L2/03-279

Subject: UTC Agenda & doc: Ethiopic Word Space U+1361
From: "Mark Davis" <mark.davis@jtcsv.com>
Date: Fri, 22 Aug 2003 07:26:52 -0700
To: <l2doc@unicode.org>
CC: <lisa@unicode.org>, "Cathy Wissink" <cwissink@microsoft.com>

Please enter the following as a document and on the Agenda.
==============

These are forwarded messages on the issue of U+1361. My take is that it should be added to MidLetter in http://www.unicode.org/reports/tr29/#Word_Boundaries, but I haven't heard any more from Daniel.

Mark

----- Forwarded message from Daniel Yacob <perl@geez.org> -----

Subject: Change Request for U+1361 Handling
From: Daniel Yacob <perl@geez.org>
Date: Fri, 11 Apr 2003 15:10:18 -0400
Message-Id: <E1943uk-0001X8-00@geez.org>
To: perl-unicode@perl.org

Greetings,

Often when creating variable and method names the underscore char is used as a word separator (->get_something). It is a natural choice because it is visually similar to a space.

I found out today that the Ethiopic word space (U+1361) can not be used in name formation. InterCaps is not an option since, well, Ethiopic doesn't have caps. Using underscore suddenly became visually odd, mentally disruptive even. No other printable char would be more natural than U+1361 as a word seperator in an Ethiopic sequence (U+1361 has no other purpose).

So... I don't know what the issues are, but can U+1361 be added to the list of legal chars for variable, function, etc names. That is, may it be treated under the same rules that apply to _ ?

thanks!

/Daniel

----- End forwarded message -----

--
Jarkko Hietaniemi <jhi@iki.fi> http://www.iki.fi/jhi/ "There is this special biologist word we use for 'stable'. It is 'dead'." -- Jack Cohen

----- Forwarded message from Jarkko Hietaniemi <jhi@iki.fi> -----

Subject: Re: Change Request for U+1361 Handling
From: Jarkko Hietaniemi <jhi@iki.fi>
Date: Sat, 12 Apr 2003 01:09:17 +0300
Message-ID: <20030411220917.GA5689@vipunen.hut.fi>
To: Daniel Yacob <perl@geez.org>
Cc: perl-unicode@perl.org
Reply-To: jhi@iki.fi
In-Reply-To: <E1943uk-0001X8-00@geez.org>
User-Agent: Mutt/1.4i

> So... I don't know what the issues are, but can U+1361 be added to the list of legal chars for variable, function, etc names. That is, may it be treated under the same rules that apply to _ ?

I don't think making exceptions like that is a good path to start down on. Currently the U+1361 is classified as Punctuation. While I do understand the Ethiopian word space is a bit special since it is a word separator, I am afraid there are plenty of punctuation characters other people might argue would be useful to allow in variable names. If you think the U+1361 should be part of the the set of "word characters" (characters eligible for variable names and the like), I propose you contact the Unicode consortium. (After all, that way the issue would get fixed also for other software, not just Perl.)

--
Jarkko Hietaniemi <jhi@iki.fi> http://www.iki.fi/jhi/ "There is this special
biologist word we use for 'stable'. It is 'dead'." -- Jack Cohen

----- End forwarded message -----

----- Original Message -----
From: "Daniel Yacob" <locales@geez.org>
To: <mark.davis@jtcsv.com>
Cc: <locales@geez.org>
Sent: Thursday, April 17, 2003 17:49
Subject: Re: Posix-Style Properties

Dr Davis,

Firstly, thank you for pointing me to the TRs concerning word and linebreak properties. I will be looking through them wrt U+1361 in the near future (probably a summer project).

The ICU support for Unicode regular expressions is something that I have had a lot of interest in but no time to investigate, I hope to sometime this year. I've experimented with extending POSIX style character classes for syllabaries since 1997, its gotten to the point where the utility of the extensions is so great now that I fully depend on them and there is no turning back for me. Naturally I think that they would be invaluable to other people working daily with syllabaries.

In some form I'd to eventually incorporate them into ICU and into any developing RE standards for Unicode based character classes. I'll need to print and review related TRs as well. I can imagine you shaking your head as you read this :) if you do however think syllabic classes can become a part of Unicode REs in this area, I can make some effort to write some documentation in the style of the applicable TRs.

If you use Perl, I have a package that overrides the Perl RE interpreter to implement the classes for Ethiopic. It provides some example usage, two other pacakages at CPAN now depend on this package. In a few months I expect to extend the package for Canadian Aboriginal Sylalbics, Cherokee, Hiragan, Katakana and Yi.

Some material:

Overview:

http://syllabary.sourceforge.net/Articles/PatternMatching/

Perl 5.8.x Package:

http://search.cpan.org/author/DYACOB/Regexp-Ethiopic-0.07/

Table data from package (the "Overview" notation is now syncronized with the notation here):

http://search.cpan.org/src/DYACOB/Regexp-Ethiopic-0.07/doc/index.html

cheers,

/Daniel