# Word boundaries and line breaks

Norbert Lindenberg, 2024-02-15

The UTC has asked me in action item 176-A101 to propose improved wording for the last paragraph of the introduction of Section 4 of UAX 29, referring to document L2/23-160 item 4.12.

As L2/23-160 item 4.12 documents, this action item relates to feedback I provided on two paragraphs in the introduction of Section 4. The discussion in item 4.12 says that PAG "agreed to make a change to figure 2 for Unicode 15.1". I don't know what that change was meant to be be; I don't see one when comparing revisions 41 and 43 (or even 44) of UAX 29; and I don't see a problem with that figure.

My feedback related to the first and third paragraphs following that figure. The comment on the third paragraph has been partially addressed in revision 43 by changing the last sentence from "That means that satisfactory treatment of languages like Chinese or Thai requires special handling." to "The relationship of line break and word break boundaries is script-specific and may require special handling for satisfactory treatment."

I propose to move the second paragraph to first place, and then replace the other two with largely new text:

> As with the other default specifications, implementations may override (tailor) the results to meet the requirements of different environments or particular languages. For some languages, it may also be necessary to have different tailored word break rules for selection versus Whole Word Search.
>
> Whether the default word boundary detection described here is adequate, and whether word boundaries are related to line breaks, varies between scripts. The style of context analysis in line breaking (see UAX 14 section 3.1) used for a script can provide some rough guidance:
>
> - For scripts that use the Western style of context analysis, default word boundaries and default line breaks are usually adequate. A default line ~~boundary~~ break opportunity is usually a default word boundary, but there are exceptions such as a word containing a SHY (soft hyphen): it will break across lines, yet is a single word. Tailorings may find additional line break opportunities within words due to hyphenation. Scripts in this group include Latin, Arabic, Devanagari, and many others; they can be identified by having letters with line break class AL.

- For scripts that use the East Asian or Brahmic styles of context analysis, the default word boundary detection is not adequate; it needs tailoring. The default line breaks, on the other hand, are usually adequate. Word boundaries are irrelevant to line breaking. Scripts in this group include Chinese, Japanese, Brahmi, Javanese, and others; they can be identified by having letters with line break class ID, AK, or AS.
- For scripts that use the South East Asian style of context analysis, neither the default word boundaries not the default line breaks are adequate. Both need tailoring. The reason is that line breaks should only occur at word boundaries, but there's no demarcation of words. Scripts in this group include Thai, Myanmar, Khmer, and others; they can be identified by having letters with line break class SA.

Hangul is treated as part of the first group for default word boundary detection; as part of the second group for default line breaking. Some scripts may be treated as being part of the first group only because not enough information is available for them.

The proposed text also obsoletes the following note in section 4.1.1:

- ~~For Thai, Lao, Khmer, Myanmar, and other scripts that do not typically use spaces between words, a good implementation should not depend on the default word boundary specification. It should use a more sophisticated mechanism, as is also required for line breaking. Ideographic scripts such as Japanese and Chinese are even more complex. Where Hangul text is written without spaces, the same applies. However, in the absence of a more sophisticated mechanism, the rules specified in this annex supply a well-defined default.~~