ISO/IEC JTC1/SC2/WG2 N4639
Date: 2014-09-23

Title: Theory and Rules of Nushu Character Unification
Source: China
Action: FYI

Nyushu Character Unification——Theory and Rules

Liming Zhao (Tsinghua University)

In recent years, Nyushu has been a spotlight for both research and public. Due to the possible economic benefits and academic reputation, some people make up many fake characters and use them in different situations. This poses an even worse effect to the now endangered scripts because fake characters may be misleading to more scholars and the public who really want to study Nyushu. Therefore, it is essential to verify the original characters from fake ones in order to do further research on Nyushu in the future.

Professor Liming Zhao and her group of Nyushu rescuing project from Tsinghua University collected and organized around 1,000 original documents written in Nyushu in total and translated 640 readable files from them character by character into Chinese. The scanned pictures of these 640 files and their 220,000-character translation in Chinese are published as the 5-volume ***Chinese Nyushu Collection*** by Zhonghua Book Company in 2005.

Definitionand Description of Nyushu

Nyushu is exclusively used by peasant women around Jiangyong County, Hunan Province, China in Xiao River Basin. To talk about the real miseries in life is the main social function of Nyushu to its writers. Nyushu characters are written in italic, which look like the Chinese Hanzi "多". These characters are derived from Chinese Hanzi by their appearance. Nyushu is a syllabic writing system. Each writer usually uses 4 to 5 hundred characters in total which are enough to keep down all the syllables in local Chinese dialect spoken by these women. Therefore, the local Jiangyong dialect can be well recorded by two writing systems: Chinese Hanzi used by men, and Nyushu used by women.

There are three main features of Nyushu:

1. Nyushu is a folk script used by peasant women. It is a writing system originally created and used by foot-bound women in rural areas. Therefore, it has internal linguistic and literary rules which cannot be coined by other people outside their community. This is the basic

principle for any research on Nyushu.

Although Nyushu is only used by women, it is an open, flexible and positive script rather than cipher and mystery. Local women who write Nyushu live a normal peasant life in China (different from the rumor that there was an isolated and mysterious “female society”). That women read and sing Nyushu poems written in paper or paper fans is an open act. Men know it too but they didn’t use to pay any attention to it.

2. Nyushu is a phonographic script that records the local Chinese dialect. One character is used to keep down several same or similar syllables. (Zhao, 1987, 1990, 2004). It takes only several hundreds of unique characters for a person to write down all the syllables in spoken language. Nyushu is a syllabic phonogram. The fact that one Nyushu character may have several Chinese Hanzi alternatives is one of its most important feature. Therefore, the current Nyushu writing system which consists of several hundreds of characters are enough for use. It is unscientific and irresponsible to coin fake characters and even collect them into a dictionary.

3. Nyushu is a variation of Chinese Hanzi. To be specific, it is a variation of Hanzi regular script[1]. Based on the 350 characters used in the old Nyushu works by anonymous writers in Ming and Qing dynasty[2] and the 300 characters used in Nyushu works by YANG Huanyi, 95% percent of them are derived from Chinese Hanzi.

Grapheme theory——theory for character unification

Grapheme theory (character-based theory)

Grapheme is the smallest unit that can bring about a change of meaning in writing systems. It is the basic sematic unit in written languages which is the smallest contractive unit recognized by the whole community. Therefore, grapheme theory is based on and represents the social perception of scripts.

Grapheme is analogous to phoneme in spoken languages. Grapheme theory is a practical theory for writing systems in which characters may have several variant forms. It is called for during the process of dealing with variants of one same character. It is one of the basic and universal theory in terms of philology.

Grapheme is defined as the smallest distinguishing unit in semantics. One grapheme may contains several different characters which are used to represent the same meaning, and these characters are considered as variants of one grapheme. For example in Chinese Hanzi, character

[1] Regular script(楷书) is one of theHanzi script style that originated in around AD 200 and was recognized as the official style since AD 700. It is the only official style of Hanzi writing in all countries where Chinese characters are used.

[2] AD 1368-1912.

回、囲 and 囘 have exactly the same meaning thus they are of one same grapheme. The different ways of writing are caused in the historical transformation of Hanzi writing system. During the transformation, character 回 turned out to be the most popular one and hence the canonical character of this grapheme, while other two characters are perceived as variants.

Grapheme theory in the process of Hanzi characters

In the history of Hanzi, abundant experience and knowledge were accumulated on variant characters and the process of them. However in an information age, coding of canonical characters is essential to this issue. Luckily, we already have much experience on variant characters, especially on different characters used in Chinese mainland, Hong Kong, Taiwan, Japan and Korea. For example, GB2312 code, the official and standard code table for simplified Chinese characters used in Chinese mainland, was issued by Standardization Administration of China (SAC) in May 1, 1981. This standard is also adopted in other countries and regions like Singapore.

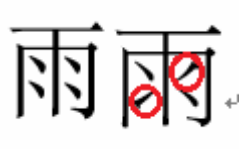
With more and more international communication and data exchange nowadays, different codes used in different areas has been a barrier to efficient communication. There are more and more files written in multiple languages, which makes it impossible to record information from all over the world using a simple code point. Therefore, the standard Unicode is introduced to address the problem. Unicode is an international coding standard for the universal coding of multiple languages.

Hanzi characters in the latest CJK (China-Japan-Korea) code points in Unicode are picked from the “character source” in several countries and regions. Some main rules of unification (unifying same character/ variant character from different sources into one grapheme) and selection are as followed:

a) Characters that have different Hanzi origins (unrelated in their historical transformation) shall not be unified. for example, character 土 and 士 have different origin and transformation history, therefore they are not variant forms of one grapheme and shall not be unified.

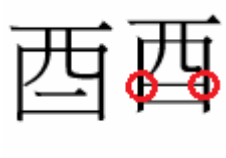
Characters on conditions below are considered as variants (share one grapheme and shall be unified):

b) Different direction of points, e.g. 雨,雨

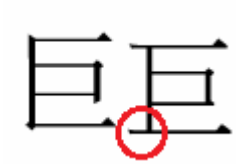


c) Whether two strokes are touched or crossed, e.g.不, 𣎴

d) Whether two strokes are parted or touched, e.g.酉, 酉



e) Whether a stroke is divided into two at a turn, e.g. 巨,𠀋



f) Whether a stroke is straight or has a turn, e.g. 西, 𠂢

g) Whether a long vertical stroke has an ending hook, e.g. 朱, 𠂆

h) Whether a long horizontal stroke has a starting dot, e.g. 丈, 𠂇

i) Whethera leaning stroke has a starting dot, e.g.𠂆， 八

j) Any combination of the rules above.

Besides, characters come from more authorized dictionaries are preferred. The priority of dictionaries comes as: ***Kangxi Dictionary, Chinese-Japanese Dictionary, Chinese Dictionary, Character Origin*** and etc[3].

Variant characters in Nyushu

However, variant characters in Nyushu are different from that in Hanzi. Since Hanzi is a logogram, variant characters of one grapheme in Hanzi have the same pronunciation and meaning but different strokes. Although Nyushu is originated in Hanzi, it is a phonogram in which every character is a symbol of a set of similar syllables in local dialect. People cannot know the exact meaning of each Nyushu character until it is read out in context.

Another difference between Hanzi variants and Nyushu variants is that in Hanzi variants are caused by different dialects while in Nyushu, variants are caused by different characters used by different writers. Since each Nyushu character represents several similar syllables, one syllable in local dialect may have several equivalent Nyushu characters which are preferred by different writers. Therefore, one syllable may be wrote down by several variants. With the definition of grapheme, our group clustered characters as variants of one grapheme according to rules below:same meaning; same Hanzi origin; same basic structure; same perception by different writers. For variant characters of one grapheme, we usually use the most frequently written character as the canonical character of this cluster.

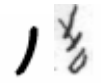
Based on our analysis, there are four kinds of variant characters in Nyushu:

a) Different Hanzi origins. Since Nyushu is derived from Hanzi, a writing system that is rich in homophonies and variant characters, Nyushu characters that have same or similar

[3]《康熙字典》《大汉和字典》《中华大字典》《大字源》, according to ***UTF-8 character table UCS*** ISO/IEC 10646:2003IDT issued by standardization administration of PRC.

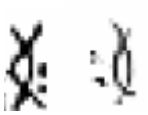
pronunciation may be originated from different Hanzi characters. Those variants in Nyushu have no semantic difference in practice and writers usually can replace them by each other freely. For example.

我[ŋu42]/[ie13] 

一[i5] 

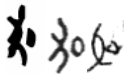
For this kind of variants, we consider them as two graphemes. Variant characters belonging to this type are not very common, and take up only 1.5% of the 398 characters. Although they are considered as two graphemes according to grapheme theory and our rules, these characters have no functional difference in use and are freely chosen by writers. Even one writer will use characters from both graphemes to make their writing diverse.

b) Different ways to write a curve. A curve in Nyushu can be both written as one strokes or two. And this kind of writing difference makes no semantic change to writers. For example:

娘[ŋiaŋ42] 

刻[kʰɯ5] 

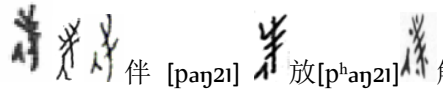
c) Different side particles. Some characters have same main particle, basic structure and meaning but different side particles (usually dots or small circles). These characters have same meaning and function to Nyushu writers. For example:

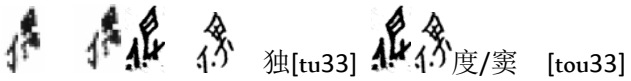
如[y42] 

离 [la42] 

鸟 [liu35] 

d) Similar writing. Some characters have similar structure or appearance so that people sometimes use them as same characters in context. For example,





Methodology

Data

Based on several years of field work, our group collected and translated more than 90% of all the existing documents written in Nyushu. These materials add up to 640 articles and 220,000 characters in total which are published as ***Chinese Nyushu Collection*** in 2005.

In the process of writing ***Chinese Nyushu Collection***, our group wrote ***Nyushu Character Table***, ***Comparison of Nyushu Characters***, and built a database for Nyushu characters. The ***Canonical Nyushu Characters and Their Hanzi Origin Verification*** based on the ***Collection*** and ***Character Table*** is therefore a trustful Nyushu dictionary. In the ***Verification***, counts of every character we found in the ***Collection*** as well as the Hanzi origins of each Nyushu character are included, which are very useful for both philological and historical linguistics studies. Here are some rules of how data used in ***Canonical Nyushu characters and their Hanzi Origin Verification*** are collected:

1) Character. To be as authentic as possible, characters proposed were taken from old books written by anonymous writers which may date back to early Qing dynasty[4]. These handscripts add up to around 30,000 characters. There are several reasons why we use characters from old files rather than others. Firstly, they are written naturally in a female community which wasn't affected by outside world and a pursuit of benefit. Secondly, they are classical documents written by Nyushu masters which only used the most canonical characters in original Nyushu.

2) Pronunciation. Nyushu is used to record local dialect——the Jiangyong dialect. Although dialects differ a little in different villages where Nyushu is used, Nyushu is read according to the local "traditional Chinese"——the dialect spoken in the town instead of dialects in different villages. Therefore, the pronunciation used were from dialect in downtown of Jiangyong County according to Study in Jiangyong Dialect by Huang (1993).

3) Hanzi origin. Based on their relationship with Hanzi origins, Nyushu characters were created through three basic means: glyph borrowing, transformation and reproduction. Hanzi origin of every Nyushu character is included in the ***Verification*** although for 4 characters, their

[4]AD 1644-1840

origins remain unclear.

We also referred to ***Character Table of YANG Huanyi*** when compiling the dictionary. Yang (1909-2004) was the only officially recognized inheritor of Nyushu script when our group conducted the field work. Yang couldn't write Hanzi so her writing wasn't affected by later Hanzi writing. Therefore, her works are also considered as an authentic source for academic study.

Unification rules

According to grapheme theory, Nyushu characters that meet the standards below are considered as variants of one grapheme:

1) Same Hanzi Origin;

2) Same or similar structure;

3) Same pronunciation;

4) No semantic difference.

Homophonies that satisfy the rules above are considered as variants of one grapheme. Therefore, it is clear that in the four types of variant characters discussed above, characters in type a) are considered as different graphemes while other three kinds of variant characters are variants of one grapheme.

With the standards above, we clustered 220,000 characters' raw data into hundreds of Nyushu grapheme.

Canonical characters

After the unification, our group also define the canonical character of each grapheme according to their frequency. The chart below is a part from Comparison of Nyushu Characters that showcased how a canonical character is chosen.

For example, for the syllable [tɕie44], the most frequently used glyph is  , therefore it is defined as canonical character by us. The least frequently used glyph is  , which is considered as a variant. Other glyphs like  and  are temporarily used because of their similar pronunciation. It is better to perceive them as writing mistakes instead of variants.

Pronunciation	Hanzi origin	Hanzi translation	Nyushu characters				
			Anonymous	GAO Yinxian	YI Nianhua	YANG Huanyi	HE Yanxin
tɕie44	珍/金	金真襟针斟徵蒸珍巾贞侦斤筋	金 5 真 46 襟 47 金 5 真 40 襟 38 针 16 怎真 1	 金 156 真 110 襟 20 针 24 斟 4 徵 3 珍 16 斤 11 筋 4 金 8 襟 5 斤 1 针 1 金 1 襟 2 真 1 蒸 1	 金 85 真 82 襟针 13 斟 122 徵 9 珍 8 贞 4 斤 11  斟 1	金 48 真 43 襟 11 针 13 珍 1 巾 4 斤 1  金 19 真 34 襟 4 针 7 珍 3 斤 2  金 1 真 2 真 1	金 75 真 73 襟 10 针 16 斟 4 珍 4 巾 6 斤 10 真 1
suoɯ	山/生	山生牲笙甥衫	山 3 生 13 山 3 生 9 牲 1 生 10 山 1 生 9甥 1	 山 123 生 213 牲 10 笙 6 甥 30 山 2 山 1 生 3 生 1	 山 32 生 298 牲 2 笙 1 甥 27  牲 1	 生 123 笙 3 甥 1 山 38 山 5  山 3	山 63 生 128 牲 3 甥 8 衫 1 山 30 牲 1 衫 1

Results

The main results of Nyushu character unification and canonicalization are as followed:

1. According to the analysis of 220,000-character raw data in ***Chinese Nyushu Collection***, we concluded that each writer used about 500 distinctive characters (variants included). After the unification and canonicalization, 398 canonical characters are clustered finally. Details of characters and documents written by each author are listed below:

	Number of canonical characters	Number of characters
Anonymous	358	34800
GAO	334	62100
YI	362	49700
YANG	304	36000
HE	380	39600
TOTAL	398	220000

2. According to 35,000 characters used in 110 articles in ***Collected Works of YANG Huanyi***(2004) and ***Character Table of YANG Huanyi***, Yang used 437 distinctive characters (variants included). After the unification, there are 304 graphemes in her works.

3. Japanese professor Endo proposed 466 characters (variants included, of which one character repeated for twice so there should be 465 characters in total) from the 27,000-character works by HE Yanxin. After comparing them from the 39,000-character works by He that we translated in Chinese Nyushu Collection, we concluded that 380 graphemes were used in He's works.

Based on the results above, we are able to draw the conclusion that Nyushu is not a complex writing system. As long as a peasant woman is able to speak the local dialect and masters 300 to 400 phonetic symbols——the Nyushu characters, she will be able to keep down the local dialect, write about her thoughts and even translate literature works written in Chinese. That is the real practical usage of Nyushu characters.

Discussion

As an intangible cultural heritage, it is urgent for linguists and scientists to work together to protect the endangered script. Therefore, conclusions based on abundant field work, authentic materials and scientific theory are essential for further studies. Made-up or fake materials will do nothing but to disturb our efforts to protect it.

In an Information Age, coding of characters is a very important way to protect endangered script. However, variants have been main barrier during the coding of character-based writing systems. For these characters, the grapheme theory proved to be very useful especially when dealing with writing systems that consist many characters and individual writing preference. Therefore, it can be used as a universal theory on character-based script.

In terms of Nyushu, despite that we conducted an intensive study in it, there are still many problems remaining. Further studies on phonetics, philology and even computational process are expected on this topic.

Reference

[1] 陈瑾．试析女字形音义的特点[J]．妇女文字和瑶族千家峒，1986．

[2] 赵丽明．女书造字法初探[J]．妇女文字和瑶族千家峒，1986．

[3] 《女书的文字学价值[J]．华中师范大学学报，1989,6，转载《新华文摘》1990．

[4] 黄雪贞．江永方言调查[M]．社会科学文献出版社 1993．

[5] 赵丽明、李蓝、唐功炜．江永土话的内部差异与女书的规范读音[J]．汉字的应用与传播，华语教学出版社,2000.

[6] 曹志耘、赵丽明．从方言看女书[J]．中国社会语言学，第 2 期，2004．

[7] 李蓝.《从语言学的角度研究女书》载《女书的历史与现状》，中国社会科学出版社，2005．

[8] 赵丽明.《女书基本字与字源考》．中国社科院女书国际研讨会，2004．后载《女书用字比较》知识产权出版社，2006．

[9] 中国女书合集[M]．中华书局，2005．

[10] 女书用字比较[M]．知识产权出版社，2006．

[11]女书字数统计与异体字处理[J]．女书用字比较，2006．

[12] 李蓝．论江永女书的来源[J]．汉藏语言研究，2006 年 3 月．

[13] 国家质监总局．信息技术通用多八位编码字符集 UCS．ISO/IEC10646 :2003IDT 报批稿．

[14]《汉字编码问题》http://www.china-askpro.com/msg6/qa41.shtml 等．

[15] 全如璊．《术语的理论与实践》引言 http://www.cnterm.com 中国术语信息网．2001-10-26.

[16]赵丽明．《阳焕宜女书基础字考》载《汉字传播与文化交流》，国际文化出版公司 2004 年 9 月.

[17]远藤枝织《中国女文字研究》，日本明治书院 2002 年.