L2/19-015

Proposed Overhaul of kZVariant Data in the Unihan Database

John H. Jenkins
7 January 2019

This is a heavily-revised version of L2/15-266.

The `kZVariant` data in the Unihan database is known to be of uneven quality. I recommend we resolve this problem as follows:

1) Amend the description of the field so that the first paragraph reads: *"The Unicode value(s) for known z-variants of this character, that is, variants which would ordinarily be treated as unifiable. This includes cases where unifiable variants have been separately encoded due to the source separation rule or non-cognate rule, as well as cases where unifiable variants have been separately encoded owing to errors in the unification process."*

Other rephrasings have been suggested and should be taken into account.

2) Completely replace the existing data. The new data should consist of the examples of non-unification given in ISO/IEC 10646 Annex S, section S.3 (Source code separation examples), plus a list of known unification errors. Eiso Chen, Henry Chan, and Richard Cook all have lists of duplicates and unification errors which can be used.

The data needs to be further curated so that it is both transitive and reflexive; that is, if A is a z-varaint of B and B is a z-variant of C, then A is a z-variant of C, and if A is a z-variant of B, then B is a z-variant of A.

3) Redefine the regular expression for the field. I would recommend something like:

U\+[23]?[0-9A-F]{4}(\[DINSU\])?(\{[GHJKMPU]+\})?(<k[A-Za-z0-9]+(,k[A-Za-z0-9]+)*)?

Breaking this down in human-readable fashion, the pieces are:

U\+[23]?[0-9A-F]{4}
: The code point

(\[DINSU\])?
: The kind of z-variant.
  - D: Duplicate
  - I: IRG Example from Annex S
  - N: Non-cognate rule
  - S: Source-code separation
  - U: Unifiable

(\{[GHJKMPU]+\})?
: The locale(s) for which this pair represents an actual z-variant.
: The locale is just a list of IRG sources, with P instead of KP.
: If no locales are listed, then it applies to all locales.

(<k[A-Za-z0-9]+(,k[A-Za-z0-9]+)*)?
: The UAX #38 sources for this data (if any).

It may be desirable to allow multiple type/locale pairs.

The action item would be on me to coordinate with the Unihan Subcommittee to generate data, then once that's done to upload it into the Unihan database. This would be for Unicode 13.0.