L2/03-298

From: "Mark Davis" <mark.davis@jtcsv.com>
Date: 2003-08-25 18:13:04 -0700
To: "Rick McGowan" <rick@unicode.org>
Subject: Fw: UTS #18, U+1361 Comments

Rick, can you make this a doc for the meeting.

Mark

_____
http://www.macchiato.com
  ªEppur si muoveº

----- Original Message -----
From: "Daniel Yacob" <locales@geez.org>
To: <jarkko.hietaniemi@nokia.com>; <kenw@sybase.com>; <locales@geez.org>;
<mark.davis@jtcsv.com>; <rick@unicode.org>
Cc: <magda@unicode.org>
Sent: Monday, August 25, 2003 21:51
Subject: UTS #18, U+1361 Comments

> Greetings,
>
> Since Mark suggested doing a write up U+1361 with respect to TRs 14 and 29
> I've been unable to address it during the summer. I should be able to do so
> in September and collect input from colleagues. The following then is a less
> detailed summary with a focus on applicable character class properties.
>
>
> U+1361 is the building block for all other Ethiopic punctuation and is
> itself a printed space character that in typesetting will have spacing rules
> applied to it. U+1361 is also an overloaded character which naturally leads
> to confusion over its use. In part its use is in the locale of the beholder.
>
>
> Context:
>
>   Word Separator (\s, \p{Zs}) : default
>   General Punctuation (\p{Po}): in *_ER locales
>   Joining Punctuation (\p{Pc}): software source files (java,perl,xml,etc)
>
>
> Some Other UnicodeData.txt comments:
>
> 1) U+1368 should be Zp, unless Z characters must also be whitespace. Some
>   caution as a paragraph separator is required. Multiple U+1368 symbols
>   on the same line indicate only a single collective paragraph break.
>
> 2) Labeling the Ethiopic digits as Nd can be misleading. They are only
>   radix-10 like in isolation and can not be combined as Section 4.6 of
>   the Unicode Handbook describes. That is U+136aU+1369 is undefined,
>   it is not the value 21. So field 6 of UnicodeData.txt should be left
>   empty for the digits.
>
>
> The following is some expansion on U+1361 use. I'm omitting most of what
> could be said for its type setting properties since the context here is its

> character class. I hope this will help illuminate properties that I have
> overlooked, the idea of a "MidLetter" is new to me, I don't believe U+1361
> qualifies as such.
>
>
> Use as Word Separator
>
> This is original and expected use of the punctuation. Since the 1940s the
> U+0020 has taken over in typeset materials (U+1361 is still the default
> space for anything hand writeen -pen or pencil + paper). Religious works,
> in any language using Ethiopic script, continue to use U+1361. In modern use
> U+1361 is more common in poetry and in a quoted passage in a document that is
> otherwise using U+0020.
>
> It is important that software treat U+1361 with whitespace rules (and
> treat the left and right side kerning whitespace as elastic). It is an
> annoyance when a word processor treats an entire sentence as a single long
> string. Line breaks should be allowed to the right of the character.
>
>
> Use as Comma
>
> In modern Eritrean conventions U+1361 will be used in place of Ethiopic
> Comma (U+1363). I've noticed this convention just beginning to creep into
> modern Ethiopian Tigrigna through the work of a single Ethiopian poet. I
> have yet to see the practice go further, honestly I hope it doesn't. A
> space character will also follow U+1361 under this usage. Eritrean religious
> materials will still use U+1361 as a word separator and embedded examples
> can still be expected.
>
>
> Lazy Use (Misuse) of U+1361
>
> There is a lazy use of U+1361 that misrepresents the accepted use of the
> character. Lazy use will be found in electronically composed documents
> where the author does not want to change keyboard drivers into English
> and back. This leads to U+1361 used for ASCII colon, for example in time
> formats. Also it may be the case that the font system used does not provide
> (or the corresponding keyboard system makes difficult to compose) U+1362
> and/or U+1366 and so these letters are composed with two U+1361 symbols
> and U+1361 with hyphen respectively.
>
>
> Use as a Joining Punctuation
>
> Since bringing up the issue on the perl-unicode list I've kept my eye
> out for U+1361 used as a joining character. I have yet to find a clearly
> unambiguous example.
>
> When one would think to put a printed space between two Ethiopic words
> U+1361 is the natural choice, analogous to _ (U+005F, underscore) for
> Roman-based script which looks out of place, even visually confusing,
> when used with Ethiopic. This is the best argument I can make at this
> time for U+1361 to be used as a joining character for strings.
>
> I think the use is as valid as "_" in the same context where "_" is used.
> Outside of a software related context I can not find "_" used in English.

> Maybe I haven't thought long and hard enough, but I'm just can't think
> of any examples. "_" as memory serves is a hold over from type writers
> which used the character to create underlined text -now done graphically
> by word processors. It is now used as a type of false-space to avoid
> parsing a single string as two or more strings. In these software
> contexts where "_" has the word joining property, so should U+1361.
>
>
> /Daniel
>
>
> ==========
>
> From: "Mark Davis" <mark@macchiato.com>
> To: <mark.davis@macchiato.com>,
> "Daniel Yacob" <unicode@geez.org>
> Subject: Re: UnicodeData Tweak for U+1368
> Date: Fri, 11 Apr 2003 21:21:40 -0700
>
> There is no chance that the General_Category property would become a list;
that
> would break compatibility.
>
> What the UTC has done in some cases is to add additional properties to reflect
> attributes that cannot be captured in the GC property. These tend to be based
on
> the particular algorithm involved. For example, linebreak has a much different
> organization of properties.
>
> Perhaps you can start by describing the different contexts that U+1361 can be
> used in, and what its behavior is as far as wordbreak, linebreak, etc. You
might
> also reference the following in particular.
>
> http://www.unicode.org/reports/tr14/tr14-13.html
> and
> http://www.unicode.org/reports/tr29/tr29-4d4.html
>
> Mark
>
> ----- Original Message -----
> From: "Daniel Yacob" <unicode@geez.org>
> To: <mark.davis@macchiato.com>
> Cc: <unicode@geez.org>
> Sent: 2003 Apr 11 Fri 16:58
> Subject: UnicodeData Tweak for U+1368
>
>
> >
> > Greetings Dr Davis,
> >
> > I was just chasing a problem into the UnicodeData.txt file and
> > noticed that U+1368 was marked with a "Po" while a "Zp" definition
> > might be more suitable. I'm assuming "Zp" had not been defined
> > at the time at the time Ethiopic was proposed for Unicode. Unless
> > the "Z" class is implied to be a printed type.
> >

> > The problem I was investigating was why Perl 5.8.0 would not
> > accept U+1361 in place of underscore (U+005F) in variable names.
> > U+1361 may be processed as a Zs, Pc, Pd or Po depending on the
> > context. The multi-context issue is likely common for a good
> > number of punctuation characters.
> >
> > Has the suggestion come up (and been shot down) previously that
> > the "General Category" field become a prioritized list of classes?
> > Using my notes on U+1361 as an example the Unicode Data definition
> > might become:
> >
> > 1361;ETHIOPIC WORDSPACE;Zs,Pd,Pc,Po;0;L;;;;;N;;;;;
> >
> > thanks,
> >
> > /Daniel
> >
>