L2/14-140R

CTT: Remove Most Cyrillic Contractions

2014-aug-05
Markus Scherer

Proposal

Remove from the Common Template Table all Cyrillic contractions, except for й (and Й) which are used in Russian and many other languages.
(й=U+0439, Й=U+0419: `Cyrillic Small/Capital Letter Short I`)

For example, change mappings like

```
<U04D9> <S04D9>;<BASE>;<MIN>;<U04D9> % CYRILLIC SMALL LETTER SCHWA
<U04D8> <S04D9>;<BASE>;<CAP>;<U04D8> % CYRILLIC CAPITAL LETTER SCHWA
<U04DB> <S04DB>;<BASE>;<MIN>;<U04DB> % CYRILLIC SMALL LETTER SCHWA WITH DIAERESIS
<U04D9_0308> <S04DB>;<BASE>;<MIN>;<U04DB> % CYRILLIC SMALL LETTER SCHWA WITH DIAERESIS
<U04DA> <S04DB>;<BASE>;<CAP>;<U04DA> % CYRILLIC CAPITAL LETTER SCHWA WITH DIAERESIS
<U04D8_0308> <S04DB>;<BASE>;<CAP>;<U04DA> % CYRILLIC CAPITAL LETTER SCHWA WITH DIAERESIS
```

to

```
<U04D9> <S04D9>;<BASE>;<MIN>;<U04D9> % CYRILLIC SMALL LETTER SCHWA
<U04D8> <S04D9>;<BASE>;<CAP>;<U04D8> % CYRILLIC CAPITAL LETTER SCHWA
<U04DB> <S04D9>;"<BASE><TREMA>";"<MIN><MIN>";<U04DB> % CYRILLIC SMALL LETTER SCHWA WITH
                                                       DIAERESIS
<U04DA> <S04D9>;"<BASE><TREMA>";"<CAP><MIN>";<U04DA> % CYRILLIC CAPITAL LETTER SCHWA WITH
                                                       DIAERESIS
```

and remove the then-unused contraction strings like

```
collating-element <U04D8_0308> from "<U04D8><U0308>" % decomposition of CYRILLIC CAPITAL
                                                       LETTER SCHWA WITH DIAERESIS
collating-element <U04D9_0308> from "<U04D9><U0308>" % decomposition of CYRILLIC SMALL
                                                       LETTER SCHWA WITH DIAERESIS
```

and remove their primary weights from the `% First-level weight assignments`

```
<S04DB> % CYRILLIC SMALL LETTER SCHWA WITH DIAERESIS
```

Rationale

The presence of these contractions makes collation of Cyrillic text slower for all implementations of UCA and ISO 14651, due to the required lookahead. For example, it slows down Cyrillic-text collation by 20-30% in ICU, which is by far the most prevalent implementation of the Unicode Collation Algorithm and the Common Template Table.

Performance of collation is vital because collation is used in many processes on the internet and in databases, including sorting, searching in a sorted table, full-text search, range selection, etc. In most applications, collation is done incrementally, on demand, rather than by using sort keys.

The original rationale for including contractions for many Cyrillic accented letters in the CTT and the DUCET was to provide more accurate default (untailored) collation for many languages. However, almost all languages require tailoring anyway; and most implementations are based on the Unicode CLDR data which tailors its default table and provides further tailorings for many languages.

Because the Cyrillic contractions slow down collation of Cyrillic text in all languages and all implementations, almost all of them are suppressed. For example, for Russian, all but the й and Й contractions are suppressed. The Unicode CLDR project is in the process of removing the Cyrillic contractions (except for й and Й) from its default table, and adding only the relevant ones to the per-language tailorings.

The Cyrillic contractions complicate the generation of the CTT and the DUCET, complicate implementations by having to suppress them, and complicate the creation and maintenance of per-language tailorings by having to find irrelevant contractions to remove rather than relevant contractions to add.

Therefore, the removal of most of the Cyrillic contractions from the CTT and the DUCET would simplify the tools and processes used for generating and maintaining the default table, and would simplify tailorings and clarify their relationship to the default table.

CTT contractions

Most of the CTT Cyrillic contractions are for non-slavic letters (and an archaic letter) that are each used in very few languages with relatively small numbers of speakers. Four others are used in one or two languages each.

ӑӓӛѓӗӝӟӥїйӧӫќўӱӳӵӹӭѷ + uppercase forms & decompositions/equivalents
See the equivalent DUCET chart: http://www.unicode.org/charts/collation/chart_Cyrillic.html

Occur in CLDR exemplar characters (these are slavic letters): й (Russian, Ukrainian, Azerbaijani, Belarusian, Bulgarian, Kazakh, Kyrgyz, Mongolian, Ossetic, Sakha, Tajik, Uzbek) ў (Belarusian) ѓ (Macedonian) ќ (Macedonian) ї (Ukrainian, Rusyn)

(Further accented Cyrillic letters occur in CLDR exemplar characters: ё ӣ ӯ)

Do not occur in CLDR exemplar characters (these are non-slavic letters): ӑ (Chuvash) ӓ (Khanty, Sami, Mari) ӛ (Khanty) ӗ (Chuvash) ӝ (Udmurt) ӟ (Udmurt) ӥ (Udmurt) ӧ (Altay, Khakas, Komi, Kurdish, Mari, Udmurt) ӫ (Even, Khanty) ӱ (Altai, Khakass, Khanty, Mari) ӳ (Chuvash) ӵ (Udmurt) ӹ (Mari) ӭ (Sami)

Archaic Cyrillic letter does not occur in CLDR exemplar characters: ѷ (Church Slavonic)