Subject: UTC Agenda Item: Corrections to Script property values
From: Mark Davis
Date: Thu, 29 Jan 2004 08:57:11 -0800

L2/04-043

Please enter this as a document in the registry and on the agenda.

Note: the document 2003-03427 should be printed for the meeting.

There is an email trail below, but the main items are:

Process:
1. Add a request for script categorization on the standard character proposal forms.
2. Provide update of the status of ISO 15924

For 4.0.1:
3. Change the script of greek san (03FA and 03FB) to be, well, Greek.
4. Adopt most of the changes recommended in http://www.unicode.org/L2/L2003/03427-script-codes.txt. My recommendations are to accept them except the following:

- Arabic accents are used in Syriac also, and should not be assigned to Arabic alone. If we think that the only other script they'll be used in is Syriac, we could have an joint category (Arabic-Syriac)
- SC;0964;DEVANAGARI # DEVANAGARI DANDA
SC;0965;DEVANAGARI # DEVANAGARI DOUBLE DANDA
SC;0970;DEVANAGARI # DEVANAGARI ABBREVIATION SIGN
Are definitely pan-Indic. They must either remain Common, or be changed to a joint category (Indic).
-SC;302A;HAN # IDEOGRAPHIC LEVEL TONE MARK
..SC;302F;HANGUL # HANGUL DOUBLE DOT TONE MARK
Not sure about these, whether they can only really be used with the scripts Michel suggests.
-SC;FD3E;ARABIC # ORNATE LEFT PARENTHESIS
SC;FD3F;ARABIC # ORNATE RIGHT PARENTHESIS
These are definitely not just Arabic -- they are dingbatty parentheses.
- > 02B0;MODIFIER LETTER SMALL H;Lm -- in UCD but not in Latin in Scripts
  (while there are other Lms in Scripts)
I think the modifier letters might be 'pan' script. Ken says they should be Latin.

Ken also suggests that we allow composition in Scripts.txt. I think that might be a breaking change for some people, and that since the combinations are rather limited, we should continue to have a separate code for them: but that we could make the composition clear in the name, so that people could have an orthogonal API if they wish.

Mark

http://www.macchiato.com
► शिष्यादिच्छेत्पराजयम् ◄

----- Original Message -----
From: "Mark Davis" <mark.davisjtcsv.com>
To: "Deborah W. Anderson" <dwanderspacbell.net>; "'Kenneth Whistler'" <kenwsybase.com>
Sent: Fri, 2004 Jan 16 09:52
Subject: Re: Greek Collation questions > new FAQs on collation?

1. This is based on the availalble information. Script proposals should contain
that information but often don't. (actually, I'll put in an agenda item on that)

2. Allkeys can be modified by decision of the UTC. We would need a proposal from
someone (you?) for the next meeting describing the desired changes and reasons.

It takes a while for that to come out in a new version of UCA, so in the meantime tailoring is the way to go.

Mark

_____

http://www.macchiato.com

► शिष्यादिच्छेत्पराजयम् ◄

----- Original Message -----
From: "Deborah W. Anderson" <dwanderspacbell.net>
To: "'Kenneth Whistler'" <kenwsybase.com>
Cc: "'Mark Davis'" <mark.davisjtcsv.com>
Sent: Fri, 2004 Jan 16 09:26
Subject: FW: Greek Collation questions > new FAQs on collation?

Ken,
I didn't hear anything back from Mark or you on this, but I ought to at least make one response to the list I am on re:collation fairly soon. Any feedback would be appreciated. Rick had suggested I contact you two about this.

The questions below do raise a couple of points that aren't specifically addressed in the Collation FAQs, and might be helpful to add to the FAQs the two questions below, though I'll leave this up to you two to decide.

Potential additional FAQs:

1. How is the default collation for a specific script determined? (I assume this is partly based on what a Unicode script proposal offers -- if it does indeed provide a sorting order.)

2. Can the collation order defined in allkeys.txt be changed or modified? What evidence is needed in order to recommend a change in collation (and to whom should the request be addressed)? Or should I instead provide a tailored collation for my own uses?

Debbie
Deborah Anderson
Researcher, Dept. of Linguistics
UC Berkeley

-----Original Message-----
From: Deborah W. Anderson [mailto:dwander@pacbell.net]
Sent: Monday, January 12, 2004 9:40 PM
To: 'Mark Davis'; 'Ken Whistler'
Subject: Greek Collation questions

Dear Mark and Ken,
There have been a couple of posts on another list about Greek collation. I've distilled the comments and questions down to the following (with longer postings for the final question about collation and the koppas). I will forward the answers to the list. Peter Kirk was the person who wrote in with the Unicode "point-of-view."

Mark: The Greek collation chart at http://www.unicode.org/charts/collation/ is still missing GREEK
CAPITAL LETTER SAN and GREEK SMALL LETTER SAN (03FA and 03FB), in case you've forgotten.

1. Why is Greek letter san to be sorted after Bactrian sho (which follows sampi and is after omega)? In the document http://www.tlg.uci.edu/final/san.pdf (page 3), it recommended san be placed after pi. Did it have anything to do with the fact that san and sampi are historically related? "Sampi's usage as a number makes its position after omega reasonable."

[DA: The basic issue seems to be: how is the order determined for allkeys.txt?]

After someone suggested that san be moved in its historical position (after

pi),
> Peter Kirk responded:
>> The alternatives might be to put san with sampi on the doubtful >grounds that sampi is a variant of san; or to put san with sigma on >the grounds that in modern transcription it is usually replaced by >sigma. See http://www.tlg.uci.edu/~opoudjis/unicode/nonattic.html >and http://www.tlg.uci.edu/~opoudjis/unicode/numerals.html.
>
>> Other changes that might be worth considering are "unifying" >digamma and stigma, as well as archaic and modern koppa, as >variants with the same top level collation. See the same >references.

> 2. (From P Kirk) Why are the five Greek small capital letters, U+1D26 to 1D2A in the Phonetic Extensions block, listed as separate at the highest level? "This is most certainly a mistake; but it is a widespread mistake because small capital Latin letters are also separate at the highest level."

> 3. Doesn't sorting archaic koppa [q-shaped koppa] after modern koppa [z-shaped koppa] mean that the letters hold a primary difference to each other and thus will never be equivalent in comparisons?

> [Background to the following posts re: koppas: The z-shaped koppa is used numerically in modern Greek, the q-shaped koppa is the archaic Greek alphabetic letter which can also be used numerically in older Greek texts. Some fonts, including Arial Unicode MS, put the q-koppa in what is now the z-shaped koppa's place.]

> From J. March:
>> I stated earlier that it would be a negative consequence that they > [the archaic and modern koppas] could not be compared as
>> equivalents if they retain a primary difference from each other.
>> But, on the other hand the only real reason that they would ever >need to be treated equivalently is if (as we all fear will be the case) >someone misuses one for the other.
>> But, if those working in modern Greek just use the z-shaped >koppa, and those working in ancient Greek just use the archaic >koppa, and those quoting an ancient Greek text in a modern >Greek commentary use the z-shaped koppa in the body of their >work and the archaic koppa in the quotations or when specifically
>> referring to the archaic koppa then there will never be a need to >perform a search with the expectation that the modern and archaic >koppas will be treated as equals. One will either be looking for >one or the other--never both. Its only when an author uses them >incorrectly that one would want the ability to treat them >equivalently.
>
>> In an academic setting, which will
>> almost always be the case with archaic koppa, I think we have to >demand precision on the part of the author. Though in truth I don't >think it will be that difficult: modern Greek writers can surely >recognize the modern koppa--its their native language after all, >and anyone writing about archaic Greek had better be able to >differentiate an archaic koppa from a z-shaped koppa--they are >completely different shapes.
>
>> ... But I'm leaning
>> more and more toward keeping the koppas the way they are now: >sorted one after the other and with a primary difference between >them--there shouldn't be a need for a search to treat them as >equivalent. And if an author does use them incorrectly, then its >that text that should be corrected. Allowing the koppas to be >compared equivalently tacitly encourages imprecise usage by
>> leaving the boundary between them permanently vague.

> Peter Kirk responded:
>> There is another issue here. ... there are potentially many texts
>> which
>> were encoded, correctly at the time, with modern koppa in place
>> of archaic koppa. As a general Unicode principle, it is not >required that such texts be recoded, and so modern koppa will for >ever be a legal alternative for archaic koppa. While that does not >imply that they must be

collated together, it does suggest that it >would be a good idea.

Also, there are widely distributed fonts which have only one
koppa, including Arial Unicode MS, installed on most Windows >systems, which has an archaic koppa glyph at the modern koppa >code point. At the time these fonts were designed this was not an error. While they are around, writers are likely to use the modern koppa code point for archaic koppa, and readers are likely to be confused if they don't get an expected match. Again, this implies that collation together is a >good idea.

See http://www.tlg.uci.edu/~opoudjis/unicode/numerals.html#koppa
for more details.

I wonder, is there a counter-argument for collating the two
koppas separately, apart from its rather ineffective role in >enforcing a proper
distinction? I note that such enforcement will cause difficulties for
readers but very little for writers; it's a bit like punishing the
victims of a crime rather than the perpetrators.

Debbie
Deborah Anderson
Researcher, Dept. of Linguistics
UC Berkeley

----- Original Message -----
From: "Mark Davis" <mark.davisjtcsv.com>
To: "Deborah W. Anderson" <dwanderspacbell.net>
Cc: <unicoreunicode.org>
Sent: Fri, 2003 Nov 21 15:59
Subject: Re: When will the Greek collation charts be updated with new characters?

That's a good question; as it turns out, the reason is that the collation charts
group by scripts (for the index in the left navbar). And those characters are not marked as Greek in http://www.unicode.org/Public/UNIDATA/Scripts.txt. So they collate correctly, but the chart just doesn't show them.

However, this does show a defect in the Scripts.txt data. Might be a good time for others to scan through that file to see if there are any other outliers.

Mark

_____
http://www.macchiato.com
► शिष्यादिच्छेत्पराजयम् ◄

----- Original Message -----
From: "Deborah W. Anderson" <dwanders@pacbell.net>
To: "'Mark Davis'" <mark.davis@jtcsv.com>
Sent: Fri, 2003 Nov 21 14:55
Subject: When will the Greek collation charts be updated with new characters?

Hi Mark,
I just received a question: When will the Greek collation charts be updated on the Unicode website with the new characters? I see san is missing (03FA and 03FB)

Thanks,
Debbie

Deborah Anderson
Researcher, Dept. of Linguistics
UC Berkeley

----- Original Message -----
From: "Kenneth Whistler" <kenwsybase.com>
To: <mark.davisjtcsv.com>
Cc: <kenwsybase.com>; <michelsumicrosoft.com>; <jarkko.hietanieminokia.com>
Sent: Thu, 2003 Dec 04 18:30
Subject: Re: When will the Greek collation charts be updated with new characters?

> Mark,
>
> > Michel has made a proposal for changing some of the assignments (and also introducing a Kana script value for characters used in both Hiragana and Katakana, which I favor).
>
> Kana is a good idea, but as you note below, there are other instances of multiple scripts associated with a character.
>
> I think the right thing to do here is to bite the bullet and define, for Scripts.txt, a script composition method -- basically some moral equivalent of a set unioning nomenclature. So for U+30FC you would get:
>
> 30FC;HIRAGANA+KATAKANA # KATAKANA-HIRAGANA PROLONGED SOUND MARK
>
> or something similar. Then implementations could do as they please in terms of applying category labels to such unions of usage of scripts, rather than having Scripts.txt be responsible for coming up with a new category to label every possible combination which turns up.
>
> This is the natural way to deal with usage of accents in Arabic and Syriac, or pan-Indic punctuation, and so on. And it gives a simpler mechanism for people to tweak usage when they find local anomalies in usage, instead of us giving them precooked sets of categories.
>
> > We will be reviewing this in the next meeting (we didn't have time in the last one). It looks like it would be productive for us all to compare notes about what should be done to come up with a unified approach, before the meeting. I'm including a copy of Michel's document below.
> >
> > The main constraining issue is that characters that can be used (normally) in multiple scripts should not have a specific script assignment (unless we do something like the Kana approach).
> >
> > My initial take on Michel's list is:
> > - Arabic accents are used in Syriac also, and should not be assigned to Arabic alone. If we think that the only other script they'll be used in is Syriac, we could have an joint category (Arabic-Syriac)
> > - SC;0964;DEVANAGARI # DEVANAGARI DANDA
> > SC;0965;DEVANAGARI # DEVANAGARI DOUBLE DANDA
> > SC;0970;DEVANAGARI # DEVANAGARI ABBREVIATION SIGN
> > Are definitely pan-Indic. They must either remain Common, or be changed to a joint category (Indic).
> > -SC;302A;HAN # IDEOGRAPHIC LEVEL TONE MARK
> > ..SC;302F;HANGUL # HANGUL DOUBLE DOT TONE MARK
> > Not sure about these, whether they can only really be used with the scripts Michel suggests.
> >
> > - > 02B0;MODIFIER LETTER SMALL H;Lm -- in UCD but not in Latin in Scripts
> > (while there are other Lms in Scripts)
> > I think the modifier letters might be 'pan' script. Ken?
>
> No, they are basically Latin. Even the Greek and Cyrillic modifier letters are just adaptations of those letters for use with predominantly Latin transcription systems. And that would be the more useful designation for them, if any designation would be useful at all in this regard.

--Ken