L2/14-106

DUCET: Remove Most Cyrillic Contractions

2014-apr-29
Markus Scherer

Proposal

For the next UCA version after 7.0.0, remove from the DUCET all Cyrillic contractions, except for й (and Й) which are used in Russian and many other languages.

Rationale

Most of the DUCET 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.

The presence of these contractions makes Russian collation slower (e.g., by about 20-30% in ICU).

It is far more understandable to have per-language tailorings add relevant contractions than to remove irrelevant ones.

The CLDR TC has agreed to remove these Cyrillic contractions from the CLDR root collation. (CLDR ticket #7246) Tailoring the CLDR root collation compared with the DUCET is non-trivial, and keeping the number of differences small is desirable.

DUCET contractions

http://www.unicode.org/charts/collation/chart_Cyrillic.html
ӑӓӛѓӗӝӟӥїйӧӫќўӱӳӵӹӭѷ + uppercase forms & decompositions/equivalents

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)