L2/09-276

Date: August 5, 2009

To: Unicode Technical Committee

From: Mark Davis and Peter Edberg

Subject: Note on LineBreak property values

1. The LineBreak property value AI is used for a rather odd mixture of characters. It is described as follows:

> **AI: Ambiguous (Alphabetic or Ideograph)**
>
> Some characters that ordinarily act like alphabetic or symbol characters (which have the AL line breaking class) are treated like ideographs (line breaking class ID) in certain East Asian legacy contexts. Their line breaking behavior therefore depends on the context…
>
> As updated, the AI line breaking class includes all characters with East Asian Width A that are outside the range U+0000..U+1FFF, plus the following characters:
>
> 24EA CIRCLED DIGIT ZERO
> 2780..2793 DINGBAT CIRCLED SANS-SERIF DIGIT ONE..DINGBAT NEGATIVE CIRCLED SANS-SERIF NUMBER TEN

(To see the full set, use http://unicode.org/cldr/utility/list-unicodeset.jsp?a=\p{lb%3Dai})

This set of characters is based on notions that are becoming much less relevant, such as the East Asian Width property.

2. The LineBreak property value SG is already deprecated:

> **SG: Surrogates (XP) (Non-tailorable)**
>
> Line break class SG comprises all code points with General_Category Cs. The line breaking behavior of isolated surrogates is undefined. In UTF-16, paired surrogates represent non-BMP code points. Such code points must be resolved before assigning line break properties. In UTF-8 and UTF-32 surrogate code points represent corrupted data and their line break behavior is undefined.
>
> > *Note:* The use of this line breaking class is deprecated. It was of limited usefulness for UTF-16 implementations that did not support characters beyond the BMP. The correct implementation is to resolve a *pair* of surrogates into a supplementary character before line breaking.

However, LineBreak.txt still assigns the value SG to D800..DFFF.

3. Both AI and SG are reassigned by LineBreak rule LB1:

> … *Resolve AI, CB, SA, SG, and XX into other line breaking classes depending on criteria outside the scope of this algorithm.* In the absence of such criteria, it is

> recommended that classes AI, SA, SG, and XX be resolved to AL, except that characters of class SA …

4. For Unicode 6.0 we may want to consider changing the use of these values. They can lead to confusion; users of the standard may not realize that characters designated AI are in fact treated exactly as if they were AL by the default line break algorithm.

We could consider eliminating any assignment of code points to these property values, instead assigning all of the code points currently designated AI to AL, and all of the code points currently designated SG to XX. It would be useful to have some initial discussion on this now so as to get a general direction from the UTC.