L2/21-151

Update to the Syntax for the Unihan Database's `kTotalStrokes` Field

John H. Jenkins

9 August 2021

Summary

The current syntax for the `kTotalStrokes` field in the Unihan Database allows specifying up to two values. Specifically, the description for the field says: "When there are two values, then the first is preferred for zh-Hans (CN) and the second is preferred for zh-Hant (TW). When there is only one value, it is appropriate for both."

While the total strokes of a unified ideograph is more useful for Chinese than other languages, this is still unnecessarily sinocentric and unintuitive. We recommend here an extension to the existing syntax to address this.

Background

It was pointed out in the public feedback to the Unicode 14.0 beta that there is a problem with the `kTotalStrokes` field for `U+537F` 卿 (see the recommendation *Unihan-UTC168-R12* in **L2/21-129**). This is illustrated by the Unicode 13.0 code charts:

In this case, the G-, H-, T-, and V-glyphs are identical and have a stroke count of ten, whereas the J- and K-glyphs have a stroke count of 12. *Unihan-UTC168-R12* recommended that the `kTotalStrokes` value for `U+537F` 卿 be changed to "10 12" to reflect this. This cannot be done given the field's current description.

Proposal

We suggest that the syntax for the `kTotalStrokes` field be altered to accommodate different IRG source-specific values. This involves tagging a value with a series of IRG source identifiers. We recommend that the identifiers be either a single upper-case letter ("G," "H," "M," "T," "J," "K," "P," "V," "U," and "S") or a single upper-case letter followed by a single lower-case letter ("Uk"). This makes it easy to find the boundaries of the source identifiers without resorting to misleading single-letter identifiers or encumbering the field with additional punctuation. This format uses "P" instead of "KP" so that it can be consistent with the format of `kIICore` and `kUnihanCore2020`.

Because these IRG source identifiers are used in several fields, we recommend that a new section be added to UAX #38 documenting them. Following §3.8 seems a reasonable place. The descriptions for the `kIICore` and `kUnihanCore2020` fields would then be updated appropriately.

The syntax for the `kTotalStrokes` field would become something like:

```
\d+(:([GHJKMPSTUV]|Uk)+)?
```

The field's description would be updated to:

> The total number of strokes in the character (including the radical). Each value consists of a decimal value followed by an optional series of IRG-source identifiers.
>
> The IRG source identifiers indicate the IRG sources for which a particular value is preferred. If a particular IRG source is not present in any value, then the preferred kTotalStrokes value for this IRG source is unspecified.
>
> IRG source identifiers may not be repeated, and every IRG source identifier must correspond to a defined kIRG_*Source value.
>
> The preferred value is the one most commonly associated with the character in modern text using customary fonts.
>
> This field is targeted specifically for use by CLDR collation and transliteration. As such, it is subject to considerations that help keep pinyin-based Han collation (and its tailorings) and transliteration reasonably stable.

Although it is possible algorithmically to convert from the old syntax to the new, this may not be desirable, to minimize the impact on existing parsers. An umarked number by itself would match a `GT` IRG source identifier, and two unmarked numbers would then match a `G` source identifier followed by a `T` source identifier. That is, "`10`" would be equivalent to "`10:GT`", and "`10 12`" to "`10:G 12:T`".

Under this new syntax, `U+537F` 卿 could be assigned the `kTotalStrokes` value "`10:GHTV 12:JK`".

It should be noted that if a character has no `kIRG_*Source`, the corresponding string must not occur in the `kTotalStrokes` IRG source identifiers. The value "`4:GHT`" is valid for `U+4E95`, whereas the value "`4:GHTU`" is not; `U+4E95` has no `kIRG_USource` value.

Moreover, the values in the `kTotalStrokes` field may not be repeated. The field value "`4:GHT 5:G`" is invalid.

We *do not* require that if a `kIRG_*Source` value is defined for a given character, then the corresponding identifier is used in the `kTotalStrokes` value. This is because of the practical impossibility of generating this data in any reasonable length of time. For `U+4E95`, for example, the `kTotalStrokes` value "`4:GT`" is acceptable, but the value "`4:GHTJKPV`" would be ideal.