L2/13-083

# Feedback on the Proposed Update for UAX#14 (unidode.org/reports/tr14-31.html)

*Asmus Freytag*

April 28, 2o13

## Rule LB21a

I have a problem with the content of the rule added a few revisions ago:

> LB21a  Don't break after Hebrew + Hyphen
>
> HL (HY | BA) ×

The effect of this rule is to prevent **any and all** line breaks following a HYPHEN MINUS, HYPHEN (and a number of other, script specific hyphens), if they are preceded by a Hebrew letter (and not followed by a space or ZWSP).

I have tried to find out what actual requirements this rule is trying to address and have been unsuccessful in getting any details from either the proposers or the editors.  The documentary evidence in the register is limited to the meager statement in document: L2/11-141R (http://www.unicode.org/L2/L2011/11141r-seg-linebreak.pdf) which characterizes the issue as "With <Hebrew hyphen non-hebrew>, there is no break on either side of the hyphen."

The complete lack of evidence and documentation for the requirement, makes it difficult to evaluate whether "non Hebrew" should literally encompass any other character in Unicode, or just typical word characters in some likely non-Hebrew script.

In particular, it does not address the issue whether the unknown source originating that "requirement" did contemplate characters, such as ideographs, which normally provide for a linebreak opportunity.

Also, the new rule prevents a line break in the case <Hebrew hyphen Hebrew>, which is **not covered** by the stated requirement.

In a recent private communication I was informed that "the origin of this was some info from Hebrew experts at Apple".

I would find it very unlikely that linebreak feedback received from native Hebrew users would take into account cross-script interaction with every script in Unicode, unless these users were also experts at internationalized text layout with actual experience in a wide range of other scripts. (And, in my experience, few users can correctly translate the handling of a particular edge case into a workable general rule, and even fewer can correctly write a UAX#14 style rule) That's why I'm very leery about taking L2 11/141 at face value without any actual evidence.

An additional reason why I question this particular rule, is that the statement is unusual in its scope for an alphabetic script. In alphabetic scripts one can typically count on spaces being present, so one doesn't worry about letter-punctuation boundaries or letter-ideograph boundaries the same way as one would do for an ideographic script.

The more common concerns in an alphabetic script would be in keeping "words" together, that is, to prevent an accidental separation of letter character (or digits) by word-internal punctuation. In alphabetical scripts that use spaces it is really unusual to make a punctuation bind more tightly to the following character than a letter character, yet that is the effect of rule LB21a as written.

In the absence of more detailed justification, I strongly suspect therefore the intent of the rule is

> "treat hyphen as part of a Hebrew word and don't allow a line break inside the Hebrew word (even if continued in another script)".

I am not convinced that "non-hebrew" necessarily was intended to cover things like opening punctuation, ideographs or many other characters before which a linebreak is allowed in cases where spaces aren't used to separate words.

I further suspect that a reformulation ("treat hyphen following a Hebrew letter as a Hebrew letter") is in fact a more accurate representation of the requirement. (I'm not going to quibble right now about equating BA and HY here, although that may be overbroad as well, and is acknowledged as such in L2/11-141).

If, as I suspect, the reformulation expresses what was intended for linebreak then the difference shows up in these scenarios

HL HY ID

HL HY OP

etc.

In all of these cases I would have expected there to be a line break after the HY, even if it the HY occurs at the end of a Hebrew word. Hebrew embedded in Chinese would be less likely to use spaces than Hebrew surrounding fragments in Chinese; therefore Chinese users are disproportionately affected by adding such strongly acting a Hebrew-specific rule in the default algorithm.

In the absence of better evidence, I propose that LB21a be changed to a "treat as" rule similar to LB9.