L2/08-344

PROPOSAL TO THE UNICODE TECHNICAL WORKING GROUP OF THE UNICODE CONSORTIUM TO REVERT TO THE ORIGINAL PROPERTIES OF THE ZERO WIDTH SPACE CHARACTER (ZWSP)

6 September 2008

Javier Solá

Characters affected: u200B ZERO WIDTH SPACE

Object the proposal: *Reversal of fundamental changes introduced in the definition of the ZWSP character in July 2008 through an ERRATA to the standard. Revert to ZWSP previous behavior as a word boundary, reversing the error introduced in 2003 by changing the ZWSP to be a format character, and not taking this change into account in AUX #29.*

The ZERO WIDTH SPACE

In some South East Asian languages, such as Thai, Myanmar, Khmer (Cambodian) and Lao, words are written without spaces between them. The space -when used- signals a pause in the speech, as a comma would in most western languages.

Most computer applications that deal with text (word processors, web browsers, search engines) use spaces as word boundaries and line-breaking opportunities, and - unless these applications are specially prepared - they cannot do line-breaking, word searches, word selection, spell-checking, and other functions for these languages that do not use spaces as separators.

To solve this issue, the UNICODE standard created the ZERO WIDTH SPACE character (0x200B), a character that had all the characteristics of a space, except that it had zero width, and would not be displayed in a text. By using this character between words, text written in these languages retains its visual aspect (no space between words), but has the same advantages as all other languages that use a SPACE character as a separator (selection, indexing, spell-checking).

Naturally, the ZERO WIDTH SPACE (ZWSP) was placed in the SPACES group of general punctuation, as it has the same practical characteristics of other spaces (except its visibility).

Usage of the ZWSP

The use of ZWSP has been quite successful, specially in Cambodia. Software that uses this technology was integrated and made mandatory in all schools of Cambodia, and it is also quickly replacing the old ways of typing (legacy fonts and line-breaking done by hand) in the government. UNICODE Spell-checkers for Khmer where developed, and made mandatory in schools.

Most word processing software (Ms Office, OpenOffice) already use ZWSP as a word boundary and a line-breaking opportunity. It is also used by search engines, such as Google, for searching and indexing Khmer text. Without the ZWSP (present in all the pages), it would not be possible to either present pages correctly in Internet (no line-breaking), or do indexing for this group of languages,

The National Standard Unicode Keyboard for Cambodia, issued by the National ICT Development Authority of the Cambodian government places the ZWSP character directly in the SPACEBAR, displacing the SPACE to SHIFT+SPACEBAR, as it is used much less. This choice was made based on statistical analysis that showed that ZWSP made for 16 out of every 100 strokes in the keyboard. All official Ministry of Education computer textbooks in Khmer teach how to use the ZWSP and how to do spell-checking based on it as a separator. The Khmer name for this character, as printed in the keyboards translates as " the gap that cannot be seen".

Changes to the ZWSP in the UNICODE Standard

And then ZWSP was changed by the UNICODE consortium 27 October 2003 through PRI 21.

> ***Changing U+200B Zero Width Space from Zs to Cf 2003.10.27***
>
> *There have been persistent problems with usage of the U+200B Zero Width Space (ZWSP). The function of this character is to allow a line break at positions where it normally would not be allowed, and is thus functionally a format character with a general category of Cf. This behavior is well documented in the Unicode Standard, and the character not considered a Whitespace character in the Unicode Character Database. However, for historical reasons the general category is still Zs (Space Separator), which causes the character to be misused. ZWSP is also the only Zs character that is not Whitespace. The general category can cause misinterpretation of rule D13 Base character as allowing ZWSP as a base for combining marks.*
>
> *The proposal is to change the general category of U+200B from Zs to Cf.*
>
> ***Resolution:** Closed. The general category of U+200B will be changed from Zs to Cf in Unicode version 4.0.1.*

The PRI only considered the line-breaking property of the character, while the standard clearly established that it was also a word boundary. To correct the fact that Format characters are not considered as line-breaking characters, ZWSP was explicitly included in the line-breaking rules in UAX# 29, but **it was NOT included in the Word Boundary rules**.

> Note: AUX# 29 states that "Format characters are also ignored by default, because these characters are normally irrelevant to such boundaries..."

When - due to regressions in word processing software, this inconsistency was detected in 2008, and pointed out – instead of doing the necessary corrections to revert the regression caused by this change in the UNICODE standard, THE WORD-BOUNDARY PROPERTY OF THE CHARACTER WAS RETROACTIVELLY ERASED FROM THE UNICODE STANDARD with a simple ERRATA note in May 2008, to make the ZWSP consistent with UAX# 29, instead of considering changing UAX# 29 to be consistent with the properties of the character and its usage. The errata implies that what the standard said before, what made software work for these languages, and what and was applied by ALL Office software was... just a an error.

> *In certain locations in the text of the standard, there are erroneous statements implying that use of a U+200B ZERO WIDTH SPACE (ZWSP) character indicates a "word break", when what are in question are actually line break opportunities. The text should be corrected to read as noted below.*
>
> *In Section 16.2, Layout Controls, on page 535 of The Unicode Standard, Version 5.0, the following text:*
>
> ***Zero Width Space.*** *The U+200B ZERO WIDTH SPACE indicates a word boundary, except that it has no width. Zero-width space characters are intended to be used in languages that have no visible word spacing to represent word breaks, such as Thai, Khmer, and Japanese.*
>
> *should be replaced by this text:*
>
> ***Zero Width Space.*** *The U+200B ZERO WIDTH SPACE indicates a line break opportunity, except that it has no width. Zero-width space characters are intended to be used in languages that have no visible word spacing to represent line break opportunities, such as Thai, Khmer, and Japanese.*
>
> *In Section 11.3, Myanmar, on p. 381 of The Unicode Standard, Version 5.0, the following text:*
>
> ***Spacing.*** *Myanmar does not use any whitespace between words. If word boundary indications are desired—for example, for the use of automatic line layout algorithms—the character U+200B ZERO WIDTH SPACE should be used to place invisible marks for such breaks.*
>
> *should be replaced by this text:*
>
> ***Spacing.*** *Myanmar does not use any whitespace between words. If explicit line break opportunities are desired—for example, for the use of automatic line layout algorithms—the character U+200B ZERO WIDTH SPACE should be used to place invisible marks for such breaks.*
>
> *Note that as for other textual errata for the text of Unicode Version 5.0, these also apply to the text of Unicode Version 5.1.*

The consequences of this change are devastating for the use of computers and the UNICODE standard in countries like Cambodia, Lao or Myanmar, specially for Cambodia, which defines UNICODE as the National standard in the Information and Communication Policy in the process of being decided by Parliament at this time.

If this change is applied, automatically Khmer text cannot be searched, selected, indexed or spell-checked. All the text that has been typed during the last five years becomes unusable, as it always includes ZWSP.

Also, all word processing software becomes obsolete, as it respects the original properties of ZWSP. The upcoming version of OpenOffice.org (3.0) has corrected the issue by explicitly ignoring the change in UNICODE and including ZWSP as a Word Boundary, as the regression is not acceptable for its users. Search engines, if they apply the change, will have to stop indexing Thai, Khmer, Myanmar and Lao. Spell-checking automatically stops working, as there are no word delimiters.

We believe that this change should be immediately reverted by the UNICODE Consortium, first by canceling the ERRATA that eliminates the WORD BOUNDARY property of the ZWSP. Such crucial change in the standard cannot be done just by a simple errata, without any consultation and without measuring the enormous consequences for those affected.

Second, the technical issue should be resolved, either including ZWSP explicitly in UAX# 29 as a word boundary (and a grapheme boundary if necessary) or by reverting its status to what the UNICODE standard's say that it is, a SPACING character.

As the change to a SPACING character might affect work done in the last years that took this technical into account, the solution of modifying UAX# 29 seems easier.

What is clear is that the change in the word boundary property of ZSWP goes against the most basic and sacred principle of the UNICODE standard: that it might not be changed. Also, it bypasses the UTC and public review by doing such crucial change through an errata.

Proposal:

1. That the May 2008 ERRATA that changes the Word Boundary property of ZWSP be canceled or eliminated.
2. That UAX# be modified to include ZWSP as a word boundary, modifying also grapheme boundaries if it was necessary.