Line breaking around quotation marks

To: PAG, UTC
From: Robin Leroy
Date: 2023-03-08

This document is a proposal for changes to Unicode Standard Annex #14, *Unicode Line Breaking Algorithm*, in order to improve its handling of « this kind » of quotation marks, and fix strange edge cases in the existing handling of quotation marks.

Contents:

Proposal

Replace LB15 with LB15a and LB15b.

> ~~*LB15 Do not break within ‘”[’, even with intervening spaces.*~~
>
> ~~QU SP* × OP~~
>
> *LB15a Do not break after an unresolved initial punctuation that lies at the start of the line, after a space, after opening punctuation, or after an unresolved quotation mark, even after spaces.*
>
> (sot | BK | CR | LF | NL | OP | QU | GL | SP) [:Pi:] SP* ×
>
> *LB15b Do not break before an unresolved final punctuation that lies at the end of the line, before a space, before a prohibited break, or before an unresolved quotation mark, even before spaces.*
>
> × [:Pf:] (SP | GL | WJ | CL | QU | CP | EX | IS | SY | BK | CR | LF | NL | eot)

Background

Unicode Standard Annex #14 has the following to say about class QU:

> If language information is available [...] the quotation marks could be tailored to either OP or CL depending on their actual usage.

ICU doesn't do that; it seems like a safe bet that this isn't very common. In the absence of such tailoring, the behaviour of the line breaking algorithm is pretty bad.

Examples:

E1. Missing break opportunity between a closing quotation mark and an opening parenthesis:
"All Gaul is divided into three parts" ×(Caes. *BGall.* 1.1.1)

E2. Undesired break opportunities inside of French or Vietnamese quotation marks:

« ÷Ceci tuera cela. Le livre tuera l'édifice. ÷»

« ÷Người ta sinh ra tự-do và bình-đẳng về quyền-lợi, và phải luôn luôn được tự-do và bình đẳng về quyền-lợi. ÷»

(*The Unicode Standard*, Version 14.0.0, recommends the use of NBSP for that, saying "this choice helps line breaking algorithms". However, the line breaking algorithm is, by design, smart enough to deal with ordinary spaces before French :!?;, so it would make sense to do a better job here; NBSP also poses interchange issues on the web, where it gets turned into a regular space when copied around.)

How did we end up here?

The standard on which UAX #14 was originally based, JIS-X-4051:1995, had no equivalent to QU; in that standard, the characters " and " were instead treated as opening and closing respectively (or Western, depending on context).

UAX #14 says of QU:

> The default is to treat them as both opening and closing.

This sentence dates back to Revision 4 of then-draft UTR #14 (1998). It was true in that revision, which did not have a distinction between prohibited and indirect breaks. It is however not true in today's UAX #14, nor has it ever been true in any version of Unicode. CL forbids many breaks that QU doesn't, and OP forbids many breaks that QU doesn't.

Specifically, LB14 forbids breaks after OP (even after spaces):

> *LB14 Do not break after '[', even after spaces.*
>
> OP SP* ×

while LB13 forbids breaks before CL (even with preceding spaces, as it is applied before spaces break in LB18):

> *LB13 Do not break before ']' or '!' or ';' or '/', even after spaces.*
>
> × CL
>
> × CP
>
> × EX
>
> × IS

× SY

The current rules pertaining to QU (LB15 cited above & LB19, indirect break either side of QU) can instead be rephrased as the following heuristics for the resolution of QU:

H1. when applying rule LB14, treat QU as OP in the sequence QU SP+ OP
H2. when applying rule LB14, treat QU as OP in the sequence QU [^SP]
H3. when applying rule LB13, treat QU as CL in the sequence [^SP] QU

where H1 does not apply recursively, so that we have " ÷" ×[.

Heuristics H2 and H3 are fairly self-explanatory: a quotation mark adjacent to a nonspace might have its contents on that side; they work nicely for most styles of quotation marks; they are implemented by LB19.

> **Note:** It hasn't escaped our attention that rule LB19 also addresses non-quotation-mark usages of QU characters.
>
> **Example:** "Any quoted text in English"

Heuristic H1, which arises from LB15, is strange. It correctly identifies the opening quotation mark in cases such as the following:

> « [La Loi] doit être la même pour tous, soit qu'elle protège, soit qu'elle punisse. »

However, in a quotation that does not contain brackets, the quotation marks are unresolved (see example E2 above), and the closing equivalent is not resolved either:

> « ×[Le livre] tuera [l'édifice] ÷»

It also has false positives as in example E1 above.

Improving heuristics

The difficulty is that one cannot simply treat Initial Punctuation (", «, etc.) as opening and Final Punctuation (", », etc.) as closing as JIS-X-4051:1995 did, because our scope includes languages wherein Final Punctuation is opening, and languages wherein Initial Punctuation is opening.

Examples of such languages include "Swedish" (also »like this») and »Danish«.

However, the key realization is that these languages do not use spaces inside of their quotation marks; this suggests the following heuristics:

1. Treat Initial Punctuation as opening under either condition a or condition b:
    a. unless it is preceded by something that can be expected at the end of quoted text.
    b. if it is preceded by something that can be expected before quoted text.
2. Treat Final Punctuation as closing under either condition a or condition b:
    a. unless it is followed by something that can be expected at the beginning of quoted text.
    b. if it is followed by something that can be expected after quoted text.

The difference between the conditions a. and b. in each case is the direction in which they err in any overlap between the "something that".

We propose changing the heuristics above as follows (note: [:Pi:] ∪ [:Pf:] ⊊ QU).

H1′. when applying rule LB14, treat QU as OP in the sequence QU [^SP]
H2′. when applying rule LB13, treat QU as CL in the sequence [^SP] QU
H3′. when applying rule LB14, treat [:Pi:] as OP in the sequence
(sot | BK | CR | LF | NL | OP | QU | SP | GL) [:Pi:]
H4′. when applying rule LB13, treat [:Pf:] as CL in the sequence
[:Pf:] (SP | GL | WJ | CL | QU | CP | EX | IS | SY | BK | CR | LF | NL | eot)

The big groups can be explained as follows:

- (sot | BK | CR | LF | NL | OP | SP | GL) [:Pi:]
  - (sot | BK | CR | LF | NL) [:Pi:]: initial quotation mark at the start of a line.
  - (OP | QU) [:Pi:]: initial quotation mark at the beginning of a parenthetical, or nested quotation (a QU preceding a [:Pi:] would be treated as OP by H1′).
  - SP | GL [:Pi:]: initial quotation mark after spaces, including no-break spaces, *e.g.*, in

    «GL«SP

- [:Pf:] (SP | GL | WJ | CL | CP | EX | IS | SY | BK | CR | LF | NL | eot)
  - [:Pf:] (BK | CR | LF | NL | eot): final quotation mark at the end of a line.
  - [:Pf:] (WJ | CL | QU | CP | EX | IS | SY): final quotation mark before a prohibited break; this includes:
    - CL | QU: nested quotations.
    - CP: quotation at the end of a parenthetical.
    - IS | EX: quotation before various trailing punctuation (.:;, etc.)
  - SP | GL: spaces.

This is equivalent to the proposed changes to the regular expression rules.

Why only now?

The line breaking algorithm was designed to properly cope with spaces before most French punctuation; one might wonder why it does not handle quotation marks. The reason is that it used to be implementable with a pair table with special handling of spaces; this does not provide enough context to resolve quotation marks.

Since Unicode Version 9 and the regional indicator rule LB30a, the pair table has been abandoned, so rules can have more context.

Collateral damage

In some edge cases involving weird usage of smart quotes in Swedish or Danish style that start with punctuation that would normally be found at the end of a sentence, this forbids line breaks that might have been OK:

- tom sträng ×””
- punkt ×».»
- punktum ×».«

Effect on real text

- Swedish:
  - ”...”: https://sv.wikisource.org/wiki/Dumt_f%C3%B4lk
    - test: no effect.
  - »...»: https://sv.wikisource.org/wiki/Synn%C3%B6ve_Solbacken/Kapitel_1
    - test: no effect.
- French:
  - « ... » in prose: https://fr.wikisource.org/wiki/Les_Trois_Mousquetaires/Chapitre_4
    - test
  - « ... » within verse: https://fr.wikisource.org/wiki/La_L%C3%A9gende_des_si%C3%A8cles/Apr%C3%A8s_la_bataille
    - test
  - » ... » in a multi-paragraph letter: https://fr.wikisource.org/wiki/Le_Rouge_et_le_Noir/Chapitre_LXVI
    - test
- Vietnamese « ... »:
  - https://vi.wikisource.org/wiki/Tuy%C3%AAn_ng%C3%B4n_%C4%90%E1%BB%99c_l%E1%BA%ADp_Vi%E1%BB%87t_Nam
    - test
  - https://vi.wikisource.org/wiki/Th%E1%BB%81_non_n%C6%B0%E1%BB%9Bc_(t%E1%BA%ADp_truy%E1%BB%87n_ng%E1%BA%AFn)/Th%E1%BB%81_non_n%C6%B0%E1%BB%9Bc/I
    - test