TO: UTC
FROM: Deborah Anderson
DATE: 3 November 2006
RE: Line-breaking information as requested in L2/06-224 (on commas and full stops for sundry scripts)

The following are responses from various experts on those line-breaking questions explicitly listed in document L2/06-224. Read through the comments carefully, lest the suggested action – if given -- be incorrect. (Note that useful background information is contained in some responses, for example, on Ethiopic punctuation and on Buginese PALLAWA.)

1. N'KO COMMA U+07F8
Current linebreaking property: IS
Suggested action: No change; retain linebreaking property IS

From Michael Everson and Mamady Doumbouya:
> ARABIC COMMA is used more often as a decimal separator in N'Ko, but
> NKO COMMA can sometimes be used as a decimal separator.

2. MONGOLIAN COMMA U+1802, MONGOLIAN MANCHU COMMA U+1808, MONGOLIAN FULL STOP U+1803, AND MONGOLIAN MANCHU FULL STOP U+1809
Current linebreaking property: BA (for all of the above)
Suggested action: See below.

From Andrew West <andrewcwest@gmail.com>:
>On 02/06/06, Asmus Freytag <asmusf@ix.netcom.com> wrote:
> 1802;BA # MONGOLIAN COMMA
> 1808;BA # MONGOLIAN MANCHU COMMA
>
> -verify that Mongolia commas may appear at the beginning of a line
> following a space

Mongolian and Manchu commas never appear at the beginning of a line (whether or not following a space).

> 1803;BA # MONGOLIAN FULL STOP
> 1809;BA # MONGOLIAN MANCHU FULL STOP
>
> - verify that Mongolian and Coptic full stops may appear at the
> beginning of a line following a space, and that they don't require a
> space following them to break a line.
>

Mongolian and Manchu full stops never appear at the beginning of a line (whether of not following a space). They do not require a following space to break.

3. COPTIC FULL STOP U+2CFE and COPTIC OLD NUBIAN FULL STOP U+2CF9
Current linebreaking property: BA
Suggested action: See comments below.

From Steven Emmel <emmstel@nwz.uni-muenster.de>:

I think I understand the point of the two questions [below], and if I understand them rightly, then both times the answer is "no" (certainly for Coptic, and I would expect also for Old Nubian, although I am more or less just guessing in the latter regard).

>a) Can the Coptic full stop [and Old Nubian full stop] occur at the beginning of a line >following a space?

(Of course anything is possible, but as a rule) any space before a Coptic full stop should be treated as a "hard space," inseparable either from the following stop itself or from a preceding character.

>b) Does the Coptic full stop [and Old Nubian full stop] require a space following it in >order to break a line?

It should always be possible (but not mandatory) to break a line after a full stop, even if the very next character in the string is not a space.

4. ETHIOPIC COMMA U+1363 and ETHIOPIC FULL STOP U+1362 (with info on other Ethiopic punctuation marks)
Current linebreaking property: AL
Suggested action: Retain AL property

From Daniel Yacob <dyacob@gmail.com>:

Note 1:
I'll read the UAX and think about it. I remember that the Ethiopic punctutation were sub-optimally treated and I've wanted to put some info together for that. For instance "ETHIOPIC COLON" is an unfortunate name, it is really a glyph variant of ETHIOPIC COMMA.

There are two styles that punctuation gets used in. In the traditional style, punctuation is used in the same was as Ethiopic wordspace, that is letters appear immediately left and right and breaks occur to the right of the punctuation. In modern use a space will follow

a punctuation mark, but never an ethiopic wordspace. A line should never begin with a punctuation symbol. Ethiopic punctuation can follow or be followed by non-ethiopic punctuation such as parenthesis and quoation marks

Note 2:
In modern writing, a white space should follow a comma and other punctuation. In classic writing (where Ethiopic Wordspace is used) there would be no white space after punctuation. No punctuation is used with Ethiopic numerals (except in numbering Bible verses), but Ethiopic comma can be used with western number as a separator in the same way as with English practices. This can be found easily enough but I don't know what the *proper* rules are here, it may be the case that the writer was using an old style font where Ethiopic comma replaces English comma and the writer didn't care enough to switch the font back to Latin.

I have never seen an Ethiopian reference that indicates different usage for Ethiopic Colon, and the references that do mention it will equate it to Ethiopic Comma -its just another visual style (a third is 3 dots in a triangular arrangement). A person typically uses whatever their software offers (both symbols are not usually given) or what they find easiest to type. Peter Daniels' "World Writing Systems" mentions this equivalence also (if I remember right, his info comes from Getachew Haile who is most highly regarded in this area).

In modern Eritrean writing, Ethiopic Wordspace will be used in place of Ethiopic Comma -I've noticed this practice begin to creep into Ethiopia Tigrigna practices also.

The rules that Asmus states look right, that a line break occurs after a comma unless used to partition western numbers. This would apply in both cases where white space or Ethiopic word space are the default space symbol.

A line break should follow Ethiopic Full Stop. In good typesetting no Ethiopic punctuation should ever appear at the start of a new line. The Ethiopic Full Stop is only used to end a sentence, it is not used in other contexts.

Note 3:
A space is not required after Ethiopic full stop to break a line. A space wouldn't be required after any punctuation. In the traditional practice, where the Ethiopic wordspace is used, no white spaces (or ascii spaces) are expected to follow punctuation before letters begin again. In modern practice space after punctuation is still not required.

Some more background that might help. Classic Ethiopic writing is done on "Branna" which is treated cow or goat skin, which is produced very laboriously, but whole books are done this way. Because the writing material production is then very costly, all available space is used to its fullest and as many letters as can be are crammed onto a

line. This is likely how the Ethiopic space came about, as a simple symbol to minimally separate the crammed together words such that the eye can distinguish them.

Also to conserve space, the hyphenation rules are essentially non-existent. So a word can be broken across two lines at any point -it helps that every letter is also a syllable. No hyphenation symbol is used, or needed, you know it is the same word going across the line because no wordspace or other ethiopic punctuation has arrived to indicate otherwise. This hyphenation system (or lack thereof) is still used today, even when ethiopic wordspace is not used, not easy for computers to process, but people know the continuation via context.

5. ARABIC COMMA U+060C and ARABIC FULLSTOP U+ 06D4
Current linebreaking property: EX
Suggested action: Review comments below

> The questions involve 060C ARABIC COMMA and 06D4 ARABIC FULL STOP. Do
> these stick together with the other letters at the end of a line? Is
> it true that these characters don't require a space following them to
> break a line?

From Tom Milo < t.milo@chello.nl>:

060C ARABIC COMMA is a borrowing from Latin typography and behave exactly like its model. Historical Arabic documents use scriptio continua enhanced with occasional final forms and space to mark paragraphs or periods.

I have no experience using 06D4 ARABIC FULL STOP.

Obviously you recognize the typical scriptio continua pattern of early alphabetic writing in what I described about Arabic. In Arabic the unit of writing is a group of connected archigraphemic letters separated by spaces - or the end of the line. In such cases where a space or line-break is preceded by a final form of a letter, then the word-separating status of that position is clear. But words in Arabic frequently end in letters that disconnect regardless, in which case the space or line breaking remains ambiguous.

From Jonathan Kew <jonathan_kew@sil.org>:

I don't think there's any real evidence either way. The notion of a "space" is a bit vague in Arabic script, which has often been written without spaces between words, the boundaries being indicated only by the non-joining of adjacent letters (which is of course ambiguous in the case of those that never join to the left anyway).

Consistent use of spaces is, I think, a fairly recent development under the influence of word processing, and where spaces are used (at all), they would always be used after comma and full stop. So I doubt we'll find any firm evidence as to whether a line could break after one of these marks, in the absence of a space. And it doesn't really matter which way the default Unicode rules go on this; provided the behavior is consistent across systems, typists will quickly learn what to type in order to get the results they want.

Probably the simplest and most natural approach is to treat the Arabic comma like the Western one (whichever property that has); users of Arabic script will not have any clearly different expectations, so uniformity of behavior will be easiest to understand.

As for 06D4 "arabic full stop", this is one of those oddities that I think shouldn't have been encoded. Much Arabic-script text uses the standard U+002E PERIOD mark, and this is simply that character in a style based on Urdu Nastaliq writing. So it shouldn't be a separate character at all, just a font design choice. But as it's in the Standard, we're stuck with it, and people are indeed using it in some cases. So its line-break properties should probably be the same as U +002E, for consistency between the two characters being used for the same purpose.

6. BUGINESE END OF SECTION U+1A1F (with info on BUGINESE PALLAWA U+1A1E)
Current linebreaking property: AL (but BUGINESE PALLAWA is BA)
Suggested action: See comments

From "Sirtjo Koolhof" <koolhof@kitlv.nl>:

The end of section mark you ask about is extremely rare in Bugis manuscripts. I only know of a few examples, all from early 19th century mss, and from the printed books published in Singapore by the missionary Thomsen. They occur for example after a short introduction, before the main text begins. No line breaks occur after this, or the pallawa. At the end of a text it is sometimes used to mark the end of the main text and the beginning of a colophon.

I hope this information is sufficient, if not don't hesitate to contact me again.

Sirtjo Koolhof
Librarian

KITLV/Royal Netherlands Institute of Southeast Asian and Caribbean Studies PO Box 9515 2300 RA Leiden Netherlands Tel +31 (0)71 5272456 Fax +31 (0)71 5272638

7.1 ARMENIAN COMMA U+055D
Current linebreaking property: AL
Suggested action: No change, retain AL

a) From: "Santoukht Mikaelian" <Santoukht@berkeley.edu>

In answer to your first question, and on the basis of what I know, the comma stays together with the other letters. No space is required between the comma and the preceding letters. However, a space is required after the Armenian comma. As far as I know, the Armenian comma functions just the same as the English comma.

I have a source at home that I can check, and if I find anything different, I'll let you know.

b) From: "Cowe, Peter" <cowe@humnet.ucla.edu>

The comma in Armenian would stay ith the other letters on the same line and not move down to the next one. Also, no space is required after the comma to break the line. I should be glad to answer any other specific queries you have on the subject.

7.2 ARMENIAN FULL STOP U+0589
(Still verifying this is used numerically)

8. CANADIAN SYLLABICS FULL STOP U+166E
Current linebreaking property: AL
Suggested action: No change, retain AL

From Chris Harvey <chris@languagegeek.com>:

> Does the CANADIAN SYLLABICS FULL STOP, U+ 166E stay with the other letters
> when they it appears at a line end (i.e., the full stop does not end up at the beginning
> of a new line)? Is a space required after the full stop in order to cause a
> line-break?

The full stop works just like the Latin period. A space would be required to cause a line-break; it does stay with the other letters at a line-end.

9. REVERSED SEMICOLON U+204F
(Still working on this one.)