Title: UAX #29 should mention CLDR/ULI Sentence Break Suppressions
Source: Steven Loomis (IBM Corporation), Mark Davis (Google, Inc.), Peter Edberg (Apple, Inc.), Laurenţiu Iancu (Microsoft Corporation)
Status: Individual contribution
Action: For consideration by the Unicode Technical Committee
Date: 2015-02-03
URL: http://bit.ly/HrHamster

Proposal

This document proposes amending UAX #29 to mention ULI segmentation suppression data (found in the CLDR data) as a resource which implementations can use to improve the quality of language-specific sentence breaking.

Referenced documents

This is an additional document in response to Action Item [138-A94]. An additional document [Iancu, et al.] before the UTC addresses the case of “Mr.Hamster” (no space between words, which is the case mentioned in [PRI240]) and its relation to other types of break opportunities. These two documents and approaches should be considered complementary to each other.

Problem statement

For situations such as (English) “Mr. Mouse” or (German) “Hr. Maus” (with a space between), UAX #29 Sentence Boundary Rules would provide a break opportunity before Mouse/Maus. However, the FULL STOP at the end of “Mr.” and “Hr.” indicates an abbreviation and not a sentence terminator.

ULI Segmentation Suppression data

The Unicode Localization Interoperability technical committee [ULI-TC] developed and maintains a set of entries for different languages where a Sentence Boundary should be suppressed. This data is hosted and distributed by the Common Locale Data Repository [CLDR]. See [TR35] section 9.2 “Segmentation Suppressions” for a description of the data format:

> *“The segmentation **suppressions** list provides a set of cases which, though otherwise identified as a segment by rules, should be skipped (suppressed) during segmentation. For example, in the English phrase "Mr. Smith", CLDR segmentation rules would normally find a Sentence Break between "Mr" and "Smith". However, typically, "Mr." is just an abbreviation for "Mister", and not actually the end of a sentence.”*

For English, the ULI suppression data might contain the list “Maj.”, “Mr.”, “Lt.Cdr.”, and other entries.

An online demo is available at http://demo.icu-project.org/icu-bin/icusegments which demonstrates the use of ULI suppression data. To use it, change the “Locale” popup from “English (non-ULI)” to “English (ULI)”. Note that the break between “Mr.” and “Weston” is suppressed when “(ULI)” is chosen.

- “Mr. |Weston” (**Without** break suppressions, a break opportunity after “Mr.”)
- “Mr. Weston” (**With** break suppressions, no break opportunity)

Recommended Text: UAX #29 section 5

[UAX #29] section 5 currently has the following note:

> *Note: As with the other default specifications, implementations are free to override (tailor) the results to meet the requirements of different environments of particular languages.*

The following text or similar could be added to that note:

> For example, Locale-sensitive boundary suppression specifications can be expressed in LDML [UTS35] and specific sentence boundary suppressions are contained in the Common Locale Data Repository [CLDR] and may be used to improve the quality of boundary analysis.

Recommended Text: UAX #29 section 2

[UAX #29] section 2 currently has the following note:

> *Note: Locale-sensitive boundary specifications can be expressed in LDML [UTS35] and be contained in the Unicode Locales project [CLDR]. The repository already contains some tailorings, with more to follow.*

This text could be amended as follows (including an update for the current name and status of CLDR):

> Note: Locale-sensitive boundary specifications (including boundary suppressions) can be expressed in LDML [UTS35] and tailorings are available in the Common Locale Data Repository [CLDR].

References

- [L2/15-068 Iancu et al.], Preventing sentence breaks within words like “Mr.Hamster”, February 2015. http://www.unicode.org/cgi-bin/GetMatchingDocs.pl?L2/15-068
- [138-A94] Action Item for Mark Davis et al., Investigate changes to sentence break to prevent words from spanning sentences in UAX #29, February 2014, http://www.unicode.org/L2/L2014/14026.htm #138-A94
- [PRI240] Feedback on Public Review Issue #240, Proposed Update UAX #29, Unicode Text Segmentation, April 2013, http://www.unicode.org/review/pri240/
- [UAX29] Mark Davis, Unicode Standard Annex #29, Unicode Text Segmentation, Version 7.0.0, June 2014, http://www.unicode.org/reports/tr29/tr29-25.html
- [ULI] The Unicode localization interoperability project, http://uli.unicode.org
- [UTS35] UTS #35: *Unicode Locale Data Markup Language (LDML), version 26* http://www.unicode.org/reports/tr35/tr35-37/tr35.html
- [CLDR] Unicode Common Locale Data Repository http://cldr.unicode.org/