JTC1/SC2/WG2/IRG
Date :21/04/2002
N 906

ISO/IEC JTC1/SC2/WG2/IRG

Ideographic Rapporteur Group

(IRG)

Source/Contribution Identifier : Dr. Lu Qin

Meeting : 19th IRG Meeting in Macau

Title : Reference on Ideograph Variants
Status : Personal Contribution

Requested Action: Due to recent resolution in Unicode, ideograph variant definition and handling needs to be revisited. The enclosed paper, which will be presented in Unicode Conference No 21, gives an overview on ideograph variants and the feasibility of using variation selectors to describe variants. This can serve as reference material for variant handling in Ext. C.

# Ideograph Variants-What They Are and How to Handle Them

Lu Qin

Department of Computing, The Hong Kong Polytechnic University,
Hung Hom, Hong Kong
csluqin@comp.polyu.edu.hk

**香港理工大學電子計算學系**

**Abstract**

Ideographic variants have been created and used since the early invention of the ideographic characters for many different reasons, and different people have different understandings of the variants. Some variants have been created purely for different calligraphy styles, and some due to extensions in meanings. The use of ideograph variants causes confusion in learning Chinese. It introduces more uncertainty into the computerization of Chinese information. Unicode in its early stage had a lot of problems with variants, and a set of unification rules was developed to avoid giving unnecessary code points to variants. However, some variants still got their way into the standard. With the possible Extension C in ISO 10646, it is inevitable that some more variants will be introduced into the standard. Thus there is a need for a systematic study and handling of ideograph variants.

This paper first gives the definition and classification of Chinese variants. It then discusses the issue of variants in relation to character coding standards. It further investigates into the feasibility of using variant symbols introduced in Unicode for the representation of ideograph variants. Finally, it goes through the character decomposition scheme to explore an algorithmic way to represent variant information of different characters.

**Keywords**: Character decomposition, ideograph variants, new applications

## 1. What are Chinese Character Variants

Characters are the smallest units in the written form of the Chinese language. Each Chinese character is associated with three types of features, namely, glyph, pronunciation, and meaning. None of the three features needs to be unique. However, under a specific text context, the meaning and pronunciation are usually unique.

Some Chinese characters are associated with more than one glyph. Generally speaking, *Chinese character variants*(“異體字”), or *variant forms of a Chinese character (多形字)*[1, 2] refer to a set of Chinese characters having exactly the same pronunciation and meaning, but varying in glyph shapes. In other words, variants can be used to replace one another in any text

without changing the meaning of that sentence. There is no rule in the choice of which variant form to use. For example, the two glyphs "沒" and "? "are considered variants even though the first one is commonly used in Taiwan, and the second in mainland China. The character "bone" can also have different glyphs, "骨", "骨", "骨", which are considered variants. Some variants can be drastically different in their glyph shapes, such as '臆" and '窗", and "焊" and "釬".

Variants have been used since the invention of Chinese characters. In the ancient oracle script, the character "goat, 羊" had at least 32 different variant forms and the character "phoenix, 鳳" 50 different variants. Over the long evolution of the Chinese writing system, character glyph shapes have stabilized to the current forms and the number of variants has also tended to decrease. Due to the natural affinity of people for simplification, some variants have been created along the way. In fact, according to the definition of variants, a simplified Chinese character and the traditional form of the same character are also considered variants. Others have been created because of changes of writing style. Figure 1 shows how the Chinese character "爲" has changed from the oldest oracle script to its current forms and, in the process, it ended up with three variant forms (shown in the last 3 columns).

| Style name | Oracle script 甲骨文 | Jin script 金文 | Xiao Zhuan 小篆 | Li style 隸書 | Kai style 楷體 | Simplified 簡體 |
|---|---|---|---|---|---|---|
| |  |  |  |  |  |  |

**Figure 1** An example of a Chinese character Glyph Change

Chinese characters are formed by strokes. The length of strokes, the angles at which the strokes are written, the distances between different strokes, the thickness and the height-width ratio are all considered variant features, which can change without affecting our recognition of a character. It is the invariant features that are normally used to distinguish different characters/glyphs. Firstly, the types of strokes, which are used as fundamental units to form Chinese characters, are considered an invariant feature. For example, the character "big" is formed by three strokes, "橫, 一", "撇, 丿", and "捺, 乀". Secondly, the relative positions of the strokes in a character are considered an invariant feature and normally do not change. Thirdly, the relative positions of the components are also an invariant feature. For example, the character "旮" has the component "九" on the top and "日" at the bottom. If the two components switch positions to form "旯", it is considered a different character altogether.

Variant forms of Chinese characters have created a lot of problems, even before the computer age, causing confusion in learning Chinese. With computerization, variants giving different code points also create additional problems for searching and sorting. For instance, if one wants to search the keyword "陸勤" over the Internet, even if "陆勤" exists, such information

cannot be located, unless additional mapping information on both "陸" and "陆" is maintained by the search engine.

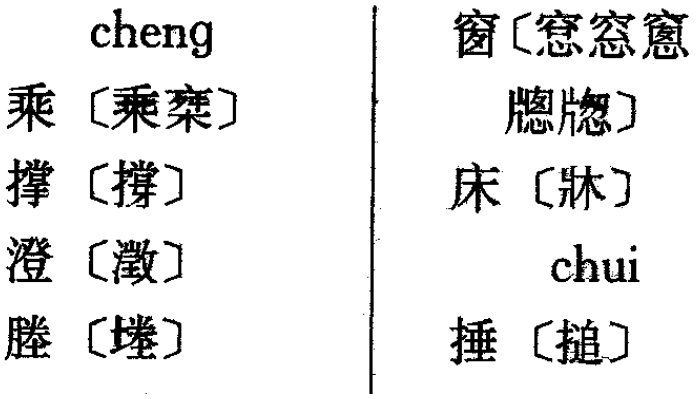


**Figure 2** A Sample page of the variant table published by the Chinese Government

The first serious effort on the unification and normalization of Chinese character variants was started in the earlier 1950s by the Chinese Government. The first standard on variants was published on Dec. 22, 1955, in which 810 groups of variants were specified[2]. Figure 2 shows a sample page of the variant table. The first glyph in each group is considered the canonical form, and the glyphs inside the brackets are considered "obsolete" and can be used only in limited circumstances[4]. There were three revisions to [2], in which some of the "abolished variants" were reinstated as regular characters. Reasons for the reinstatement can be found in [4]. Table 1 provides some statistics on [2]. Note that the majority of variants has only two variant forms in each group. There are only 2 groups of characters having 6 variant forms. Another interesting part is that [2] was published to help with the literacy movement. Characters listed in [2] are considered "common" characters and not some rarely used characters. Out of the 810 so-called commonly used characters, 1,055 non-canonical variants are "abolished". This is a very good indicator as to the scale of Chinese character variants existing in the Chinese language. By considering more traditional forms of the commonly used characters, we would find even more variants. Taking the character "difficult/difficulty" "難" as an example, there are a total of 9 variant forms listed in the KangXi Dictionary.

| Groups | Total number | Total abolished characters |
|---|---|---|
| 2 in a group | 609 | 609 |
| 3 in a group | 167 | 334 |
| 4 in a group | 26 | 78 |
| 5 in a group | 6 | 24 |
| 6 in a group | 2 | 10 |
| Grand Total | 810 | 1,055 |

**Table 1** Summary of the Variant Table

Although the formation rules of Chinese character variants seems quite arbitrary, there are still some regularities[5]. In order to discuss them, we must introduce two subclasses of components used commonly in Chinese character formation: the phonetic component(聲符) and the ideographic component(形符). Examples of phonetic components are "袁" in "猿" and "我" in "鵝". On the other hand, "犭" in "猿" and "鳥" in "鵝" are ideographic components which are actually classification indicators, not directly denoting the specific meaning of these characters. There are seven major types of variant formation sources:

1. **Different ideographic components**: variants may be formed due to the use of different ideographic components that have similar meanings. For example, since both ideographic components "隹" and "**鳥**" symbolize "bird", the character for "chicken" takes both "雞" and "鷄" as two variant forms.

2. **Different phoneme components**: variants may be formed due to the use of different phonetic components having the same pronunciation. For example, "袁" and "爰" are different components, but are of the same pronunciation when serving as phonetic components in character formation. Two variant characters, "猿" and "猨", are thus generated.

3. **Different placements of radicals**: variants may be formed due to different ways of placing the radicals. For example, for the character "hill top", there are two variants, "峰" and "峯". The first one has the radical "mountain, 山" placed on the left hand side; the second has it placed on top of the other components. However, this rule cannot be generalized. Sometimes, the change of the relative positions of components will lead to completely unrelated characters. For example, "yin"("**吟**") and "han" ("**含**") do not seem to be semantically related.

4. **Use of different character formation rules**: variants may be formed due to the use of different character formation rules. Examples include the character "thread" having both variant forms of "線" and "綫".

5. **Change of writing style**: variants may be formed due to the change of writing style. Historically, during the period Chinese characters were changed from the smooth cursive "Xiao Zhuan" (小篆) style to the more modern stroke based "Li"(隸書) style, many character glyphs were changed, thus leaving a hefty number of variants. Examples are "  " from "Xiao Zhuan" (小篆) and "  " as a result of the "Li"(隸書) style transformation. This type of transformation often produces variants that are drastically different in shapes, as was indicated in this example.

6. **Simplification**: As a result of simplification, co-existence of both the traditional form and the simplified form creates variant forms. The most systematic simplification is given in the document published in 1964[6] and its revised version was also published in 1986[7]

7. **Normalization/stroke variants**: more variant forms are created due to the different choices of strokes, or small variation in the relative positions of strokes. These variants in many cases are the same characters which have slight change of forms because of a mixture of different writing styles. We refer to these stroke variants as *macro level variants.* For instance, the character "strong" has two variant forms "強" and "强", the character "corner/horn" can be written either as "角" or "角". Variants under this class are very close in shape, topologically speaking. More variant forms are created as a result of the normalization of Chinese character glyphs published in 1965[8]. The normalized styles are also called the "new style" glyphs. The character "bone" used in the mainland of China taking the glyph of "骨" is an example of a new glyph shape created in modern times. Other variants may be due to slight variation in the choice of strokes. Many more examples of stroke variants due to different styles can be found in the CJK unification document[9].

**2. Unification in ISO 10646/Unicode**

Variants can cause problems in electronic processing of Chinese. Since variants are supposed to have the same meaning, if different people use different variants, searching and sorting of information on-line will become more difficult and less unified. For this reason, issues related to variants have been seriously discussed, and the result of that discussion leads to the development of the unification rules and the unification process for CJK Ideographs in ISO 10646[9].

**Figure 3** Unification Examples in ISO 10646

The analysis of unification is done in two levels. Firstly, the component structures of two characters are examined. If two characters are considered to have different component structures, especially at the base level, they will not be unified. For instance, the two characters "峰" and "峯" will not be unified because the first one is described by the IDC ⿰ (U+2FF0) and the second by ⿱ (U+2FF1). The unification rules are given by examples. Figure 3 shows a sample section of the unification examples. If we examine carefully the seven types of variants given in Section 1 against the examples given in [9], it is not difficult to conclude that the first six types of variants are not being unified in ISO 10646. In other words, for the first six types of variants, the mapping information can only be provided through some mapping tables.

The Unicode Consortium has requested for the mapping information to be made available to vendors[10]. However, it should be pointed out that the mapping table is not universal because variants are country/region dependent. Characters defined in ISO 10646 that are considered variants in one place may well be considered as different characters in another. For instance, the character "village" has historically taken two variant forms "村" and "邨", which are still considered variants in China today. However, "邨"is no longer considered a variant of "村" in Hong Kong because "邨" specifically refers to rent-controlled, large government-managed residential estates. Due to the deviation in meaning, they are no longer considered variants.

Variants subjected to unification rules are applied to only those so-called macro level variants in the CJK repertoire. That is, CJK unification is applied mostly to ideograph components serving at the same structural position only. Variants can be of the change-of-stroke types, as in "**內**" and "内" where the variations are in "人" and "入". Macro level variants sometimes can also differ in the number of strokes, as in "呂" and "呂" where the second component variant has an extra stroke.

### 3. How to Handle New Variants in ISO 10646/Unicode

It is not practical to think that we can eliminate all variant forms other than the canonical form in ISO 10646, unless we can remove all the six forms of variant formation. It should be noted that the Chinese definition of ideograph variants cannot be generalized easily to other ideograph characters in the CJK repertoire both in terms of pronunciation and meaning. We should also understand that the so-called "abolished" variants in one place may still have to appear in the CJK repertoire because they may be used in different countries and regions.

There have been suggestions in recent ISO/IEC ITC/SC2/WG2/IRG meetings[10] to use the sixteen *variation selectors* defined in ISO 10646 in the range of U+FE00 to U+FE0F to represent newly submitted characters that can be considered variants. For simplicity reasons, we use the short form *VS-1*, *VS-2*, … *VS-16* to denote the variation selectors, respectively. Unicode has also accepted another 240 variation selectors in the range of U+E0110 to U+E01FF[11], denoted by VS-17 to VS-256 in this paper, respectively. One of the main

purposes of this collection is to represent catalogued ideographs in plain text without their requiring separate code points. UTC is said not to approve any explicit encoding of variant characters with their own code points in the future unless there is a significant reason why this must be done. For all practical purposes, for this new *variant-character principle* to be implemented, the term "variant characters" may have to be redefined or clearly specified under this context.

We must remember that one fundamental characteristic of ideograph variants is that variant forms should have the same meaning, whereas pronunciation and glyph shapes are secondary. In fact, pronunciation is totally disregarded in CJK unification. As stated earlier, CJK unification rules are applied mostly to characters similar in shape with only component variants and they are considered cognate though not necessarily the same in meaning(as the meaning part cannot be verified easily across different countries/regions). This new variant-character principle for character acceptance would definitely warrant some serious discussion on the definition of variants in this context. One of the reasons for the debate is that the CJK unifications were mostly applied to characters submitted by different sources and thus the meaning issue did not take up a very important role. However, in new submissions, the potential "variant characters", which are considered macro level variants, are more likely to be from the same source(country/region). This creates a big problem for other members in the IRG to argue whether the "variant characters" do have the same meaning or not. In fact, all members in the IRG understand that their submissions should be characters, not variants in the traditional sense of the word character variants.

One possible way to define "variant characters" in this context is to completely drop the traditional criterion on meaning. Rather, we can concentrate purely on glyph shapes. In other words, we can examine the new characters completely based on the analysis of component variants. The author will not attempt to give a new definition for the term variant character here. It has to be agreed by all IRG members and approved by WG2.

However, it should be pointed that describing character variants in plain text is not an easy job and sometimes can be ambiguous. For instance, if we want to use "強"(U+5F37) as the principal character and use a variant selector (<強,VS-1>) to describe the variant "强"(U+5F3A), a possible description following ISO 10646's style in describing other variants can be given as follows:

| Sequence (UID notation) | Description of variant appearance |
|---|---|
| <5F37,FE00> | 5F37(強) with 53E3(口) on the top right corner being replaced by 53B6(厶) |

Without the glyphs of the character and the components given in the brackets, this sentence is quite confusing. Sometimes, the description on the relative positions of the involved

components can be ambiguous especially when there is an addition or a removal of strokes instead of a replacement.

If we use the ideograph description sequence(IDS)[12] to describe variants, the description of the variant appearance would be more precise and elegant. Basically, the description has two parts. Firstly, it can use the IDS to describe the decomposition of the principal character so that the variant components can be located precisely. Secondly, we describe the variant component using plain text with reference to the component only, as shown below:

| Sequence (UID notation) | Description of variant appearance |
|---|---|
| <5F37,FE00> | 5F13 (弓)  <53B6(厶), FE00(VS-1) > 866B(虫) where<br><53B6(厶), FE00(VS-1) > is 53B6(厶) replaced by 53B6(厶) |

Note that in the above expression, the description of a variant character of 5F37(強), is transformed to (1) pinpointing where the variant component is located, and (2) indicating what the component is being changed to. It should be noted that the ideograph composition scheme in [12] must be extended to include variant selectors. The extension is straight forward as the only change involves the replacement of an ideographic component by a UID notation, and thus the details are not described here.

**4. Conclusion**

This paper gives an overview of the types of ideograph variants and what kinds of ideograph variants are unified in ISO 10646. With the newly proposed variant character principles, it is understood that the possibility of using variant selectors to describe new characters instead of giving new code points is entirely possible. The definition of variant characters must be revisited and confirmed before Extension C work can proceed smoothly. It is also the author's belief that plain text description should be augmented by an ideograph description sequence which can describe the variant forms more elegantly.

**Acknowledgement:**

This work is partially supported by a grant from the Innovation and Technology Fund, Hong Kong SAR Government.

**References**:

[1] GF 1001-2001,《**國家語言文字工作委員會 語言文字規範 第一批異形整理表**》(The First Series of Standardized Forms of Words with Non-standardized Variant Forms)**, 國家語言文字工作委員會，**2001 **年** 12 **月** 17 **日**.

[2] 《**國家語言文字工作委員會 關于發布第一批異體字整理表的聯合通知**》，**國家語言文字工作委員會，**1955 **年** 12 **月** 22 **日**.

[3] **〈漢字拓撲結構分析〉，劉連元，《語文研究》**(*The Topological Analysis of*

*Chinese Characters* by Liu Lian Yuan)，1990 年第四期.

[4] 《字的基礎知識》王鼎吉編著，中國和平出版社，1996 年 4 月?

[5] 《古今字》洪成玉著，語文出版社, 1995 年 10 月.

[6] 《國家語言文字工作委員會 關于簡化字的聯合通知》，國家語言文字工作委員會，1964 年 3 月 7 日.

[7] 《國家語言文字工作委員會 關于重新發表<簡化字總表>的聯合通知》，國家語言文字工作委員會，1986 年 10 月 10 日.

[8] 《國家語言文字工作委員會 印刷通用漢字字形表》，國家語言文字工作委員會，1965 年 1 月 30 日.

[9] Annex S (Informative) Procedure for the Unification and Arrangement of CJK Ideographs, ISO/IEC 10646-1:2000(E), International Standard: Information Technology-Universal Multiple-Octet Coded Character Set(UCS) - Part 1: Architecture and Basic Multilingual Plane, ISO/IEC, 2000

[10] Unicode Consortium, "Variation Data for Existing and Future Ideographs in ISO/IEC 10646", Document No. N850, working document of ISO/IEC JTC1/SC2/WG2/IRG meeting #18, Tokyo, Dec. 2001

[11] Unicode Consortium, "Proposed New Characters: Pipeline Table", http://www.unicode.org/unicode/alloc/Pipeline.html

[12] Qin LU, "The Ideographic Composition Scheme and Its Application in Chinese Text Processing", Proceedings of the 18th International Unicode Conference, Hong Kong, April, 2001