L2/12-038

Title: Grapheme-break related properties for U+0E2F THAI CHARACTER PAIYANNOI'

Source: IBM (Nattapong Siralappanich, natta@th.ibm.com and V.S. Umamaheswaran, umavs@ca.ibm.com)

Date: 2012-01-26

The Proposal:

The proposal is to give the character **U+0E2F (ฯ) THAI CHARACTER PAIYANNOI** the Grapheme_Cluster_Break property values of both a '**Prepend**' and an '**Extend**', changing from the current '**Default**'

Problem statement:

The character U+0E2F (**ฯ**) is mainly used to abbreviate certain words. Most of the time, it is a suffix of an abbreviation. Sometime it appears both as a prefix and a suffix of an abbreviation. For example:

- กรุงเทพ**ฯ** (used as a suffix)
- **ฯ**ล**ฯ** (used as a prefix and a suffix)
- **ฯ**พณ**ฯ** (used as a prefix and a suffix)
- โปรดเกล้า**ฯ** (used as a suffix)

Basically, the function of this character is to abbreviate a word Very few words may have both a prefix and a suffix. The two examples shown above are the only two we have found so far. If a Paiyannoi is found as prefix of abbreviation, the Paiyannoi as suffix is also is required at the end of the abbreviated word. Otherwise, the prefix Paiyannoi is invalid. In some rare cases, it is used as a symbol to render Thai lunar calendar. This type of calendar is mainly used by some oracles and monks. For example:

However such usage is very rare, and is not expected to be represented in plain text.

There is no 'Grapheme_Clusgter_Break' property (other than the 'Default') assigned to this character, either in Table 2 under section 3.1 in UTR 29, or in the corresponding data file http://www.unicode.org/Public/6.0.0/ucd/auxiliary/GraphemeBreakProperty.txt .

The problem we have found is that in a line wrapping operation the suffix Paiyannoi character (examples 1 and 4 above) appears at the beginning of a line instead of being kept with the grapheme cluster it belongs to in the abbreviated word. It should not be split from the last grapheme cluster in the word. Similarly the prefix Paiyannoi in the second and third examples should not be at the end of a line split from the grapheme cluster it forms with the characters following it. The following is an example of running text – it is a modified sentence from news at: http://www2.narathiwat.go.th/narathiwat/news_poc/aspboard_Question.asp?GID=837

ทรงพระกรุณาโปรดเกล้าฯ ให้ ฯพณฯ พลเอกสุรยุทธ์ จุลานนท์ องคมนตรี
เป็นผู้แทนพระองค์ไปในการแข่งขันเรือกอ ฯลฯ ชิงถ้วยพระราชทาน กับชมการแสดง แสง เสียง
และสื่อผสมที่กรุงเทพฯ

Current Grapheme Clustering:

ท ร ง พ ร ะ ก รุ ณา โป ร ด เก **ล้า ฯ** ให้ **ฯ พ ณ ฯ** พ ล เ อ ก สุ ร ยุ ท ธ์ จุ ลา น น ท์ อ ง ค ม น
ต รี เป็ น ผู้ แท น พ ระ อ ง ค์ ไป ใน กา ร แข่ ง ขั น เรื อ ก อ **ฯ ล ฯ** ชิ ง ถ้ ว ย พ ระ รา ช ทา น กั
บ ช ม ก า ร แส ด ง แส ง เสี ย ง และ สื่ อ ผ ส ม ที่ ก รุ ง เท **พ ฯ**

The white space before or after the Paiyannoi in the above example is current grapheme breaks involving the Payannoi. The bold characters show what it should be clustered with. The following shows the correct grapheme clustering.

Correct Grapheme Clustering:

ท ร ง พ ร ะ ก รุ ณา โป ร ด เก **ล้าฯ** ให้ **ฯพ ณฯ** พ ล เ อ ก สุ ร ยุ ท ธ์ จุ ล า น น ท์ อ ง ค ม น ต รี เป็ น ผู้ แท น พ ระ อ ง ค์ ไป ใน กา ร แข่ ง ขั น เรื อ ก อ **ฯลฯ** ชิ ง ถ้ ว ย พ ระ รา ช ทา น กั บ ช ม ก า ร แส ด ง แส ง เสี ย ง และ สื่ อ ผ ส ม ที่ ก รุ ง เท **พฯ**

By current Default property, the character is split off from the grapheme cluster it should be kept with. The character itself can not be used as standalone character (unless the surrounding context is white space character). With the proposed property change It will be included as part of the appropriate grapheme cluster, except for the degenerate case.