Subject: Re: [tei-council] Hyphenation discussion
From: Kevin Hawkins <kevin.s.hawkins@ultraslavonic.info>
Date: Sun, 16 Jan 2011 12:21:34 -0500
To: TEI Council <tei-council@lists.village.Virginia.EDU>

This is excellent. A few notes are made in red below.

I like your plan of action but also would very much like to see the note in the definition of lb clarified regarding use of inWord, noBreak, and mayBreak. Could this note also be added to the definitions of <cb/> and <pb/>?

I also notice that you wrote "noBreak" below, but the current note in the definition of lb says "nobreak". Do we need to keep "nobreak" for backwards-compatibility?

—Kevin

On 1/15/2011 12:54 PM, Lou Burnard wrote:

> ```
> <!--
>
> As promised earlier, I have now written an extended discussion of
> the issues around hyphenation, which I present for Council's
> consideration below. I am less sure than I was that the best place to
> insert this in the Guidelines would be the current section 3.2
> (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#COPU)
> since that section is really just a summary, but I can't think of a
> better place. If this text is felt to be useful, I think I would
>
> (a) delete the third sentence of 3.2 ("Thus, for example,
> different...concerned") since it duplicates what is said in my new
> section;
>
> (b) Introduce a new subsection titled "Functions of punctuation"
> following the second para "We discuss some typical cases below".
> This subsection would contain the remainder of the existing 3.2.
>
> (c) add the following new subsection following it.
>
> Other suggestions would be welcome however.
>
> If people are happy with this text I will also go ahead and modify the
> existing discussion of <lb/> to be consistent with what is said here.
>
> -->
>
> <div ><head>Hyphenation</head>
>
> <p>Hyphenation as a phenomenon is generally of most concern when
> producing formatted text for display in print or on screen: different
> languages and systems have developed quite sophisticated sets of rules
> about where hyphens may be introduced and for what reason. These
> generally do not concern the text encoder, since they belong to the
> domain of formatting and will generally be handled by the rendition
> ```

```
software in use. In this section, we discuss issues arising from the
appearance of hyphens in pre-existing formatted texts which are being
re-encoded for analysis or other processing. Unicode distinguishes
three visually similar characters for the hyphen, although it also
retains the undifferentiated hyphen-minus (U+002D) for compatibility
reasons. The hard hyphen (U+2010) is distinguished from the minus sign
(U+2212) which should be used only in mathematical expressions, and
also from the soft hyphen (U+00AD) which may appear in <soCalled>born
digital</soCalled> documents to indicate places where it is acceptable
to insert a hyphen when the document is formatted. </p>

<p> Historically, the hard hyphen has been used in printed or
manuscript documents for two distinct purposes. In many languages, it
is used between words to show that they function as a single syntactic
or lexical unit. For example, in French, <mentioned>est-ce
que</mentioned>; in English <mentioned>body-snatcher</mentioned>,
<mentioned>tea-party</mentioned> etc. It may also have an important
role in disambiguation (for example, by distinguishing say a
<mentioned>man-eating fish</mentioned> from a <mentioned>man eating
fish</mentioned>). Such usages, although possibly problematic when a
linguistic analysis is undertaken, are not generally of concern to
text encoders: the hyphen character is usually retained in the text,
because it may be regarded as part of the way a compound or other
lexical item is spelled. Deciding whether a compound is to be
decomposed into its constituent parts, and if so how, is a different
question, involving consideration of many other phenomena in addition
to the simple presence of a hyphen. </p>

<p> When it appears at the end of a printed or written line however,
the hard hyphen generally indicates that — contrary to what might be
expected — a word is not yet complete, but continues on the next line
(or over the next page or column or other boundary). The hyphen
character is not, in this case, part of the word, but just a signal
that the word continues over the break. Unfortunately, few languages
distinguish these two cases visually, which necessarily poses a
problem for text encoders. Suppose, for example, that we wish to
investigate a diachronic English corpus for occurrences of "tea-pot"
and "teapot", to find evidence for the point at which this compound
becomes lexicalized. Any case where the word is hyphenated across a
linebreak, like this: <eg><![CDATA[tea-
pot]]></eg> is entirely ambigous: there is simply no way of deciding
which of the two spellings was intended.
</p>

<p>As elsewhere, therefore, the encoder has a range of choices:
<list>
<item>They
may decide simply to remove any end-of-line hyphenation from the
encoded text, on the grounds that its presence is purely a secondary
matter of formatting. This will obviously apply also if line endings
are themselves regarded as unimportant.</item>
<item>Alternatively, they may decide to record the presence of the
hyphen, perhaps on the grounds that it provides useful morphological
information; perhaps in order to retain information about the visual
appearance of the original source. In either case, they need to decide
whether to record it explicitly, by including an appropriate punctuation
character in the encoding [Can we be more explicit than "in the encoding"? Perhaps
"in the text data"?], or implicitly by supplying an appropriate
attribute value on the <gi>lb</gi> element used to record the fact of
the line division. Use of the <att>type</att> attribute to provide morphological
```

information is discussed below, but the rend attribute may be used instead if visual appearance of the hyphen is more important.</item>
</list>
A similar range of possibilities applies equally to the representation of other common punctuation marks, notably quotation marks, as discussed in <ptr target="#COHQQ"/>.</p>

<p> The <soCalled>text data</soCalled> of which XML documents are

composed is decomposable into smaller units, here called <term>orthographic tokens</term>, even if those units are not explicitly indicated by the XML markup. The ambiguity of the end-of-line hyphen also causes problems in the way a processor identifies such tokens in the absence of explicit markup. If token boundaries are not explicitly marked (for example using the

<gi>seg</gi> or <gi>w</gi> elements), ~~in~~ for most languages a processor will rely on character class information to determine where they are to be found: some punctuation characters are considered to be word-breaking, while others are not. In XML, the newline character in text data is a kind of white space, and is therefore word breaking. XML mixed-content rules are notoriously confusing on this issue. [The previous sentence feels misplaced, and it's not clear which issue you're discussing (whether the newline character is word-breaking?)] However, it is generally unsafe to assume that whitespace adjacent to markup tags will always be preserved, and it is decidedly unsafe to assume that markup tags themselves are equivalent to whitespace. </p>

<p> The <gi>lb</gi>, <gi>pb</gi>, and <gi>cb</gi> elements are notable exceptions to this general rule, since their function is precisely to represent (or replace) line, page, or column breaks, which, as noted above, are generally considered to be equivalent to white space. These elements provide a more reliable way of preserving the lineation, pagination, etc of a source document, since the encoder should not assume that (untagged) line breaks etc. in an XML source file will necessarily be preserved. </p>

<p>In cases where the <gi>lb</gi> element does not in fact correspond with a token boundary, the <att>type</att> attribute should be given a special value to indicate that this is a "non-breaking" line break. The values proposed by these Guidelines are <val>noBreak</val> or (for compatibility with existing recommendations) <val>inWord</val>. A value <val>mayBreak</val> is also available, for cases where the encoder does not wish (or is unable) to determine whether the orthographic token concerned is broken by the line ending or not.</p>

<p>As a final complication, it should be noted that in some languages, particularly German and Dutch, the spelling of a word may be altered in the presence of end of line hyphenation. For example, in Dutch, the word <mentioned>opaatje</mentioned> (<gloss>granddad</gloss>), occurring at the end of a line may be hyphenated as <mentioned>opa-tje</mentioned>, with a single letter a. An encoder wishing to preserve the original form of this orthographic token in a printed text while at the same time facilitating its recognition as the word <mentioned>opaatje</mentioned> will therefore need to rely on a more sophisticated process than simply removing the hyphen. This is however essentially the same as any other form of normalization

```
accompanying the recognition of variations in spelling or morphology:
as such it may be encoded using the <gi>choice</gi> element discussed
in <ptr target="#COED"/>, or the more sophisticated mechanisms for
linguistic analysis discussed in chapter <ptr target="#AI"/>.
</p>
</div>


_______________________________________________
tei-council mailing list
tei-council@lists.village.Virginia.EDU
http://lists.village.Virginia.EDU/mailman/listinfo/tei-council
```