L2/19-013

UAX 14 Improve Line Breaking Around Numbers

2019-01-07
Andy Heninger

This is a proposal for UAX 14 rule changes to improve line breaking around numbers.

Because we are late in the cycle for Unicode 12, I suggest that it be implemented first as a tailoring in CLDR & ICU 64, and promoted to UAX 14 after we have some experience with it.

Background Docs

[147-A76] Action Item for Andy Heninger: Propose a solution to the problem he documented in the PRI #322 feedback of April 20, re linebreaking around numbers.
http://www.unicode.org/cgi-bin/GetL2Ref.pl?147-A76

From http://www.unicode.org/review/pri322/

> **Date/Time:** Wed Apr 20 15:19:33 CDT 2016
> **Name:** Andy Heninger
> **Report Type:** Error Report
> **Opt Subject:** UAX 14 feedback, PRI #322
> The UAX-14 line breaking of numbers beginning with a decimal point can be bad.
> Consider the string "`start .789 end`".
>
> With the default rules there will only be one break, "`start .789 |end`".
> Rule LB13, "x IS" will prevent a break before the number.
>
> With the tailoring of numbers from example 7 of section 8.2 there will be an unexpected break after the full stop, yielding "`start .|789 |end`", because the regular expression for numbers does not allow a character of class IS to precede the first digit.
>
> How this might be fixed will require some thought.
>
> This problem was originally reported by Bernhard Fey in an ICU bug report,
> https://unicode-org.atlassian.net/browse/ICU-12017

The correct line breaking for this example would be `"start |.789 |end"`

Considerations:

The IS (Infix Numeric Separator) line break class includes more than just numeric decimal separators.

The IS characters are:

002C	COMMA
002E	FULL STOP
003A	COLON
003B	SEMICOLON
037E	GREEK QUESTION MARK (canonically equivalent to 003B)
0589	ARMENIAN FULL STOP
060C	ARABIC COMMA
060D	ARABIC DATE SEPARATOR
07F8	NKO COMMA
2044	FRACTION SLASH
FE10	PRESENTATION FORM FOR VERTICAL COMMA
FE13	PRESENTATION FORM FOR VERTICAL COLON
FE14	PRESENTATION FORM FOR VERTICAL SEMICOLON

French Spacing Rules

French requires spaces before and after colons and semicolons, which are included in the Line Break IS class along with period and comma. This adds a constraint to this problem. Well-formed French punctuation should not be separated from its preceding word, which eliminates the simple solution of simply breaking before IS characters that are preceded by a space.

Proposal

Changes or additions to the existing UAX 14 text are highlighted.

Modify LB13 and LB14 to break before an IS (decimal point) if it looks like the beginning of a number.

With the existing rule LB13 there is no break before an IS in any context.

***LB13** Do not break before ‘]’ or ‘!’ or ~~‘;’~~ or ‘/’, even after spaces.*

× CL

× CP

× EX

~~× IS~~

× SY

***LB14** Do not break after ‘[’, even after spaces.*

OP SP* ×

***LB14a** Break before a numeric separator that begins a number and follows a space*

SP ÷ IS NU

***LB14b** Do not break before numeric separators, even after spaces.*

× IS

In #14 section 8 (Customization) modify the regular expression for numbers to permit a leading IS (decimal point).

8 Customization

Example 7. Regular expression-based line breaking engines might get better results using a tailoring that directly implements the following regular expression for numeric expressions:

(PR | PO) ? (OP | HY) ? IS ? NU (NU | SY | IS) * (CL | CP) ? (PR | PO) ?

I have prototyped these changes in ICU. They are implementable, and appear to work as expected.