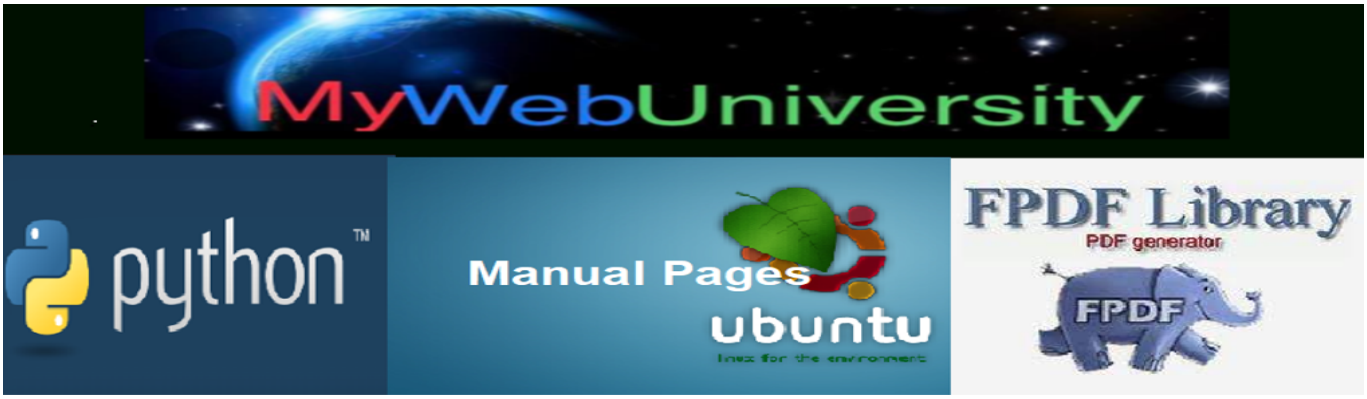

Rocky Enterprise Linux 9.2 Manual Pages on command 'Unicode::Collate::Locale.3perl'

$ man Unicode::Collate::Locale.3perl

Unicode::Collate::Locale(3perl) Perl Programmers Reference Guide Unicode::Collate::Locale(3perl)

NAME

Unicode::Collate::Locale - Linguistic tailoring for DUCET via Unicode::Collate

SYNOPSIS

```
use Unicode::Collate::Locale;
#construct
$Collator = Unicode::Collate::Locale->
    new(locale => $locale_name, %tailoring);
#sort
@sorted = $Collator->sort(@not_sorted);
#compare
$result = $Collator->cmp($a, $b); # returns 1, 0, or -1.
```

Note: Strings in @not_sorted, $a and $b are interpreted according to Perl's Unicode support. See perlunicode, perluniintro, perlunitut, perlunifaq, utf8. Otherwise you can use "preprocess" (cf. "Unicode::Collate") or should decode them before.

DESCRIPTION

This module provides linguistic tailoring for it taking advantage of "Unicode::Collate".

Constructor

The "new" method returns a collator object.

A parameter list for the constructor is a hash, which can include a special key "locale" and its value (case-insensitive) standing for a Unicode base language code (two or three-letter). For example, "Unicode::Collate::Locale->new(locale => 'ES')" returns a collator tailored for Spanish.

$locale_name may be suffixed with a Unicode script code (four-letter), a Unicode region (territory) code, a Unicode language variant code. These codes are case-insensitive, and separated with '_' or '-'. E.g. "en_US" for English in USA, "az_Cyrl" for Azerbaijani in the Cyrillic script, "es_ES_traditional" for Spanish in Spain (Traditional).

If $locale_name is not available, fallback is selected in the following order:

1. language with a variant code
2. language with a script code
3. language with a region code
4. language
5. default

Tailoring tags provided by "Unicode::Collate" are allowed as long as they are not used for "locale" support. Esp. the "table" tag is always untailorable, since it is reserved for DUCET.

However "entry" is allowed, even if it is used for "locale" support, to add or override mappings.

E.g. a collator for Spanish, which ignores diacritics and case difference (i.e. level 1), with reversed case ordering and no normalization.

```
Unicode::Collate::Locale->new(
    level => 1,
    locale => 'es',
    upper_before_lower => 1,
    normalization => undef
)
```

Overriding a behavior already tailored by "locale" is disallowed if such a tailoring is passed to "new()".

```
Unicode::Collate::Locale->new(
    locale => 'da',
    upper_before_lower => 0, # causes error as reserved by 'da'
)
```

However "change()" inherited from "Unicode::Collate" allows such a tailoring that is reserved by "locale". Examples:

```
new(locale => 'fr_ca')->change(backwards => undef)
new(locale => 'da')->change(upper_before_lower => 0)
```

```
new(locale => 'ja')->change(overrideCJK => undef)
```

Methods

"Unicode::Collate::Locale" is a subclass of "Unicode::Collate" and methods other than "new" are inherited from "Unicode::Collate".

Here is a list of additional methods:

"$Collator->getlocale"

Returns a language code accepted and used actually on collation. If linguistic tailoring is not provided for a language code you passed (intensionally for some languages, or due to the incomplete implementation), this method returns a string 'default' meaning no special tailoring.

"$Collator->locale_version"

(Since Unicode::Collate::Locale 0.87) Returns the version number (perhaps "/\d\.\d\d/") of the locale, as that of Locale/*.pl.

Note: Locale/*.pl that a collator uses should be identified by a combination of return values from "getlocale" and "locale_version".

A list of tailorable locales

locale name	description
af	Afrikaans
ar	Arabic
as	Assamese
az	Azerbaijani (Azeri)
be	Belarusian
bn	Bengali
bs	Bosnian (tailored as Croatian)
bs_Cyrl	Bosnian in Cyrillic (tailored as Serbian)
ca	Catalan
cs	Czech
cu	Church Slavic
cy	Welsh
da	Danish
de__phonebook	German (umlaut as 'ae', 'oe', 'ue')
de_AT_phonebook	Austrian German (umlaut primary greater)

dsb	Lower Sorbian
ee	Ewe
eo	Esperanto
es	Spanish
es__traditional	Spanish ('ch' and 'll' as a grapheme)
et	Estonian
fa	Persian
fi	Finnish (v and w are primary equal)
fi__phonebook	Finnish (v and w as separate characters)
fil	Filipino
fo	Faroese
fr_CA	Canadian French
gu	Gujarati
ha	Hausa
haw	Hawaiian
he	Hebrew
hi	Hindi
hr	Croatian
hu	Hungarian
hy	Armenian
ig	Igbo
is	Icelandic
ja	Japanese [1]
kk	Kazakh
kl	Kalaallisut
kn	Kannada
ko	Korean [2]
kok	Konkani
lkt	Lakota
ln	Lingala
lt	Lithuanian
lv	Latvian
mk	Macedonian

ml Malayalam

mr Marathi

mt Maltese

nb Norwegian Bokmal

nn Norwegian Nynorsk

nso Northern Sotho

om Oromo

or Oriya

pa Punjabi

pl Polish

ro Romanian

sa Sanskrit

se Northern Sami

si Sinhala

si__dictionary Sinhala (U+0DA5 = U+0DA2,0DCA,0DA4)

sk Slovak

sl Slovenian

sq Albanian

sr Serbian

sr_Latn Serbian in Latin (tailored as Croatian)

sv Swedish (v and w are primary equal)

sv__reformed Swedish (v and w as separate characters)

ta Tamil

te Telugu

th Thai

tn Tswana

to Tonga

tr Turkish

ug_Cyrl Uyghur in Cyrillic

uk Ukrainian

ur Urdu

vi Vietnamese

vo Volapu"k

wae Walser

wo Wolof

yo Yoruba

zh Chinese

zh__big5han Chinese (ideographs: big5 order)

zh__gb2312han Chinese (ideographs: GB-2312 order)

zh__pinyin Chinese (ideographs: pinyin order) [3]

zh__stroke Chinese (ideographs: stroke order) [3]

zh__zhuyin Chinese (ideographs: zhuyin order) [3]

Locales according to the default UCA rules include am (Amharic) without "[reorder Ethi]", bg (Bulgarian) without "[reorder Cyrl]", chr (Cherokee) without "[reorder Cher]", de (German), en (English), fr (French), ga (Irish), id (Indonesian), it (Italian), ka (Georgian) without "[reorder Geor]", mn (Mongolian) without "[reorder Cyrl Mong]", ms (Malay), nl (Dutch), pt (Portuguese), ru (Russian) without "[reorder Cyrl]", sw (Swahili), zu (Zulu).

Note

[1] ja: Ideographs are sorted in JIS X 0208 order. Fullwidth and halfwidth forms are identical to their regular form. The difference between hiragana and katakana is at the 4th level, the comparison also requires "(variable => 'Non-ignorable')", and then "katakana_before_hiragana" has no effect.

[2] ko: Plenty of ideographs are sorted by their reading. Such an ideograph is primary (level 1) equal to, and secondary (level 2) greater than, the corresponding hangul syllable.

[3] zh__pinyin, zh__stroke and zh__zhuyin: implemented alt='short', where a smaller number of ideographs are tailored.

A list of variant codes and their aliases

variant code alias

---

dictionary dict

phonebook phone phonebk

reformed reform

traditional trad

big5han	big5
gb2312han	gb2312
pinyin	
stroke	
zhuyin	

Note: 'pinyin' is Han in Latin, 'zhuyin' is Han in Bopomofo.

INSTALL

Installation of "Unicode::Collate::Locale" requires Collate/Locale.pm, Collate/Locale/*.pm, Collate/CJK/*.pm and Collate/allkeys.txt. On building, "Unicode::Collate::Locale" doesn't require any of data/*.txt, gendata/*, and mklocale.

Tests for "Unicode::Collate::Locale" are named t/loc_*.t.

CAVEAT

Tailoring is not maximum

Even if a certain letter is tailored, its equivalent would not always tailored as well as it. For example, even though W is tailored, fullwidth W ("U+FF37"), W with acute ("U+1E82"), etc. are not tailored. The result may depend on whether source strings are normalized or not, and whether decomposed or composed. Thus "(normalization => undef)" is less preferred.

Collation reordering is not supported

The order of any groups including scripts is not changed.

Reference

locale	based CLDR or other reference
af	30 = 1.8.1
ar	30 = 28 ("compat" wo [reorder Arab]) = 1.9.0
as	30 = 28 (without [reorder Beng..]) = 23
az	30 = 24 ("standard" wo [reorder Latn Cyrl])
be	30 = 28 (without [reorder Cyrl])
bn	30 = 28 ("standard" wo [reorder Beng..]) = 2.0.1
bs	30 = 28 (type="standard": [import hr])
bs_Cyrl	30 = 28 (type="standard": [import sr])

ca 30 = 23 (alt="proposed" type="standard")

cs 30 = 1.8.1 (type="standard")

cu 34 = 30 (without [reorder Cyrl])

cy 30 = 1.8.1

da 22.1 = 1.8.1 (type="standard")

de__phonebook 30 = 2.0 (type="phonebook")

de_AT_phonebook 30 = 27 (type="phonebook")

dsb 30 = 26

ee 30 = 21

eo 30 = 1.8.1

es 30 = 1.9.0 (type="standard")

es__traditional 30 = 1.8.1 (type="traditional")

et 30 = 26

fa 22.1 = 1.8.1

fi 22.1 = 1.8.1 (type="standard" alt="proposed")

fi__phonebook 22.1 = 1.8.1 (type="phonebook")

fil 30 = 1.9.0 (type="standard") = 1.8.1

fo 22.1 = 1.8.1 (alt="proposed" type="standard")

fr_CA 30 = 1.9.0

gu 30 = 28 ("standard" wo [reorder Gujr..]) = 1.9.0

ha 30 = 1.9.0

haw 30 = 24

he 30 = 28 (without [reorder Hebr]) = 23

hi 30 = 28 (without [reorder Deva..]) = 1.9.0

hr 30 = 28 ("standard" wo [reorder Latn Cyrl]) = 1.9.0

hu 22.1 = 1.8.1 (alt="proposed" type="standard")

hy 30 = 28 (without [reorder Armn]) = 1.8.1

ig 30 = 1.8.1

is 22.1 = 1.8.1 (type="standard")

ja 22.1 = 1.8.1 (type="standard")

kk 30 = 28 (without [reorder Cyrl])

kl 22.1 = 1.8.1 (type="standard")

kn 30 = 28 ("standard" wo [reorder Knda..]) = 1.9.0

ko 22.1 = 1.8.1 (type="standard")

kok 30 = 28 (without [reorder Deva..]) = 1.8.1

lkt 30 = 25

ln 30 = 2.0 (type="standard") = 1.8.1

lt 22.1 = 1.9.0

lv 22.1 = 1.9.0 (type="standard") = 1.8.1

mk 30 = 28 (without [reorder Cyrl])

ml 22.1 = 1.9.0

mr 30 = 28 (without [reorder Deva..]) = 1.8.1

mt 22.1 = 1.9.0

nb 22.1 = 2.0 (type="standard")

nn 22.1 = 2.0 (type="standard")

nso [*] 26 = 1.8.1

om 22.1 = 1.8.1

or 30 = 28 (without [reorder Orya..]) = 1.9.0

pa 22.1 = 1.8.1

pl 30 = 1.8.1

ro 30 = 1.9.0 (type="standard")

sa [*] 1.9.1 = 1.8.1 (type="standard" alt="proposed")

se 22.1 = 1.8.1 (type="standard")

si 30 = 28 ("standard" wo [reorder Sinh..]) = 1.9.0

si__dictionary 30 = 28 ("dictionary" wo [reorder Sinh..]) = 1.9.0

sk 22.1 = 1.9.0 (type="standard")

sl 22.1 = 1.8.1 (type="standard" alt="proposed")

sq 22.1 = 1.8.1 (alt="proposed" type="standard")

sr 30 = 28 (without [reorder Cyrl])

sr_Latn 30 = 28 (type="standard": [import hr])

sv 22.1 = 1.9.0 (type="standard")

sv__reformed 22.1 = 1.8.1 (type="reformed")

ta 22.1 = 1.9.0

te 30 = 28 (without [reorder Telu..]) = 1.9.0

th 22.1 = 22

tn [*] 26 = 1.8.1

```
    to              22.1 = 22
    tr              22.1 = 1.8.1 (type="standard")
    uk               30 = 28 (without [reorder Cyrl])
    ug_Cyrl          https://en.wikipedia.org/wiki/Uyghur_Cyrillic_alphabet
    ur              22.1 = 1.9.0
    vi              22.1 = 1.8.1
    vo               30 = 25
    wae               30 = 2.0
    wo             [*] 1.9.1 = 1.8.1
    yo               30 = 1.8.1
    zh               22.1 = 1.8.1 (type="standard")
    zh__big5han       22.1 = 1.8.1 (type="big5han")
    zh__gb2312han     22.1 = 1.8.1 (type="gb2312han")
    zh__pinyin        22.1 = 2.0   (type='pinyin' alt='short')
    zh__stroke        22.1 = 1.9.1 (type='stroke' alt='short')
    zh__zhuyin        22.1 = 22    (type='zhuyin' alt='short')
  ----------------------------------------------------------------------
```

[*] http://www.unicode.org/repos/cldr/tags/latest/seed/collation/

AUTHOR

The Unicode::Collate::Locale module for perl was written by SADAHIRO Tomoyuki, <SADAHIRO@cpan.org>.

SEE ALSO

Unicode Collation Algorithm - UTS #10

<http://www.unicode.org/reports/tr10/>

The Default Unicode Collation Element Table (DUCET)

<http://www.unicode.org/Public/UCA/latest/allkeys.txt>

Unicode Locale Data Markup Language (LDML) - UTS #35

<http://www.unicode.org/reports/tr35/>

CLDR - Unicode Common Locale Data Repository

<http://cldr.unicode.org/>

Unicode::Collate

Unicode::Normalize