Collation and Correcting Errors in the Unicode Character Database

Author: Richard Wordingham
Date: 24 July 2012

I understand from a review of old discussions that the program used to generate default collation tables for both 14651 and the UCA collations (well, DUCET at least) , "sifter", uses a modified version of UnicodeData.txt. One of the modifications is a *non-standard* version of compatibility decomposition including bespoke decomposition types such as 'sort'.

We have recently been made aware that sifter uses the numeric value and type fields, or at least the value field, in determining the collation of characters of general categories Nl and No. (Compatibility decomposition may play a role in stopping compatibility Roman numerals sorting as numbers.) When the numeric values were corrected for U+1240F CUNEIFORM NUMERIC SIGN FOUR U to U+12414 CUNEIFORM NUMERIC SIGN NINE U (from 4 to 9 to 40 to 90), this stopped them being collated as secondary variants of positional decimal digits. Only U+1240F and U+12410 can be considered sexagesimal digits. This unplanned change to collation was reversed by modifying the sifter program itself.

Ken Whistler has formally proposed that this modification to sifter be removed by reversing the correction to UnicodeData.txt. Would not a better approach be for UnicodeData.txt to be correct and keep the incorrect values, possibly with a tag, in the file used by sifter? There are other errors in numeric values in UnicodeData.txt:

1) Of the cuneiform numbers and punctuation, I can confirm that only the members of the DISH, ASH and ASH TENU series truly have the values in the range 1 to 9. The DISH series are sexagesimal digits, not mere numbers, so I believe they should have numeric type "digit". The other series have other values, being multiples of 10, 60, 600, 3,600, 36,000 or 216,000.

This remark also applies to the following six numbers proposed in L2/12-207 (a.k.a. ISO/IEC JTC1/SC2/WG2/N4277). Rather than

```
12469;CUNEIFORM NUMERIC SIGN FOUR U VARIANT FORM;Nl;0;L;;;;4;N;;;;;
1246A;CUNEIFORM NUMERIC SIGN FIVE U VARIANT FORM;Nl;0;L;;;;5;N;;;;;
1246B;CUNEIFORM NUMERIC SIGN SIX U VARIANT FORM;Nl;0;L;;;;6;N;;;;;
1246C;CUNEIFORM NUMERIC SIGN SEVEN U VARIANT FORM;Nl;0;L;;;;7;N;;;;;
1246D;CUNEIFORM NUMERIC SIGN EIGHT U VARIANT FORM;Nl;0;L;;;;8;N;;;;;
1246E;CUNEIFORM NUMERIC SIGN NINE U VARIANT FORM;Nl;0;L;;;;9;N;;;;;
```

they should actually be

```
12469;CUNEIFORM NUMERIC SIGN FOUR U VARIANT FORM;Nl;0;L;;;;40;N;;;;;
1246A;CUNEIFORM NUMERIC SIGN FIVE U VARIANT FORM;Nl;0;L;;;;50;N;;;;;
1246B;CUNEIFORM NUMERIC SIGN SIX U VARIANT FORM;Nl;0;L;;;;60;N;;;;;
1246C;CUNEIFORM NUMERIC SIGN SEVEN U VARIANT FORM;Nl;0;L;;;;70;N;;;;;
1246D;CUNEIFORM NUMERIC SIGN EIGHT U VARIANT FORM;Nl;0;L;;;;80;N;;;;;
1246E;CUNEIFORM NUMERIC SIGN NINE U VARIANT FORM;Nl;0;L;;;;90;N;;;;;
```

I suggest giving them compatibility decompositions to U+1240F to U+12414 – it would certainly make sense to collate them as compatibility variants.

2) U+1D369 COUNTING ROD TENS DIGIT ONE to U+1D371 COUNTING ROD TENS DIGIT NINE are digits in a decimal place value system, so they should have numeric type "digit" and values 1 to 9.

3) The alternating between digit sets is also seen in the Telugu fraction digits, U+0C78 to U+0C7E. It is not clear to me why these have numeric type "numeric" rather than "digit".

4) U+3021 HANGZHOU NUMERAL ONE to U+3029 HANGZHOU NUMERAL NINE are digits in a decimal place value system, so they should have numeric type "digit". (I believe this will not affect collation.)

It makes sense to me to correct UnicodeData.txt and temporarily have erroneous old values in the data file used to generate DUCET. For all I know, there may even already be a notation to handle temporary discrepancies, as when general categories change. I say temporarily, because the CLDR root locale has been correcting (or rationalising, at least) the default collation orders tailored for use with human languages, and as part of the process moved letter-like and other numbers from amongst the symbols to join the 'decimal digits' and those with secondary differences from them. It will be a small logical step to rationalise the orders within scripts, even if numeric sorting is not supported properly. At some point, ISO 14651 will either be scrapped or aligned with the CLDR root locale, and then DUCET will follow step.