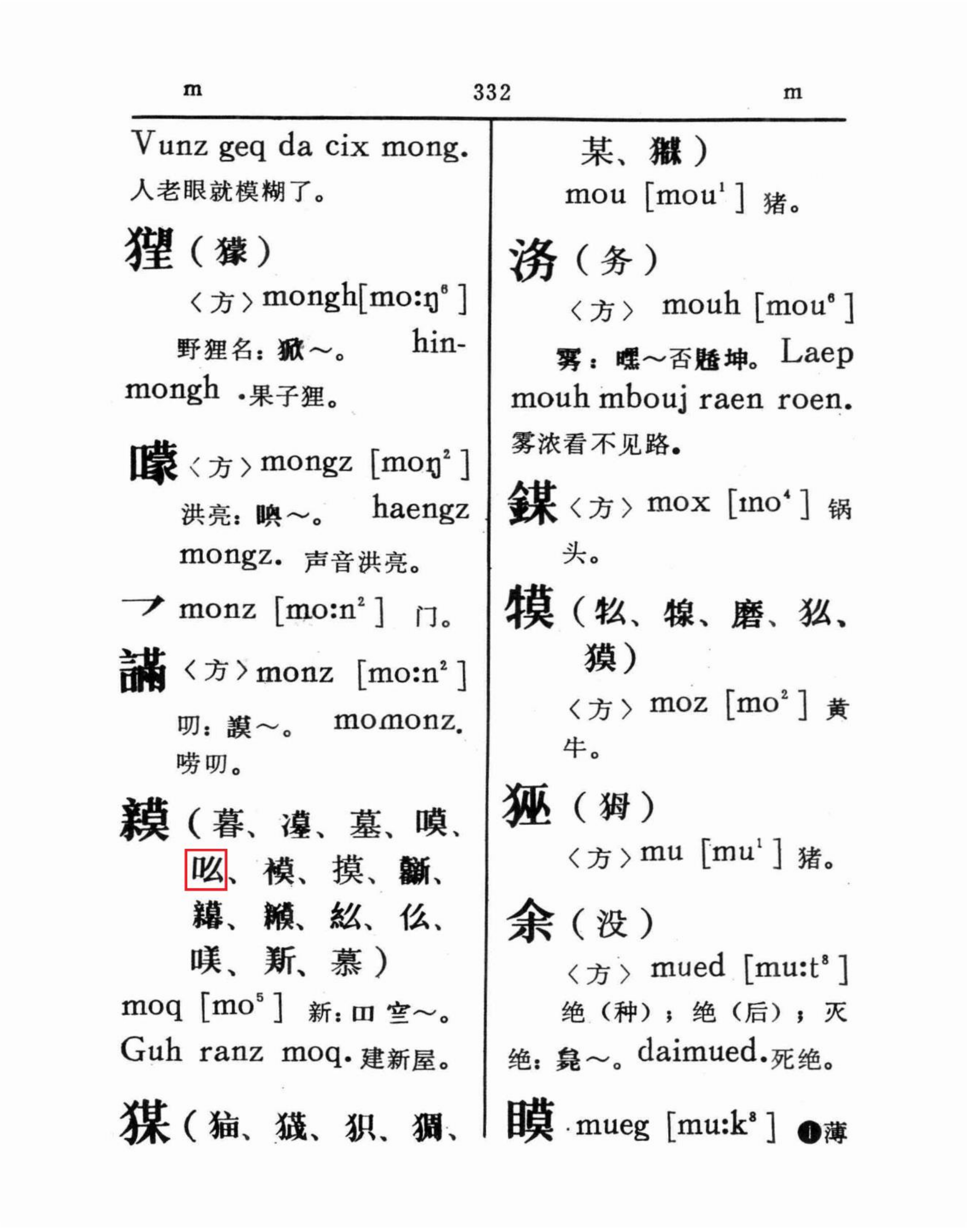
The U+2F83B is unified to U+5406 吆 currently:



2F83B
口 30.3
吆 T3-233C
吆 H-9AC8
≡ 5406 吆
~ 5406 FE00

However, we found that this ideograph is completely different from U+5406 吆 sometimes:

(1) Sawndip



Vunz geq da cix mong. 人老眼就模糊了。

⿰犭望（獴）〈方〉mongh[mo:ŋ^6] 野狸名：⿰犭敬～。 hinmongh .果子狸。

⿰口蒙〈方〉mongz [moŋ^2] 洪亮：⿰口兴～。 haengz mongz. 声音洪亮。

乛 monz [mo:n^2] 门。

⿰言满〈方〉monz [mo:n^2] 叨：⿰言莫～。 momonz. 唠叨。

⿰禾莫（暮、⿰氵墓、墓、嗼、吆、⿰衤莫、摸、⿰鬲斯、⿰米暮、⿰米模、⿰糸么、仫、⿰口美、⿰莫斤、⿱莫小）moq [mo^5] 新：⿰口 ⿱宀至～。Guh ranz moq. 建新屋。

⿰犭某（猫、⿰犭戈、⿰犭只、⿰犭周、某、⿰犭某）mou [mou^1] 猪。

⿰氵务（务）〈方〉 mouh [mou^6] 雾：嘿～否⿰目番坤。Laep mouh mbouj raen roen. 雾浓看不见路。

⿰金某〈方〉mox [ɪno^4] 锅头。

⿰牛莫（⿰牛么、⿰牛象、磨、⿰犭么、獏）〈方〉moz [mo^2] 黄牛。

⿰犭巫（⿰犭母）〈方〉mu [mu^1] 猪。

余（没）〈方〉 mued [mu:t^8] 绝（种）；绝（后）；灭绝：臬～。daimued. 死绝。

瞙 mueg [mu:k^8] ❶薄

(From 《古壮字字典》 page 332)

When used in Sawndip, it's pronunciation is moq, which is near to 麽 in standard Chinese. It's phonetic element is 么(simplified form of 麽), completely different from that of 吆(幺).

(2) Dialect

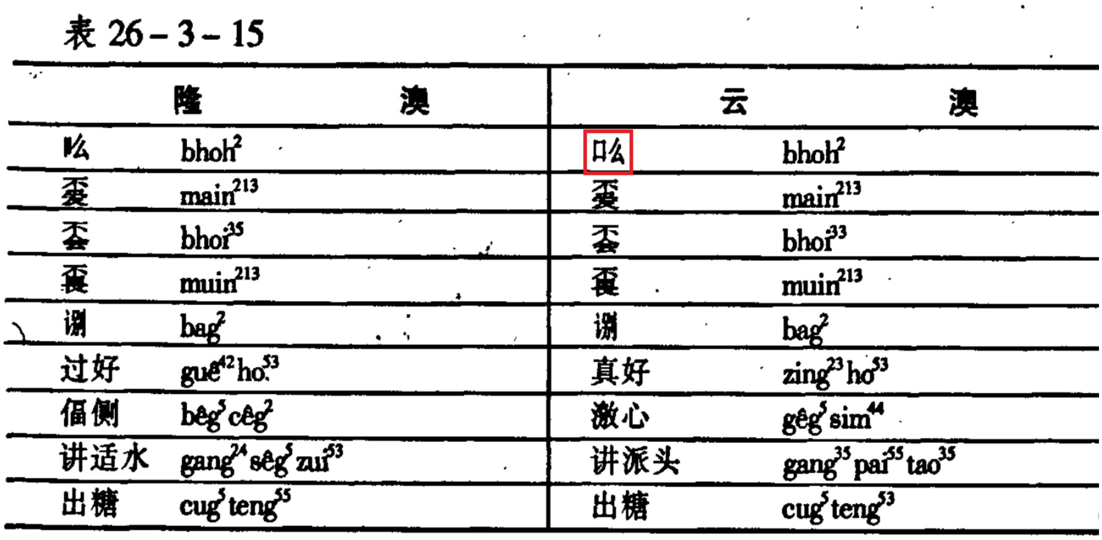


表 26－3－15

隆	澳	云	澳
吆	bhoh^{2}	⿰口么	bhoh^{2}
嫑	main^{213}	嫑	main^{213}
𠀾	bhoi^{35}	𠀾	bhoi^{33}
[illegible]	muin^{213}	[illegible]	muin^{213}
谢	bag^{2}	谢	bag^{2}
过好	gue^{42}ho^{53}	真好	zing^{23}ho^{53}
偪侧	bêg^{5}cêg^{2}	激心	gêg^{5}sim^{44}
讲适水	gang^{24}sêg^{5}zui^{53}	讲派头	gang^{35}pai^{55}tao^{35}
出糖	cug^{5}teng^{55}	出糖	cug^{5}teng^{53}

(From 南澳县地方志编纂委员会 编:《南澳县志》, 北京: 中华书局, 2000 年 10 月第 1 版, 2000 年 10 月第 1 次印刷, ISBN 7-101-02542-0, p725)

When used in Longao &Yunao dialect, it's pronunciation is bhoh^{2} , which is near to 麼 in standard Chinese. It's also different from 吆, and should not be unified.

2. Comments

This ideograph, ⿰口么, is frequently used in many textbooks & dictionaries. It's because its high frequency that we found the different usage from 吆. And it was misunified to 吆 for over 15 years, which misleads users and is not acceptable. Finally, it still need to be encoded separately, and when used as variant of 吆, the new code point for ⿰口么 is duplicate with original U+2F83B, which causes trouble to users. If we encode them separately at beginning, these problems would not occur and also acceptable for users.

Luckily, this situation is not even the worst. What if when ⿰口么 is not cognate with 吆, but users use U+2F83B instead? What if an ideograph A is registered as IVD of B, but later A was found different from B, but users are already used to use IVD for A, even when A is not cognate with B? This problem may never be solved completely, even A was later separately encoded from B. If we encode them separately at beginning, this problem would not occur, either. Some may ask, what should we do when A is used cognate with B? Will encode them separately confuse users in this situation? Well, we can do some work on bibliographic search system & electronic database, which is much more flexible than IVD & UCS standard, and they can be changed at any time to solve problems unlike specific standards, whose some part cannot changed forever. And bibliographic search system & electronic database is more convenience, nearer to users than IVD & UCS standard. So these problems can be solved & avoided completely.

The IVD is not a bad solution, however, it's not a good solution to Han variants, either. It can only recommended when A is the variant of B **at 100% situations**. When we cannot sure A must be the variant of B in whatever situations, then the IVD should not be used and we should simply encode them separately to solve problems completely. Otherwise, problems will remain and may never be solved completely, which causes troubles to users.

This IVD problem is unavoidable and should be solved as soon as possible to avoid chaos. Therefore, I propose to recommend to use IVD only when **an ideograph doesn't like to have possibility to be non-cognate with target ideograph**. This may be shocking and impractical, however, a clear and rigorous red line must be lined to best avoid and solve problems. Abusing IVD may cause many of these problems that may never be solved, which wouldn't occur if encoding separately as explained above. We should reduce controversial unification rules & actions to make Han encoding process more constructive & instructive. If we relax unification rules, the Han encoding process will in a mess and cause trouble to users. Just reducing number of Han ideographs despite users is controversial with our initial aims for constructive & instructive process of Han encoding.

Like ⿰口么(assume that we don't know it's different from 吆 in some situations), this kind of ideographs are from authoritative sources and greatly-needed, and it's likely to be non-cognate at some situations (for example, 么 is likely to be non-cognate with 幺 when used as simplified form of 麼). So it's not suitable for them to be added to IVD as explained above. They can be separately encoded at beginning, since they are greatly-needed, and to best avoid problems when later found that they are non-cognate as mentioned at paragraph 2.

Meanwhile, the IRG should avoid relaxing unification rules to refrain the problems (the more rules refer to a pair that is likely to be non-cognate added, the more frequent this problem will be, the IRG should avoid adding these rules which identify this kind of pairs as unifiable variant). This decision may be unreasonable, however, **the non-cognate problem should be concerned**, and **it's even worse for users** than problems of too many encoded variants. This doesn't mean that all this kind of rules should be obsoleted, but just avoid being relaxed for abuse. And for some rules, open green lights for a few special conditions only. Moreover, variants which from authoritative sources and greatly-needed are in fact not too much, opening green light to them for separate encoding will not boost the increase of encoding variants, but helpful to instructive process of Han encoding and solve the non-cognate problem. This is the reasonable attitude for digitization. If the unification rules are rigorous and instructive, the problem of too many variants and non-cognate problems can be completely solved simultaneously.

3. Proposed Actions:

(1) Disunify U+2F83B and assign a new code point for ⿰口么 in URO.

(2) Add a rule which states that **"IVD should only be used when an ideograph doesn't likely to be non-cognate with target ideograph under whatever circumstances, otherwise this ideograph should not be added to IVD without any further discussions for IVD allowed. This ideograph can be separately encoded if it's form authoritative sources and greatly-needed."**

(End of Document)