Title: Comments on Nushu in WG2 N4484, /IEC 10646:2014/PDAM1

Status: Individual Contribution

Source: suzuki toshiya, Hiroshima University, Japan

After the interrupt from 2009 to 2012 for the standardization of Nushu, I appreciate the efforts by China NB, SEI and the experts from US, UK, Ireland, Finland and TCA. Japanese expert contribution by Endo Orie, WG2 N3705, was almost ignored in the reopened discussion, but the efforts by Deborah Anderson and Michael Everson improved some questionable parts greatly. However, Nushu charset in PDAM text (WG2 N4484) still needs more improvement, or clarification. This document lists the questionable parts left in PDAM1 code chart.

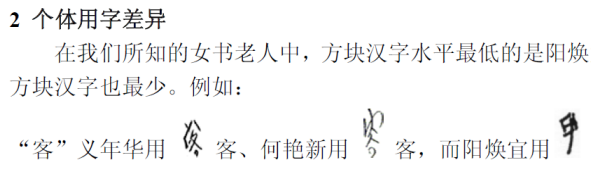
In general, following points are expected to be discussed for further improvement.

- Stability of the statistical investigation
  The idea choosing a referential glyph by the statistical frequency is reasonable, but the current statistical results seem to be unstable to define the normative and uncancellable attributes (like character names). Therefore, although the charset in Nushu Duben (ISBN 978-7-5438-5282-2) could be useful to define the basic charset, it would not be appropriate to define the normative and uncancellable attributes (like character names).
- Stroke counts and phonetic value in the character names
  Some glyphs in PDAM text have different stroke counts from the section where they are placed. In addition, it seems that sometimes Nushu Duben prioritizes source Hanzi than the phonetic value. Considering that the separated indexing in Nushu Duben character is the most important criteria of the charset, and some Nushu characters are difficult to determine their referential phonetic value by the statistical result, the inclusion of the phonetic value in the character names should be reconsidered.

Considering 2 points in above, I ask Nushu expert group or China NB for a referential document including ***mutually associated*** data to review the proposed Nushu code chart; the proposed glyph shape, raw glyph shape(s), its semantics, its variants and the statistical results for them. The documents previously submitted from China NB included most of them, but they are not mutually associated. Taking WG2 N3598 as an example, the proposed glyph shapes are presented in the proposed code chart (page

9-20), the raw glyph shapes are presented in the comparison of 3 dictionaries (Appendix C, page 31-49), the variants and frequencies are presented in the source Hanzi table (page 72-91). They are no associated via shared identifier. As Endo Orie commented in WG2 N3705 and Deborah Anderson commented in WG2 N4442 and N4451, when a reviewer finds a discrepancy, it is difficult to guess the difference is acceptable one, mistakenly differentiated one, or anything else.

For example, Appendix F, the section 女書交際用字研究 (study of the Nushu glyph for interchange) picks the glyph variants to mean guest (客);


2 个体用字差异

在我们所知的女书老人中，方块汉字水平最低的是阳焕宜，因此她用字最少，而且出现方块汉字也最少。例如：

“客”义年华用 客、何艳新用 客，而阳焕宜用 客；

Figure 1: Variants of “guest” glyph (WG2 N3598, Appendix section 3-2-2, p.98)

But I could find none of them in the proposed charset in PDAM, or the frequency table in WG2 N3598. I’m not sure why, they were unified with other characters? Or, their frequencies were too low to be coded in the basic charset? The changes since the last official proposal from China NB should be summarized.

1. Stability of the statistical investigation submitted to WG2

As the proposals from China NB states, the proposed charset is not designed to be the complete set, but the basic representative glyph set. Anyway, the encoding architecture would be expected to be stabilized at the first basic charset, to avoid the ad-Hoc changing of the encoding architecture in future extensions. In original China NB proposal, the character identity was decided by the statistical investigation (e.g. the last China NB proposal, WG2 N4341, chapter I, section 2 “the judgment and collation of Nyushu basic and variant characters”). In PDAM1, although the character identity discussion is simplified as the indexing characters in Nushu Duben (according to WG2 N4461), the identity of the indexing character in the book is supposed to be based on similar statistical investigation. But checking the content carefully, the stability of the statistical investigation is questionable.

One reason would be that the preliminary classification of the glyphs before counting their frequencies. It would be described in following subsections.

Another reason would be that the number of the authors of the Nushu document collection is not so large; according to WG2 N3598 chapter II, section 1 “the collection and collation of Nushu characters”, the materials by 4 authors (Gao Yinxian, Yi Nianhua, Yang Huanyi and He Yanxin) document is 85%. Therefore, the statistical

result may reflect individual preferences. Originally, Nushu Yongzi Bijiao presented per-author statistics. Choosing a general purpose (or most neutral) referential glyph from per-user referential glyph would be more stable approach.

1.1. Statistical summarization is questionable

As Endo Orie had pointed in WG2 3705 (2009), the statistical investigation result seems to be inappropriately presented, and questionable for its reliability in as-is manner. For example, the process how the representative glyph for Y42 is determined was described in WG2 N4341 section 2-B.

B. **Prior to basic characters according to their frequency**

That is to say, within the characters of the same character cell, we chose the character with the highest frequency. The character with the highest frequency is called basic representative and others as allograph. The final result of the study and the statistics can be referred to '*Nyushu basic characters and their origin verification*' (published in *Nyushu character-comparison, 2006*) .

Take the following table to explain:

y42	如余餘 儒虞娛	如105 如10 如8	如131 余10 如54 如2 儒1	如213 余2 如3 如1	如75 余3 如11 如11 如4	如141 余3 如11

There are three main graphs for the sound of y42 如590-89-12, which has the same pronunciation and the meaning and has general similar structures. The first one has a rather high frequency. So is the representative glyph.

Figure 2: Description of the frequency statistics (WG2 N4341, section 2-B)

It notes that there are 3 major glyphs frequently used (590 times), (89 times), (12 times). But how the frequency could be calculated from the table? In the table, the 1st glyph frequency might be counted as 如105 + 如131 余10 + 如213 余2 + 如141 余3 = 590. But the 2nd glyph frequency is questionable; 如8 + 如54 + 如3 + 如11 + 如11 = 87 times (not 89 times). The 3rd glyph is found 75 times (如75 余3), not 12 times. Although the conclusion " is the most frequently used glyph" is occasionally same, I have a concern that the summarized frequency is not

reliable even if it is in published material. There are some glyphs whose frequency is less than 10, using the materials including such errors as an authorized reference would be inappropriate to stabilize the encoding architecture.

1.2. Statistical investigation seems to be under development

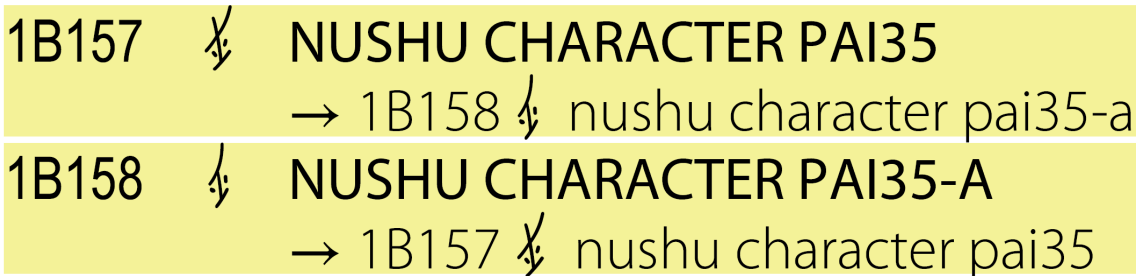


Figure 3: Variants of PAI35 in PDAM (WG2 N4484)

	WG2 3337 (2007)	WG2 3598 (2009)
U+1B157 glyph	（本） pai^{35} 本 ^{176}	（本 ） pai^{35} 本 ^{34} piu^{35} 表 ^{16} tɕioŋ^{35} 整 ^{5} 颈 ^{2}
U+1B158 glyph	（正/ pou^{35} 打 ^{289} tɕyn^{21} 转 $^{139\text{-}2\text{-}0}$ 卷 ^{10} tɕioŋ^{21} 正 ^{43} 镜 ^{31} 敬 ^{21} 政 tsoŋ^{42} 渐 ^{39} pau^{35} 宝 ^{15} 保 ^{10} tɕioŋ^{35} 整 ^{5} tɕioŋ^{33} 警 ^{2} tsoŋ^{42} 惭 ^{2} tsiaŋ^{21} 帐 ^{2} tɕiŋ^{44} 肩 ^{1} p‘ai^{35} 品 ^{1} ɕyn^{35} 选 $^{0\text{-}0\text{-}7}$ piu^{35} 表 $^{0\text{-}2\text{-}16}$	（正/本） pai^{35} 本 $^{212\text{-}178\text{-}9}$ pou^{35} 打 ^{289} tɕioŋ^{21} 正 ^{151} 镜 ^{31} 政 ^{24} 敬 ^{21} 竟 ^{1} tɕyn^{21} 转 $^{139\text{-}2\text{-}0}$ 卷 ^{10} 眷 ^{8} tɕioŋ^{44} 正（~月） ^{50} 京 ^{9} 惊 ^{9} tsoŋ^{33} 渐 ^{41} pau^{35} 宝 ^{15} 保 ^{10} tɕioŋ^{35} 整 $^{11\text{-}40\text{-}3}$ 颈 $^{5\text{-}2\text{-}2}$ 景 $^{0\text{-}2}$ tɕie^{35} 诊 ^{3} 拯 ^{1} tsioŋ^{35} 井 ^{4} tsoŋ^{42} 惭 ^{2} tɕioŋ^{33} 警 ^{2} tɕiaŋ^{21} 帐 ^{2} p‘ai^{35} 品 ^{1} piu^{35} 表 $^{0\text{-}2\text{-}0}$

Table 1: Statistical results for PDAM PAI35 variants

There are 2 characters with same phonetic values, and similar shapes are coded separately. This pair was originally coded for different pronunciation (WG2 N4341 designed U+1B157 glyph for PAI, U+1B158 glyph for POU). According to WG2 N4472R, it was decided that the phonetic values were determined by the top entry of the phonetic values in Nushu Duben, and these 2 characters are now considered as same phonetic

value characters. However, tracking the older proposals, it seems that the assignment of U+1B157 glyph to PAI35 and of U+1B158 glyph to POU35 was reasonable.

In WG2 N3337, the top frequency of the phonetic values for U+1B158 glyph was not PAI35 but POU35. The frequency U+1B157 for PAI35 (本) was originally 176 (in WG2 N3337 (2007)) but decreased to 34 (in WG2 3598 (2009) in later. If we had discussed in 2007, the phonetic value for the character name of U+1B158 should be POU35. On the other hand, some frequency was moved from U+1B158 to U+1B157; the frequency of U+1B158 glyph for PIU35 (表) is decreased from 0-2-16 to 0-2-0, and the decreased frequency seems to be moved to U+1B157 glyph. By such changes, I have a concern that the glyph identification during the statistical investigation was under development. Is it reasonable to expect that the most frequent phonetic values in Nushu Duben are already stabilized?

1.3. Some of rarely used character are supposed to be daily-used

In the statistical result of WG2 N3598, there are some characters at low frequency, less than 10. Some of them are supposed to be very daily used words, like, 尿, 炭, 田, 酒, etc.

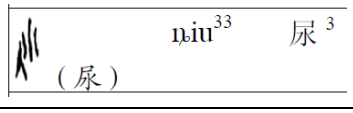
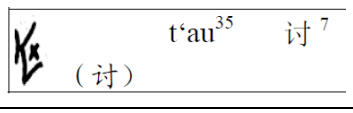
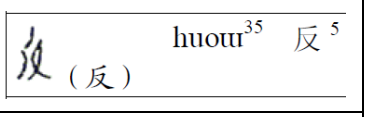
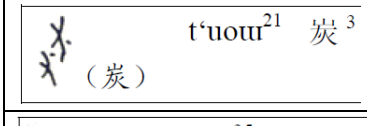
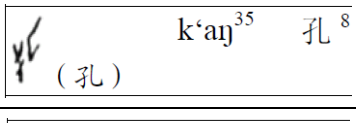
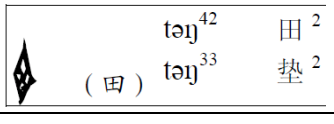
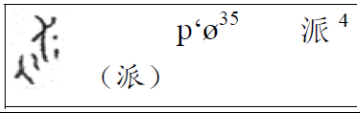
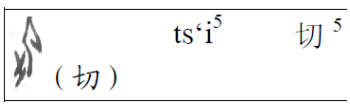
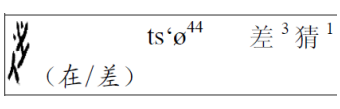
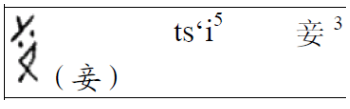
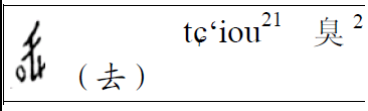
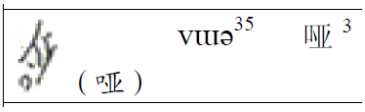
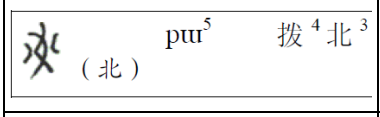
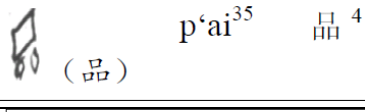
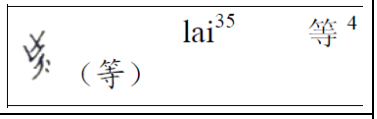
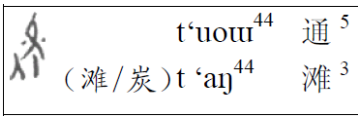
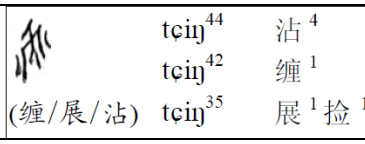
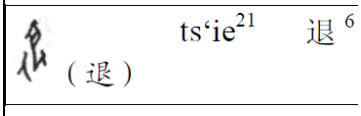
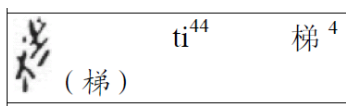
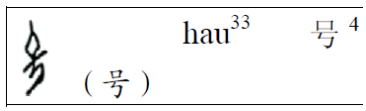
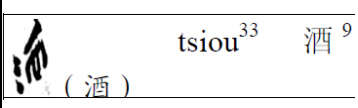
 ȵiu^{33} 尿 ^{3} （尿）	 t‘au^{35} 讨 ^{7} （讨）	 huoɯ^{35} 反 ^{5} （反）
 t‘uoɯ^{21} 炭 ^{3} （炭）	 k‘aŋ^{35} 孔 ^{8} （孔）	 təŋ^{42} 田 ^{2} təŋ^{33} 垫 ^{2} （田）
 p‘ø^{35} 派 ^{4} （派）	 ts‘i^{5} 切 ^{5} （切）	 ts‘ø^{44} 差 ^{3} 猜 ^{1} （在/差）
 ts‘i^{5} 妾 ^{3} （妾）	 tɕ‘iou^{21} 臭 ^{2} （去）	 vɯə^{35} 哑 ^{3} （哑）
 pɯ^{5} 拨 ^{4} 北 ^{3} （北）	 p‘ai^{35} 品 ^{4} （品）	 lai^{35} 等 ^{4} （等）
 t‘uoɯ^{44} 通 ^{5} （滩/炭）t ‘aŋ^{44} 滩 ^{3}	 tɕiŋ^{44} 沾 ^{4} tɕiŋ^{42} 缠 ^{1} (缠/展/沾) tɕiŋ^{35} 展 ^{1} 捡 ^{1}	 ts‘ie^{21} 退 ^{6} （退）
 ti^{44} 梯 ^{4} （梯）	 hau^{33} 号 ^{4} （号）	 tsiou^{33} 酒 ^{9} （酒）

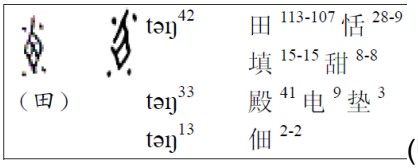
Table 2: Sample of rarely used characters in statistical result (WG2 N3598)

It is reasonable to consider a situation that some characters are low frequency because there is different character with same phonetic values. However, checking the variant for the character that are supposed for daily used words, the situation is different.

a) U+1B150 (尿, NJIU33, urine) has no different character with same phonetic value in WG2 N3598 nor PDAM. So the word of “尿” is used only 3 times in 220000 character documents?

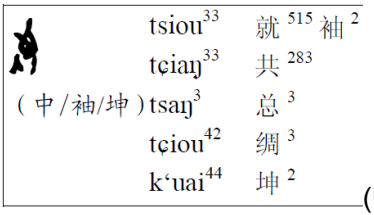
b) U+1B194 (炭, THUOW21, charcoal) has no different character with same phonetic value in WG2 N3598 nor PDAM. So the word of “炭” is used only 3 times in 220000 character documents?

c) U+1B1A0 (田, TENG42-A, rice field) has a character with same phonetic value (U+1B25A, TENG42), but its source Hanzi is again 田. What is the rational to handle U+1B1A0 separately? In addition, the phonetic value TENG42 is appropriate? The frequency of TENG33 for U+1B1A0 is same with TENG42.



(U+1B25A utilization statistics)

d) U+1B232 (酒, TSIOU33-A, wine) has a character with same phonetic value U+1B1FB (袖/坤, TSIOU33). However, the usage of TSIOU33 does not include the usage of the glyph to mean "酒". Does it mean the investigated Nushu document collection rarely use the word "酒"?



(U+1B1FB utilization statistics)

From the cases a), b), d), it is supposed that the investigated Nushu document collection might be biased. From the case c), it is supposed that the distinction of U+1B1A0 and U+1B25A was unexpectedly introduced by the stroke number counting for unstable glyph shapes.

1.4. Some glyph shapes are not the most frequent glyph in the statistical result

According to the statistical result for HANG44, the glyph shape in older China NB proposals and PDAM is not the most frequent shape in the statistics result in WG2 N3598. It seems that the proposed glyph shape is same with the first proposal on 2007; it was the most frequent glyph shape at that time (but the phonetic value was different, it should be HANG21). In later, the existence of the shape difference was recognized (in WG2 N3337, only one shape was counted, and the frequency was 32. in WG2 3598, the frequency was divided into to 2 shapes; 15 + 17 = 32), and the most frequent glyph shape and phonetic value was also changed. However, the proposed glyph shape was not changed.

1B238  NUSHU CHARACTER HANG44

WG2 N3337 (2007)		WG2 N3598 (2009)	
 (欢)	han^{21} 汉 ^{32} han^{44} 欢 ^{15} han^{33} 唤 ^{15} han^{33} 换 ^{14} 焕 ^{11} 汗 ^{7} 唤 ^{6} 翰 ^{6} han^{13} 旱 ^{13} $k'an^{44}$ 糠 ^{1}	 (欢)	han^{44} 欢 $^{305\text{-}164}$ 荒 $^{5\text{-}2}$ han^{21} 汉 $^{17\text{-}15}$ han^{33} 换 ^{14} 焕 ^{11} 汗 ^{7} 唤 ^{6} 翰 ^{6} han^{13} 旱 ^{13} $k'an^{44}$ 糠 ^{1}

Table 3: Historical development of U+1B238 glyph frequency

There are 2 glyphs with different phonetic values and quite similar shapes. U+1B193 is classified in the 6-stroke characters, and U+1B1B2 is in the 7-stroke characters. In the earliest proposal, it seems that these glyphs are unified (see the statistic table in WG2 N3337). In later statistics, the shape difference might be found (thus, LEW44 glyph is classified in the 6-stroke, and OE44 glyph is in the 7-stroke) and counted the frequencies separately. OE44 glyph in WG2 N3337 was different from LEW44 glyph. However, in PDAM1 code chart, OE44 glyph is still similar to LEW44 glyph. If U+1B1B2 glyph shape is preferred, the stroke count should be reconsidered. If WG2 N3598 OE44 glyph is preferred, the glyph should be fixed.

1B193  NUSHU CHARACTER LEW44
→ 591A 多

1B1B2  NUSHU CHARACTER OE44

	WG2 N3337 (2007)	WG2 N3598 (2009)
LEW44	 （多） ləɯ^{44} 多 ^{542} 知 ^{14} 闹 ^{2} ləɯ^{33} 落 ^{126} 洛 ^{3} ø^{44} 衣 ^{93} luoɯ^{44} 单 ^{11}	 （多） ləɯ^{44} 多 ^{542} ləɯ^{33} 落 ^{126} 洛 ^{3} luoɯ^{44} 单 ^{11} 丹 ^{11} nəɯ^{33} 闹 ^{2} tsʻɯə^{21} 刺 ^{2}
OE44	(not found)	 （衣） ø^{44} 衣 ^{93}

Table 4: Historical development of LEW44 and OE44 frequency

The glyph looking like U+1B105 for POE5 was not found in older China NB proposals (furthermore, it is not listed as a variant), although China NB always proposed the glyph for POE5. Also the existence of the glyph looking like U+1B105 was found in Endo Orie's He Yanxin glyph list in 中国女文字研究 (ISBN 4-625-48300-X) p.292-332. It is supposed some editorial errors exist in the statistical result table.

1B105  NUSHU CHARACTER POE5

1B106  NUSHU CHARACTER SWF2

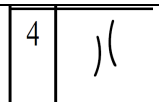
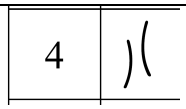
2007	WG2 3598 (2009)
 (WG2 N3287)	
 （八） pø^{5} 八 ^{326} pəŋ^{44} 拔 ^{1} (WG2 N3337)	 （八） pø^{5} 八 ^{326} pəŋ^{44} 拔 ^{1}

Table 5: Incompatible glyphs between proposed chart and frequency result for POE5

2. Stroke counts and phonetic value

The Nushu is recognized as the syllabic writing system, thus each character name includes its phonetic value. However, the identity of the character is not determined by its phonetic value, and the source Hanzi is supposed to be the main factor to consider the separated encoding.

2.1. How the indexing item in the statistical result is allocated?

As described in the first proposal, WG2 N3287, section 3 rule a) described, the basic rule to give the identity to Nushu glyph is the recognization of the source Hanzi. Even if the phonetic value and semantics are same, the glyphs derived from different Hanzis are coded separately. For example, U+1B101:U+1B14E are separately coded although the unification of them does not cause the semantic loss.

1B101 NUSHU CHARACTER I5 → 1B14E nushu character i5-a → 4E00 一	 (一) i^{5} 一 1547 i^{33} 叶 1
1B14E NUSHU CHARACTER I5-A → 1B101 nushu character i5	 (壹) i^{5} 一 1064 i^{21} 以 3 i^{33} 叶 2

Table 6: example of separation by source Hanzi difference for the semantically equivalent pair.

On the other hand, it seems that the variant with significant shape difference would not be coded separately, if the source Hanzi is same. The variants of U+1B188 (门, MAI42) have significant shape difference, but not coded separately.

1B188 NUSHU CHARACTER MAI42 → 1B251 nushu character mai42-a	 (门) mai^{42} 门 816-5-2 又读 mai^{21} 闻（耳~） 112 民 96 mai^{21} 闻（嗅） 30 mau^{42} 毛 18 mai^{42} 眉 4

Table 7: example of unification by source Hanzi identity for the different shapes.

But there are confusing exceptions. As mentioned in above, U+1B1A0 and U+1B25A are coded separately, although their source Hanzis are same, and the shape difference is subtle than those in U+1B188 variants. If U+1BA0 shape is far different and its frequency 4 is sufficient to separate, why the second candidate glyph in U+1B25A

statistics is not coded?

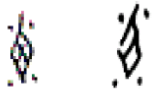
U+1B1A0	U+1B25A
 təŋ^{42} 田 ^{2} （田） təŋ^{33} 垫 ^{2}	 təŋ^{42} 田 $^{113\text{-}107}$ 恬 $^{28\text{-}9}$ 填 $^{15\text{-}15}$ 甜 $^{8\text{-}8}$ （田） təŋ^{33} 殿 ^{41} 电 ^{9} 垫 ^{3} təŋ^{13} 佃 $^{2\text{-}2}$

Table 8: example of separation by shape difference for semantically equivalent and same source Hanzi.

The current status of the separation of U+1B157 and U+1B158 (see Table 2) would be similar case.

2.2. Stroke number is stable?

As I presented in above, I have a concern that the frequency in the statistical result is under development, and it is inappropriate to determine the uncancellable attributes of the characters. Therefore I have a concern that the number of stroke is not stabilized yet, therefore, the name of the blocks should not use the classification by the stroke number, even if the content is roughly ordered by the stroke number.

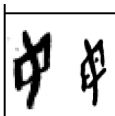
	mu^{13}	母 $^{232\text{-}224}$ 马 $^{72\text{-}32}$
	mu^{33}	木 $^{36\text{-}3}$ 目 ^{12} 墓 $^{4\text{-}4}$
（母）	mu^{42}	麦 $^{2\text{-}1}$
	mou^{13}	亩 $^{2\text{-}1}$ 牡 $^{5\text{-}2}$

Figure 4: Frequencies of the MU13 variants (WG2 N3598)

For example, the frequencies of 2 shapes U+1B152 MU13 (母) are almost comparable. When the most frequently used glyph is known to be the (currently) 2nd one, the number of strokes would be changed from 5 to 7.

As shown for U+1B1B2 OE 44 (衣), the representative glyph is already inconsistent with the number of strokes to be used its code point. Also the representative glyph shape  for U+1B13C CYA5 (出) is difficult to recognize its stroke number as 5 (5 is for the 2nd frequently used glyph  ?). When more documents are investigated, the most frequently used glyph could be changed. It should be noted that CYA5 glyph was not found in the code chart in WG2 N3287 and the statistical results in WG2 N3337 (also it should be noted that Endo Orie's He Yanxing glyph list published in 2002 had already included the 2nd frequently used glyph for CYA5, but the proposals in 2007 included no characters for the phonetic value CYA5).

3. Summary

As presented in above, the stability of the statistical investigation of Nushu document is supposed to be still under development. Therefore, a document including the mutually associated data (code chart glyph, raw glyph, variant shapes, pronunciation, semantic Hanzi, frequency, author, etc) is needed to review the proposed code chart from the viewpoint of their stability. Also the changes from the last official proposal, WG2 N4341, should be summarized with the rationales.

Also considering that the representative glyph, its number of strokes, its phonetic value could be changed by the future investigation, the character naming convention should be reconsidered. CJK Unified Ideograph does not include any information except of its codepoint, or, Egyptian Hieroglyphs and Linear B Ideograms use the indexing number of their referential materials. Such conventions would be considerable options.

(end of document)