Title: Proposal to accept the submission to register the “PanCJKV” IVD collection
Source: Ken Lunde (Adobe)
Status: IVD Registrar contribution
Action: For consideration by UTC
References: UTS #37 & UAX #38
Date: 2016-03-10

Abstract

The proposed “PanCJKV” IVD collection—intended for Pan-CJK font implementations that distinguish glyphs on a per-region basis, to include implementations that serve a limited set of regions, such as Pan-Chinese ones—is designed to be general-purpose and future-proof. Single-region fonts, which represent the vast majority of modern CJK fonts, do not benefit from this IVD collection as all of the glyphs in such fonts are already designed according to the conventions of the target region. Furthermore, while it is imaginable that some font implementations may opt to include a mixed selection of glyphs that are not defined by a particular standard, regional or otherwise, such implementations are out of the scope of the proposed IVD collection, because the selection of a particular font is not be expected to be maintained in “plain text” environments.

With that stated, the situation surrounding the registration of this IVD collection is somewhat unique in that I am serving dual roles: the IVD (Ideographic Variation Database) Registrar and the registrant of the proposed IVD collection. The specific concerns are that 1) in the former role I am authorized by the UTC to accept IVD collection registration submissions and issue an associated PRI, meaning that UTC approval is not technically necessary; and 2) in the latter role I completely understand that this IVD collection would result in a large and unprecedented number of IVSes (Ideographic Variation Sequences), which suggests that some amount of UTC discussion and possibly approval—at least in terms of accepting the submission itself—would be a prudent measure. I prefer to ward off any possible “conflict of interest” feelings given the dual roles.

Proposal

This document proposes that the “PanCJKV” IVD collection submission be accepted for registration and processed in the usual manner, which first entails the IVD Registrar issuing a PRI, followed by the IVD collection registrant responding to any and all feedback prior to the registration of the IVD collection and of the sequences in that collection.

Introduction

The idea of the “PanCJKV” IVD collection parallels the development of the Adobe-branded Source Han Sans and Google-branded Noto Sans CJK projects, which represent the same Pan-CJK typeface design and are identical other than their names. The fundamental intent of this IVD collection is to represent regional variation of CJK Unified Ideographs in “plain text” environments.

What is a Pan-CJK font? Put simply, a Pan-CJK font is based on a uniform typeface design that is intended to serve more than one East Asian language or region, and therefore requires more than one glyph for a non-zero number of CJK Unified Ideographs.[*] Of course, the main benefit of a Pan-CJK font is to share a significant number of glyphs across the supported regions.

There are currently two common methods for representing regional variation in Pan-CJK fonts, which are described below:

- Through the use of separate font resources whose default glyphs adhere to the conventions of a particular region. This method involves specifying the appropriate font for the selected text.

- Through the use of language-tagging and the OpenType 'locl' (*Localized Forms*) feature, which requires that an application supports this feature, and that the text is properly language-tagged. This method is used to override the default glyphs.

Unfortunately, neither method allows regional variation to survive in "plain text" environments, such as when copying text that is either selected with a specific font whose default glyphs are intended for a particular region, or text that has been carefully language-tagged.

The way in which variation selectors are assigned for new sequences involves registering the next available one for a particular base character. However, given that the proposed IVD collection would involve registering eleven IVSes for every CJK Unified Ideograph, doing so in a uniform manner would lead to a greater degree of predictability. The current distribution of variation selectors also suggests that consuming the last eleven ones, VS246 through VS256, would make the most sense. Nothing written in UTS #37 would prevent this, as variation selector assignments are made at the discretion of the IVD Registrar.

The table below lists the variation selectors that would be completely consumed by this IVD collection, along with their code points and the two-letter ISO 3166-1 alpha-2 country code that corresponds to the intended region:

Variation Selector	**Code Point**	**Region Code**
VS246	U+E01E5	VN (Việt Nam)
VS247	U+E01E6	KP (DPRK)
VS248	U+E01E7	KR (ROK)
VS249	U+E01E8	JP (Japan)
VS250	U+E01E9	MY (Malaysia)
VS251	U+E01EA	MO (Macao SAR)
VS252	U+E01EB	HK (Hong Kong SAR)
VS253	U+E01EC	TW (ROC)
VS254	U+E01ED	SG (Republic of Singapore)
VS255	U+E01EE	CN (PRC)
VS256	U+E01EF	XK (Kāngxī)—pseudo-region

[*] While the same is true of other characters, particularly punctuation whose shape and position within the em-box can vary according to regional conventions, those characters are out of the scope of the IVD.

There are 80,388 CJK Unified Ideographs in Unicode Version 8.0, which translates into 884,268 IVSes when supporting all eleven regions. There is some degree of correspondence between most of these regions and the columns of the multiple-column CJK Unified Ideograph code charts, or the IRG sources, but it is neither absolute nor complete. Two good examples are SG (Republic of Singapore) and MY (Malaysia): the former is covered under the kIRG_GSource source field; the latter is not yet covered, but could potentially be covered under the same.

In order to accommodate the possibility of support for additional regions in the future, the variation selector assignments are in seemingly reverse order, starting from the final variation selector, VS256 (`U+E01EF`), which is assigned to the XK pseudo-region that corresponds to so-called Kāngxī forms, from the Kāngxī dictionary (康熙字典 *kāngxī zìdiǎn*).*

The proposed sequence identifiers take the form of the standard glyph name for the base character that uses well-established "`uni`" (BMP) or "`u`" (non-BMP) prefixes, an underscore (`U+005F`), and finally an uppercase two-letter region code. The form of `U+4E00` as used in Japan (JP) would thus use "`uni4E00_JP`" as the sequence identifier.

The open source PanCJKV IVD Collection project that I set up on GitHub serves three purposes: 1) to provide a preview of the IVD's standard *IVD_Collections.txt*, *IVD_Sequences.txt*, and *IVD_Stats.txt* data files as they would look if the "PanCJKV" IVD collection were to be registered; 2) to eventually serve as the permanent site that describes this IVD collection, if registered; and 3) to provide tools and implementation examples for the benefit of developers (described later in this document).

Regional Variation Versus Language Variation

There is some amount of controversy as to whether CJK Unified Ideographs vary according to language or region (aka country), but in the context of actual implementations, meaning typeface designs and the fonts based on them, by far the most common approach is for the glyphs to adhere to the conventions of a particular country, meaning that regional variation is the norm. This is the fundamental reason why the "PanCJKV" IVD collection specifies regions not languages in the sequence identifiers.

Why Register IVSes For All Regions?

While it may appear wasteful to register IVSes for all regions and for all CJK Unified Ideographs, doing so provides to implementers the freedom and flexibility to supply glyphs that they deem to be appropriate for the supported regions. Doing so also makes the IVD collection future-proof in the sense that any possible horizontal extension is already accommodated by a registered IVS.

Furthermore, the degree to which glyphs can vary according to regional conventions depends on the typeface style (such as sans serif, serif, script, and so on) and the specific typeface design. The "PanCJKV" IVD collection accommodates all common typeface styles and the typeface designs based on them, and thus alleviates the need to register separate Pan-CJK IVD collections for each typeface style or each typeface design.

* The order of the characters of the XK region tag is intentional, which is simply a transposed form of the more intuitive KX. This was done to ensure that a valid private use ISO 3166-1 alpha-2 country code is used. While the possibility of KX becoming a valid country code in the future is relatively low, it is not outside the realm of extreme possibilities and non-zero, hence the use of XK.

Relationship to Unicode Versions

Unicode Version 8.0 includes 80,388 CJK Unified Ideographs. As CJK Unified Ideographs are added to new versions of Unicode, additional "PanCJKV" IVSes will need to be registered to accommodate them. This process is largely mechanical and thus non-controversial.

UTS #37 Considerations

There are two requirements imposed by UTS #37 (*Unicode Ideographic Variation Database*) that may require special treatment or consideration for the "PanCJKV" IVD collection:

- IVD collection submissions, to include submissions of additional sequences for a registered IVD collection, should not specify the variation selector. The IVD Registrar is tasked to assign the variation selectors to the registered sequences.
- A representative glyph is required for each registered sequence.

The first requirement is easily satisfied by the IVD registrant (aka me) simply not including variation selectors in the submission, but to instead request to the IVD Registrar (aka also me) to uniformly assign variation selectors to each of the eleven regions.

The second requirement is trickier, but can perhaps be satisfied by deferring to the Unicode code charts that correspond to CJK Unified Ideographs, along with the following statement: *The "PanCJKV" IVSes for each CJK Unified Ideograph are expected to be displayed according to the conventions and limitations of a particular implementation, in terms of which particular regions are supported and the glyphs that are supplied for them , and that there is no guarantee that characters will display according to the Unicode code charts nor according to regional conventions.*

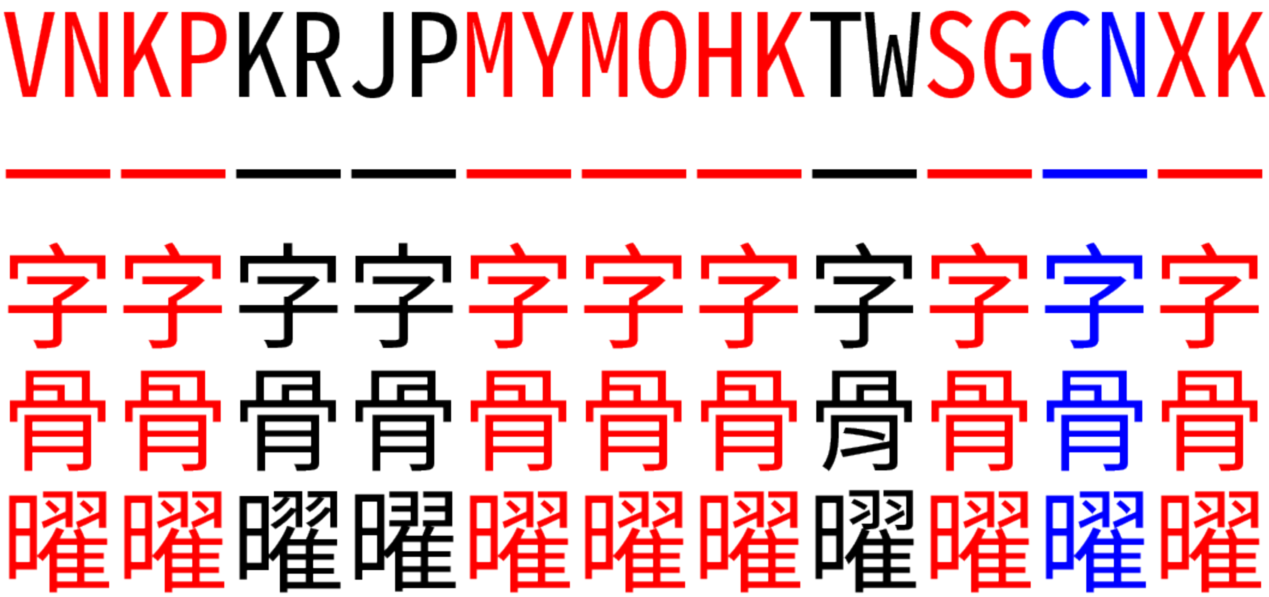
Code Point Versus Regional Glyph Coverage

While purely an implementation detail, and completely independent of the "PanCJKV" IVD collection, typical Pan-CJK fonts include two types of coverage: code points and the (regional) glyphs to which they map. The former is necessarily a superset of the latter, meaning that some code points may lack a glyph that is appropriate for one or more of the supported regions, and will instead display using the default glyph in the case when there is only a single glyph available, or the closest approximation in the case when multiple glyphs are available (it is up to each implementation to determine the closest approximation). Some characters that are truly single-source, such as the many thousands of PRC simplified forms, arguably require only a single glyph, and are expected to display as such regardless of the selected region.

Implementation Examples

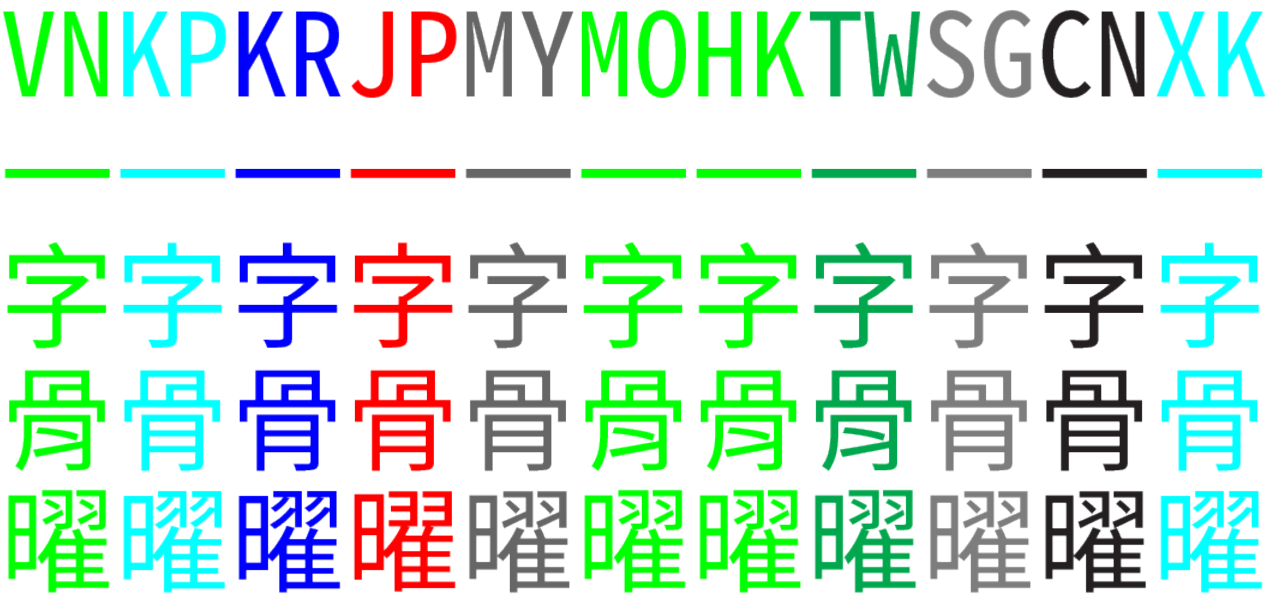
Two implementation examples have been developed, in the form of fully-functional fonts that are based on Source Han Sans, whose IVS coverage differs only in which regions are supported.

The first implementation example, *SourceHanSansR04-Regular.otf*, supports the same four regions that are formally supported in the current version of Source Han Sans (Version 1.004 as of this writing), specifically CN, TW, JP, and KR. Given that 29,777 CJK Unified Ideograph code points are supported, this translates into 119,108 IVSes. The image below shows how 一 (`U+4E00`), 字 (`U+5B57`), 骨 (`U+9AA8`), and 曜 (`U+66DC`) are handled via this example font, using blue to indicate the default (CN) glyph, black to indicate the glyphs for supported regions (KR, JP, and TW), and red to indicate the glyphs for unsupported regions that necessarily default to the default (CN) glyph:



The way in which unsupported regions are handled is not particularly helpful, mainly because some of them could more appropriately follow the conventions of another supported region rather than the default one, which takes us to the second and more compelling implementation example.

The second implementation example, *SourceHanSansR11-Regular.otf*, supports all eleven regions, and aliases the seven unsupported regions to the four formally supported ones in a way that most accurately adheres to regional conventions. Malaysia and the Republic of Singapore more closely follow the conventions used in PRC, hence MY and SG alias to CN. Hong Kong SAR, Macao SAR, and Việt Nam more closely follow the conventions used in ROC, hence HK, MO, and VN alias to TW. DPRK and the Kāngxī dictionary more closely follow the conventions used in ROK, hence KP and XK alias to KR. 327,547 IVSes are supported in this implementation. The image below shows how unsupported regions are aliased to supported ones, using lighter versions of blue (KR), black (CN), and green (TW) to indicate regions that are being aliased:



Both implementation examples and their sources, along with a Perl script that is used to create their UVS (*Unicode Variation Sequence*) definition files from the *IVD_Sequences.txt* file, are provided in the open source PanCJKV IVD Collection project on GitHub.

Conclusion

The proposed “PanCJKV” IVD collection, though designed to be general-purpose and future-proof, is also meant only for Pan-CJK implementations that distinguish glyphs on a per-region basis, to include ones that are intended to support CJK Unified Ideographs for a limited set of regions. And, like other IVD collections, this one can be partially implemented, in terms of which base characters and regions are supported, as demonstrated by the two implementation examples.

Unless the UTC has a good reason to reject its submission, I request that it be handled in the usual manner, meaning that a PRI be issued.

That is all.