L2/20-087

Proposal to fix Hebrew in UAX #14 by ruling out LB21a..b

For consideration by the Unicode Technical Committee

2020-03-27
Marcel Schneider (charupdate@orange.fr)

> *We should always tell what we see.*
> *Above all we should always*
> *—which is more difficult—*
> *see what we see.*
>
> Charles Péguy, 1910

This proposal is a part of the set designed to replace the *Proposal to re-engineer spaces and punctuation in UAX #14* announced in L2/20-005 *Proposal to make material changes to UAX #14,* that it is a complement of.

This and the related *Proposal to adjust space characters in UAX #14* form a minimal submission and won't be followed for now by the missing proposals of the set, due to the on-going pandemic.

1 Problems

1.1 Internal and external inconsistency of LB21a

In the *Unicode Line Breaking Algorithm,* rule LB21a is formalized with a regex that does not match the rule's title:

Do not break after Hebrew + Hyphen

HL (HY | BA) ×

Nor does this regex match the original requirement as stated in item #1 of proposal L2/11-141R *Segmentation & Linebreak*:

With <hebrew hyphen non-hebrew>, there is no break on either side of the hyphen.

More in detail, the following inconsistencies are found in rule LB21a:

1. The line breaking class BA contains also almost all breaking spaces, but rule LB21a is not overtly about spaces, nor does the original proposal L2/11-141R, item #1, mention spaces.
2. The script-specificness of the character following the hyphen is dropped in the rest of L2/11-141R, item #1, and in rule LB21a. Nevertheless, the Hebrew hyphen *maqaf* is not mentioned, despite its alteration per rule LB21a is expected to raise concern.
3. Class AL was split into AL and HL (or class HL was broken out of AL) for the purpose of implementing item #1 of L2/11-141R, but class BA was not split. The proposal considered splitting class BA in two, but only "because of either Soft Hyphen or Hyphenation point", not with respect to spaces.

1.2 Alteration of the Hebrew *maqaf* per LB21a

Rule LB21a applies also in a plain Hebrew context. If Hebrew had a preference for non-breaking hyphens, U+05BE HEBREW PUNCTUATION MAQAF would be non-breaking; yet it has been assigned the line breaking class BA. If that did not meet user expectations, changing class to GL would be the way to go.

The maqaf massively occurs at line end in Biblical Hebrew [1]. Its line breaking behavior should be kept as intended.

The Original Proposers' intent was probably not to make all hyphens in Hebrew non-breaking and recommend adding a ZWSP to make a breaking hyphen up. Yet that is the effect of LB21a. In that, this rule is harmful to Hebrew itself, that it was designed to serve.

1.3 Alteration of spaces in Hebrew per LB21a

Unlike rule LB12a, this rule is only about hyphens, not about "spaces and hyphens". Hence while rule LB12a uses [^SP, BA, HY] to exclude both spaces and hyphens, rule LB21a cannot use [HY | BA] to refer to hyphens only, since BA is a mixed class encompassing both hyphens and spaces. By using BA to refer to hyphens, rule LB21a inflicts some collateral damage to breaking fixed-width spaces. This results in fixing the narrower fixed-width spaces U+2004..U+200A by making them non-breaking, but also in breaking the wider fixed-width spaces U+2000..U+2003 by making them non-breaking altogether.

More precisely, the unspoken yet assumed goal was most probably to provide Hebrew with a full range of non-breaking fixed-width spaces, that all other scripts had been deprived of, regardless of backwards compatibility. In the process, Hebrew is deprived of breaking em and en spaces. That is another way this rule does harm to Hebrew.

1.4 Inappropriateness of rules LB21a and b

1.4.1 LB21a

As of **LB21a,** I cannot help noting that UAX #14 was used to push undocumented features through, in a relative opacity and with undesirable side effects on the very Hebrew script that LB21a was meant to fix, so I consider LB21a as a "quick-and-dirty" solution.

With respect to this topic, the date of the process is interesting in that it took place over a decade after encoding NARROW NO-BREAK SPACE for all scripts, not for Mongolian alone. (About this, please see page 3 of L2/19-112 *Proposal to define a space character as a group separator*.) By the time of the process, NNBSP was going to be available in virtually all relevant fonts. So visibly in Hebrew, NNBSP was not enough to meet user expectations regarding non-breaking fixed-width spaces.

1.4.2 LB21b

LB21b in turn is not needed, since its purpose "Do not break between SOLIDUS and Hebrew letters" is addressed by the SY class specification (quoted from version Unicode 13.0.0 of UAX #14; **bold added**):

> URLs are now so common in regular plain text that they need to be taken into account when assigning general-purpose line breaking properties. Slash (solidus) is allowed as an additional, limited break opportunity to **improve layout of Web addresses.** As a side effect, some common abbreviations such as "w/o" or "A/S", which normally would not be broken, acquire a line break opportunity. The recommendation in this case is for **the layout system not to utilize a line break opportunity allowed by SY unless** the distance between it and the next line break opportunity **exceeds an implementation-defined minimal distance.**

Small constructs are already handled per this recommendation, while preventing breaks in large constructs—should they occur in Hebrew, namely in internationalized URLs—is anyway not desirable.

2 Solutions

2.1 Pro-forma solution

*T*here would be a way of making LB21a match its alleged design goal. This solution is however not recommended, as it makes the rule de facto pointless and fails to address the underlying issue.

Hence this is not a true solution and is provided only pro forma. The point of this is to show that making the rule match its alleged design goal was so easy that not doing it resulted forcibly from a design decision.

2.1.1 The part after the hyphen

Restricting the part after the hyphen to non-Hebrew is done by adding a negated class:

HL (HY | BA) × [^HL]

2.1.2 The hyphen part

Restricting the hyphen pattern to hyphens is done by removing the breaking spaces from class BA, so as to catch no more than hyphens and other punctuation characters that have been appropriately assigned class BA. This removal is suggested in the simultaneously submitted *Proposal to adjust space characters in UAX #14.*

2.2 Actual solution

I'd suggest giving Hebrew and all interested scripts a proper solution today, belated as it is. The addressed problems can be solved for all scripts at once, so making a special case for Hebrew is not needed.

2.2.1 Delete rule LB21a

Since the goal of adding rule LB21a for Hebrew was actually to fix the breaking fixed-width spaces issue, not to tweak hyphens, I'd suggest beginning with deleting rule LB21a as a part of solving the problem.

2.2.2 Delete rule LB21b

This rule "Do not break between solidus and Hebrew letters" / SY × HL / can be deleted alongside, because of the findings reported in subsection 1.4.2 above.

2.2.3 Cancel the Hebrew letter class

In the wake, the HL class can be merged with AL again, since it was created—by breaking it out of AL—for the sole purpose of adding rule LB21a.

As a consequence, regexes in UAX #14 containing the string "AL | HL" (10 instances) become more streamlined again.

2.2.4 The rest of the solution

The rest of the solution is about fixing the space characters. This is suggested in the simultaneously submitted *Proposal to adjust space characters in UAX #14.*

3 Rationale

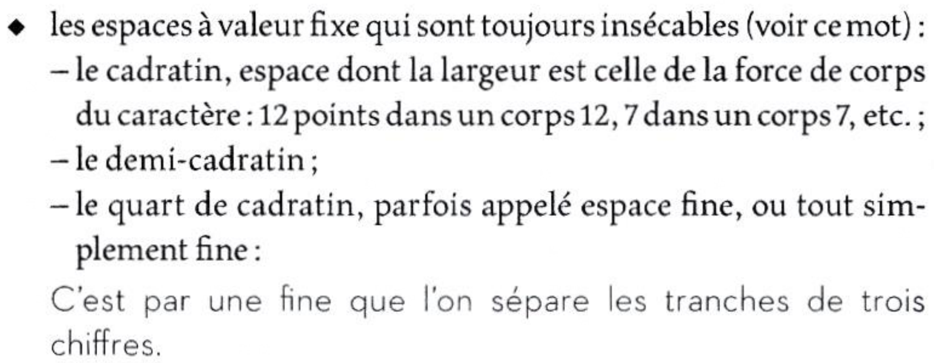
This rationale tries to reinforce the point in making a case for changing focus and solving the actual problem. Instead of focusing on Hebrew alone, Unicode should focus on all scripts, at least on those that are interested in using non-breaking fixed-width spaces. The actual problem is how to get non-breaking fixed-width spaces for all interested scripts.

3.1 Interest in non-breaking fixed-width spaces

First there is a need to bring evidence that scripts other than Hebrew are actually interested in non-breaking fixed-width spaces. The *Problem* section works out how LB21a should be read so as to be prevented from missing the actual point of the rule. When applying that lecture, we understand that rule LB21a showcases the interest of Hebrew user communities in non-breaking fixed-width spaces.

Yet this interest is not proper to Hebrew. Please see the simultaneously submitted *Proposal to adjust space characters in UAX #14* for a larger set of evidence, including the one cited below in an abbreviated form.

Latin user communities likewise are interested in a full range of non-breaking fixed-width spaces. For example, one of the most renowned and widely taught and followed French style guides [2] states on page 90:



◆ les espaces à valeur fixe qui sont toujours insécables (voir ce mot) :
– le cadratin, espace dont la largeur est celle de la force de corps du caractère : 12 points dans un corps 12, 7 dans un corps 7, etc. ;
– le demi-cadratin ;
– le quart de cadratin, parfois appelé espace fine, ou tout simplement fine :
C'est par une fine que l'on sépare les tranches de trois chiffres.

The first line translates to **(bold added)**: "**the fixed-width spaces** that **are always non-breaking** (see this word):" The second-level bullet list enumerates the em space, the en space, and the four-per-em space, equated to the thin space, that is used as a group separator in numbers, the example sentence at the end of the snippet states.

Since Hebrew user communities are okay that U+2000..U+200A are all non-breaking (not only U+2007), while Hebrew has nothing particular with respect to these spaces that Cyrillic, Greek, Latin and so on would not have, rule LB21a is strong evidence that **these spaces can change line breaking class,** except perhaps for the 2—actually 4—largest ones U+2000..U+2003.

3.2 Issue history

Adding rule LB21a was a part of the proposed update of UAX #14 L2/11-385 (also revision 27). It was first published in January 2012 for Unicode 6.1.0. But soon the new rule became controversial, as we can see in L2/13-083 *Feedback on the Proposed Update for UAX#14* (about revision 31) from former UAX #14 editor Asmus Freytag, concluding that he was "very leery" with respect to L2/11-141, and that he was not inclined to take the proposal "at face value without any actual evidence." (Page 1, bottom lines.)

3.2.1 Backtracking the issue

1. In L2/10-427 *Comments on Public Review Issues (August 3, 2010 - October 27, 2010),* item "Fri Oct 22 11:50:10 CDT 2010" suggested to solve a problem with EM DASH in Spanish.
2. Out of that, UTC #125 defined an Action Item 125-A99 about "the general issue of breaks after dashes".

3. That resulted also in L2/11-141 *Segmentation & Linebreak* according to the Recently-closed-action-items list L2/12-053. But in L2/11-141R, the issue with EM DASH is only item #4. I'm unable to backtrack items #1..#3 about hyphen in Hebrew and Finnish.
4. At meeting #127, UTC took action only on item #6 about Japanese (127-A23), while taking no action on the other items.

I didn't find any record of further decisions.

3.3 Rule implementation

The industry seems to actually implement LB21a. For example, here Word Online makes *maqaf* behave like a non-breaking hyphen, and HYPHEN, EM SPACE, THIN SPACE all are non-breaking after Hebrew letters. As soon as I start typing Latin letters before any of these characters (HEBREW PUNCTUATION MAQAF, HYPHEN, EM SPACE, THIN SPACE), that one becomes breaking.

That scheme and behavior makes of course for **a very disruptive user experience,** one way or the other way, and does in my opinion not deserve support from Unicode.

3.4 Rule context

Losing a breaking *maqaf* for a non-breaking one is not so much of an issue in Hebrew, where most hyphenated words have a short side; and making maqaf breaking again takes only a setting in TeX [1].

Obviously in Hebrew, the advantage of having a full set of non-breaking fixed-width spaces far outweighs the loss of breaking hyphens. That insight probably triggered a lobbying effort whose visible part are the "Hebrew experts at Apple" mentioned as initiators by the author of L2/13-083 *Feedback on the Proposed Update for UAX#14*, who had been informed that "the origin of this was some info from Hebrew experts at Apple" and suspected that the intended target was not Hebrew followed by Non-Hebrew.

It is probably safe to assume that complaints from Hebrew communities were not about breaking *maqaf*, as that would have been fixed by changing the *maqaf*'s line breaking class assignment only. Rather, these complaints seem to focus on breaking fixed-width spaces.

Despite its widespread use, Hebrew is considered a "minority language". This status of a very important minority language obviously prompted Apple to pay attention and to lend an ear to the feedback, that was picked up by Apple's representative to Unicode, co-author of L2/11-141.

The result could be called a Hebrew-specific default tailoring of breaking spaces to non-breaking spaces, as a way of fixing—for a single script—the issue about the wrong line breaking class assignment of BA instead of GL to U+2002..2006 and U+2008..U+200A. About this issue, please see L2/19-115 *Proposal to ensure usability of fixed-width spaces.* (As shown in the simultaneously submitted *Proposal to adjust space characters in UAX #14,* the line breaking class assignment to U+2000..U+2001 was wrong too, as it should have been SP, not BA.)

References

[1] Morris Roger, David Purton, *“hyphenation” in Biblical Hebrew*, TeX Stackexchange, 2018-08-22..2018-08-23, <https://tex.stackexchange.com/questions/447194/hyphenation-in-biblical-hebrew>

[2] Louis Guéry, *Dictionnaire des règles typographiques,* 5th edition, ediSens, Paris, 2019, ISBN 978-78235113-357-6.