L2/14-010

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: 2014-01-15

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 (correctly) assigned STerm but not Terminal_Punctuation [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

This assignment appears to be 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. Assign the binary property value Terminal_Punctuation = Yes to the six characters enumerated above, in PropList.txt for Unicode 7.0, and
2. Add an invariant test to check that STerm $\subset$ Terminal_Punctuation.