L2/05-093

Some comments on Unicode line-breaking, mirroring, and bidi line processing

Kent Karlsson
2005-04-18

Line-break table in UAX 14

Here's a suggested tidied up table (table 2) for the line breaking. I've permuted the entries, so that the table looks "nicer" (to me anyway), and coloured some entries for emphasis. I've merged adjacent equivalent columns, and adjacent equivalent rows. Even though there still are equivalent rows in the table, I haven't merged them, since then the order between the property values in the rows and in the columns would then become different. I've added AI, which, when not definitely resolved to ID should behave like AL. (QU also really needs some kind of resolution to OP or CL, but QU is still included in the table.) So far, this is only editorial.

Suggestions for changed behaviour for two of the entries (marked in red) are made below the table. So is a change for the entire column for SY.

	OP	BB,PR	B2	PO	IN	ID	NU	AI,AL	CM	NS	BA,HY,QU,GL	WJ,SY,IS,EX,CL	ZW
OP	^	^	^	^	^	^	^	^	@	^	^	^	^
BB	%	%	%	%	%	%	%	%	#	%	%	^	^
PR	%	_	_	_	_	%	%	%	#	%	%	^	^
B2	_	_	^	_	_	_	_	_	#	%	%	^	^
PO	_	_	_	_	_	_	_	_	#	%	%	^	^
IN	_	_	_	_	%	_	_	_	#	%	%	^	^
ID	_	_	_	%	%	_	_	_	#	%	%	^	^
NU	_	_	_	%	%	_	%	%	#	%	%	^	^
AI,AL,CM	_	_	_	_	%	_	%	%	#	%	%	^	^
NS,BA	_	_	_	_	_	_	_	_	#	%	%	^	^
HY	_	_	_	_	_	_	%	_	#	%	%	^	^
QU	^	%	%	%	%	%	%	%	#	%	%	^	^
GL,WJ	%	%	%	%	%	%	%	%	#	%	%	^	^
SY	_	_	_	_	_	_	%	_	#	%	%	^	^
IS	_	_	_	_	_	_	%	%	#	%	%	^	^
EX	_	_	_	_	_	_	_	_	#	%	%	^	^
CL	_	_	_	%	_	_	_	_	#	^	%	^	^
ZW	_	_	_	_	_	_	_	_	_	_	_	_	^

^ denotes a *prohibited break*: B ^ A is equivalent to **B SP*** × **A**; in other words, never break before A and after B, even if one or more spaces intervene.

> COMMENT: That seems a bit strong, as well as too weak sometimes. I would suggest limiting (and extending) this to "**B** SP**?** × **A**, that is, don't break between A and B even if there is a (single) SPACE between them. In some display modes several spaces, tabs, or even NLFs (including FORM FEED) are collapsed to a single space. For such display modes it is the single resulting SPACE that is counted for the line break determination." (with a corresponding change to the rules).

Further, if the ^ for CL-NS is changed to % (which seems reasonable, a space should allow for a break then), then NS and BA are the same in the table (and the NS column can be merged with the BA,HY,QY,GL column.

Similarly, if the ^ for QU-OP is changed to % (which is also quite reasonable, given that QU-AL, QU-NU, etc. are %, and there is no reason to assume that the QU is really OP in this case), then QU and GL are the same in the table (and the QU row can be merged with the GL,WJ row).

In addition, SY has a "^"-column. I'm not sure why it is not a "%"-column, making it the same as HY. Is SPACE before SOLIDUS really used in any proper spelling? And even if so, should that prevent a line break at that point?

@ denotes a *prohibited break for combining marks:* B @ A is equivalent to **B SP*** × **A**, where A is of class **CM**. For more details see >Section 7.5, *Combining Marks*.

COMMENT: Same as for ^: limit to a single SPACE between, but that single SPACE may sometimes be collapsed whitespace (and hence really include tabs, and even NLFs)

Line-break property datafile

Line-break properties for Thai/Lao/Philippine characters

0E2F;SA # THAI CHARACTER PAIYANNOI
0EAF;SA # LAO ELLIPSIS

PAIYANNOI is a danda (full stop-like), and even used in abbreviations in the same way as full stop is (I'm not sure if other dandas are used for abbreviations, but the Khmer danda (KHAN) apparently is). Both PAIYANNOI and LAO ELLIPSES are apparently used as ellipes, and should allow line break after.

All other dandas have the BA line-break property, including the Khmer ones. So should PAIYANNOI and LAO ELLIPSIS.

The following two are also dandas:
1735;AL # PHILIPPINE SINGLE PUNCTUATION
1736;AL # PHILIPPINE DOUBLE PUNCTUATION

The Thai and Lao repeat marks:
0EC6;SA # LAO KO LA
0E46;SA # THAI CHARACTER MAIYAMOK
These should have the linebreak property NS, just as the Khmer repeat mark:
17D7;NS # KHMER SIGN LEK TOO

These two "format controls"
17B5;SA # KHMER VOWEL INHERENT AA
17B4;SA # KHMER VOWEL INHERENT AQ

are invisible control (Cf) characters, and as most Cf, these should have line-break property **CM**.

One might wonder why the Khmer “apostrope”
17DC;AL # KHMER SIGN AVAKRAHASANYA
does not have line-break property **SA**, though... Which I think it should have.

The two very similar script specific "bullet" characters:
17D9;NS # KHMER SIGN PHNAEK MUAN
0E4F;AL # THAI CHARACTER FONGMAN
should have the same line-break property. I would guess **BB**, or **OP**. (As I would for 2022;AL # BULLET; i.e. BULLET should also have line-break property **BB**, or, maybe better, **OP**.)

Line-break vs. bidi properties

Line-break treats (by default) almost all C0 control characters as if they were "combining marks". For some there are exceptions. However, some do not get excepted from that, even though that would be expected:
000B;CM # <control>
001C;CM # <control>
001D;CM # <control>
001E;CM # <control>
001F;CM # <control>

U+000B, line tabulation (VT), should have the **BK** line break property, just as FORM FEED has.

U+001C, U+001D, and U+001E should (by default) have the **BK** property, since bidi (by default) considers these to be paragraph boundary (B) characters.

U+001F should (by default) have the **BA** property, since bidi (by default) considers this to be a segment boundary character (like HT/tab).

Alternatively, the (default) bidi property for the four latter (the ISn) should be **BN**, keeping their **CM** line-break property. Line tabulation should still get the line-break property **BK**.

(Side remark: I did encounter a range of printers (IBM SureMark) that uses GS (IS3, U+001D) as an additional esc-sequence introducer (they also use ESC for other printer commands). So I would not mind if U+001C-U+001F all got default bidi property **BN** instead of B or S.)

Some other line-break (and bidi) comments

One would also expect
0089;<control>;Cc;0;BN;;;;;N;CHARACTER TABULATION WITH JUSTIFICATION;;;;
to have the (default) bidi property **S** (instead of BN), and the line break property **BA** (instead of CM), but at least there is no apparent conflict (and maybe nobody cares about U+0089 anymore).

One would also expect
0082;<control>;Cc;0;BN;;;;;N;BREAK PERMITTED HERE;;;;
to have the (default) line break property **ZW** (instead of CM), just as ZWSP, since U+0082 can be used for ZWSP in some legacy character encodings.

Likewise,
0083;<control>;Cc;0;BN;;;;;N;NO BREAK HERE;;;;
should (by default) have the (default) line break property **WJ** (instead of CM), just like WORD JOINER and ZERO WIDTH NO-BREAK SPACE, since U+0083 may be used as ZW no-break space/word joiner in some legacy character encodings.

Also, these two:
001A;<control>;Cc;0;BN;;;;;N;SUBSTITUTE;;;;
FFFD;REPLACEMENT CHARACTER;So;0;ON;;;;;N;;;;;
should have more similar properties, both for bidi (**ON**) and for linebreak (**AI**), since U+001A can be used as replacement character for many legacy character encodings.

I would have assumed that
3006;ID # IDEOGRAPHIC CLOSING MARK
should have line-break property **NS**.

These two:
3164;ID # HANGUL FILLER
FFA0;AL # HALFWIDTH HANGUL FILLER
are actually control characters (not letters, Lo, and they are NOT really related to any of the the Jamo fillers, despite their compatibility decomposition), used to mark the start of a Hangul letter composition sequence (plus being a trail(!) consonant filler). (Not that this way of composing Hangul syllables should be used anymore, now that the conjoining Jamos are encoded.) But other lead Cf characters (for Arabic) have lb prop AL, so I guess that letting these Cfs (erroneously having general category Lo) retain their lb properties is no real error.

0F7F;BA # TIBETAN SIGN RNAM BCAD
Not sure why this visarga (and combining character) is singled out as allowing a break after. It could in principle be followed by more combining marks (well, so could other BA chars...), but does Tibetan visarga only occur at the end of a word? All other visargas have line break property **CM**.

Should not the following have line-break property **ID** (rather than AL):
2630...2637 AL TRIGRAM FOR HEAVEN...TRIGRAM FOR EARTH
268A...268F AL MONOGRAM FOR YANG...DIGRAM FOR GREATER YIN
1D300...1D356 AL MONOGRAM FOR EARTH...TETRAGRAM FOR FOSTERING

I would also suggest that the following ones are to have the same lb property. E.g. NS, or perhaps better, **EX** for consistency with other exclamation/question marks:
203C NS DOUBLE EXCLAMATION MARK
203D AL INTERROBANG
2047 AL DOUBLE QUESTION MARK
2048 AL QUESTION EXCLAMATION MARK
2049 AL EXCLAMATION QUESTION MARK

The following two should have line break property **OP**, since they are opening punctuation. ¡"Spanish", really! does not allow a line-break just after the ¡, assuming the "" is resolved to OP.
00A1 AI INVERTED EXCLAMATION MARK
00BF AI INVERTED QUESTION MARK

General category and line-break for some quote marks

275B HEAVY SINGLE TURNED COMMA QUOTATION MARK ORNAMENT
275C HEAVY SINGLE COMMA QUOTATION MARK ORNAMENT
275D HEAVY DOUBLE TURNED COMMA QUOTATION MARK ORNAMENT
275E HEAVY DOUBLE COMMA QUOTATION MARK ORNAMENT
should not be So, but instead be Pi/Pf, as for similar quote marks.

Shouldn't the following be Ps/Pe rather than Pi/pf?
2E02 LEFT SUBSTITUTION BRACKET Pi
2E03 RIGHT SUBSTITUTION BRACKET Pf
2E04 LEFT DOTTED SUBSTITUTION BRACKET Pi
2E05 RIGHT DOTTED SUBSTITUTION BRACKET Pf
2E09 LEFT TRANSPOSITION BRACKET Pi
2E0A RIGHT TRANSPOSITION BRACKET Pf
2E0C LEFT RAISED OMISSION BRACKET Pi
2E0D RIGHT RAISED OMISSION BRACKET Pf
2E1C LEFT LOW PARAPHRASE BRACKET Pi
2E1D RIGHT LOW PARAPHRASE BRACKET Pf
And then the corresponding change to line-break property.

The "ordinary" ones here have lb property IS, these ones should too:
FE54 NS SMALL SEMICOLON
FE55 NS SMALL COLON
FF1A NS FULLWIDTH COLON
FF1B NS FULLWIDTH SEMICOLON

Mirroring (in the main datafile)

Consider this pair of parentheses:
FD3E;OP # ORNATE LEFT PARENTHESIS
FD3F:CL # ORNATE RIGHT PARENTHESIS

These parentheses do not have the bidi mirror property (which most parentheses have). Presumably because they are only used for Arabic. However, given the sample glyphs in the charts, this appears to make these to have either wrongly turned glyphs, or have the opposite general category/line-break property to what one would expect. So I guess these should be mirrored after all. Or have their sample glyphs mirrored, or have their g.c./lb properties swapped. Or, maybe these really are supposed to look backwards...

The following, currently *NON-mirrored*, characters are not supposed to be used with Arabic/Hebrew:
FE59 OP SMALL LEFT PARENTHESIS
FE5A CL SMALL RIGHT PARENTHESIS
FE5B OP SMALL LEFT CURLY BRACKET
FE5C CL SMALL RIGHT CURLY BRACKET
FE5D OP SMALL LEFT TORTOISE SHELL BRACKET
FE5E CL SMALL RIGHT TORTOISE SHELL BRACKET

But that is the case also for the following *mirrored* characters:
FF08 OP FULLWIDTH LEFT PARENTHESIS
FF09 CL FULLWIDTH RIGHT PARENTHESIS
FF3B OP FULLWIDTH LEFT SQUARE BRACKET
FF3D CL FULLWIDTH RIGHT SQUARE BRACKET
FF5B OP FULLWIDTH LEFT CURLY BRACKET
FF5D CL FULLWIDTH RIGHT CURLY BRACKET
FF5F OP FULLWIDTH LEFT WHITE PARENTHESIS
FF60 CL FULLWIDTH RIGHT WHITE PARENTHESIS
FF62 OP HALFWIDTH LEFT CORNER BRACKET
FF63 CL HALFWIDTH RIGHT CORNER BRACKET

So I'd suggest that the "SMALL" ones here be mirrored too, or that the “FULLWIDTH” ones aren’t mirrored either. If made mirroring these quote marks should be made mirroring too:
301D OP REVERSED DOUBLE PRIME QUOTATION MARK
301E CL DOUBLE PRIME QUOTATION MARK
301F CL LOW DOUBLE PRIME QUOTATION MARK

I would suggest that for the following:
007E TILDE
00AC NOT SIGN
2310 REVERSED NOT SIGN
2319 TURNED NOT SIGN
2AEC DOUBLE STROKE NOT SIGN

2AED REVERSED DOUBLE STROKE NOT SIGN
FF5E FULLWIDTH TILDE
FFE2 FULLWIDTH NOT SIGN
all are to have general category **Sm** (also 2310 and 2319), and all to be **mirrored** (like 2AEC and 2AED already are).

Bidi, line leading spaces, and SHY shaping in bidi

Line (segment, really) leading spaces

Section separators (anywhere on the line) and white-space at the end of a line are reset to the paragraph embedding level. However, line leading white-space characters still may end up at a "visually non-leading" position. For example, (with the same convention as used in UAX 9) for right to left paragraphs, the line "...CAR" can come out as "RAC..." (were . is space), while the "...car" can come out as "...car", if the embedding level has become increased by bidi controls (e.g. for a quotation within a paragraph). If the spaces are used as (admittedly simplistic, but used in plain text) line start indentation, the latter is then wrong (for right to left paragraphs), and should be "car...". This resetting should be done not only for characters with bidi property WS, but also for NBSP and NNBSP (that are bidi CS, and has other treatment when *between* numbers).

E.g., the input (where . marks space, and uppercase marks Arabic/Hebrew):

```
CAR.CAR.RACK<ls>
..."<lre>eng.eng.english<ls>
...quote<pdf>"<ls>
RACE
```

(where the line breaks are LS-es, the bidi controls are
correctly INSIDE of the quote marks) should turn out as:

```
          KCAR.RAC.RAC
eng.eng.english"...
          "quote...
               ECAR
```

NOT as (it is currently, unless I'm too confused):

```
          KCAR.RAC.RAC
eng.eng.english"...
          "...quote
               ECAR
```

----------suggested replacement text for within L1 of UAX 9----------
3. any sequence of white-space (WS), NBSP, and NNBSP characters preceding or succeeding a segment separator or paragraph separator, and

4. any sequence of white-space (WS), NBSP, and NNBSP characters at the beginning of the line and at the end of the line.

[...] this means that leading white-space will appear at the visual beginning of the line or segment and trailing white-space will appear at the visual end of the line or segment (in the paragraph direction). Tabulation will always have a consistent direction within a paragraph.

Paragraph default embedding level

Furthermore, to get the default paragraph direction better chosen (if the default is used), all digits should be considered strong left-to-right in rules P2 and P3. I.e., look also for EN and AN as strongly directional L, which, if first strongly directional character, should set the (default, if there is no higher level protocol saying otherwise) paragraph level to 0. This affects the (default) treatment of leading and trailing spaces in each "segment", esp. for (almost) purely numeric "paragraphs" or "columns", possibly followed by a unit (in Arabic, I've seen "cm" transcribed into Arabic) (logically) after the digits. If one has tried to do poor-man's alignment (which is what one does in plain text, which by definition has no higher level protocol for alignments) by using leading spaces, those spaces should be on the left side of the digits.

Bidi and SHY shaping

Shaping is said to logically occur after all the steps of the bidi algorithm. While that has problems in general, in particular in relation to line breaking, I would just like to point out that SHY should be "shaped" after line breaking but before rearrangement (L2) (so that the determination of whether it is at a line end is not messed up). Thus the input "CAR RUN bil<shy>buren TURN" may be displayed (for an rtl paragraph)

```
   bil- NUR RAC
     NRUT buren
```

Note also that the SHY must NOT be removed at step X9, even though it is (currently) BN, since it may actually be visible ("shaped" to one or other actual hyphen later on). (Indeed, I'm not at all keen on this removal at all, and would rather see the non-removal variety be the normative one, and not mentioning the removal variety.)