L2/21-226

Proposed Supplement to the Unihan Database's `kTotalStrokes` Field

John H. Jenkins

2 November 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 a new supplementary field in the Unihan database 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 14.0 code charts:

In this case, the G-, H-, T-, and V-source glyphs are virtually identical and have a stroke count of ten, whereas the J- and K-source 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 a new field be added to the Unihan database, provisionally named `kIRGStrokeCount`. The data and syntax of the `kTotalStrokes` field would be left as-is.

We recommend that the syntax for the `kIRGStrokeCount` field involve tagging a value with a series of IRG source identifiers, and that the identifiers be single uppercase letters ("B", "G," "H," "M," "T," "J," "K," "P," "V," "U," and "S"). This uses "B" instead of "UK" for the United Kingdom and "P" instead of "KP" for North Korea, so that it can be consistent with the format of `kIICore` and `kUnihanCore2020`.

Because many of these single-letter 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 `kIRGStrokeCount` field would become something like:

`(\d+:[BHJKMPSUV]+)|-`

The field's description would be updated to:

> The total number of strokes in the character (including the radical). Each value consists either of a decimal value followed by one or more IRG source identifiers (see <insert reference here>), or of the special value "-".
>
> The IRG source identifier indicates the IRG sources for which a particular value is preferred. The source identifiers "G" and "T" are not used in this field, as these IRG sources are fully covered by the kTotalStrokes field.
>
> The stroke count value is the one for the glyph as shown in the code charts.
>
> Multiple stroke counts are listed in increasing numeric order. Stroke counts may not be repeated.
>
> If there is a single kTotalStrokes value for a character, the IRG sources sharing this stroke count should not be explicitly listed. If all IRG sources share this stroke count, then the value of "-" is used. The kIRGStrokeCount value for U+4E95 is therefore "-" instead of "4:HJKPV".

> The kIRGStrokeCount "-" value may not be used where there are two kTotalStrokes values for a character. Thus, the kIRGStrokeCount value for U+9AA8 骨 is "10:HJKPV".
>
> For IRG sources which do not include a source reference, the kIRGStrokeCount field should not have a corresponding value.
>
> Unlike the kTotalStrokes field, the data in this field is not to be taken as exhaustive. Where it is defined for a character, however, it includes explicit or implicit values for all IRG sources containing the character.

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

It should be noted that if a character has no `kIRG_*Source`, the corresponding string must not occur in the `kIRGStrokeCount` IRG source identifiers. The value "`12:JK`" is valid for `U+537F`, whereas the value "`12:JKU`" is not; `U+537F` has no `kIRG_USource` value.

Moreover, the source identifiers in the `kTotalStrokes` field may not be repeated. The field value "`4:H 5:H`" is invalid.

We *do not* require that if a `kIRG_*Source` value other than G and T is defined for a given character, that there be a corresponding `kIRGStrokeCount` value. This is because of the practical impossibility of generating this data in any reasonable length of time.

To limit clutter, we include the following for situations where there is but one `kTotalStrokes` value (meaning a stroke count shared by the G and T sources):

The `kIRGStrokeCount` field should not contain a value for IRG sources whose stroke counts agree with the `kTotalStrokes` field value. For `U+537F`, the `kIRGStrokeCount` value would therefore be "`12:JK`" instead of "`10:HV 12:JK`".

If *all* the IRG sources agree on a stroke count, then the special value of "`-`" is used for the `kIRGStrokeCount` value. `U+4E95` 井 could then have the `kIRGStrokeCount` value of "`-`" instead of "`4:HJKPV`".