ISO/IEC JTC1/SC2/WG2 N4562

Title: **Feedback To ~~N4446~~ N4556 (Chinese Reply To N4513 and N4533).**

Source: **suzuki toshiya**

Date: **2014/02/25 (wrong document number corrected)**

Status: **Individual Contribution**

1. Background

To the Nushu draft charset in PDAM1 ballot, WG2 N4484, I and Orie Endo submitted a document N4513 just before PDAM1 ballot, and submitted additional document to UTC#138 (L2/14-050 = WG2 N4533). Japan NB referred N4513 and voted negative with comment for PDAM1 ballot. China NB had submitted a reply to N4513 and N4533, but I think it is too short to reply to answer both of N4513 and N4533. In this document, I summarize the outlines of N4513 and N4533, and which parts are replied by ~~N4446~~ N4556, to clarify the issues to be discussed in future. Also a few notes on Nushu Duben are supplied.

2. WG2 N4513

WG2 N4513 is basically a list of questions suspecting the stability and the reliability of the statistics used PDAM1 charset.

2.1 Questions on Numerical Errors in the Statistics

Section 1.1: **The introductory part has already numerical error.** The introductory part describing the statistical method has already error. It was already found by Endo in 2009 and asked why, but no reply.

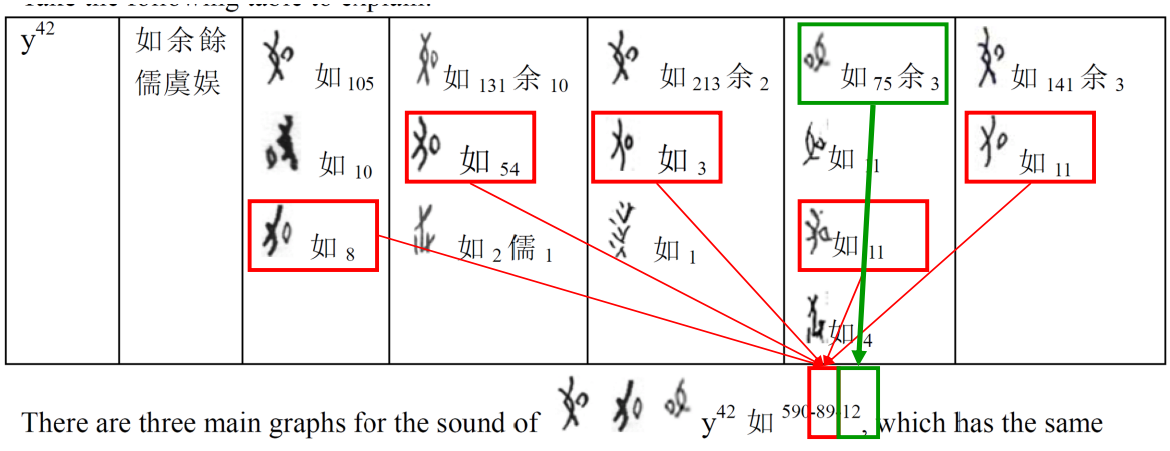

Figure 1: Sample of Elementary Mistake in the Frequency Counting (WG2 N4513)

Section 1.2: **The statistic value is changed during 2007 and 2009, is it stable?** The statistic values are changed between the proposal in 2007 and that in 2009. Considering that the number of Nushu native users is not increased, it is difficult to think the difference is caused by the introduction of new materials.

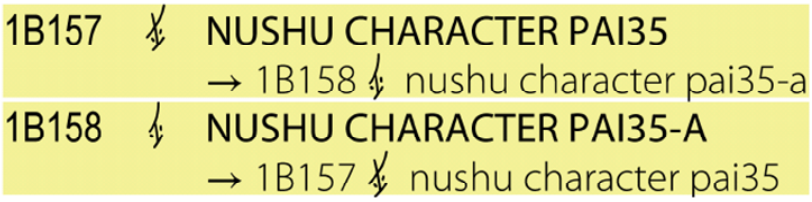

Figure 3: Variants of PAI35 in PDAM (WG2 N4484)

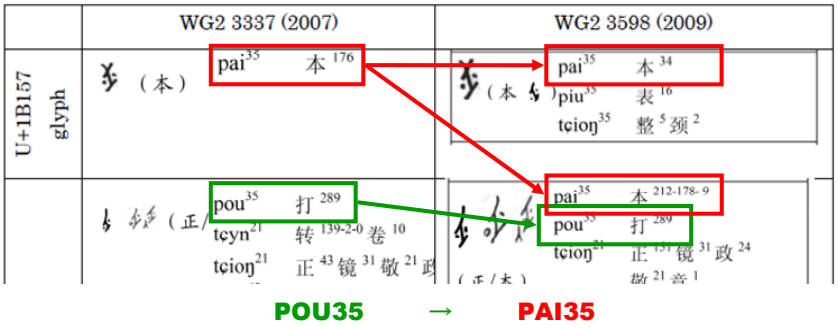

Figure 2: Sample Case that Unstable Frequency Changed Representative Phonetic Value (WG2 N4513)

Section 1.3: **The characters for daily-used words are not found in the statistics.** The occurrences of the characters like urine, charcoal, rice field, wine, are less than 10 in the collection of 220,000 characters.

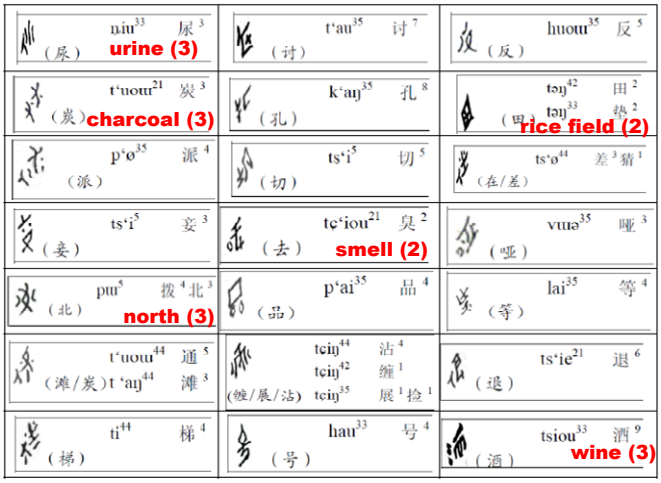

Figure 3: Sample of Infrequent Usage of the Daily-Used Words (WG2 N4513)

Section 1.4: **Some representative glyphs are not the most frequently used.** Some representative glyphs are chosen by the statistics on 2007. By the statistics on 2009, the most frequently used shapes are different. Furthermore, some glyphs are not find the statistic table at all.

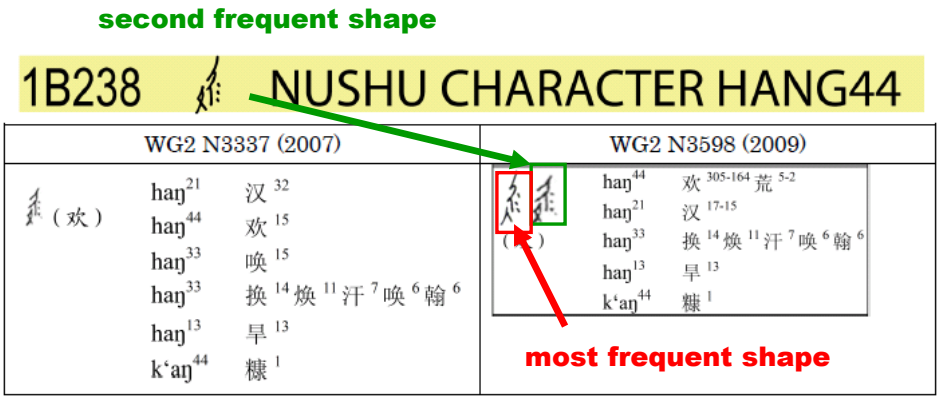

Figure 4: Sample of the Case that Most Frequently Used Glyph is not Used (WG2 N4513)

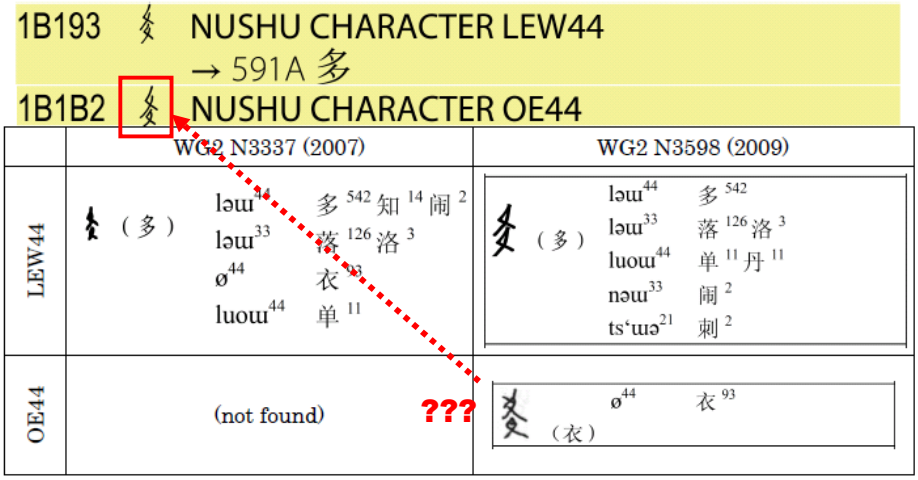

Figure 5: Sample of the Representative Glyph Missing in Statistics (WG2 N4513)

2.2 Questions on the Stability of Stroke Counting

Section 2.1: Stroke counting without unification rule is difficult to use. The rule to count the stroke is briefly described, but no rule to unify/disunify the characters with same phonetic value and semantics is given, thus difficult to determine "this character is not found in the representative glyph in the chart but unified" or "this character is not coded yet".

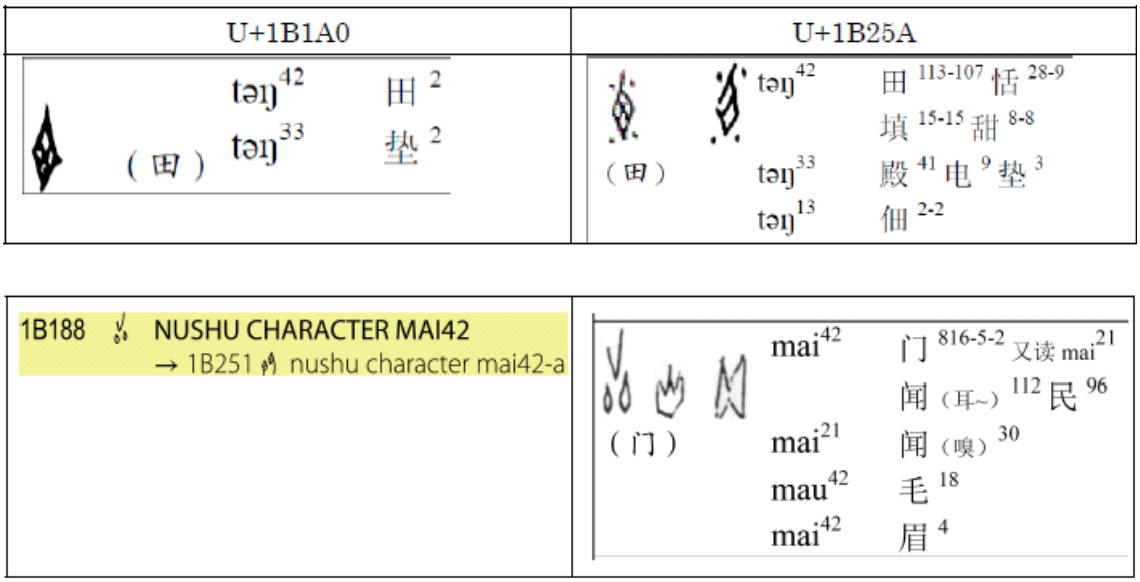

Figure 6: Sample of Same-Semantic-But-Disunified-By-Shape and Unified-But-Significant-Different-Shapes (WG2 N4513)

Section 2.2: Stroke counting is stable? Some glyphs are assigned to the stroke number which seems to be different from the number counted by the described rule. If the author of the proposal makes a mistake in stroke counting, the users may have more difficulties to find the glyphs they are looking for.

3. WG2 N4533

WG2 N4513 is a list of the questions and almost no proposal to solve the difficulty. After the submission of WG2 N4513, I and Endo discussed about the possible solutions. WG2 N4533 is a memo of the discussion, about

- How to choose a stable representative glyph
  the frequency evaluation without the consideration of the author is not good idea, choosing one author is better for the consistency.
- How to give a stable name for the characters
  the inclusion of the phonetic value is not good idea, because the statistics is unstable, and the author-dependency is not fully considered.
- How to sort the characters for users/researchers convenience
  sorting by the shape similarity would be useful. Also a serial number in some referential material is considerable option, although it would not be so easy.

Also, PDAM1 charset (380 chars) and He Yanxin glyph list (467) were compared, with the consideration of the unification (the comparison was not simple 1-by-1 check). The cross section was 317 chars, 63 chars are only found in PDAM1 charset, and 150 chars are only found in He Yanxin glyph list. The unification rule to minimize this gap should

be defined. At present, PDAM1 charset could not cover the Nushu text, even if we restrict our scope to the material by only one author (as Zhang agrees, the variety of He Yanxin glyph is not the largest one).

3. Table of Unmappable He Yanxin Glyphs (150 Glyphs)

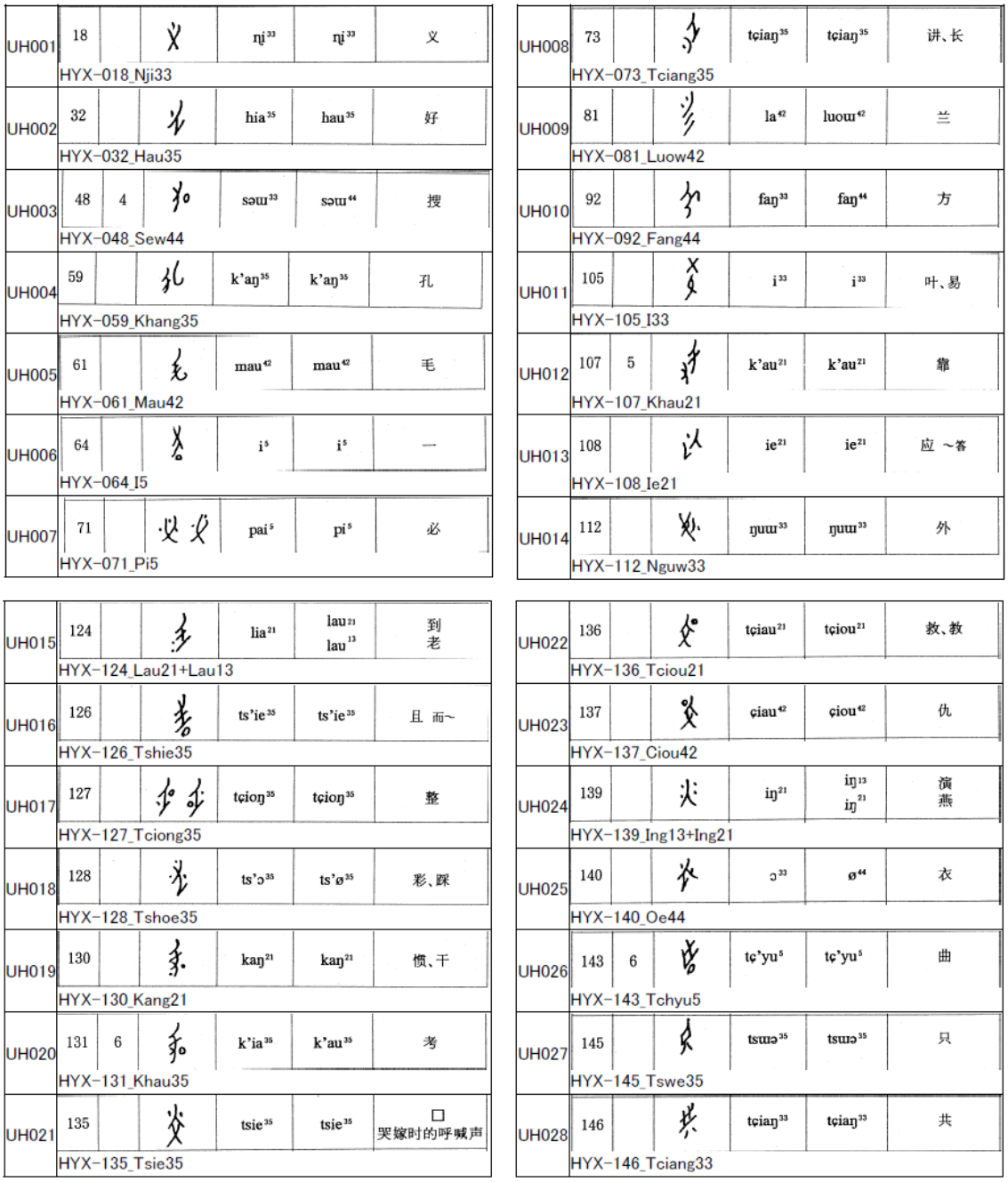

Figure 7: Excerpt of He Yanxin glyph which is not found in PDAM1 charset (WG2 N4533)

4. WG2 ~~N4446~~ N4556

This document is titled as "Reply to N4513 and N4533 (Suzuki and Endo's comments on

Nushu Encoding)", but the content deals only the ideas in N4553. The questions in N4513 are not answered at all. And, some replies seem to be based on the misunderstandings of the point of the questions.

- Representative glyph

  ~~N4446~~ N4556 replies the representative glyph is chosen by the frequency. N4513 had already given the questions the reliability of the frequency in the previous submissions, but no answer for the questions. In addition, "Nushu Yongji Bijiao" is published on 2006, thus, the submissions in 2007 and 2009 are post-survey submissions, but the statistical values are different. There is no explanation why we could choose the most frequently use glyph is stable today.

- Collation

  ~~N4446~~ N4556 replies how PDAM1 chart was compiled briefly, and no comment about the advantage / disadvantage between the PDAM1 chart collation and the idea proposed by Endo. N4513 had already concerned the insufficient description of the rule is a barrier to identify the character, but no answer is given.

- Naming

  ~~N4446~~ N4556 replies that the naming issues are discussed in preceding WG2 meeting, and no necessary to discuss again. It means that the raised questions (the statistics seem to be unstable and the most frequently used phonetic value would be unstable either, etc) are already discussed and resolved to ignore the instability? Or, it simply says as "it is too late to discuss again". There is no clarification of the essential requirement to include the phonetic value in the character names. Also I should emphasize that the numerical error in the statistics was already reported by Endo on 2009 (WG2 N3705), and there was no reply. I think there was sufficient time to reply, or issue a corrigendum.

- Source

  N4513 and N4533 assumes that the PDAM1 charset is primarily referring Nushu Duben, and the content is different from Nushu Yongji Bijiao (and the difference would be caused by the errors in the summarization of the frequency count from Nushu Yongji Bijiao to WG2 submissions). ~~N4446~~ N4556 replies without the name of Nushu Duben, but says as if PDAM1 charset is based on Nushu Yongji Bijiao too. Does it mean that the switching the referential material from Nushu Yongji Bijiao is acceptable?

5. Additional Comment on Nushu Duben

After the submission of WG2 N4513, Endo is researching how Nushu Duben is widely used in the region where most Nushu users are living. Unfortunately, yet she has not

got any information about the educational use of Nushu Duben. Therefore, at present, I have a concern to classify the Nushu Duben as the primal reference of the standardization of Nushu. Maybe Nushu Duben was the latest snaphot of Zhao Liming's survey when China updated their submission on 2009, but it does not mean that Nushu Duben is the final stabilized report of her survey. By the progress of her research, the statistics would be changed in future.

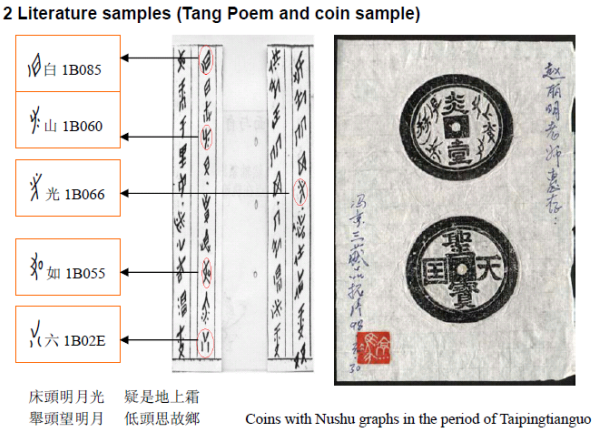

Figure 8: The coin image used in the submissions after Nushu Duben (WG2 N3463)

Endo had ever concerned that the submissions since 2008 (after the publishing of Nushu Duben) include the inappropriate pictures of the coin(s) of Nushu (see N3463 p. 23). The submission tells as if the coin is the evidence of the official usage of Nushu in Taipingtianguo (1850-1864). But the coin was quite controversial (all known comments by the experts were saying it was forged). As Endo wrote in WG2 N3705, Zhang Tie Bao (张铁宝), the expert from Taipingtianguo museum in Nanjing concluded the coin was a forged object. But WG2 N3426 wrote as if he concluded the coin was reliable evidence. Recently Zhang found the document and criticized the wrong conclusion about the coin. Considering such background, although the inappropriate picture itself is not the part of the international standard, referring Nushu Duben as the primary reference would not be good idea.

(end of document)