Subject: unexpected line break within abc.123
From: Stanisław Hodur
Date: 2024-02-14, 03:52 AM

Dear Sirs,
there is an issue with the line breaking algorithm. While “abc.abc”, “123.123” and “123.abc” are all unbreakable, yet “abc.123” allows break after the dot.

In common language, only 123.123 is used (with the decimal dot in some languages and for dates). Anything like abc.123, 123.abc or abc.abc can be used as identifiers, like Internet address (unicoder.org, 64.182.27.164) or document clause (II.1.A of 2006/42/EC). Breaking inside such an identifier is annoying and can be misleading. Consider the last example (clauses) and the following text:

> According to 2006/42/EU, the declaration should specify the manufacturer, as stated in II.1.
> A. See the attachment for details. As for the partly completed machinery, see II.1.
> B. The details follow.

Compare it to a well formed version below. You will notice that the **meaning** of the above text has changed due to improper line breaking.

> According to 2006/42/EU, the declaration should specify the manufacturer, as stated in
> II.1.A. See the attachment for details. As for the partly completed machinery, see
> II.1.B. The details follow.

Many projects use your (Unicode.org) line breaking algorithm, like Pango. And many other projects base on those projects, like Gedit and WeasyPrint. It is best to fix the error at the source. Maybe you can attach the issue to https://www.unicode.org/review/pri490/?

Best regards
Stanislaus Hodur

PS1. I prefer to describe the problem in simple words rather than with your formal notation, as I am not sure whether I can follow it correctly.

PS2. The statements below are taken from the Pango support forum, https://gitlab.gnome.org/GNOME/pango/-/issues/776.

Luca Bacci@lb90· 3 weeks ago
DeveloperAdd reactionMore actions

Indeed the table at https://www.unicode.org/Public/UCD/latest/ucd/auxiliary/LineBreakTest.html#table has IS÷NU, so Pango follows the spec correctly.

I think it's a defect in example 7 of UAX#14. Perhaps we can ask for clarification on the Unicode mailing list or report an RFC to https://www.unicode.org/reporting.html

Peng Wu@pwu· 4 weeks ago
ContributorAdd reactionMore actions

To pass the LineBreakTest.txt tests, pango uses the customized LB25 Rule.

If the customized LB25 Rule doesn't have `ISxNU`, pango allows line break by default with LB31 Rule.

Peng Wu@pwu· 4 weeks ago
ContributorAdd reactionMore actions

I think pango supports Rule LB25 with Example 7 of Customization.

The regular expression is `( PR | PO) ? ( OP | HY ) ? NU (NU | SY | IS) * ( CL | CP ) ? ( PR | PO) ? .`

URL: https://www.unicode.org/reports/tr14/#Examples

Luca Bacci@lb90· 1 month ago
DeveloperAdd reactionMore actions

Here are the rules Pango implements: https://www.unicode.org/reports/tr14/#LB1. We have to find which rule prohibits `abc.123` from breaking on two lines.

- abc is classified as AL (alphabetic)
- . is classified as IS (numeric infix separator)
- 123 is classified as NU (numeric)

I believe it's LB25, which lists `ISxNU`