L2/25-272

Title: Descriptions vs. usage of the “GU”, “TU” and former “G9” Unihan source prefixes
Date: 2025-12-05
Source: HarJIT a.k.a. Harriet Riddle
Action: For consideration by CJK WG and IRG

1. Abstract

Out of all of the Unihan source references with prefixes documented as “the source reference for this ideograph has been moved”, only 30% strictly fit this description. By specific source prefix, all `HU`, `KU` and `KPU` source references fit that description, as do 97% (all but one) of the `TU` source references, but only 4% of all `GU` source references do. Amending the description of the `GU` source prefix may be called for; (re-)introducing a source-reference notation for GBK 1.0 may also be helpful.

2. Table of Contents

3. Background

In UAX #38, the GU, HU, TU, KU and KPU source prefixes are documented as follows:[1]

> "The source reference for this ideograph has been moved; the value is its code point."

The same descriptions are listed on the "IRG Source Prefixes" page.[2]

As such, back in October, the descriptions in Wikipedia's coverage were similarly changed from "may have been moved" to "has been moved" to better align the descriptions with UAX #38.[3] This change was correct per Wikipedia policy, since it aligns the descriptions with a reputable source as opposed to original research.

However, a majority of the source references with these prefixes do not in fact fit this description, almost all of which have the GU prefix. These include:

- 67 CJK Extension B characters whose original source references simply read either G4K or GBK without a suffix, and which could not be located in the sources in question;
- 28 source references to GBK 1.0 (the character encoding, not the GBK source) which were lost when the G9 prefix was discontinued and replaced with GU;
- 1 place where a CJK Unified Ideograph was added without a source reference; and
- 6 other places where the original source reference has been removed without moving it elsewhere.

All of these cases are discussed in more detail below.

Table 1: Statistics of "moved" versus "removed" origins of self-referencing source references

	GU-	HU-	TU-	KU-	KPU-	Σ
Moved	4	1	32	4	2	43
Removed	100	0	1	0	0	101
Missing	1	0	0	0	0	1
Σ	105	1	33	4	2	145

1 https://www.unicode.org/reports/tr38/#kIRG_GSource

2 https://www.unicode.org/irg/prefixes.html

3 https://en.wikipedia.org/w/index.php?title=CJK_Unified_Ideographs&diff=prev&oldid=1314877114

4. Specific cases

4.1. Unresolvable 四庫全書 or 中國大百科全書 references

The following 43 Extension B characters had a `G4K` source reference without a suffix, and could not be located in the source in question: `U+20746`, `U+210AA`, `U+21510`, `U+2165E`, `U+218D3`, `U+2195F`, `U+21E2D`, `U+2209B`, `U+22197`, `U+22FEF`, `U+235C5`, `U+23B48`, `U+23BD4`, `U+23CFA`, `U+23DB4`, `U+23EF4`, `U+24403`, `U+244D8`, `U+2471F`, `U+24AAF`, `U+24FC1`, `U+25145`, `U+2528C`, `U+25A77`, `U+25F53`, `U+25F68`, `U+26308`, `U+26625`, `U+26B4D`, `U+26E3D`, `U+270AC`, `U+27841`, `U+278DD`, `U+28759`, `U+2877A`, `U+28F29`, `U+29334`, `U+29429`, `U+29874`, `U+29AF5`, `U+29AFB`, `U+29CD7` and `U+2A29D`.

The following 24 Extension B characters had a `GBK` source reference without a suffix, and could not be located in the source in question: `U+2057C`, `U+212FB`, `U+212FC`, `U+223CB`, `U+23414`, `U+23636`, `U+23762`, `U+23823`, `U+23C86`, `U+23DA7`, `U+23DB3`, `U+23FE8`, `U+240C8`, `U+241C2`, `U+241C5`, `U+242C8`, `U+25618`, `U+256AE`, `U+2620F`, `U+289B9`, `U+29596`, `U+29794`, `U+2980D` and `U+299FD`.

These source references were not "moved" to another codepoint, but were rather never unique, nor (by the current format) well-formed, and thus removed.

2 Extension E characters (`U+2C28D` and `U+2C7FE`) had `GBK` source references which are similarly no longer used for any codepoint (`GBK-1008.47` and `GBK-1014.71`).

4.2. GBK 1.0 (IRGN0278) source characters

Prior to the introduction of the `GU` source prefix, the `G9` source prefix had been used for similar purposes. This had not been ideal: the GB18030 representation of a CJK Unified Ideograph codepoint outside of the original URO is usually four bytes long, meaning a source-reference suffix of eight hexdigits in length, while the `GU` prefix takes UTF-32 representations of only five hexdigits in length.

However, certain characters had had `G9` source references with suffixes only four hexdigits in length, since they had been sourced from GBK 1.0 (IRGN0278).[4] Due to the existence of the `GE` source, this has no effect on characters in the original URO, but does affect 6 characters appended to the URO, all 12 Unified Ideographs in the Compatibility Ideographs block, and 2 Extension B characters, as well as 8 Compatibility Ideographs.

4 https://www.unicode.org/irg/docs/n0278-GBKv1.pdf#page=79

Table 2: 20 Unified Ideographs from GBK 1.0 (IRGN0278)

U+9FB5	G9-FE61
U+9FB6	G9-FE66
U+9FB7	G9-FE67
U+9FB8	G9-FE6D
U+9FB9	G9-FE7E
U+9FBB	G9-FEA0
U+FA0E	G9-FE42
U+FA11	G9-FE44
U+FA13	G9-FE45
U+FA14	G9-FE46
U+FA18	G9-FE47
U+FA1F	G9-FE48
U+FA21	G9-FE4A
U+FA23	G9-FE4B
U+FA24	G9-FE4C
U+FA27	G9-FE4D
U+FA28	G9-FE4E
U+FA29	G9-FE4F
U+20089	G9-FE52
U+241FE	G9-FE91

Table 3: 8 Compatibility Ideographs from GBK 1.0 (IRGN0278)

U+F92C	G9-FD9C
U+F979	G9-FD9D
U+F995	G9-FD9E
U+F9E7	G9-FD9F
U+F9F1	G9-FDA0
U+FA0C	G9-FE40
U+FA0D	G9-FE41
U+FA20	G9-FE49

4.3. Preferred representation already has a G-source

The original source reference of U+31F68 is GXM-00175. The preferred Unicode representation of GXM-00175 is now considered to be U+26C25. However, because U+26C25 already has a G-source, namely GHZ-53205.02, the GXM-00175 source reference was not "moved" there, but rather ceased to be used for any codepoint.

4.4. Pseudo-GB1 and pseudo-GB8 source references

The original source reference of U+58ED is G1-7D47. This is a "virtual" GB/T 12345 position (kPseudoGB1), hence its removal. The published/unextended version of GB/T 12345 leaves this position blank; hence, the source reference was not "moved" anywhere, but rather removed altogether.

The original source reference of U+8780 is G8-2D78. This position is not allocated by GB/T 8565.2 itself, but rather an extension made by ITU T.101 Annex C (ISO-IR-165), hence its removal. Hence, the source reference was not "moved", but rather removed altogether. Per IRGN2808, its source reference will be changed to G7-3E21 in Unicode 18.

4.5. Original source reference removed in error?

U+2B089 is an interesting case study: its original source reference is TD-5278, but it acquired a source reference of UCI-00939 in Unicode 6.1, which possibly uses the wrong source prefix—the UCI prefix was only supposed to be used if all other sources had been removed; hence, the "correct" source prefix in this case should probably have been UTC. In the following Unicode version (Unicode 6.2), the TD-5278 source reference was removed for no clear reason (CNS 11643 still maps 13-5278 to U+2B089).[5] In Unicode 13, the U-source prefix was changed to UTC, and the TU-2B089 source reference was added. The JMJ-060019 source reference was added in Unicode 16, followed by GCA-J0126 in Unicode 17.

It is possible that the TD-5278 reference was removed in error due to the presence of a UCI source reference which should have used the UTC prefix. Notably, if it is restored, the "has been moved" description for the TU prefix will become fully accurate.

5 https://www.cns11643.gov.tw/wordView.jsp?ID=873080

4.6. Has no “original” source reference at all

U+24FB9 never had an “original” source reference: the Unicode 3.1 Unihan database lists no source-reference property for it. The only reference listed for it in SuperCJK is a “virtual” Kangxi Dictionary position,[6] which is included because Extension B was ordered by Kangxi Dictionary position rather than using First Residual Stroke.

It has subsequently borne various “placeholder” source references: namely, GKX-0000.00 (Unicode 5.2–6.1), UCI-00942 (Unicode 6.2–11), G9-96389B35 (Unicode 12.0–12.1), and eventually the current GU-24FB9 (since Unicode 13); it has thus far never had a valid, non-self-referencing, G-source (although it is also TA-6675).[7]

5. The remaining GU source references

This leaves only four GU source references which actually reflect a “moved” reference: GHZ-21072.07 (from U+2180C to U+5AB2), GKX-0903.28 (from U+25D89 to U+7C51), GKX-1325.26 (U+28BBA to U+9459) and GGH-1004.37 (U+2BDA4 to U+2AA56).

6. Possible courses of action

Retaining the status quo would mean that the source-prefix description, especially of GU, would remain misleading.

For the characters originating from GBK 1.0, a separate source prefix might be helpful, either by reviving the G9 prefix with a revised definition, or by introducing an e.g. GGBK prefix (taking care not to collide with the existing GBK prefix).

However, given the significant number of characters where the G-source was incomplete from the beginning and cannot be located (the Extension B non-unique GBK and G4K sources), or even missing from the beginning (U+24FB9), the existence of a significant number of GU source references where the original source reference cannot be accurately described as “moved” seems unavoidable.

Altering the description of the GU prefix to state e.g. “moved or removed” might be the simplest solution (introducing a separate source prefix for “removed” would not provide much benefit). It would not address TU-2B089, but that could be addressed by reïnstating TD-5278, if TCA are in favour (and if it was in fact removed in error).

6 https://www.unicode.org/irg/docs/no802-SuperCJKv14.pdf#page=1027

7 https://www.cns11643.gov.tw/wordView.jsp?ID=681589