L2/12-345

Title: Default property values for unassigned code points in the Currency Symbols block
Source: Laurențiu Iancu, Microsoft Corporation
Status: Individual contribution
Action: For consideration by the Unicode Technical Committee
Date: 2012-10-29

1. Abstract

Adding support for newly-introduced currency symbols in software can be costly when such symbols are accelerated into new versions of the Unicode Standard. The cost can be high especially in servicing previous releases of software products, which require component updates where UCD properties are used to determine the behavior of such symbols. To reduce the impact on shipped products and improve forward compatibility, this proposal is to set default property values to the unassigned code points in the Currency Symbols block, in particular default Bidi_Class and Line_Break property values.

2. Motivation

Examples of recently encoded currency symbols that have widespread usage include U+20B9 INDIAN RUPEE SIGN, accelerated into the Unicode Standard Version 6.0, and U+20BA TURKISH LIRA SIGN, the single new assigned character in Version 6.2. Other modern currency symbols may be added in the future, such as symbols for the Azerbaijani manat or potentially the Russian ruble. In such event, implementation and servicing costs can be reduced if the unassigned code points where new currency symbols are allocated carry default property values that are typical for currency symbols, rather than generic values that are replaced when currency symbols are encoded.

The default Bidi_Class and Line_Break property values of unassigned code points (in ranges not listed in the headers of the UCD files DerivedBidiClass.txt and LineBreak.txt), including unassigned code points in the Currency Symbols block, are Bidi_Class = L and Line_Break = XX. Those property values differ from the typical property values of currency symbols, Bidi_Class = ET and Line_Break = PR. Until a shipped product is serviced to update its character properties, the bidirectional and line-breaking behavior for new currency symbols is incorrect. If the default property values match the typical, then correct behavior is ensured when similar new currency symbols are introduced.

This proposal is to default to Bidi_Class = ET and Line_Break = PR the unassigned code points in the Currency Symbols block. A precedent exists to define default Bidi_Class and Line_Break property values for unassigned code points in certain ranges and the Currency Symbols block is well scoped in terms of expected future character assignments to reduce the risk of churn. Figure 1 illustrates the Bidi_Class and Line_Break property values for the code points in the Currency Symbols block, contrasting the current and proposed defaults for unassigned code points, as of Version 6.2 of the Unicode Standard.

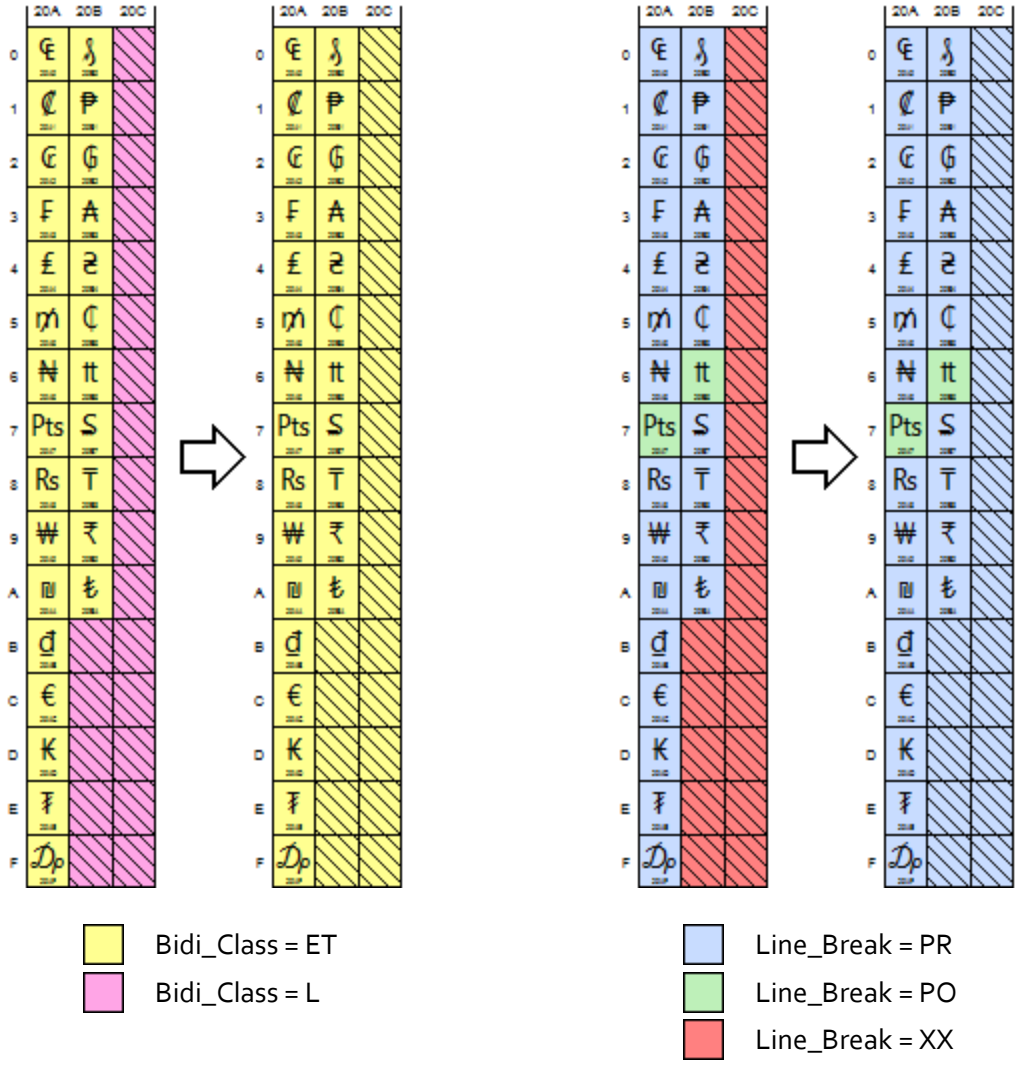

Figure 1: Current and proposed Bidi_Class and, respectively, current and proposed Line_Break property values for the code points in the Currency Symbols block. Code charts created with UniBook™ [1].

3. Behavior

Consider the scenario of an accelerated, newly encoded currency symbol, for which a user may be able quickly enough to find a font containing a glyph for the new symbol. Apart from glyph rendering, layout processes such as bidirectional reordering and line breaking will not function properly until the responsible software component is serviced to support the character according to its assigned property values. The inadequate behavior is due to the software component using default property values for code points that were unassigned at the time of its release.

The following examples illustrate inadequate behavior of a new currency symbol, denoted by '¤', having the assigned property values Bidi_Class = ET and Line_Break = PR, but handled according to the default property values Bidi_Class = L and Line_Break = XX (resolved to AL) for unassigned code points.

Input sequence of bidi classes		Correct bidi reordering with ¤ having Bidi_Class = ET	Actual bidi reordering due to ¤ having the default Bidi_Class = L
Sequence	Paragraph		
AL_1 ¤ EN AL_2	LTR	AL_2 EN ¤ AL_1	AL_1 ¤ EN AL_2
	RTL	AL_2 EN ¤ AL_1	AL_2 ¤ EN AL_1
AL_1 ¤ AN AL_2	LTR	AL_2 AN ¤ AL_1	AL_1 ¤ AL_2 AN
	RTL	AL_2 AN ¤ AL_1	AL_2 ¤ AN AL_1

Input sequence of line-breaking classes	Correct line-breaking behavior with ¤ having Line_Break = PR	Actual line-breaking behavior due to ¤ having the default XX
AL ¤ NU	AL ÷ ¤ NU	AL % ¤ NU
CM ¤ NU	CM ÷ ¤ NU	CM % ¤ NU
IS ¤ NU	IS ÷ ¤ NU	IS % ¤ NU
¤ ID	¤ % ID	¤ ÷ ID
¤ H2 (or H3, JL, JV, JT instead of H2)	¤ % H2 (and similarly for H3, JL, JV, JT)	¤ ÷ H2 (and similarly for H3, JL, JV, JT)
¤ IN	¤ ÷ IN	¤ % IN

Figure 2: Inadequate bidirectional and line-breaking behavior of a currency symbol '¤' resulting from it being handled according to the default property values for unassigned code points. The differences from the correct behavior are more pronounced in bidi. For line breaking, '÷' denotes a direct line-break opportunity and '%' an indirect one, i.e., a break opportunity only when intervening spaces are present.

4. Property values of currency symbols

When new currency symbols are encoded, the default Bidi_Class property value L of the previously unassigned code points where they are allocated usually becomes ET. This has been the case, for example, for the nine currency symbols encoded from Unicode Version 4.1 to Version 6.2, U+20B2 GUARANI SIGN through U+20BA TURKISH LIRA SIGN. In fact, all 27 characters in the Currency Symbols block assigned as Unicode 6.2 are Bidi_Class = ET, but the Bidi_Class defaults are explicitly documented in the UCD file DerivedBidiClass.txt starting with Version 4.0. The recently accepted currency symbol U+20BB NORDIC MARK SIGN also has a suggested Bidi_Class = ET in its proposal [2].

Similarly, of the eleven currency symbols encoded between Unicode Version 3.2 and Version 6.2, U+20B0 GERMAN PENNY SIGN through U+20BA TURKISH LIRA SIGN, ten were assigned the Line_Break property PR from the default XX documented in LineBreak.txt since Unicode 3.0. One of them, U+20B6 LIVRE TOURNOIS SIGN, was assigned Line_Break = PO when introduced in Version 5.2.

Given the typical Bidi_Class and Line_Break property value assignments of currency symbols, the amount of change from the current defaults would be greatly reduced if the defaults matched the typical values in the first place. Such defaults would, in turn, help reduce the cost of servicing software products previously released.

5. Summary of proposed changes

This proposal is to set the default property values for the unassigned code points in the Currency Symbols block to Bidi_Class = ET and Line_Break = PR. These default assignments incur updating the UCD files DerivedBidiClass.txt and LineBreak.txt as outlined below (in their respective formatting and excluding U+20BB NORDIC MARK SIGN which was accepted for encoding [2]).

Suggested updates to DerivedBidiClass.txt (resulting when the file is generated):

```
# The unassigned code points in the Currency Symbols block default to "ET":
#     Currency Symbols:  U+20A0  -  U+20CF
...
# Bidi_Class=European_Terminator
...
20BC..20CF    ; ET # Cn  [20] <reserved-20BC>..<reserved-20CF>
```

Suggested updates to LineBreak.txt:

```
# The unassigned code points in the Currency Symbols block default to "PR":
#     Currency Symbols:                        U+20A0..U+20CF
...
20BC..20CF;PR # <reserved-20BC>..<reserved-20CF>
```

6. References

[1] ASMUS, Inc., *UniBook™ Character Browser*, http://www.unicode.org/unibook/.

[2] Nina Marie Evensen, Deborah Anderson, *Proposal for one historic currency character, MARK SIGN*, ISO/IEC JTC1/SC2/WG2 N4308, http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4308.pdf, and INCITS L2/12-242, http://www.unicode.org/L2/L2012/12242-mark-sign.pdf, July 2012.