

DATE:94/08/26

To: Experts of GB_WG, CJK_IRG,
ISO-IEC JTC1/SC2/WG2,
Unicode Inc. and

Whom it may concern

From: Zhang_Zhoucai 张轴材

Former Chief Editor & Convenor of
Editorial Working Group, CJK-JRG

TEL: (861)847.1418 (O)

(861)841.6971 (H)

FAX: (861)841.1360

Address: 54,Xisanhuan Bei Rd. Beijing 100044

Subject: Vertical Unification of CJK Ideographs

中日韩汉字的垂直认同

Status : Individual Contribution

=====

The attached paper was presented and discussed at the 1st International Symposium of The National Language Research Institute, Japan, held on 20-21 of January, in Tokyo.

《中日韩汉字的垂直认同》一文曾于今年一月在东京举行的日本国语研究所第一届国际会议上宣讲并讨论。

In request of my ex-colleagues in CJK-JRG, I would like to submit this paper to CJK-IRG, ISO/SC2/WG2 and Unicode Inc. for reference and discussion as the topic is somehow related to the mission of those organizations.

应前CJK-JRG的老朋友的要求, 现将此文呈送CJK-IRG, ISO/SC2/WG2 和 Unicode 供参考与讨论, 毕竟该课题与上述组织的使命有这样那样的联系。

Lee Collins (Taligent), gave significant assistance in English expressing of this paper. He proposed to use " Vertical Identification " instead of " Vertical Unification " in order to avoid possible confusion. This suggestion is reasonable and acceptable. It is only because of the consideration for symmetry of Horizontal Unification and Vertical Unification, the term of Vertical Unification remains unchanged. Nevertheless, the

Identification is more precise to expresse my intention in this paper.

Taligent的康立论曾对本文的英文表述给予重要帮助。他曾建议使用 " Vertical Identification " 而不要用 " Vertical Unification " 以避免混淆。作者认为他的建议是中肯的、可接受的；但 仅是由于对称性的考虑 (Horizontal Unification / Vertical Unification) ,该措辞才没有改变。尽管如此，作者仍认为 Identification 一词能更加确切地表达本文的意图。

During the meeting in Tokyo, the following points were discussed and/or emphasized:

在上述东京会议上，曾讨论或强调过以下诸点：

1. No code change with ISO 10646 or Unicode shall be made due to the Vertical Unification or Identification.

垂直认同不会引起 ISO 10646 或 Unicode 的代码改变。

2. Rules is needed to avoid over_unification (such as 一一壹) or under_unification if the link between 联/GB-連/JIS is missing).

The unification will be done mainly at script level, while the unification at semantics level is inevetable because of the context dependency.

垂直认同的规则确有必要，以避免认同过头（如把“一一壹”认同）或认同不足（比如，忘了把“联/GB-連/JIS”认同）。

垂直认同主要在文字一级进行，但由于语境相关性，语义一级的认同也是不可避免的。

3. In addition to the 6 pairs of tables (see Page.12), below 4 essential tables should be established first:

G <--> G (within GB)

T <--> T (within CNS)

J <--> J (within JIS)

K <--> K (within KSC)

which were omitted in my attached paper.

除了第十二页的六对表格之外，下列四个基本表格应当首先建立：即在 GB、CNS、JIS、KSC 之内先进行垂直认同。这一点在文中忽略了。

4. The usage of the Vertical Unification for multi-entry information retrieval in multi-lingual environment was indicated.

For example, different readers may successfully retrieve up the papers/books written by 松岡榮志, no matter he/she keys in 松岡榮志, 松岡榮志, or 松岡榮志, if the vertical unification is built in the software system. Of course, 松岡榮志、松岡榮志、or 松岡榮志 will be considered as strange and invalid string.

垂直认同用于多文种环境中多入口信息检索, 在会议中亦曾论及。比如, 不同国家/地区的读者键入松岡榮志, 松岡榮志, 或松岡榮志都可以成功地检索到松岡榮志撰写的文章或书籍, 只要垂直认同的机制嵌入了软件系统; 当然, 松岡榮志, 松岡榮志, 松岡榮志则会认为是奇怪的非法字符串。

=====

Zhang, Zhoucai 張軸材

DATE:93/12/10

CCID: Research Center of Computer & Microelectronics Industry Development

LIDI: Languages Informatics Development Inc.

54,Xisanhuan Bei Rd. Beijing 100044, China

中日韓漢字的垂直認同 Vertical Unification of CJK Ideographs

Content

1. Introduction and Definitions
 - 1.1 Horizontal Unification (HU) of CJK Ideographs
 - 1.2 Vertical Unification (VU) of CJK Ideographs
 2. Scope and Limitations
 3. Classification of Vertical Unification
 - 3.1 Homonyms 同形異義字
 - 3.2 Z_Variants 
 - 3.2.1 Pure Z_Variants 純 Z_變體
 - 3.2.1 Quasi Z_Variants 準 Z_變體
 - 3.3 Orthographic vs Variant Ideographs 正體與異體
 - 3.4 Old_glyph vs New_glyph 新舊字形
 - 3.5 Simplified vs Unsimplified Ideographs 簡體與繁體
 - 3.5.1 One to One relations 一對一關係
 - 3.5.2 One to Many relations 一對多關係
 - 3.6 Proposed VU Classification
 4. Characteristics of VU
 - 4.1 Language Context Dependency 語境相關性
 - 4.2 Direction Dependency 方向相關性
 - 4.3 Coupling Intensity 耦合強度
 5. Case Study - An Implementation of VU
 6. Conclusion : VU Rules Needed
-

1. Introduction and Definitions

1.1 Horizontal Unification (HU) of CJK Ideographs

【Definition】 Horizontal Unification of CJK Ideographs refers to the task that has been done by CJK-JRG according to its Unification Rules, the output of which is the CJK Unified Ideographs placed in the Basic Multilingual Plane of ISO-IEC 10646.

The Chinese Hanzi/Japanese Kanji/Korean Hanja are listed in four columns on the code table. If unified, then they have the same ISO code and appear in the same horizontal line:

I S O	C	J	K
HexCode	G-Hanzi-T	Kanji	Hanja
4E00	一	一	一
4E01	丁	丁	丁
:			
4E0E	与	与	与
:			
4E30	丰	丰	丰
:			
58F9	壹	壹	壹
:			
8207	與	與	與
:			
8C50	豐	豐	豐
:			
91D8	釘	釘	釘
:			
9488	釘		
:			

In other words, the above mentioned unification is based on the so-called X/Y/Z module, where X is the MEANINGS (the Signified), Y implies GLYPHS (the Signifier) and Z represents the Variants having minor differences of stroke forms. The unification was conducted along the Y axis.

In order to vividly depict the unification, we call it Horizontal Unification (水平認同、橫向認同), which is abbreviated to "HU".

As shown in the codetable, the HU is done mainly but not absolutely by SHAPE. On one hand, even some ideographs have the same meaning(s) and the common generic shapes with just very minor difference in their stroke forms (such as 丢/丟) are NOT unified due to the "Source Separation" of the CJK Unification Rules; On the other hand, the non_cognate ideographs are not unified either, although they may have quite similar shape (such as 士/士). Apparently, the meaning of the ideograph was taken into consideration while the HU was being done.

In fact, although the HU is glyph_persistent, the horizontally unified CJK ideographs statistically signify the same, or the similar, or the related meaning(s) with only very few exceptions.

1.2 Vertical Unification (VU) of CJK Ideographs

【Definition】 The Vertical Unification of CJK Ideographs is defined as the one which is based on the meaning (or identity) of ISO-IEC 10646 ideographs, in contrary of HU, which is primarily shape based.

VU is the abbreviation.

VU may also be considered as the extension and the complement to HU. On the code table, VU would be rendered as the up/downward pointers linking different codes to each other, which is shown as below:

	ISO HexCode	C G-Hanzi-T	J Kanji	K Hanja	
?	4E00	一	一	一	Vertical Unification
	4E01	丁	丁	丁	
	:				
	4E0E	与	与	与	
	:				
	4E30	丰	丰	丰	
	:				
	58F9	壹	壹	壹	
	:				
	8207	與	與	與	
	:				
	8C50	豐	豐	豐	
	:				
	91D8	釘	釘	釘	
	:				
	9488	釘			
	:				

The VU is expected to be conducted mainly, but only at an extent, by meaning(s) of CJK Ideographs. The VU 一 ↔ 壹 would be a kind of "over-unification", while 丁 ↔ 釘 seems linguistically (in Japanese) reasonable, but technically unnecessary.

One thing should be emphasized in advance, that VU does not imply any intent to re_unify or re_encode the CJK Ideographs. Everything with VU shall be done based on the HU result without any code change.

2. Scope and Limitations of VU

The VU discussed in this paper might be applicable for ISO-IEC 10646-1/CJK Ideographs in its applications for glyph normalization, and/or character repertoire converging within a language context, or conversion / translation among different language contexts.

To facilitate the discussion and to focus on the most useful CJK Ideographs in information technology, this paper takes a subset of CJK Ideographs that covers the ideographic characters in the following source sets:

- | | | |
|--------|------------------------|----------------|
| (1) G0 | GB 2312-80 | used in PRC |
| (2) T1 | TCA-CNS11643/1st Plane | used in Taiwan |
| (3) T2 | TCA-CNS11643/2nd Plane | used in Taiwan |
| (4) J0 | JIS x 0208-90 | used in Japan |
| (5) K0 | KS C 5601-87 | used in Korea |

As a result of subsetting, the matrix to be dealt with will become more sparse like this :

I S O HexCode	C G-Hanzi-T		J Kanji	K Hanja
4E00	一	一	一	一
4E01	丁	丁	丁	丁
:				
4E0E	与	与	与	
:				
4E30	丰	丰	□	
:				
58F9	壹	壹	壹	壹
:				
8207	□	與	與	與
:				
8C50	□	豐	豐	
:				
91D8	□	釘	釘	釘
:				
9488	釘			
:				

Note: The □ means that the ideograph is out of the subset.

Moreover, in many cases, the characters in K0 are the same as the ones in T1 or T2, hence, the K column is omitted on some diagrams for simplicity in this paper.

G, T, J and K are also denoted as respective contexts in this paper, besides those character sets.

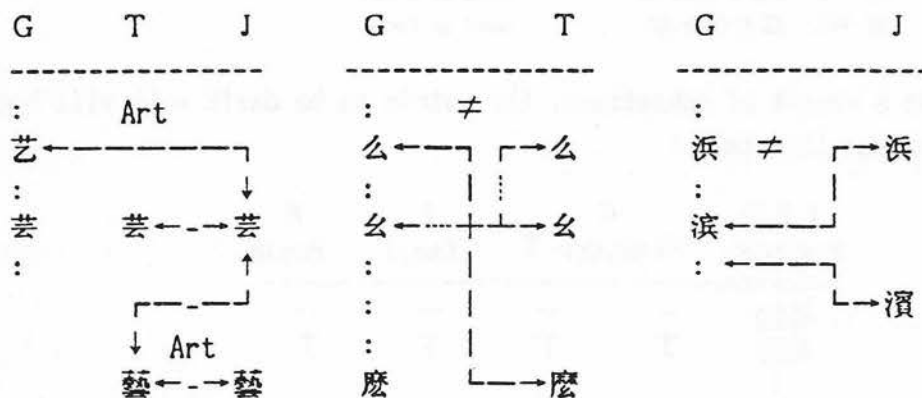
3. Classification of Vertical Unification

3.1 Homonyms 同形異義字

A few horizontally unified CJK Ideographs are to be separated vertically by their meanings.

Generally speaking, VU is to link more ideographs meaningfully. However, a few exceptions indicate a noticeable diversity of languages, which leads somewhat of splitting instead of unification of CJK ideographs.

Below are three typical examples :



Nevertheless, it should be emphasized again that

- ① they constitute only extrimely small percentage in the CJK Ideographs repertoire ;
- ② but they are most frequently used in modern CJK languages, to which attention must be paid.

3.2 Z Variants Z 變體

【Definition】 Z_Variants refer to those ideographs that have identical meaning(s) and the same generic shape but they are encoded in different cells in ISO 10646 because they had existed in a source character set as separate code points ("Source Separation").

Or, Z_Variants are those separately coded ideographs which should have been unified in CJK Ideographs if only the generic shape was in consideration .

3.2.1 Pure Z Variants 純 Z 變體

【Definition】 The Z_Variants having the same number of strokes.

Many of the Pure Z_Variants are listed below, but they are **NOT** exhausted:

兑	丢	尔	内	凜	刊	别	呐	囟	匀	吞	册	画
兌	丟	尔	内	凜	刊	別	呐	囟	匀	吞	冊	画
娛	户	彦	悦	术	查	毀	禿	稅	脫	蛻	稟	吳
娛	戶	彦	悅	朮	查	毀	禿	稅	脫	蛻	稟	吳

This kind of Z_Variants are the most important ones which may confuse people in applications, even cause failure in text searching/retrival if the process is made without care.

3.2.1 Quasi Z Variants 準 Z 變體

【Definition】 The Z_Variants having one or two differences from each other in stroke numbers.

Examples:

并	冲	强	凉	吕	换	刹	厦	奥	粤	抛	厢	刊	对	压	隆	厅	兹
並	冲	強	涼	呂	換	刹	厦	奧	粵	拋	廂	刊	对	压	隆	厅	茲

3.3 Orthographic vs Variant Ideographs 正體與異體

In different contexts, an "orthographic" ideograph in a country/region would be viewed as a variant in another country/region. In a multilingual and international environment, it is very difficult to determine whether or not an ideograph is the orthographic one because of political and/or cultural reasons. The following examples show the orthographic ideographs in G (the variants in T/J) vs the orthographic ideographs in T and J (the variants in G) :

G :	并	采	耻	冲	吊	仿	挂	炮	异	游	韵	咏	烟	灾	册
J T:	並	採	恥	冲	弔	倣	掛	砲	異	遊	韻	詠	煙	災	冊

It can be seen that the orthographic ideographs/variants and the Z_variants are overlapped each other.

3.4 Old glyph vs New glyph 新舊字形

The situation is similar to 3.5 that the "Old" vs "New" are relatively political, cultural and opposite concepts which are very difficult to be dealt with.

Note that, the followings do NOT belong to simplified/unsimplified relations, but the Old_glyph/New_glyph from the point of view of a Chinese:

吕	侶	喚	換	吳	宮
呂	侶	喚	換	吳	宮

3.5 Simplified vs Unsimplified Ideographs 簡(略)與繁

3.5.1 One to One relations (1:1) 一對一

In the Unified CJK Ideographs, over 2000 CJK Ideographs have such relations between each other, which are denoted as "1:1". Almost all of simplified ideographs have their unsimplified counterparts. The reverse is not true, i.e., some of unsimplified ideographs do not have their corresponding simplified forms in the repertoire.

Examples: Simplified Hanzi/Kanji vs. Unsimplified Ones

宝	蚕	痴	点	独	断	国	会	号	旧	礼	猫	声	区	双
寶	蠶	癡	點	獨	斷	國	會	號	舊	禮	貓	聲	區	雙

Examples: Hanzi and Kanji simplified in different ways

G	J	G	J	G	J	G	J
对-對-対	厅-廳-庁	单-單-単	画-畫-画				
压-壓-圧	辩-辯-弁	证-證-証	团-團-団				
边-邊-辺	带-帶-帯	举-舉-挙	弹-彈-弾				
齿-齒-歯	处-處-処	恶-惡-悪	发-發-発				
丰-豐-豊	广-廣-広	继-繼-継	两-兩-両				
觉-覺-覚	济-濟-済	齐-齊-斉	荣-榮-栄				
气-氣-気	卖-賣-売	脑-腦-脳	实-實-実				
图-圖-図	围-圍-囲	释-釋-釈	泽-澤-沢				
亚-亞-亜	应-應-応	营-營-営	县-縣-県				
传-傳-伝	价-價-価	从-從-従	听-聽-聴				

Examples: Simplified Hanzi only, no simplified Kanji exist

说	纺	饭	锻	车	轴	鸟	鹏	风	麦	麸	鱼	爱	币	飞	汤
說	紡	飯	鍛	車	軸	鳥	鵬	風	麥	麩	魚	愛	幣	飛	湯

Examples: Simplified Kanji only, no simplified Hanzi exist

州	知	予	炎	希	黑	毒	德	喝	突	仏	拔	尙	仮	隆	増
洲	智	豫	焰	稀	黑	毒	德	喝	突	佛	拔	壹	假	隆	増

3.5.2 One to Many relations 一對多

There are two kinds of One_to_Many relations.

- (1) One ideograph in a certain context corresponds to more than one ideographs in another context, which is denoted as (1:M) . e.g.

開	頭
发: 發	髮
钟: 鐘	鍾
	錶 情

- (2) One ideograph in a certain context replaces more than one ideographs including itself in another context, which is denoted as (1:M+1) . e.g.

前	皇
后: 後	后
注: 注	註
	意 解
樹	若
干: 榦	乾 - 幹 - 干
	淨 部
里: 里	裡 - 裏
	程 面 面
制: 制	製
	度 造

3.6 Proposed VU Classification:

Among the approaches noted above for classifying ideographs which are Vertically Unifiable, overlap and dependency can be found. In order to eliminate or reduce the dependency of VU classification, from the the viewpoint of X/Y/Z Model, the VU objects should be organized in the proposed frame:

		1:1	1:M 1:M+1
Z	Pure Z_Variants	丢 / 丟 兑 / 兌	
Y	①Quasi Z_Variants	吕 / 呂 冲 / 沖 奥 / 奧 厢 / 廂	注 / 注註
	②Y_Variants	韵 / 韻 烟 / 煙	
	③Simplified	断 / 斷 独 / 獨	发 / 發髮
	Unsimplfd.	贝 / 貝 鱼 / 魚 旧 / 舊 宝 / 寶	后 / 後后 钟 / 鐘鍾
X	X_Variants = Homonyms	芸 / 芸 浜 / 浜	

From the viewpoint of a software developer, the attribute (1:1, 1:M, or 1:M+1) makes more sense than the other attributes for defining the data structure of mapping tables.

4. Characteristics of VU

4.1 Language Context Dependency 語境相關性

VU's result depends on which language contexts are related. The examples below show how different it is when VU is done between G T vs between G J :

G	T	G	J
联	聯	联	聯
连	連	连	連
讽	諷	讽	諷
风	風	风	風
诀	訣	诀	訣
决	決	决	決

4.2 Direction Dependency 方向相關性

VU's result also depends on the direction along which the vertically unifiable ideograph points to its counterpart from one context to another.

The examples below show how different it is when VU is done from G to J vs. from J to G :

G	J	G	J
参	參	参	參
叁	參	叁	參
三	三	三	三

4.3 Coupling Intensity 耦合強度

The tie between ideographs linked by VU has its " Intensity ", which reflects the necessity, possibility, and frequency to link one ideograph with the others.

The coupling Intensity relies on contexts and VU directions too.

For example, the coupling intensities between 一 and 壹 may be set to

zero due to unnecessary; from G to J, may be set an amount for 叁 → 三
and set zero for 三 → 叁;
from 几 to 幾 and 几, the first intensity is greater than the second
one according to the statistics from a Corpus.

5. Case study - An Implementation of VU

One of the applications of VU is to convert text files between (Chinese) simplified context to unsimplified context.

A conversion software has been developed by Eten Information System Co, named as BMT : " The Bridge between Mainland and Taiwan " (兩岸橋) .

BMT enables the coexistence of simplified/unsimplified/variant Hanzi in a single computer system, the bi_directional conversion between GBcode and Big5, as well as the (1:M)/(1:M+1) handling.

The implementation of the BMT indicates that, the classification of VU is preferably to be done in terms of (1:1)/(1:M) or (1:M+1). When orthographic/variant, new_glyph/old_glyph and simplified/un_simplified Hanzi are mixed together, more non_one_to_one relations are derived, which made the process more complicated. A special routine with a sophisticated data structure handles each mapping relations. For non_one_to_one relations, the unified candidates are properly ordered, the one with higher frequency is chosen first, then checked/corrected by users. The (1:M)/(1:M+1) relations not only exist from G to T, but also appear from T to G, e.g., 著 in T has two counterparts (著著) in G which appears very often in text converting. Some unsimplified Hanzi (e.g. 麼後) in G makes confusion likely. Thus it would be wiser to ignore them and take simplified ones (么后). In general, G context uses traditional (傳承字) and simplified Hanzi only, which is different from J context in which mixed simplified/unsimplified Kanji are allowed.

6. Conclusion : VU Rules Needed

Vertical Unification of CJK Ideographs is a significant and complicated task. With the close collaboration of Chinese, Japanese and Korean scholars, it will be done best. Like the Horizontal Unification the Vertical Unification needs Rules before it is conducted formally. Some of the factors discussed in this paper have to be taken into account

for establishing of a reasonable and workable rules.
In particular, the VU's classification and the quantification of VU attributes is the first thing to be determined. Accordingly, the following 6 pairs of tables shall be developed respectively:

G → T / T → G
G → J / J → G
G → K / K → G
T → J / J → T
T → K / K → T
J → K / K → J

As a result, a new database is expected to be organized and shared by CJK ideograph users to promote the multilingual exchange of culture, science and technology.