ISO/IEC JTC1/SC2/WG2 IRG N2867R

DATE: 2025-10-15

Title: Multi-Syllabic Characters and Abbreviation Characters

Source: Witty Wen (文辰奘)

Status: Individual Contribution to IRG #65

Action: For Consideration by China and IRG

R — Summary of Updates

This revision updates the treatment of Gōryaku-gana (合略假名，kana ligature abbreviations), clarifies the placement of all-katakana and all-hiragana shaped cases (e.g., 来 and ノ), and reiterates source-prioritization for multi-syllabic abbreviation characters. Specifically:

- Gōryaku-gana are abbreviation characters. They function as condensed forms of longer readings and should not be encoded into the CJK Unified Ideograph. Such items are better handled in an independent block. These previous encoded cases are oversights.
- “Han-based sources” are defined and prioritized. In evaluating encoding for multi-syllabic abbreviation characters composed entirely of Han components, proposals substantiated by Han-based sources—i.e., submissions from China, Hong Kong SAR, Macao SAR, or TCA—should receive priority, reflecting real usage in Han environments.

I. Introduction

In addition to the recently discussed script-hybrid ideographs, we have also observed a class of already-encoded characters within the CJK Unified Ideographs (CJKUI) repertoire—so-called multi-syllabic characters (复音字)—that serve as abbreviations but are composed entirely of Han components. Throughout both historical and modern usage, there have been instances where a single ideograph was created to represent a multi-syllable word or compound, typically by merging elements of the component characters. Crucially, these do not incorporate any non-Han script letters; they remain entirely within the Han script domain.

The decisive boundary lies in the intended identity of each component: if any part is meant to represent a non-Han letter (e.g., Latin, kana, hangul, or Zhuyin), the character should be classified as a script-hybrid; if all parts are Han, it should not. Such cases must therefore be evaluated differently from script-hybrid ideographs.

To clearly distinguish between script-hybrid characters—which were originally referred to as “abbreviations” by Kojitani Gen—and multi-syllabic characters discussed in this proposal, both of which can be considered subtypes of abbreviations, the following sections will present representative cases, typological distinctions, and encoding policy considerations.

II. Multi-Syllabic Characters and Script-Hybrid Characters

A representative subclass of multi-syllabic characters consists of ideographs coined for measurement units. For example, 𠫓 (U+515B) is a compound of 克 (“gram”) and 千 (“thousand”), serving as an abbreviation for 千克 (qiānkè, kilogram). 𠦂 is another ideograph historically used for “kilowatt” or sometimes “kilogram,” created by combining “石” and “千”. Similarly, 籽 represents “kilometer,” derived from “米” and “千”. Other cases include 糎 (U+7CCE, “centimeter”) and 粍 (U+7CCD, “millimeter”), as well as capacity-related characters like 竡, 竍, 竚, and 竕, all of which follow similar formation logic—using base units (such as 米 or 升) combined with numerical or scale-indicating elements.

Another type of multi-syllabic character involves stacking or juxtaposing two (or more) entire Han characters to form a single graph, without decomposition into radicals. A notable example is U+2EDEE, included in Extension I, which represents the compound surname 相里. It was submitted as part of the population information character set from China. Functionally, it is multi-syllabic, but typologically, it is better understood as a proper-name ligature, rather than a general lexical abbreviation.

A different kind of example comes from experimental shorthand: a character resembling 门 with a Latin “T” inside, used as a simplified form for 问题/問題 ("question" or "issue"). While it abbreviates a two-syllable expression, the embedded T is Latin by design, making the character a clear example of a script-hybrid. If such a character were to be encoded, it would be preferable to replace the Latin T with a Han-like component. For instance, one could substitute 丅 (U+4E05), a component visually similar to “T”. The key point is that if a proposed abbreviation character can be converted such that all of its parts belong to the Han script, then it falls within the IRG’s

scope for encoding as a normal ideograph. China has indicated willingness to support encoding of such Han-converted cases. Accordingly, if a script-hybrid form can reasonably be restructured into Han components, I would recommend that China consider mapping it in GB 18030, rather than leaving a code-point gap.

Side Note: I believe it is necessary to point out that the encoding of 𫠜 was a review oversight as mentioned in IRG N2866. Given that the character is now part of the standard, it may be worth considering how such cases will be reflected in China's implementation of GB 18030.

III. Clarifying Terminology and Categories

Throughout this discussion, multiple related terms have appeared, and it may be helpful if IRG experts can clarify their usage to avoid confusion. In this proposal, the following distinctions are observed:

- "缩略字" (abbreviation characters) refers to the broad category of single characters used to represent longer words or phrases. This includes both characters composed entirely of Han components and script-hybrid ideographs that incorporate elements from other writing systems. For instance, 𫟹 K and 𫟹 O are abbreviations of a two-kanji name, while 𬜬 is an abbreviation of a two-character unit term. are "abbreviations" in function, but structurally and typologically different.
- "混合文种字" (script-hybrid characters) specifically denotes ideographs that combine Han components with non-Han elements such as Latin letters, kana, hangul, or Zhuyin. These forms cross script boundaries and should, ideally, be encoded outside the CJK Unified Ideographs collection—in a separate block specifically dedicated to mixed-script characters.
- "复音字" (multi-syllabic characters) are characters that represent a multi-syllabic expression (usually disyllabic) condensed into a single graph. Many abbreviation characters fall into this category. The key distinction lies in whether the components are all Han: if so, such characters may be treated as ordinary ideographs under existing IRG procedures.

In evaluating such multi-syllabic abbreviation characters, the following principles may be helpful:

- **All components should be Han Radicals/Components.** If any part of a proposed character is derived from a non-Han script—such as Latin letters, kana, hangul, or Zhuyin—it should be treated as a script-hybrid character rather than a pure Han

abbreviation. Even if the abbreviation is stylistic in nature, it still crosses script boundaries. For example, ⿰广K, which combines a Han radical with a Latin letter, fails this criterion and should be categorized as a script-hybrid ideograph.

- **Priority should be given to proposals supported by Han-based sources.** For multi-syllabic abbreviations, it is important to assess whether the character is actually used or recognized in Han-based environments. If a proposal is submitted by, or supported with evidence from, Han-based sources—such as China, Hong Kong SAR, Macao SAR, or TCA—it should be prioritized for encoding. In such cases, it would be reasonable for GB 18030 to include the character, rather than leave the corresponding code point unmapped. Conversely, if a character is submitted only by non-Han-based sources—such as Japan or Vietnam—and lacks any relevance to Han usage, it may be reasonable for GB 18030 to leave the corresponding code point unmapped.

While outside the scope of this proposal, a related issue may merit attention: some characters in the Extension blocks incorporate components visually identical to the Hangul letter ㅇ (a circle). These characters were primarily used for Korean transliterations. While their shape resembles the Latin letter O or the ideograph zero 〇, they are distinct in script origin and semantic function. It is important to note that all current CJK Unified Ideographs use components from within the Han script—including Kangxi radicals, stroke components, and CJK symbol blocks—in their IDS structures. Even when curly-brace notation is used to mark unencoded components, those placeholders still represent Han-based elements. No existing CJKUI IDS string includes Latin letters, kana, hangul letters, or Zhuyin directly as structural parts. In rare cases where a component may appear to derive from a non-Han script, it was generally Han-ified before being incorporated. Although such characters are not the main focus here, it may be worth considering whether China intends to conduct a systematic review of already-encoded ideographs that may be nominally script-hybrid in nature, and whether future updates to GB 18030 might reflect such classification—for example, by leaving selected code points unmapped.

In summary, multi-syllabic ideographs composed purely of Han components can continue to be encoded as CJK Unified Ideographs, provided certain precautions are taken. Proposals should be encouraged to use fully Han-based representations, avoiding any inclusion of Latin or other non-Han script elements. Distinctions should be made between abbreviations that have documented usage—especially in Han-based regions—and those that do not. This approach allows the IRG to expand the coverage of useful abbreviatory characters in Chinese, Japanese, Korean, and other Han-using contexts, while maintaining clear boundaries between pure Han ideographs and script-hybrid forms.

IV. Conclusion and Requested Actions

Although both multi-syllabic characters and script-hybrid ideographs may function as abbreviations, it is important to maintain a clear distinction between them. The acceptance of Han-only abbreviation characters—such as those representing multi-syllable expressions—should not be taken to imply acceptance of structurally distinct forms that incorporate non-Han scripts. In this regard, the term “abbreviation” should not be used to blur the boundary between typologically distinct categories. Encoding decisions should continue to be based on the structural identity of the components, not merely their function.

Action Requested:

- I recommend IRG will continue discussion on relevant terminology and categorization, and support the encoding of well-attested multi-syllabic characters.
- I also suggest that China avoid leaving code-point gaps in GB 18030 for such abbreviation characters where justified by Han-based usage.