L2/02-426
SC2/WG2 N2531

Note on soft hyphens in 10646/Unicode

Document Type: Working Group Document
Source: Kent Karlsson
Status: Expert Contribution
Date: 2002-11-27

1 Chart glyphs for soft hyphens

The chart glyphs for *all three* soft hyphens allocated so far (SOFT HYPHEN (00AD), ARMENIAN HYPHEN (058A), and MONGOLIAN TODO SOFT HYPHEN (1806)) should have a dashed box with suitable indication inside.

Side note: It is possible that one should allocate no-break hyphens for Armenian and Mongolian, see below, but this document does not suggest that.

2 Automatic hyphenation vs. manual intervention with automatic hyphenation

Automatic hyphenators (as well as higher level protocols for indication hyphenation points) should have at least three levels of preferredness for hyphenation points: word boundary (highest priority), morpheme boundary (within a word, middle priority), and syllable boundary (within a morpheme, lowest priority). This may be indicated by the author in a separate table, giving such priorities in a higher-level protocol, or in a dictionary with hyphenation indications for the assumed or declared language of the text, as well as heuristics.

Unfortunately, using system provided dictionaries and heuristics may sometimes lead to erroneous hyphenations, and the set of failures may vary between systems.

However, for interpreting the soft hyphens, which are most likely to be inserted manually at the points of failure of the automatic hyphenator, there is an ABSOLUTE need for determinism. There must be no dependence on dictionaries or heuristics in the interpretation of soft hyphens, precisely because soft hyphens are most likely to occur at the points where the dictionaries and heuristics failed. In the suggested new text on soft hyphens given below, default deterministic rules are given for soft hyphens, and they are based on the actual spell change rules that are used for composite words. Trying to "reverse" them would lead to non-determinism that cannot be reliably resolved into determinism even by the best dictionaries or heuristics.

The rules for soft hyphens should be dealt with by rendering software libraries, deleting the soft hyphen just before display (not in the source text) and doing spell changes (for the display, not in the source text), both according to the default rules below or some tailoring of them for a particular orthography. Fonts should have the appropriate hyphen glyph for the soft hyphen characters.

Automatic hyphenators may still be applied to the word parts adjoined by explicit soft hyphens, but should then only generate hyphenation points of low or middle priority, not high priority.

In contrast, the US suggestion (in their comments to the second amendment to 10646-1:2000) of letting soft hyphens always be invisible when not causing a line break, and then incur no spell changes, while all spell changes would be done when soft hyphens cause a line break, has a number of serious problems, that are unsolvable in a deterministic fashion without the use of a higher level protocol:

a. It would make it impossible, without the use of the private use area or new character allocations, to display Armenian and Mongolian soft hyphens.

b. It would make the insertion of extra letters non-deterministic even with the help of the best dictionaries, since the insertion or not of an extra letter can change the meaning (between two legitimate meanings) of a word.

c. It would make the uppercasing of a letter after the soft hyphen not only language dependent (German), but also non-deterministic even with the help of the best dictionaries, since it may still not be able to determine correctly whether (a prefix of) the rest of the word is a noun or not.

It is therefore *imperative* that the rules for soft hyphens in plain text are deterministic, given a language/orthography. There should also be good *default* rules in case the language/orthography is wrong or unknown. It would be to go too far to have the language/orthography specific exceptions listed in 10646, but the default rules should be given. For almost all texts, the default rules (given below) will be sufficient.

As shown in this paper (see the suggested text on soft hyphens below), all SHY exceptions to “just invisible vs. just hyphen plus line break” that are known to the author can be solved reliably without a higher level protocol about hyphenation, but possibly with information about which orthography, though the latter information is rarely needed.

The interpretation of the SOFT HYPHEN in programs currently varies greatly. Some programs always display it as a hyphen, some let it be invisible, but never let it cause a line break, even though the text is otherwise automatically line broken, some do allow the line to get broken at a SHY then showing it while it is otherwise invisible, and some programs completely misbehave, e.g. replacing the soft hyphen with a hard line break (no hyphen at all). Only with a good and clear as possible specification on how soft hyphens should be interpreted is there any hope to get them to work properly in a wider range of applications. While doing so, it is important to make soft hyphens work well also for Swedish, German and other languages where composite words are common, not just for English, as well as being able to write texts about hyphenation in Armenian and Mongolian. The text below does not, however, specify how soft hyphens should be interpreted if markup intervenes, e.g. *<italic>kräpp-</italic>papper*, where - here is a soft hyphen. It is in such cases probably best to render it visibly.

The US proposed new text about soft hyphen also misrepresents the ARMENIAN and MONGOLIAN soft hyphens as possible renderings of the SOFT HYPHEN, whilst they are actually soft hyphens themselves. The SOFT HYPHEN should be rendered as a HYPHEN (and NO-BREAK HYPHEN) when visible. The glyphs for ARMENIAN HYPHEN and MONGOLIAN TODO SOFT HYPHEN are different, while there are no corresponding non-soft hyphen characters for them.

The suggested text below, on the other hand, is in line with and clarifies the current text on soft hyphens in 10646.

3 Suggested new text on soft hyphens for annex F of 10646

Side remark: The typographic styles used in annexes F and P should be made the same.

SOFT HYPHEN (00AD) (SHY), ARMENIAN HYPHEN (058A) (ASHY), and MONGOLIAN TODO SOFT HYPHEN (1806) (MSHY): These characters are format characters, as well as graphic characters, that indicate preferred intra-word hyphenation and line-break points in the text. SHY is used for the Latin, Greek, and Cyrillic scripts. Note that an automatic hyphenator should use at least three levels of preferredness. An explicit soft hyphen should be interpreted as a high priority hyphenation point (just as HYPHEN should be), even if that is orthographically incorrect.

If the line *is not* broken at the point indicated by the soft hyphen character, the soft hyphen is often not rendered (see the rules below). The soft hyphen may then also suppress or change to lowercase the following character. The default rules for display and spell changes for non-line-breaking occurrences of soft hyphens are as follows:

Firstly try to match (rare):
*x*x<SHY>*x* → *xx*
*x*x<SHY>*X* → *xx*
*X*X<SHY>*X* → *XX*

Secondly try to match (common):
x<SHY>*z* → *xz*
x<SHY>*Z* → *xz*
X<SHY>*z* → *Xz*
X<SHY>*Z* → *XZ*
a<ASHY>*b* → *ab*
m<MSHY>*n* → *mn*

Otherwise (less common): Keep the soft hyphen and display it with a visible glyph.

In these rules, *X* and *Z* are uppercase Latin/Greek/Cyrillic letters without a compatibility decomposition in Unicode, *x* and *z* are the respective corresponding lowercase letters; for *x* and *X*, “letters” here also includes combining characters from the 03xx-block with the case being derived from the preceding base letter (if any); *a* and *b* are Armenian letters, and *m* and *n* are Mongolian letters (for *m*, including Mongolian variation selectors and U+18A9).

Note that the "otherwise" case is the only reliable way to give an explicitly rendered hyphen, e.g. in texts about hyphenation, for the Armenian and Mongolian hyphens.

If the line *is* broken at the point indicated, the soft hyphen is rendered visibly; before the line break in the case of SHY and ASHY, after the line break in the case of MSHY. By default there is *no* spell change if the line is broken.

For some orthographies the default rules may need to be modified slightly, as long as no dictionary or heuristics dependency is introduced in the interpretation of explicit soft hyphens. E.g., for "old" orthography German, for the particular case of ‘ck’ the rules for the line break case would be augmented by c<SHY>k -> k-/k and C<SHY>K -> K-/K, where / indicates the line break, but otherwise stay with the default of no spell change and visible rendering before the line break.

For example *biljett*<SHY>*tång* is rendered either as *biljettång* or *biljett-*/*tång*, *Schiff*<SHY>*Fahrt* either as *Schiffahrt* or *Schiff-*/*Fahrt*, *Sauerstof*<ZWJ>*f*<SHY>*Flasche* is rendered either as *Sauerstoffflasche* or as *Sauerstoff-*/*Flasche*, and for "old" orthography German *Suc*<SHY>*ker* should be rendered either as *Sucker* or *Suk-*/*ker*.