Date: 2014-Feb-04
From: Peter Edberg
To: UTC
Subject: Clarifying UAX#29 sentence break vs word break:
Action item 134-A78, handle PRI #240 feedback from Konstantin Ritt

Mark and I had action item 134-A78 to "Investigate solutions to the problem raised in the feedback of PRI #240 from Konstantin, Nov 20, 2012." (the first feedback item in http://www.unicode.org/review/pri240/).

The feedback is about a change in Unicode 5.1 which introduced the MidNumLet class for word break, moved some period-like and apostrophe-like characters into that class, and added rules which allowed characters of this class in the middle of words and numbers. These changes were prompted by the following documents from Mark —

- http://www.unicode.org/L2/L2007/07370-punct.html
- http://www.unicode.org/L2/L2007/07398-boundary-punct.html

— and the changes in the latter were accepted by UTC113 in Oct 2007 (Consensus 113-C6, action 113-A25).

Allowing period-like characters in the middle of a word was intended to allow things like "U.S.A" to be treated as a word. However, Konstantin mentions that this causes two problems:

1. It causes domain names like comments.gmane.org to be treated as a single word.
2. It causes a sequence like "Mr.Hamster" (no space) to be treated as a word, even though the sentence break rules find a break after "Mr.".

These are not big issues, and probably do not warrant a change to UAX #29 behavior.

As for #1, we explicitly do not intend UAX#29 word break to handle programmer-usage conventions (separating domain.name or xxx:yyy) — these are well-known issues, and CLDR provides the en_US_POSIX word break behavior to address programmer usage.

As for #2, the problem only occurs for sequences consisting of a lowercase letter followed by period followed by uppercase letter, with no intervening space. This is a somewhat unusual and artificial example. However, it does mean that a word can span a sentence boundary. While UAX#29 does not state that sentence breaks are also word breaks, it does say the following, which might suggest that interpretation: "[Sentence boundaries] are also used to determine whether words occur within the same sentence in database queries."

I think we need to add some language to UAX #29 to clarify that there are some cases (unusual in normal text) in which sentence boundaries might not also be treated as word boundaries.