L2/22-086

Specification updates for orthographic syllables and line breaking

Norbert Lindenberg with contributions from Aditya Bayu Perdana, Andrew Glass, முத்து நெடுமாறன் (Muthu Nedumaran), and Uli Kozok
2022-04-13

Proposal

This document proposes:

- To define the term "orthographic syllable" in section 6.1 of The Unicode Standard, as specified below in the section Orthographic syllables.
- To update the descriptions of line breaking in The Unicode Standard for the scripts Balinese, Batak, Brahmi, and Grantha, as specified below in the sections about these scripts.

A gray background in this document indicates proposed specification text for The Unicode Standard.

This document was split out from L2/22-080 Line breaking at orthographic syllable boundaries so that the proposed specification updates can be fast-tracked.

Orthographic syllables

The Unicode Standard uses the term "orthographic syllable" in the discussion of several Brahmic scripts, for example, in statements that line breaks can occur at the boundaries of orthographic syllables for the (Eastern) Cham, Batak, Javanese, and Makasar scripts.

The term "orthographic syllable" is defined in section 12.1, Devanagari, in a rather Devanagari-flavored form: "The effective unit of these writing systems is the orthographic syllable, consisting of a consonant and vowel (CV) core and, optionally, one or more preceding consonants, with a canonical structure of (((C)C)C)V. … The orthographic syllable is built up of alphabetic pieces, the actual letters of the Devanagari script. These pieces consist of three distinct character types: consonant letters, independent vowels, and dependent vowel signs. …".

This description is insufficient for many other scripts, which may use independent vowels or numbers at the core of orthographic syllables, conjunct forms for consonants that follow the base, register shifters, tone marks, final consonant marks, and more. A more general description of an orthographic syllable should be inserted into TUS section 6.1, Writing Systems, before the paragraph discussing abugida encoding models ("Because of legacy practice…"):

> In Brahmic scripts, text often needs to be interpreted as a sequence of *orthographic syllables*, each of which is a two-dimensional visual arrangement of glyphs that form a unit. At the core of an orthographic syllable is a base character, which can be a consonant, an independent vowel, or (in some scripts) a numeric character. Attached to this core may be dependent forms (such as half-forms, subjoined forms, repha forms, medial forms) of consonants or independent vowels, as well as nukta marks, virama marks, dependent vowel marks, register shifter marks, tone marks, final consonant marks, and other marks. It is common for different components of orthographic syllables to form ligatures. Orthographic syllables don't always correspond to phonological syllables; it is common for the final consonants of phonological syllables to become the base characters, or sometimes dependent forms, of subsequent orthographic syllables.

A definition of orthographic syllables should also be added to the Unicode glossary and linked to an updated definition of *aksara*.

Line breaking for Balinese

The Unicode Standard, section 17.3, Balinese, currently has a paragraph on hyphenation:

> ***Hyphenation.*** Traditional Balinese texts are written on palm leaves; books of these bound leaves together are called *lontar*. U+1B60 BALINESE PAMENENG is inserted in lontar texts where a word must be broken at the end of a line (always after a full syllable). This sign is not used as a word-joining hyphen—it is used only in line breaking.

According to Aditya Bayu Perdana, this description is not correct. *Pameneng* is not used as a hyphen, but as a filler when the text content of a line doesn't fill the available space entirely. It might occur within a word, but it is not required just because a line break occurs within a word.

The paragraph on hyphenation in section 17.3 of The Unicode Standard should be replaced with:

> ***Line Breaking.*** Line breaks may occur after any orthographic syllable. Traditional Balinese texts are written on palm leaves; books of these leaves bound together are called *lontar*. U+1B60 BALINESE PAMENENG may be inserted in lontar texts at the end of a line to fill the line.

Line breaking for Batak

The Unicode Standard, section 17.6, Batak, currently states that "Opportunities for a line break occur after any full orthographic syllable." According to Uli Kozok, this is not correct. Instead, lines can be broken before every spacing character. However, the reordering of the glyphs for vowel signs when a killer follows the

next consonant, as described in the Standard, is required even when the consonant that the vowel is attached to and the killer are on separate lines.

In Unicode-based text processing, line breaking typically occurs before and separate from font rendering, and font rendering does not have access to text that is located on a different line. It is then not possible to implement the traditional behavior. Instead, an orthographic syllable representing a final consonant should be kept together with the previous orthographic syllable, so that glyph reordering can work.

The line breaking information in section 17.6 of The Unicode Standard should be replaced with:

> ***Line Breaking.*** Traditionally, line breaks can occur before any spacing character. However, the vowel reordering described above is required even when a line break occurs between the characters involved. In typical Unicode-based implementations, this requires keeping the characters involved on the same line.

Line breaking for Brahmi

The Unicode Standard currently is silent on line breaking for the Brahmi script. According to Andrew Glass, it breaks at orthographic syllable boundaries.

The following information should be added to section 14.1 of The Unicode Standard:

> ***Line Breaking.*** Line breaks may occur after every orthographic syllable.

Line breaking for Grantha

The Unicode Standard currently is silent on line breaking for the Grantha script. According to Muthu Nedumaran, it breaks at orthographic syllable boundaries and does not use hyphens.

The following information should be added to section 15.13 of The Unicode Standard:

> ***Line Breaking.*** Line breaks may occur after any orthographic syllable. Hyphens are not used.

Acknowledgments

I'd like to thank Richard Ishida and members of the Unicode Script Ad Hoc for providing feedback on L2/22-080 Line breaking at orthographic syllable boundaries, from which this document was split out.