L2/14-010R

Title: Proposal to make STerm a proper subset of Terminal_Punctuation
Source: Laurențiu Iancu (Microsoft Corporation) and Ken Whistler (SAP AG)
Status: Individual contribution
Action: For consideration by the Unicode Technical Committee
Date: 2014-02-04

Issue

In the Unicode 6.3 Character Database, most of the characters with the binary property STerm also have the property Terminal_Punctuation. This is a logical classification, consistent with the descriptions of the two properties in UAX #44: STerm, sentence terminal, used for determining sentence boundaries following the algorithm in UAX #29, and Terminal_Punctuation, a property given to "punctuation characters that generally mark the end of textual units."

There are, however, six characters which are assigned STerm but not Terminal_Punctuation as of Unicode 6.3 [character set]:

STerm = Yes and Terminal_Punctuation = No

U+055C	ARMENIAN EXCLAMATION MARK
U+055E	ARMENIAN QUESTION MARK
U+1735	PHILIPPINE SINGLE PUNCTUATION
U+1736	PHILIPPINE DOUBLE PUNCTUATION
U+10A56	KHAROSHTHI PUNCTUATION DANDA
U+10A57	KHAROSHTHI PUNCTUATION DOUBLE DANDA

There are two problems with the STerm and Terminal_Punctuation assignments of these characters:

1. The first two, U+055C and U+055E, are not sentence terminals. They are tonal punctuation marks, "placed directly above and slightly to the right of the vowel whose sound is modified, instead of at the end of the sentence, as European punctuation marks are" [core specification].
2. The other four characters, U+1735, U+1736, U+10A56, and U+10A57, are dandas and correctly assigned STerm but are lacking the Terminal_Punctuation property. This assignment is also an error, as it seems incongruous for a punctuation mark to be sentence terminal (STerm) yet not at the same time mark the end of textual units (Terminal_Punctuation).

The anomaly was noticed during the property definition work for Unicode 7.0, while examining the terminal-punctuation properties of dandas.

Proposal

To establish and verify the proper inclusion of the set of STerm characters in the set of Terminal_Punctuation characters, the proposal is to:

1. Reset the binary property STerm from Yes to No for the two Armenian tonal punctuation marks U+055C and U+055E, in PropList.txt for Unicode 7.0,
2. Assign the binary property value Terminal_Punctuation = Yes to the Philippine and Kharoshthi danda characters U+1735, U+1736, U+10A56, and U+10A57, also in PropList.txt for Unicode 7.0, and
3. Add an invariant test to check that STerm $\subset$ Terminal_Punctuation.