L2/15-192

Word Break Suppressor

Submitted by Richard Wordingham, July 2015

Summary

A proposed change to the documented function of U+2060 will leave one of its key rôles unfulfilled. The submission argues that the change should not be made, and requests a new character of the change is to go ahead.

Current Situation

At present, ISO 10646/Unicode contains a character U+2060 which is described as follows:

ISO 10646 4th Edition:

> WORD JOINER (2060) and ZERO WIDTH NO-BREAK SPACE (FEFF): These characters behave like a NO-BREAK SPACE in that they indicate the absence of word boundaries, but unlike NO-BREAK SPACE they have no presentational width.

Unicode Version 7.00:

> U+2060 WORD JOINER behaves like U+00A0 NO-BREAK SPACE in that it indicates the absence of word boundaries; however, the word joiner has no width. The function of the character is to indicate that line breaks are not allowed between the adjoining characters, except next to hard line breaks.
>
> …
>
> The word joiner should be ignored in contexts other than word or line breaking.

Imminent Change

There is currently a plan to change the definition of WORD JOINER for Unicode Version 8.00. The proposed new wording is:

> U+2060 WORD JOINER behaves like U+00A0 NO-BREAK SPACE in that it indicates the absence of line breaks; however, the *word joiner* has no width. The function of the character is to indicate that line breaks are not allowed between the adjoining characters, except next to hard line breaks.
>
> ...
>
> The word joiner should be ignored in contexts other than line breaking. Note in particular that the word joiner is ignored for word *segmentation*. (See Unicode Standard Annex #29, Unicode Text Segmentation.)

Formally, this will be a sharp difference between ISO 10646 and Unicode. The difference will be that in the former it suppresses word boundaries, while in the latter it suppresses line break opportunities. The justification given is the claim that it was not intended that WORD JOINER affect word boundaries. No further justification has been located.

Current Usage in Scriptio Continua

Some writing systems, most significantly Thai, but also Burmese, Cambodian and also the Greek and Latin of Late Antiquity, do not visibly mark word boundaries. Where line-breaking respects word boundaries (which is not always the case), there are three approaches:

1) Users may be instructed to mark word boundaries explicitly – U+200B ZERO WIDTH SPACE (ZWSP) is available for this purpose. This is somewhat unnatural, as users will not immediately see changes as a result and this character has no counterpart in writing by hand.

2) In some writing systems, e.g. the modern Lao system, syllable boundaries may be fairly easy to detect reliably (no false detections and few missed), and breaking at syllable boundaries may be acceptable. Cf. the writing system of overwhelmingly monosyllabic Vietnamese, in which spaces mark syllable boundaries.

3) Dictionary-based line-breaking algorithms have long been used for Thai and are now available for Khmer. These algorithms cannot be 100% reliable even when the target language is known, but the characters ZWSP and U+2060 WORD JOINER (WJ) are available to override the line-breaking algorithms. (There is the possibility that the breaking algorithm may deduce that it has failed and switch off in the neighbourhood of these characters.)

For spell-checking, a line-breaking algorithm may be combined with a word-based spell-checking tool. The process is logically circular, and the handling of spelling errors depends on the line-breaking algorithm's fallback behaviour in the presence of unrecognised words. The word-breaking behaviour can often be fixed by inserting ZWSP.

However, it may be necessary to advise the algorithm that a word boundary that it has deduced does not actually exist. At present, one may hope to do that by inserting WJ at such points.

Consequence of the Proposed Unicode Change

For a line-breaking algorithm deducing breaks on the basis of line-breaking opportunities, encountering WORD JOINER where a break is initially identified no longer indicates that there is not a word boundary there; it merely indicates that the user has chosen not to allow a line break there. There is therefore no reason for the algorithm to try a different division of the text into words.

When a line-breaking algorithm is used to furnish word boundaries, the line breaks suppressed by WORD JOINER will still be valid word boundaries!

Review of Current Break Controls

The line-breaking effects of the normally non-rendering line- and word-break controls, as defined by Unicode Standard Annexes #14 'Unicode Line Breaking algorithm' Revision Revision 35 and #29 Unicode Text Segmentation Revision 27 and their supporting properties (for Unicode 8.00), are given in the table below. Note that these properties have only limited applicability to characters with line-break properties of Contingent_Break, Ideographic or Complex_Context. (The letters of non-logographic scriptio continua scripts are classified as Complex_Context.)

Character		Line-breaking Effect			Word-breaking Effect		
Code	**Abbr.**	**Value**	**Rule**	**Action**	**Value**	**Rule**	**Action**
00AD	SHY	BA	LB31	Enable	Format	WB4	None
034F	CGJ	GL	LB12	Disable	Extend	WB4	None
200B	ZWSP	ZW	LB8	Enable	Other	WB14	Break
2060	WJ	WJ	LB11	Disable	Format	WB4	None

Note that as far as the default break rules are concerned, there is no practical difference between CGJ and WJ. The Unicode standard does state that CGJ has very different semantics, but this has very little effect on its use between words in modern rendering systems. Its uses besides breaking control are:

1) Blocking canonical reordering. As a breaking control, CGJ would be the final element of a grapheme cluster, and therefore not have this effect.

2) It prevents the formation of a contraction for collation. However, contractions only occur within words.

While it is theoretically possible for CGJ to be the element of a contraction, there are no such examples in CLDR.

Therefore if a character is required that suppresses line breaking without suppressing word breaking, CGJ may be used. WJ could therefore be used with its ISO 10646 meaning of suppressing word breaks.

New Character

If, however, the UTC decides that U+034F WORD JOINER should be stripped of its power to suppress word breaks, a new character to replace it is needed. It should have the same properties as U+034F WORD JOINER, except that it suppresses word breaks. I propose:

U+2065 WORD BREAK SUPPRESSOR

The code point is chosen because it is already allocated as a default ignorable. The block appears to be suitable, for the new character will take on the ISO 10646 rôle of U+2060 WORD JOINER.

WORD BREAK SUPPRESSOR may be considered as a disunification of U+2060 WORD JOINER – U+2060 WORD JOINER will suppress line breaks but not word breaks, whereas U+2065 WORD BREAK SUPPRESSOR will suppress both line and word breaks.

I also recommend that UAX #29 Unicode Text Segmentation be amended to record what can be said about word-breaking in scriptio continua. This will require new values of the Word_Break property, both for letters and for the new control character WORD BREAK SUPPRESSOR.

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html**.**
See also http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** *Word Break Suppressor*
2. Requester's name: *Richard Wordingham*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *22 July 2015*
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *Yes*
   (or) More information will be provided later: *No*

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): *No*
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *Yes*
      Name of the existing block: *General Punctuation*
2. Number of characters in proposal: *one*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary B.1-Specialized (small collection) B.2-Specialized (large collection) *A*
   C-Major extinct D-Attested extinct E-Minor extinct
   F-Archaic Hieroglyphic or Ideographic G-Obscure or questionable usage symbols
4. Is a repertoire including character names provided? *Yes*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *Y*
   b. Are the character shapes attached in a legible form suitable for review? *N/A*
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
      Richard Wordingham
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
      Richard Wordingham (richard.wordingham@ntlworld.com)
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *No*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *No*
7. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *Yes*
   Text segmentation

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *No*
   If YES explain
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *No*
   If YES, with whom?
   If YES, available relevant documents:
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) *Common*
   Reference:
5. Are the proposed characters in current use by the user community? Yes, as U+2060
   If YES, where? Reference: *Thailand, Laos, Myanmar*
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *No*
   If YES, is a rationale provided?
   If YES, reference:
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *N/A*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference: *See text*
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character? *Yes*
    If YES, is a rationale for its inclusion provided?
    If YES, reference: *See Section xxx*
11. Does the proposal include use of combining characters and/or use of composite sequences? *No*
    If YES, is a rationale for such use provided?
    If YES, reference:
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? *Yes*
    If YES, describe in detail (include attachment if necessary)
    The proposed character will suppress both line and word breaks. The effect on line breaks will be as U+2060 WORD JOINER.
13. Does the proposal contain any Ideographic compatibility characters? *No*
    If YES, are the equivalent corresponding unified ideographic characters identified?
    If YES, reference: