L2/20-243

Proposal on material issues in UAX #14 relevant to French

For consideration by the Unicode Technical Committee

2020-09-24
Marcel Schneider (charupdate@orange.fr)

While UTC is currently processing (162-A53) the material proposal L2/20-005 (162-A51) and the formal one L2/20-006 (162-A52), and some thorny issues among the more disruptive ones have been addressed—in L2/20-088 and, for Hebrew, in L2/20-087—I proved unable to cater for other locales (see the latest batch).

This paper is an important update to L2/20005, and a late yet essential part of the UAX #14-related paperwork (that was so time-consuming it caused me to come late on multiple issues). It underscores the momentum of why we need to engage, and is an essential part of my commitment to fix UAX #14. I'm sorry that right now I can't afford to push farther nor to refactor all material suggestions.

Please let me know if to fix UAX #14, UTC needs anything more that I could deliver, eg a draft UAX #14 implementing all obvious fixes, to save Roozbeh Pournader's and Chris Chapman's time, on explicit request.

French punctuation spacing

The means provided in UAX #14 to space out question and exclamation marks, colon and semicolon as well as angle quotation characters are significantly suboptimal because the SPACE characters that are made non-breaking remain justifying. By contrast, French old-school punctuation spacing uses a non-breaking fixed-width thin space with question and exclamation marks and semicolon. Intermediate school uses it also with angle quotation marks, and new-school even with colon. French new-school punctuation spacing is used in France, then Switzerland aligned in 2015, and Quebec is ready to follow according to private email in 2019, else it is using intermediate school. These subtleties only highlight that locale- and usage-dependent punctuation spacing is to be handled on system level through input methods or postprocessors automatically or manually inserting U+202F (and, depending on user expectations, U+00A0). This data is a good fit for CLDR. UAX #14 is to provide the standard framework by recommending the use of appropriate non-breaking spaces, the same way it already does in section GL for non-breaking interword space.

Furthermore, the leading industry does not seem ready to implement rules LB13 through LB17 palliating the initial lack of *NO-BREAK THIN SPACE, or more accurately, palliating the wrong line breaking property value of THIN SPACE, and—if Unicode were to stick with XCCS—that of PUNCTUATION SPACE. In not following up, the industry respects widespread end-user expectations about the ability to control the line breaking behavior through discretionary use of breaking and non-breaking spaces.

Proposed actions

Rules LB13 through LB18 should be aligned on current best practice by raising LB18 before LB13, and by deleting "SP*" in LB14–LB17. — About documentation, please refer to the line edits suggested below.

French-style group separation

The rationale for migrating the group separator space from U+00A0 to U+202F for version 34 of CLDR is valid for all locales separating groups by a space (see list on page 6 of L2/19-112), and it is for UAX #14 in the first

place. UAX #14 is the very hot spot where Unicode support for French started to happen in 2007 under the incentive of Patrick Andries supported by Martin J. Dürst, and by Ken Whistler who had previously curated UTC's 1998 exploit of gaining a non-breaking thin space for general use out of the upcoming Mongolian Space, as backtracked on page 3 of L2/19-112. See Ken Whistler's L2/07-209R, cited on line 4 of page 3 of L2/19-115.

Unicode 1.0 was based on the assumption that the Figure Space of XCCS (0xEE 0x27) was for use *between* figures, while everybody like Donald Knuth knew that digits are separated into groups using a thin space, if not a period, a comma or another tiny graphic. Specifically in XCCS, that would be the Punctuation Space (0xEE 0x24), as there was no thin nor six-per-em alternative. Hence, touting FIGURE SPACE that way is at least surprising. If Unicode was serious about it, then it would have followed through by declaring likewise that PUNCTUATION SPACE was for use *with* punctuation.

Consistently, FIGURE SPACE was removed from its impostor status since Unicode 3.0, L2/20-008 shows from page 7 on. UAX #14 pre-dated TUS 3.0, but it had the brazenness of maintaining FIGURE SPACE's wrongly claimed use case after TUS dropped it. It would keep ghosting through all versions of UAX #14.

From TUS vanished any mention of a group separator space, despite recommending a such is up to the Core Specification, not to the Locale Data Repository. CLDR should rather be about whether there is a space or whatever graphic, not about whether a given locale is currently understanding or handling space characters as per the Standard. In the current context of international standardization, there is literally one single option for a digit-grouping non-breaking space that does not rely on vendor-specific tailoring or higher-level protocols.

Proposed actions

In section GL, the text from "NO-BREAK SPACE is the preferred character" to "keeps the number together for the purpose of line-breaking" should be line-edited. Compared to pages 3–8 of L2/20-005, the text below is considerably improved, expanded and clarified to highlight the point of using NNBSP, not NBSP.

> NO-BREAK SPACE has exactly the same behavior as SPACE in horizontal justification, but without providing any line break opportunity. It is ~~the preferred character to~~ used where two ~~non-hyphenated~~ words are to be ~~visually separated but~~ kept on the same line, as in the case of a title and a name: "Dr.<NBSP>Joseph Becker", provided that NBSP is not tailored as fixed-width. Otherwise, regular NBSP is emulated by <SP, WJ> in justified text. (In case of hyphenation, U+2011 NON-BREAKING HYPHEN is used instead.) ~~When SPACE follows NO-BREAK SPACE, there is no break, because there never is a break in front of SPACE.~~
>
> NARROW NO-BREAK SPACE (NNBSP) has ~~exactly~~ only the same line breaking behavior as NO-BREAK SPACE, while it does not stretch in justified text, and its width is less than the default width of SP and NBSP. This is the preferred space to use where a non-breaking THIN SPACE is required. Examples include grouping digits in locales using space as a group separator. Using NNBSP instead of NBSP avoids overly wide gaps in numbers occurring in justified text, and it conforms to the requirements of traditional typography.
>
> Other examples are found in French text, where NNBSP is used to set off certain punctuation characters, depending on the typographic school, and where it is called "*espace fine (insécable)*", literally "(no-break) thin space". (Parentheses are set because in French, a thin space is supposed to be always non-breaking.) Consistently with the most current fonts, NNBSP is best thought of as a non-breaking version of THIN SPACE.
>
> The MONGOLIAN VOWEL SEPARATOR […] in *Section 13.5, Mongolian* of [Unicode].

~~When NARROW NO-BREAK SPACE occurs in French text, it should be interpreted as an "espace fine insécable".~~

[table]
This character has no visible glyph […] prevent a line break.

[table]
This ~~is the preferred~~ space ~~to~~ is used, along with PUNCTUATION SPACE, ~~in~~ to indent numbers as a way of vertically aligning decimal separators or units. It has the same width as a digit and ~~keeps the number together for the purpose of line breaking~~ is thus too wide to be used in numbers. As a group separator, NARROW NO-BREAK SPACE should be used instead.