L2/09-263

Date: August 3, 2009

To: Unicode Technical Committee

From: CLDR (per meetings March 31, July 7, July 29)

Subject: Proposal to allow line break in “abc .def” or “abc .123”

Per UAX #14, there is no line break opportunity in the strings “abc .def” or “abc .123”, although there is if the FULL STOP is deleted. This is due to rule LB13, “Do not break before ‘]’ or ‘!’ or ‘;’ or ‘/’, even after spaces,’ which is formalized as follows:

× CL
× EX
× IS [includes U+002E FULL STOP]
× SY

Minimum proposal: Add a (tailorable) rule LB12b, “Allow a break between space and '.' if immediately followed by letters or digits,” as follows:

SP ÷ IS AL
SP ÷ IS NU

Details

1. The minimal proposal actually applies not just to FULL STOP, but to all characters of Linebreak class IS, Numeric Separator (Infix). This includes the following, and all seem to be good candidates for treatment similar to FULL STOP in this situation:

002C COMMA
002E FULL STOP
003A COLON
003B SEMICOLON
037E GREEK QUESTION MARK (canonically equivalent to 003B SEMICOLON)
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

2. Note that Linebreak class CL, Closing Punctuation, includes many characters similar to the above, and these should also be considered for similar treatment:

3001 IDEOGRAPHIC COMMA
3002 IDEOGRAPHIC FULL STOP
FE11 PRESENTATION FORM FOR VERTICAL IDEOGRAPHIC COMMA
FE12 PRESENTATION FORM FOR VERTICAL IDEOGRAPHIC FULL STOP
FE50 SMALL COMMA
FE52 SMALL FULL STOP
FF0C FULLWIDTH COMMA

FF0E FULLWIDTH FULL STOP
FF61 HALFWIDTH IDEOGRAPHIC FULL STOP
FF64 HALFWIDTH IDEOGRAPHIC COMMA

Linebreak class CL also includes many other characters such as the following:

0029 RIGHT PARENTHESIS
005D RIGHT SQUARE BRACKET
007D RIGHT CURLY BRACKET
…
2046 RIGHT SQUARE BRACKET WITH QUILL
207E SUPERSCRIPT RIGHT PARENTHESIS
208E SUBSCRIPT RIGHT PARENTHESIS
232A RIGHT-POINTING ANGLE BRACKET
…

We should discuss whether to extend proposed LB12b to include all CL characters, or perhaps just the first subset above (by splitting class CL into two classes):

SP ÷ CL AL
SP ÷ CL NU

For example, according to Kent Karlsson, in text that includes a European-style numeric interval such as “abc]1.5,20]” or “abc]–1,2[”, a line break opportunity should occur after the space in each example (and nowhere else).

Background & History

In Oct. 23-24 2006, the general issue of allowing a line break in "abc .def" was discussed under the subject "Issue with UAX 14", without a specific proposed change to UAX #14 rules. Points raised in that discussion include:

- Mark Davis: I think the goal was to allow for French spaciness, eg "Mon Dieu !" with a space before the "!". However, I don't think that should be done with "." or "/". *(Note: Mark was referring to fraction-slash here, which was not clear, and colored some of the later remarks. See list above.)*
- Michael Everson part1: Ellipses are often . . . spaced.... and in poetry written out unbroken, lines are often indicated by spaced / solidi.
- Kent Karlsson part1: I'd leave them as they are. A "/" at the beginning of a line may have a special meaning (and I think some people write things like "with / without" which should not be broken just before the "/"), and linebreaking "a, b, ..., last." (written with ordinary full stop) [before the ellipsis instead of after the other spaces] would be a bad idea.
- Kent Karlsson part2: Not sure if French uses a space before (partial) sentence ending ellipsis (written with ordinary spaces, of course). [Patrick Andries followed up to say no]
- Murray Sargent: FWIW, Microsoft Word obeys LB13 except for ‘/’ (× SY). Word is willing to wrap a ‘/’ if the ‘/’ is preceded by a space.

- Ken Whistler part1: I agree with Kent and Michael's general reasoning on this item. ["Due to .scriptSuite"] is actually the aberrant case -- namely using what is otherwise terminal punctuation in special case nomenclature as an *initiating* convention for words. In this particular case, of course, we are talking about the Unix convention of using initial "." to indicate hidden file names. I think the burden of special-casing should be on the tailoring of line-breaking that expects to find such elements and break before the "." for them, rather than changing to default in a way which will produce worse line-breaks for what are much more normal textual conventions in the use of "."
- Ken Whistler part2: On the other hand, if the only special cases that need to be dealt with are "SPACE . AL" and if that particular sequence can be identified as a line break opportunity "SPACE x . AL" without impacting otherwise the linebreaking of ".", then perhaps that would be a reasonable modification of UAX #14.
- John Cowan, responding to Ken Whistler part2: The Lojban community, which uses . as a quasi-letter (often at the end of words, sometimes at the beginning, rarely in the middle) would be very pleased by such a change.
- Asmus Freytag, responding to Ken Whistler part1: Precisely my sentiments… Inserting a ZWSP should cause a conformant implementation to allow a break.
- Asmus Freytag, responding to John Cowan: Implementers are certainly welcome to tailor their implementations for both Lojban and Unix spelling conventions. We've taken a deliberate decision to limit the *default* to use limited context of the kind B SP* A, in order to allow a large legacy base of implementations that follow the JIS X-4051 standard (more or less faithfully) to be conformant without complete rearchitecting. In particular, this allows pair-table based implementations to be conformant implementations. Precisely because this particular example uses aberrant conventions, there's no justifications to require the default rules to handle this case at the cost of making many implementations non-conformant. The best solution here is to document the issue and suggest tailoring for Lojban. (As well as to point out that more powerful implementation algorithms will be needed to handle such tailorings).
- Michael Everson part2: Some writers... um... distinguish between the ellipsis of hesitation, which has no preceding space, and ... the ellipsis of deletion, which has....

Based on that discussion, it was decided to address this in CLDR by modifying line break for just the en_US_POSIX locale; this resulted in cldrbug #2126. However, recent discussion in CLDR has suggested that perhaps this issue should be reconsidered by UTC as part of a change to UAX #14, perhaps along the lines suggested by Ken Whistler's part2 comment above—that is, implement a change specifically for “SPACE x . AL”. This would avoid some of the concerns raised in the above discussions, such as changing linebreak behavior for “…”. This would make pair-table implementation of UAX #14 only approximate the rules (see Asmus Freytag's concern), but that may be less of an issue in 2009 than it was in 2006.