L2/11-141

From: Peter Edberg, Mark Davis, Andy Heninger
Subject: Segmentation & Linebreak
Date: 2011-04-28

We have the following action:

125	A099	Andy Heninger, Peter Edberg, Mark Davis		Create a WD of a PU UAX #14 which addresses the general issue of breaks after dashes.

Here is a breakdown of the issues and our recommendations.

1. Hebrew linebreak: With <hebrew hyphen non-hebrew>, there is no break on either side of the hyphen.

We recommend that this be done in the following way:

Split AL → (AL | HL) with the following redefinitions
HL to be the current AL ∩ [:script=hebrew:]
AL to be the current AL ⊖ [:script=hebrew:]
Add new rule:

Don't break after Hebrew + Hyphen

LB21a HL (HY|BA) ×

Issue: BA includes the following.

U+00AD ( ) SOFT HYPHEN
U+058A ( ) ARMENIAN HYPHEN
U+1400 ( ) CANADIAN SYLLABICS HYPHEN
U+2010 (-) HYPHEN
U+2027 (·) HYPHENATION POINT
U+2E17 ( ) DOUBLE OBLIQUE HYPHEN

The reason we need BA is to get Hyphen (2010). It doesn't hurt to include Armenian or CA hyphen, or the double oblique. The question is whether we need to split BA in two, because of either Soft Hyphen or Hyphenation point. The same is true for #3 below.

2. Hebrew word break: with <hebrew quote hebrew> there is no break on either side of the quote (single or double)

The recommendation is to add “ (0022) to MidLetter, so that it behaves like Gershayim and Apostrophe, because it is often used for that. This shouldn’t cause any problems, because SA and Ideographs are not linked by it. So these would behave the same:

U+05F4 (״) HEBREW PUNCTUATION GERSHAYIM
U+0022 (") QUOTATION MARK

3. Finnish linebreak (and others): with <letter space hyphen letter>, we need to allow a break before the hyphen but not after the hyphen.

Currently, we break as follows. The → shows the proposed change. We believe that this wouldn’t disturb other usage.

1. a - b *break before and after*
2. a -b *break before and after → break before*
3. a- b ***break after***
4. a-b *break after*

Proposed Rule

LB21b
    SP (HY|BA) × !SP

Issue: we know this will not be straightforward to implement as-is in ICU. We could approximate it with ÷ (HY|BA) × !SP (that is, don’t break after if there was a break before).

4. Spanish (and others): with <letter space emdash letter>, don’t break between emdash & letter

The issue is that some languages use emdashes to set off a parenthetical, and you don’t want to break the surrounding ones from the contained text. In that usage, there is a space on the side where it can be broken. This doesn’t conflict with symmetrical usages (spaces either before or after).

Currently, we break as follows. The → shows the proposed change.

1. a — b *break before and after*
2. a —b *break before and after* → **break before, but *not* after**
3. a— b *break before and after* → **break after, but *not* before**
4. a—b *break before and after*

Proposed Rule:

LB21c

SP B2 × !SP
!SP × B2 !SP

5. Resubmit L2/09-263

6. Current Kinsoku too restrictive. Submit changes from #3571 to make the current rules “normal”.

TBD: Peter to look over the changes and recommend changes to classes to be supplied as a rev of this document.

7. Issue found while doing this document: LB indicates that the following are punctuation, but they are Alphabetic. This seems incorrect.

Line_Break=Close_Punctuation
U+1325B ( ) EGYPTIAN HIEROGLYPH O006D
...{1}...U+1325D ( ) EGYPTIAN HIEROGLYPH O006F
U+13282 ( ) EGYPTIAN HIEROGLYPH O033A
U+13287 ( ) EGYPTIAN HIEROGLYPH O036B
U+13289 ( ) EGYPTIAN HIEROGLYPH O036D
U+1337A ( ) EGYPTIAN HIEROGLYPH V011B
U+1337B ( ) EGYPTIAN HIEROGLYPH V011C

Line_Break=Open_Punctuation
U+13258 ( ) EGYPTIAN HIEROGLYPH O006A
...{1}...U+1325A ( ) EGYPTIAN HIEROGLYPH O006C
U+13286 ( ) EGYPTIAN HIEROGLYPH O036A
U+13288 ( ) EGYPTIAN HIEROGLYPH O036C
U+13379 ( ) EGYPTIAN HIEROGLYPH V011A