第１回　国立国語研究所国際シンポジウム報告書

The National Language Research Institute
First International Symposium

世界の国語研究所

－言語問題の多様性をめぐって－

National Language Institutes Around The World
- Diversity in Language Issues -

国立国語研究所
The National Language Research Institute

Zhang Zhoucai
Research Center of Computer &
Microelectronics Industry Development
China

Vertical Unification of CJK Ideographs

1. Introduction and Definitions

1.1 Horizontal Unification of CJK Ideographs

Horizontal unification of CJK (Chinese-Japanese-Korean) ideographs refers to the task that has been done by the CJK-JRG according to its Unification rules, the output of which is the CJK Unified Ideographs arranged in the Basic Multilingual Plane of ISO-IEC 10646.

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

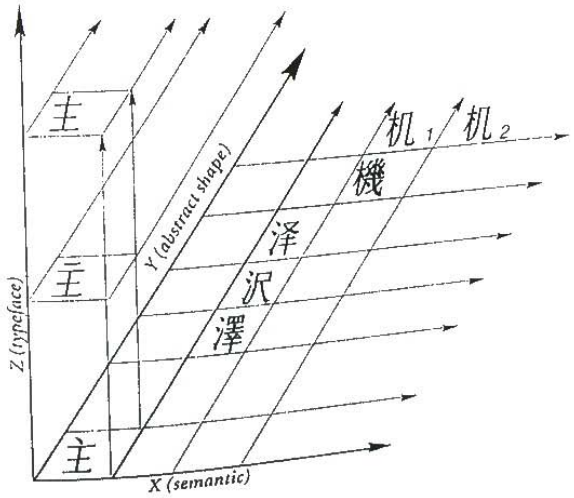
In other words, the above mentioned unification is based on the so-called X/Y/Z model, where *meanings* (the Signified) are arranged along the X axis, *glyphs* (the Signifier: character shapes) along the Y axis, and variants along the Z axis. The unification achieved in the ISO code was performed along the Y axis. This type of unification is referred to as Horizontal Unification (HU).

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

←Horizontal Unification

As shown in the codetable, HU is done mainly but not absolutely by shape. On the one hand, even some ideographs which have the same meaning(s) and common shapes, showing only minor difference in the direction of individual strokes (such as 丟 and 丢) are *not* unified due to the principle of Source Separation in the CJK Unification Rules; on the other hand, non-cognate ideographs are not unified either, although they may have quite similar shapes (such as 士 and 土). Apparently the meanings of the ideographs were taken into consideration while HU was being done.

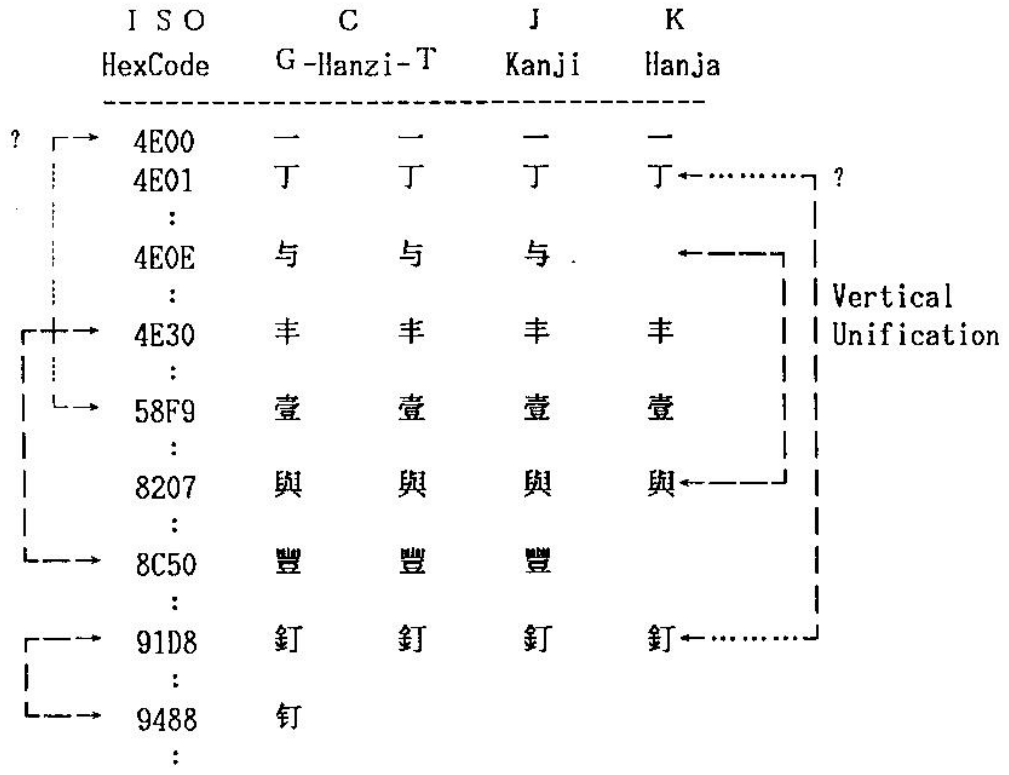
In fact, although HU stresses distinctions between glyphs, the horizontally unified CJK ideographs predominately signify the same or related meaning(s), with only a very few exeptions.



1.2 Vertical Unification of CJK ideographs.

The vertical unification (VU) of CJK ideographs is defined as the one which is based on identity in meaning among ISO-IEC 10646 ideographs, in contrast to HU, which is primarily based on shape.

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



VU should be performed so as to link characters of identical meaning, but this must be kept within limits. To unify 一 and 壱, which both indicate the number "one", would be a kind of overkill. While the unification of 丁 and 釘 might be linguistically reasonable for some languages, it would be technically unnecessary.

One thing which must be emphasized is that VU does not involve any re-unification or re-encoding of CJK ideographs. All further unification will be based on the output of HU as reflected in the ISO standard, with no further code change.

2. Scope and Limitations of VU

As discussed in this paper, VU might be applicable for ISO-IEC 10646-1 CJK ideographs in applications for glyph normalization and/or the convergence of character

repertoires within particular language contexts, 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 the PRC (mainland China)
(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, as shown below:

ISO	C		J	K
HexCode	G	-Hanzi-	Kanji	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	钉			
:				

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, so the K column is omitted in some diagrams for simplicity. G, T, J, K are used as abbreviations for these character sets.

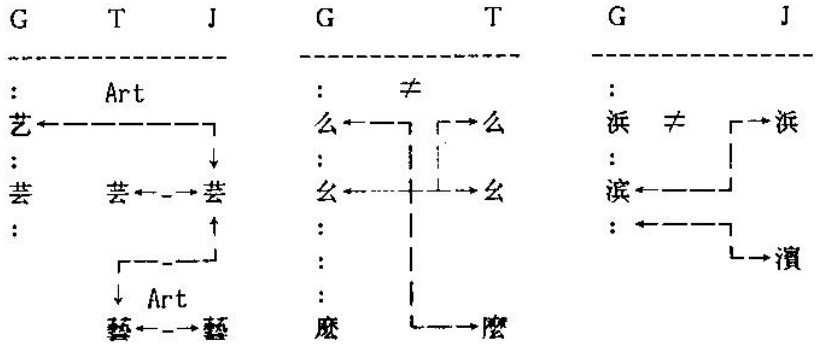
3. Classification of Vertical Unification

3.1 Homonyms

A few horizontally unified CJK ideographs need to be separated vertically by their meanings. In general, VU is intended to link ideographs or different shapes according to

their meaning. A few exceptions, however, result in splitting instead of unification of CJK ideographs. Below are some typical examples. In the first example the simplified character used in Japan with the meaning of "art" happens to be the simplified form of a very different character as used in mainland China.

It should be emphasized again that such cases constitute only an extremely small percentage of the characters in the CJK repertoire; they are however among the most frequently used characters in the modern CJK languages, and must therefore be accounted for.



3.2 Z-Variants

The term Z-Variants refers to those ideographs that have identical meanings and the same generic shape, but that are encoded in different cells in ISO 10646 because they had existed in a source character set as separate code points. The principle of Source Separation meant that this separation had to be reflected in ISO 10646 as well. We may conceive of the Z-Variants as those characters which would have been unified in the process of HU if generic shape had been the only consideration.

3.2.1 Pure Z-Variants

Pure Z-Variants are those variants which have the same number of strokes. Some examples are given below, but this is not an exhaustive list. Z-Variants of this kind are the main ones to cause confusion in the use of software applications, and may cause failure in text searching and retrieval if care is not taken.

兑 丢 尔 内 凛 刊 别 呐 囱 匀 吞 册 画
兌 丢 尔 内 凜 刋 別 吶 囪 勻 呑 冊 畫

娱 户 彦 悦 术 查 毁 秃 税 脱 蜕 禀 吴
娛 戶 彥 悅 朮 查 毀 禿 稅 脫 蛻 稟 吳

3.2.2 Quasi Z-Variants

Quasi Z-Variants are those variants which have the same generic shape, but differ in number of strokes.

并 冲 强 凉 吕 换 刹 厦 奥 粤 抛 厢 刊 对 压 隆 厅 兹
| | | | | | | | | | | | | | | | | |
並 沖 強 涼 呂 換 剎 廈 奧 粵 拋 廂 刋 対 圧 隆 庁 茲

3.3 Orthographic and Variant Ideographs

Depending on the language context, an ideograph viewed in one country or region as an orthographic ideograph (standard character) might be viewed as merely a variant ideograph in another country or region. In a multilingual and international environment, it is very difficult to determine whether or not an ideograph is an orthographic one or not due to political and/or cultural reasons. The following examples show orthographic ideographs in G (which would be viewed as variants in T and J) with their corresponding orthographic ideographs in T and J (considered to be variants in G). It can be seen that orthographic ideographs and variants, and Z-Variants overlap each other.

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

3.4 Old and New Glyphs

The distinctions between old and new glyphs is similar to that between simplified and unsimplified glyphs discussed in 3.5 below, in that "old" and "new" are relative concepts influenced by political and cultural factors, and are very difficult to deal with. Note that the following examples are seen as distinctions between old and new forms, not as distinctions between simplified and unsimplified forms (seen from the point of view of a Chinese):

吕 侣 唤 换 吴 宫
| | | | | |
呂 侶 喚 換 吳 宮

3.5 Simplified and Unsimplified Ideographs

3.5.1 One-to-One Relations

In the unified CJK ideographs, over 2,000 CJK ideographs have such relations between each other, which may be denoted as 1:1. Almost all simplified ideographs have unsimplified counterparts. The reverse is not true. That is, many unsimplified ideographs do not have corresponding simplified forms in the repertoire.

Simplified Hanzi/Kanji versus 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

Hanzi and Kanji simplified in different ways:

G	J	G	J	G	J	G	J
对	對	对	厅	廳	厅	单	單	单	画	畫	画
压	壓	压	辩	辯	弁	证	證	証	团	團	团
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

Simplified Hanzi only, no simplified Kanji exist:

说	纺	饭	锻	车	轴	乌	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

Simplified Kanji only, no simplified Hanzi exist:

州	知	予	炎	希	黒	毒	徳	喝	突	仏	抜	壱	仮	隆	増
洲	智	豫	焔	稀	黑	毒	德	喝	突	佛	拔	壹	假	隆	增

3.5.2 One-to-Many Relations

There are two kinds of one-to-many relations. In one, an ideograph in a certain context corresponds to more than one in another context, (denoted as 1:M). In the examples below, the simplified forms correspond to distinct unsimplified forms:

	開	頭						
发：	發	－	髮		钟：	鐘	－	鍾
						錶		情

In the other kind of one-to-many relation, an ideograph in a certain context corresponds to more than one ideographs, including itself, in another context (denoted as 1:M+1). In the examples below, a pre-existing ideograph is used in its original use, and as a simplified version of one or more other characters.

	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 may be unified vertically, we may propose overlap and dependency. In order to eliminate or reduce the dependency of VU classification, from the viewpoint of the X/Y/Z model the VU objects should be organized in the proposed frame. From the viewpoint of a software developer, the relational attributes 1:1, 1:M and 1:M+1 make more sense than the other attributes for defining the data structure of mapping tables.

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

4. Characteristics of VU

4.1 Language Context Dependency

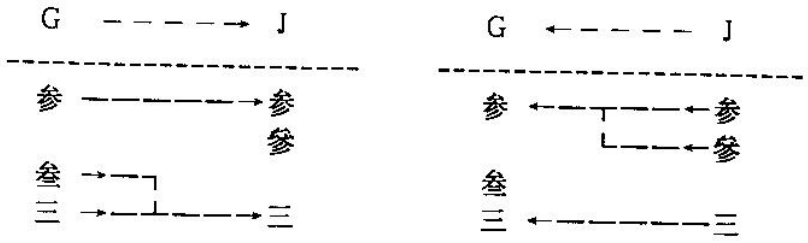
The results of VU depend on which language contexts are related. the examples below show how different it is when VU is done between GT (mainland China and Taiwan) and between GJ (mainland China and Japan). For example, the character 連 in modern Japanese is customarily used as an equivalent for 聯, so both of these must be related to the mainland Chinese character which corresponds only to 聯 in Taiwan.

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

The results of VU also depend 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, versus from J to G. In this case links are

drawn so as not to point to characters which are not in standard use, but which exist as mere variants.



4.3 Coupling Intensity

The tie between ideographs linked by VU has an "intensity" which reflects the necessity, possibility and frequency of a link to one ideograph as compared to possible links to others. The coupling intensity also reflects contexts and direction. For example, the coupling intensities between 一 and 壹 may be set to zero due to unnecessity; from G to J, the link from 叁 to 三 may be set to greater than zero, while the corresponding link from 三 to 叁 may be set to zero, since the latter character is not in standard use. Further, for the links from 几 in G to 幾 and 几 in J, the first one will have a greater intensity than the second, which may be set according to statistics from a corpus.

5. Case study: an Implementation of VU

One of the applications for VU is to convert text files between (Chinese) simplified context to unsimplified context. A conversion software has been developed by Eten Information System Co. It is called BMT for "The Bridge between Mainland and Taiwan". 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 preferable to be done in terms of (1:1)/(1:M) or (1:M+1). When orthographic characters and variants; new and old glyphs; and simplified and unsimplified characters 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 relation. For non-one-to-one relations, the unified candidates are properly ordered, the one with the higher frequency chosen first, then checked and corrected by users. The (1:M)/(1:M+1) relations not only exist from G to T, but also appear from T to G. For example, 著 in T has two counterparts, 着 and 著 in G which appear very often in text converting. Some unsimplified Hanzi (such as (麼 and 後) in G makes confusion likely. Thus it would be wiser to ignore them and adopt the simplified ones (么 and 后). In general, G contexts use

traditional and simplified Hanzi only, which differes from J contexts in which mixed simplified and unsimplified Kanji are allowed.

6. Conclusion: The Need for VU Rules

Vertical unification of CJK ideographs is a significant and complicated task. It calls for the close collaboration of Chinese, Japanese and Korean scholars. Like horizontal unification, vertical unification needs rules before it may be formally conducted. Some of the factors discussed in this paper must be taken into account to establish reasonable and workable rules. In particular, the classification and quantification of VU attributes are the first things that must be determined. Accordingly, the following six pairs of tables must be developed:

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 users of CJK ideographs, to promote the multilingual exchange of culture, science and technology.

ＣＪＫ（中日韓）統合漢字の垂直統合

（要約）

1．概論と定義

1.1 ＣＪＫ漢字の水平統合

ＩＳＯや中国，日本，韓国の漢字統合は，意味より字形からとらえている。Ｘ軸に意味，Ｙ軸に字形，Ｚ軸に変種というＸＹＺ座標系に基づき，同じ漢字はＹ軸を中心に水平統合されている。しかし，中国，台湾，日本や韓国の漢字の互換領域を考えるとき，字形重視の水平統合は，言語的文脈や簡化による差異，類義，多義性などの問題に対処できない。

1.2 ＣＪＫ漢字の垂直統合

そこで，ＣＪＫ漢字を，字形重視のＩＳＯ，水平統合における意味を柱に，1.2 の表のような垂直統合により定義つけてみる。

2．垂直統合の適用範囲と限界

ＩＳＯ，ＣＪＫ漢字表に，字形の標準化，異言語間での変換，言語的文脈による漢字のレパートリの適用による垂直統合を行ってみる。漢字を中国の GB 2312-80，台湾の TCA-CNS11643，日本の