LINEBREAK CATEGORY SUMMARY 06/07/99

Action for UTC: review and approve the data tables, with the changes noted here.

The datatable for TR 14 as posted (LineBreak.d3.txt) can be summarized as follows
The productions are of the form below and apply in any order.

Gc → LB (explicit exceptions) “comment”
Gc → LB(eaw) (explicit exceptions)
Gc → AL/ID(a) (explicit exceptions)

Here Gc → LB means, a character of General Category Gc has Linebreak class LB, Gc → LB(eaw) means the class LB applies only if the East Asian Width property of the character is (eaw). The list of exceptions in (<explicit exceptions>) overrides the general rule. AL/ID(a) means that all characters with *ambiguous* East Asian Width property are treated as AL if their width *resolves* to Narrow and ID if width resolves to FullWidth.

Cc → CM (except CR for CR, LF for LF, BA for TAB, BK for FF)
Cf → CM (except GL for ZWNBSP)
Cs → SG “surrogates”

Co → AL/ID “assuming PUA is treated as ambiguous EA width”
[This is incorrectly so listed in EastAsianWidth file]

Ll → AL(n)
Ll → AL/ID(a)
Ll → ID(f)

Lm → NS(w)
Lm → NS(h)
Lm → AL/ID(a)
Lm → AL(n) (except SA for chars 0E00..0EFF)

[30FC, 30FE, 3031..3035, FF70 are listed incorrectly in the file].

Lo → AL(n) (except SA for chars in 0E00..0EFF,1000..10FF,1780..17FF)
Lo → ID(w) (except CM for chars in 1160..11F9, NS for ‘small kana’)

[”small kana” have the word small in their name - some are listed incorrectly in the file.]

Lt → AL
Lu → AL(n)
Lu → AL/ID(a)
Lu → ID(f)

Mc → CM
Me → CM
Mn → CM

Nd → NU(n)
Nd → ID(f)

Nl → AL(n)
Nl → ID(w)

Nl → AL/ID(a)

No → AL(n)
No → ID(w)
No → AL/ID(a)
[For FullWidth superscripts getting class ID could be a mistake, PO would seem to make more sense. Maybe not a big enough issue.]

Pc → AL(n)
Pc → ID(w) (except NS for Katakana middle dot 30FB)
Pc → NS(h) "Katakana middle dots FF65"
Pc → ID(f)

Pd → BA(a) (except IN for EM DASH, AL for HORIZONTAL BAR)
Pd → BA(n) (except HY for 002D, GL for NBHY
Pd → ID(w) (except NS for WAVE DASH)
Pd → ID(f)

[EMDASH should be in the B2 class. Listed incorrectly in the file.]

Pe → CL "by design, but CL > Pe"
Pi → QU "by design, but QU > Pi + Pf"
Pf → QU "by design"
Ps → OP "by design"

Po → AL(n) (many exceptions)
Po → ID/AL(a) (many exceptions)
Po → ID(w) (many exceptions)
Po → ID(f) (many exceptions)
Po → CL(h)

[General Category Po spans almost all line breaking types. Enumeration is the only hope here.]

Explicit assignment of line breaking classes for General Category Po. This list overrides the five productions above

2027	BA	HYPHENATION POINT
0F0B	BA	TIBETAN MARK INTERSYLLABIC TSHEG
1361	BA	ETHIOPIC WORDSPACE
17D5	BA	KHMER SIGN BARIYOOSAN
1680	BA	OGHAM SPACE MARK
1806	BB	MONGOLIAN TODO SOFT HYPHEN
FF0C	CL	FULLWIDTH COMMA
FF0E	CL	FULLWIDTH FULL STOP
3001..3002	CL	IDEOGRAPHIC COMMA..IDEOGRAPHIC FULL STOP
FE50	CL	SMALL COMMA
FE52	CL	SMALL FULL STOP
FF01	EX	FULLWIDTH EXCLAMATION MARK
FF1F	EX	FULLWIDTH QUESTION MARK
0021	EX	EXCLAMATION MARK
003F	EX	QUESTION MARK
FE56..FE57	EX	SMALL QUESTION MARK..SMALL EXCLAMATION MARK
0F0C	GL	TIBETAN MARK DELIMITER TSHEG BSTAR
2025..2026	IN	TWO DOT LEADER..HORIZONTAL ELLIPSIS
2024	IN	ONE DOT LEADER
0589	IS	ARMENIAN FULL STOP
002C	IS	COMMA

002E	IS	FULL STOP
003A	IS	COLON
003B	IS	SEMICOLON
FF1A	NS	FULLWIDTH COLON
FF1B	NS	FULLWIDTH SEMICOLON
0E5A..0E5B	NS	THAI CHARACTER ANGKHANKHU..THAI CHARACTER KHOMUT
17D4	NS	KHMER SIGN KHAN
17D6..17DA	NS	KHMER SIGN CAMNUC PII KUUH..KHMER SIGN KOOMUUT
203C	NS	DOUBLE EXCLAMATION MARK
FE54	NS	SMALL SEMICOLON
FE55	NS	SMALL COLON
2030	PO	PER MILLE SIGN
2032..2033	PO	PRIME..DOUBLE PRIME
2035	PO	REVERSED PRIME
FF05	PO	FULLWIDTH PERCENT SIGN
2031	PO	PER TEN THOUSAND SIGN
2034	PO	TRIPLE PRIME
2036..2037	PO	REVERSED DOUBLE PRIME..REVERSED TRIPLE PRIME
0025	PO	PERCENT SIGN
FE6A	PO	SMALL PERCENT SIGN
005C	PR	REVERSE SOLIDUS
0022	QU	QUOTATION MARK
0027	QU	APOSTROPHE
002F	SY	SOLIDUS

Sc → PR (except cent and peseta)

Sk → AL(n) (except dictionary usage)
Sk → AL/ID(a) (except dictionary usage)
Sk → NS(w)
Sk → ID(F)

Sm → ID(w)
Sm → ID(f)
Sm → AL(h)
Sm → ID/AL(a) (except PR for PLUS-MINUS, exceptions)
Sm → AL(n) (except PR for PLUS, MINUS, MINUS-PLUS,
BA for VERTICAL LINE, NS for Fraction slash)

So → AL(n) (except ID for 303F, CB for FFFC, PR for NUMERO,
So → ID/AL(a) (except PPO for Degree, Fahrenheit, and Celsius)
So → ID(w)
So → ID(f)
So → AL(h)

Zl → BK
Zp → BK
Zs → BA(n) (except SP for SP, GL for Figure, Narrow and No-break space)
Zs → ID(w)

[The last one is listed incorrectly in the file.]

As can be seen from this, the line break behavior tracks the General Categories quite well, once the East Asian Width, esp ambiguous width, is taken into account. This is because the TR14 combines East Asian line breaking with Western style into a single scheme.

PROPOSED ASSIGNMENTS OF LB CLASS

Except for single character productions, this is the list from the previous section, sorted by LB class.

CR

000D <control> “Carriage Return”

LF

000A <control> “Line Feed”

BK

000C FORM FEED
2028 LINE SEPARATOR
2029 PARAGRAPH SEPARATOR

CM

Unless otherwise listed explicitly all characters with these General Categories

Cc + Cf + Mc + Me + Mn +

1160..11F9 HANGUL JUNGSEONG FILLER..HANGUL JONGSEONG YEORINHIEUH

SG

All surrogates, i.e. characters with General Category Cs

GL

00A0 NO-BREAK SPACE
0F0C TIBETAN MARK DELIMITER TSHEG BSTAR
2007 FIGURE SPACE
2011 NON-BREAKING HYPHEN
202F NARROW NO-BREAK SPACE
FEFF ZERO WIDTH NO-BREAK SPACE

CB

FFFC OBJECT REPLACEMENT CHARACTER

IN

2025..2026 TWO DOT LEADER..HORIZONTAL ELLIPSIS
2024 ONE DOT LEADER

HY

002D HYPHEN MINUS

SP

0020 SPACE

BA

All characters of the following combination of General Category and East Asian Width except as explicitly listed elsewhere:

Zs(n) + Pd(a,n)

0009	TAB
007C	VERTICAL LINE *
00B4	ACUTE ACCENT *
0F0B	TIBETAN MARK INTERSYLLABIC TSHEG
1361	ETHIOPIC WORDSPACE
17D5	KHMER SIGN BARIYOOSAN
2027	HYPHENATION POINT *
1680	OGHAM SPACE MARK

* dictionary usage

BB

02C8	MODIFIER LETTER VERTICAL LINE *
02CC	MODIFIER LETTER LOW VERTICAL LINE *
1806	MONGOLIAN TODO SOFT HYPHEN

* dictionary usage

B2

2014	EM DASH

NS

All characters with the following combination of General Category and East Asian Width

Sk(w) + Lm(w) + Lm(h)

plus the following characters

0E5A..0E5B	THAI CHARACTER ANGKHANKHU..THAI CHARACTER KHOMUT
17D4	KHMER SIGN KHAN
17D6..17DA	KHMER SIGN CAMNUC PII KUUH..KHMER SIGN KOOMUUT
203C	DOUBLE EXCLAMATION MARK
2044	FRACTION SLASH
301C	WAVE DASH
3041	HIRAGANA LETTER SMALL A
3043	HIRAGANA LETTER SMALL I
3045	HIRAGANA LETTER SMALL U
3047	HIRAGANA LETTER SMALL E
3049	HIRAGANA LETTER SMALL O
3063	HIRAGANA LETTER SMALL TU
3083	HIRAGANA LETTER SMALL YA
3085	HIRAGANA LETTER SMALL YU
3087	HIRAGANA LETTER SMALL YO
308E	HIRAGANA LETTER SMALL WA
30A1	KATAKANA LETTER SMALL A
30A3	KATAKANA LETTER SMALL I
30A5	KATAKANA LETTER SMALL U
30A7	KATAKANA LETTER SMALL E
30A9	KATAKANA LETTER SMALL O

30C3	KATAKANA LETTER SMALL TU
30E3	KATAKANA LETTER SMALL YA
30E5	KATAKANA LETTER SMALL YU
30E7	KATAKANA LETTER SMALL YO
30EE	KATAKANA LETTER SMALL WA
30F5..30F6	KATAKANA LETTER SMALL KA..KATAKANA LETTER SMALL KE
30FB	KATAKANA MIDDLE DOT
FE54..FE55	SMALL SEMICOLON..SMALL COLON
FF1A	FULLWIDTH COLON.. FULLWIDTH SEMICOLON
FF65	HALFWIDTH KATAKANA MIDDLE DOT
FF67..FF70	HALFWIDTH KATAKANA LETTER SMALL A..HALFWIDTH KATAKANA-HIRAGANA PROLONGED SOUND MARK

OP

All characters of General Category Ps.

CL

All characters of General Category Pe as well the following characters:

3001..3002	IDEOGRAPHIC COMMA..IDEOGRAPHIC FULL STOP
FF0C	FULLWIDTH COMMA
FF0E	FULLWIDTH FULL STOP
FE50	SMALL COMMA
FE52	SMALL FULL STOP
FF61	HALFWIDTH IDEOGRAPHIC FULL STOP
FF64	HALFWIDTH IDEOGRAPHIC COMMA

QU

All characters of General Category Pi or Pf as well as the following characters:

0022	QUOTATION MARK
0027	APOSTROPHE

EX

0021	EXCLAMATION MARK
003F	QUESTION MARK
FE56..FE57	SMALL QUESTION MARK..SMALL EXCLAMATION MARK
FF01	FULLWIDTH EXCLAMATION MARK
FF1F	FULLWIDTH QUESTION MARK

AL/ID

All characters with East Asian Width property of A (ambiguous) and the following General Category

Co(a) + Ll(a) + Lm(a) + Lu(a) + Nl(a) + No(a) + Sk(a) + Sm(a) + So(a) + Po(a)

except as explicitly listed in other places are treated as AL when their *resolved* East Asian Width is N (Narrow) and as ID otherwise.

ID

All characters of the following combinations of General Category and East Asian Width, except as explicitly listed elsewhere.

Pc(w,f) + Pd(w,f) + Po(w,f) + Ll(f) + Lo(w) + Lu(f) + Sk(f) + Sm(w,f) + So(w,f) + Nd(f) + Nl(w) + No(w) + Zs(w)

plus

303F IDEOGRAPHIC HALF FILL SPACE

NU

All narrow characters of general catgory Nd.

IS

0589 ARMENIAN FULL STOP
002C COMMA
002E FULL STOP
003A COLON
003B SEMICOLON

SY

002F SOLIDUS

AL

All characters of the following combinations of General Category and East Asian Width, except as explicitly listed elsewhere.

Ll (n) + Lm(n) + Lo(n) + Lt + Lu(n) +Pc(n) + Po(n) + Nl(n) + No(n) + Sk(n) + Sm(h,n) + So(h,n

plus

2015 HORIZONTAL BAR

PR

All currecncy symbols (General Category Sc) except as listed explicitly elsewhere and the following:

002B PLUS
005C REVERSE SOLIDUS
00B1 PLUS-MINUS
2116 NUMERO SIGN
2213 MINUS-PLUS

PO

0025 PERCENT SIGN
00A2 CENT SIGN
00B0 DEGREE SIGN
2030 PER MILLE SIGN
2031 PER TEN THOUSAND SIGN
2032..2035 PRIME..REVERSED TRIPLE PRIME
20A7 PESETA SIGN
2103 DEGREE CELSIUS
2109 DEGREE FAHRENHEIT
2126 OHM SIGN
FE6A SMALL PERCENT SIGN
FF05 FULLWIDTH PERCENT SIGN
FFE0 FULLWIDTH CENT SIGN

SA

All characters of General Category Lo or Lm in these ranges:

0E00..0EFF THAI / LAO
1780..17FF KHMER