

NAME

perlunicook - cookbookish examples of handling Unicode in Perl

DESCRIPTION

This manpage contains short recipes demonstrating how to handle common Unicode operations in Perl, plus one complete program at the end. Any undeclared variables in individual recipes are assumed to have a previous appropriate value in them.

EXAMPLES

â„ž 0: Standard preamble

Unless otherwise notes, all examples below require this standard preamble to work correctly, with the `#!` adjusted to work on your system:

```
#!/usr/bin/env perl

use utf8;          # so literals and identifiers can be in UTF-8
use v5.12;         # or later to get "unicode_strings" feature
use strict;        # quote strings, declare variables
use warnings;      # on by default
use warnings qw(FATAL utf8); # fatalize encoding glitches
use open          qw(:std :encoding(UTF-8)); # undeclared streams in UTF-8
use charnames qw(:full :short); # unneeded in v5.16
```

This *does* make even Unix programmers `binmode` your binary streams, or open them with `:raw`, but that's the only way to get at them portably anyway.

WARNING: `use autodie` (pre 2.26) and `use open` do not get along with each other.

â„ž 1: Generic Unicode-savvy filter

Always decompose on the way in, then recompose on the way out.

```
use Unicode::Normalize;

while (<>) {
    $_ = NFD($_); # decompose + reorder canonically
    ...
} continue {
    print NFC($_); # recompose (where possible) + reorder canonically
}
```

â„ž 2: Fine-tuning Unicode warnings

As of v5.14, Perl distinguishes three subclasses of UTF-8 warnings.

```
use v5.14; # subwarnings unavailable any earlier
no warnings "nonchar"; # the 66 forbidden non-characters
no warnings "surrogate"; # UTF-16/CESU-8 nonsense
no warnings "non_unicode"; # for codepoints over 0x10_FFFF
```

â„ž 3: Declare source in utf8 for identifiers and literals

Without the all-critical `use utf8` declaration, putting UTF-8 in your literals and identifiers won't work right. If you used the standard preamble just given above, this already happened. If you did, you can do things like this:

```
use utf8;
```

```
my $measure = "Ä...ngstrÄ¶m";
my @î¼soft = qw( cp852 cp1251 cp1252 );
my @á½'ï€î-ï•î¼îµî³î±î, = qw( á½'ï€î-ï• î¼îµî³î±î, );
my @é~% = qw( koi8-f koi8-u koi8-r );
my $motto = "öŸ`a öŸ'— öŸ•a"; # FAMILY, GROWING HEART, DROMEDARY CAMEL
```

If you forget `use utf8`, high bytes will be misunderstood as separate characters, and nothing will work right.

â„ž 4: Characters and their numbers

The `ord` and `chr` functions work transparently on all codepoints, not just on ASCII alone â€” nor in fact, not even just on Unicode alone.

```
# ASCII characters
ord("A")
chr(65)

# characters from the Basic Multilingual Plane
ord("îŒ")
chr(0x3A3)

# beyond the BMP
ord("ö•`>") # MATHEMATICAL ITALIC SMALL N
chr(0x1D45B)

# beyond Unicode! (up to MAXINT)
ord("\x{20_0000}")
chr(0x20_0000)
```

â„ž 5: Unicode literals by character number

In an interpolated literal, whether a double-quoted string or a regex, you may specify a character by its number using the `\x{HHHHHH}` escape.

```
String: "\x{3a3}"
Regex: /\x{3a3}/

String: "\x{1d45b}"
Regex: /\x{1d45b}/

# even non-BMP ranges in regex work fine
/[\x{1D434}-\x{1D467}]/
```

â„ž 6: Get character name by number

```
use charnames ();
my $name = charnames::viacode(0x03A3);
```

â„ž 7: Get character number by name

```
use charnames ();
my $number = charnames::vianame("GREEK CAPITAL LETTER SIGMA");
```

â„ž 8: Unicode named characters

Use the `\N{charname}` notation to get the character by that name for use in interpolated literals (double-quoted strings and regexes). In v5.16, there is an implicit

```
use charnames qw(:full :short);
```

But prior to v5.16, you must be explicit about which set of charnames you want. The `:full` names are the official Unicode character name, alias, or sequence, which all share a namespace.

```
use charnames qw(:full :short latin greek);
```

```
"\N{MATHEMATICAL ITALIC SMALL N}"      # :full
"\N{GREEK CAPITAL LETTER SIGMA}"      # :full
```

Anything else is a Perl-specific convenience abbreviation. Specify one or more scripts by names if you want short names that are script-specific.

```
"\N{Greek:Sigma}"                      # :short
"\N{ae}"                              # latin
"\N{epsilon}"                          # greek
```

The v5.16 release also supports a `:loose` import for loose matching of character names, which works just like loose matching of property names: that is, it disregards case, whitespace, and underscores:

```
"\N{euro sign}"                      # :loose (from v5.16)
```

â„ž 9: Unicode named sequences

These look just like character names but return multiple codepoints. Notice the `%vx` vector-print functionality in `printf`.

```
use charnames qw(:full);
my $seq = "\N{LATIN CAPITAL LETTER A WITH MACRON AND GRAVE}";
printf "U+%v04X\n", $seq;
U+0100.0300
```

â„ž 10: Custom named characters

Use `:alias` to give your own lexically scoped nicknames to existing characters, or even to give unnamed private-use characters useful names.

```
use charnames ":full", ":alias" => {
    ecute => "LATIN SMALL LETTER E WITH ACUTE",
    "APPLE LOGO" => 0xF8FF, # private use character
};

"\N{ecute}"
"\N{APPLE LOGO}"
```

â„ž 11: Names of CJK codepoints

Sinograms like `â€œæ±Œâ€` come back with character names of CJK UNIFIED IDEOGRAPH-6771 and CJK UNIFIED IDEOGRAPH-4EAC, because their `â€œnamesâ€` vary. The CPAN `Unicode::UniHan` module has a large database for decoding these (and a whole lot more), provided you know how to understand its output.

```
# cpan -i Unicode::UniHan
use Unicode::UniHan;
my $str = "æ•±ä°¬";
my $unhan = Unicode::UniHan->new;
for my $lang (qw(Mandarin Cantonese Korean JapaneseOn JapaneseKun)) {
    printf "CJK $str in %-12s is ", $lang;
    say $unhan->$lang($str);
}
```

prints:

```
CJK æ•±ä°¬ in Mandarin      is DONG1JING1
CJK æ•±ä°¬ in Cantonese    is dunglging1
CJK æ•±ä°¬ in Korean       is TONGKYENG
CJK æ•±ä°¬ in JapaneseOn   is TOUKYOU KEI KIN
CJK æ•±ä°¬ in JapaneseKun  is HIGASHI AZUMAMIYAKO
```

If you have a specific romanization scheme in mind, use the specific module:

```
# cpan -i Lingua::JA::Romanize::Japanese
use Lingua::JA::Romanize::Japanese;
my $k2r = Lingua::JA::Romanize::Japanese->new;
my $str = "æ•±ä°¬";
say "Japanese for $str is ", $k2r->chars($str);
```

prints

```
Japanese for æ•±ä°¬ is toukyou
```

â„ž 12: Explicit encode/decode

On rare occasion, such as a database read, you may be given encoded text you need to decode.

```
use Encode qw(encode decode);

my $chars = decode("shiftjis", $bytes, 1);
# OR
my $bytes = encode("MIME-Header-ISO_2022_JP", $chars, 1);
```

For streams all in the same encoding, don't use encode/decode; instead set the file encoding when you open the file or immediately after with `binmode` as described later below.

â„ž 13: Decode program arguments as utf8

```
$ perl -CA ...
or
$ export PERL_UNICODE=A
or
use Encode qw(decode);
@ARGV = map { decode('UTF-8', $_, 1) } @ARGV;
```

â„ž 14: Decode program arguments as locale encoding

```
# cpan -i Encode::Locale
use Encode qw(locale);
use Encode::Locale;
```

```
# use "locale" as an arg to encode/decode
@ARGV = map { decode(locale => $_, 1) } @ARGV;
```

â„ž 15: Declare STD{IN,OUT,ERR} to be utf8

Use a command-line option, an environment variable, or else call `binmode` explicitly:

```
$ perl -CS ...
or
$ export PERL_UNICODE=S
or
use open qw(:std :encoding(UTF-8));
or
binmode(STDIN, ":encoding(UTF-8)");
binmode(STDOUT, ":utf8");
binmode(STDERR, ":utf8");
```

â„ž 16: Declare STD{IN,OUT,ERR} to be in locale encoding

```
# cpan -i Encode::Locale
use Encode;
use Encode::Locale;

# or as a stream for binmode or open
binmode STDIN, ":encoding(console_in)" if -t STDIN;
binmode STDOUT, ":encoding(console_out)" if -t STDOUT;
binmode STDERR, ":encoding(console_out)" if -t STDERR;
```

â„ž 17: Make file I/O default to utf8

Files opened without an encoding argument will be in UTF-8:

```
$ perl -CD ...
or
$ export PERL_UNICODE=D
or
use open qw(:encoding(UTF-8));
```

â„ž 18: Make all I/O and args default to utf8

```
$ perl -CSDA ...
or
$ export PERL_UNICODE=SDA
or
use open qw(:std :encoding(UTF-8));
use Encode qw(decode);
@ARGV = map { decode('UTF-8', $_, 1) } @ARGV;
```

â„ž 19: Open file with specific encoding

Specify stream encoding. This is the normal way to deal with encoded text, not by calling low-level functions.

```
# input file
open(my $in_file, "< :encoding(UTF-16)", "wintext");
OR
open(my $in_file, "<", "wintext");
binmode($in_file, ":encoding(UTF-16)");
```


â„ž 24: Disabling Unicode-awareness in builtin charclasses

Disable `\w`, `\b`, `\s`, `\d`, and the POSIX classes from working correctly on Unicode either in this scope, or in just one regex.

```
use v5.14;
use re "/a";
```

OR

```
my($num) = $str =~ /(\d+)/a;
```

Or use specific un-Unicode properties, like `\p{ahex}` and `\p{POSIX_Digit}`. Properties still work normally no matter what charset modifiers (`/d /u /l /a /aa`) should be effect.

â„ž 25: Match Unicode properties in regex with `\p`, `\P`

These all match a single codepoint with the given property. Use `\P` in place of `\p` to match one codepoint lacking that property.

```
\pL, \pN, \pS, \pP, \pM, \pZ, \pC
\p{Sk}, \p{Ps}, \p{Lt}
\p{alpha}, \p{upper}, \p{lower}
\p{Latin}, \p{Greek}
\p{script_extensions=Latin}, \p{scx=Greek}
\p{East_Asian_Width=Wide}, \p{EA=W}
\p{Line_Break=Hyphen}, \p{LB=HY}
\p{Numeric_Value=4}, \p{NV=4}
```

â„ž 26: Custom character properties

Define at compile-time your own custom character properties for use in regexes.

```
# using private-use characters
sub In_Tengwar { "E000\tE07F\n" }

if (/ \p{In_Tengwar}/) { ... }

# blending existing properties
sub Is_GraecoRoman_Title { <<'END_OF_SET' }
+utf8::IsLatin
+utf8::IsGreek
&utf8::IsTitle
END_OF_SET

if (/ \p{Is_GraecoRoman_Title}/ { ... }
```

â„ž 27: Unicode normalization

Typically render into NFD on input and NFC on output. Using NFKC or NFKD functions improves recall on searches, assuming you've already done to the same text to be searched. Note that this is about much more than just pre- combined compatibility glyphs; it also reorders marks according to their canonical combining classes and weeds out singletons.

```
use Unicode::Normalize;
my $nfd = NFD($orig);
my $nfc = NFC($orig);
my $nfkd = NFKD($orig);
```



```
my $count = 0;
while ($str =~ /\X/g) { $count++ }

# OR: cpan -i Unicode::GCString
use Unicode::GCString;
my $gcs = Unicode::GCString->new($str);
my $count = $gcs->length;
```

â„ž 34: Unicode column-width for printing

Perl's `printf`, `sprintf`, and `format` think all codepoints take up 1 print column, but many take 0 or 2. Here to show that normalization makes no difference, we print out both forms:

```
use Unicode::GCString;
use Unicode::Normalize;

my @words = qw/crÃ"me brÃ»lÃ@e/;
@words = map { NFC($_), NFD($_) } @words;

for my $str (@words) {
    my $gcs = Unicode::GCString->new($str);
    my $cols = $gcs->columns;
    my $pad = " " x (10 - $cols);
    say str, $pad, " |";
}
```

generates this to show that it pads correctly no matter the normalization:

```
crÃ"me      |
creï€me     |
brÃ»lÃ@e    |
bruï,leï•e  |
```

â„ž 35: Unicode collation

Text sorted by numeric codepoint follows no reasonable alphabetic order; use the UCA for sorting text.

```
use Unicode::Collate;
my $col = Unicode::Collate->new();
my @list = $col->sort(@old_list);
```

See the *ucsort* program from the *Unicode::Tussle* CPAN module for a convenient command-line interface to this module.

â„ž 36: Case- and accent-insensitive Unicode sort

Specify a collation strength of level 1 to ignore case and diacritics, only looking at the basic character.

```
use Unicode::Collate;
my $col = Unicode::Collate->new(level => 1);
my @list = $col->sort(@old_list);
```

â„ž 37: Unicode locale collation

Some locales have special sorting rules.

```
# either use v5.12, OR: cpan -i Unicode::Collate::Locale
```

```
use Unicode::Collate::Locale;
my $col = Unicode::Collate::Locale->new(locale => "de__phonebook");
my @list = $col->sort(@old_list);
```

The *ucsort* program mentioned above accepts a `--locale` parameter.

â„ž 38: Making cmp work on text instead of codepoints

Instead of this:

```
@srecs = sort {
    $b->{AGE}    <=>    $a->{AGE}
                ||
    $a->{NAME}    cmp    $b->{NAME}
} @recs;
```

Use this:

```
my $coll = Unicode::Collate->new();
for my $rec (@recs) {
    $rec->{NAME_key} = $coll->getSortKey( $rec->{NAME} );
}
@srecs = sort {
    $b->{AGE}    <=>    $a->{AGE}
                ||
    $a->{NAME_key} cmp    $b->{NAME_key}
} @recs;
```

â„ž 39: Case- and accent-insensitive comparisons

Use a collator object to compare Unicode text by character instead of by codepoint.

```
use Unicode::Collate;
my $es = Unicode::Collate->new(
    level => 1,
    normalization => undef
);
```

```
# now both are true:
$es->eq("GarcÃ-a", "GARCIA" );
$es->eq("MÃ¡rquez", "MARQUEZ");
```

â„ž 40: Case- and accent-insensitive locale comparisons

Same, but in a specific locale.

```
my $de = Unicode::Collate::Locale->new(
    locale => "de__phonebook",
);
```

```
# now this is true:
$de->eq("tschÃ¼ÃŸ", "TSCHUESS"); # notice Ã¼ => UE, ÃŸ => SS
```

â„ž 41: Unicode linebreaking

Break up text into lines according to Unicode rules.

```
# cpan -i Unicode::LineBreak
```

```
use Unicode::LineBreak;
use charnames qw(:full);

my $para = "This is a super\N{HYPHEN}long string. " x 20;
my $fmt = Unicode::LineBreak->new;
print $fmt->break($para), "\n";
```

â„ž 42: Unicode text in DBM hashes, the tedious way

Using a regular Perl string as a key or value for a DBM hash will trigger a wide character exception if any codepoints wonâ€™t fit into a byte. Hereâ€™s how to manually manage the translation:

```
use DB_File;
use Encode qw(encode decode);
tie %dbhash, "DB_File", "pathname";

# STORE

# assume $uni_key and $uni_value are abstract Unicode strings
my $enc_key   = encode("UTF-8", $uni_key, 1);
my $enc_value = encode("UTF-8", $uni_value, 1);
$dbhash{$enc_key} = $enc_value;

# FETCH

# assume $uni_key holds a normal Perl string (abstract Unicode)
my $enc_key   = encode("UTF-8", $uni_key, 1);
my $enc_value = $dbhash{$enc_key};
my $uni_value = decode("UTF-8", $enc_value, 1);
```

â„ž 43: Unicode text in DBM hashes, the easy way

Hereâ€™s how to implicitly manage the translation; all encoding and decoding is done automatically, just as with streams that have a particular encoding attached to them:

```
use DB_File;
use DBM_Filter;

my $dbobj = tie %dbhash, "DB_File", "pathname";
$dbobj->Filter_Value("utf8"); # this is the magic bit

# STORE

# assume $uni_key and $uni_value are abstract Unicode strings
$dbhash{$uni_key} = $uni_value;

# FETCH

# $uni_key holds a normal Perl string (abstract Unicode)
my $uni_value = $dbhash{$uni_key};
```

â„ž 44: PROGRAM: Demo of Unicode collation and printing

Hereâ€™s a full program showing how to make use of locale-sensitive sorting, Unicode casing, and managing print widths when some of the characters take up zero or two columns, not just one column each time. When run, the following program produces this nicely aligned output:

```
CrÃ“me BrÃ“lÃ“e..... â„ž,-2.00
Ã‰clair..... â„ž,-1.60
FideuÃ“ ..... â„ž,-4.20
Hamburger..... â„ž,-6.00
JamÃ“n Serrano..... â„ž,-4.45
LinguiÃ“sa..... â„ž,-7.00
PÃ“tÃ“..... â„ž,-4.15
Pears..... â„ž,-2.00
PÃ“ches..... â„ž,-2.25
SmÃ“,rbrÃ“,d..... â„ž,-5.75
SpÃ“tzle..... â„ž,-5.50
XoriÃ“so..... â„ž,-3.00
Î“Î“Î“Î“Î“..... â„ž,-6.50
Ë‰Ë‰Ë‰Ë‰Ë‰..... â„ž,-4.00
Ã“Ã“Ã“,Ã“Ã“..... â„ž,-2.65
Ã“Ã“Ã“Ã“Ã“Ã“Ã“..... â„ž,-8.00
Ã“,Ã“Ã“Ã“Ã“Ã“,Ã“Ã“Ã“Ã“Ã“..... â„ž,-1.85
Ã“Ã“,..... â„ž,-9.99
Ã“Ã“Ã“..... â„ž,-7.50
```

Here's that program; tested on v5.14.

```
#!/usr/bin/env perl
# umenu - demo sorting and printing of Unicode food
#
# (obligatory and increasingly long preamble)
#
use utf8;
use v5.14;                                # for locale sorting
use strict;
use warnings;
use warnings qw(FATAL utf8);              # fatalize encoding faults
use open     qw(:std :encoding(UTF-8));   # undeclared streams in UTF-8
use charnames qw(:full :short);           # unneeded in v5.16

# std modules
use Unicode::Normalize;                    # std perl distro as of v5.8
use List::Util qw(max);                    # std perl distro as of v5.10
use Unicode::Collate::Locale;              # std perl distro as of v5.14

# cpan modules
use Unicode::GCString;                     # from CPAN

# forward defs
sub pad($$$);
sub colwidth(_);
sub entitle(_);

my %price = (
    "Î“Î“Î“Î“Î“Î“, "                => 6.50, # gyros
```

```

"pears"                => 2.00, # like um, pears
"linguiÃ§a"            => 7.00, # spicy sausage, Portuguese
"xoriÃ§o"              => 3.00, # chorizo sausage, Catalan
"hamburger"           => 6.00, # burgermeister meisterburger
"Ã©clair"              => 1.60, # dessert, French
"smÃ,rbriÃ,d"         => 5.75, # sandwiches, Norwegian
"spÃtztzle"           => 5.50, # Bayerisch noodles, little sparrows
"Ã©...Ã-."            => 7.50, # baol zi5, steamed pork buns, Mandarin
"jamÃ³n serrano"       => 4.45, # country ham, Spanish
"pÃches"               => 2.25, # peaches, French
"Ã,Ãf¥Ãf¼Ã,-ÃfªÃf¼Ãf " => 1.85, # cream-filled pastry like eclair
"Ã§%ê±,ë|¬"          => 4.00, # makgeolli, Korean rice wine
"Ã¬¿Ã,,"              => 9.99, # sushi, Japanese
"Ã•ŠÃ,,Ã;|"           => 2.65, # omochi, rice cakes, Japanese
"crÃme briÃlÃ©"        => 2.00, # crema catalana
"fideuÃ "              => 4.20, # more noodles, Valencian
                        # (Catalan=fideuada)
"pÃtÃ©"               => 4.15, # gooseliver paste, French
"Ã•Š¥¼Ã•¿,¼Ã••"      => 8.00, # okonomiyaki, Japanese
);

my $width = 5 + max map { colwidth } keys %price;

# So the Asian stuff comes out in an order that someone
# who reads those scripts won't freak out over; the
# CJK stuff will be in JIS X 0208 order that way.
my $coll = Unicode::Collate::Locale->new(locale => "ja");

for my $item ($coll->sort(keys %price)) {
    print pad(entitle($item), $width, ".");
    printf " â,-%.2f\n", $price{$item};
}

sub pad($$$) {
    my($str, $width, $padchar) = @_;
    return $str . ($padchar x ($width - colwidth($str)));
}

sub colwidth(_) {
    my($str) = @_;
    return Unicode::GCString->new($str)->columns;
}

sub entitle(_) {
    my($str) = @_;
    $str =~ s{ (?=\pL)(\S)      (\S*) }
           { ucfirst($1) . lc($2) }xge;
    return $str;
}

```

SEE ALSO

See these manpages, some of which are CPAN modules: *perlunicode*, *perluniprops*, *perlre*, *perlrecharclass*, *perluniintro*, *perlunitut*, *perlunifaq*, *PerlIO*, *DB_File*, *DBM_Filter*, *DBM_Filter::utf8*,

Encode, *Encode::Locale*, *Unicode::UCD*, *Unicode::Normalize*, *Unicode::GCString*, *Unicode::LineBreak*, *Unicode::Collate*, *Unicode::Collate::Locale*, *Unicode::UniHan*, *Unicode::CaseFold*, *Unicode::Tussle*, *Lingua::JA::Romanize::Japanese*, *Lingua::ZH::Romanize::Pinyin*, *Lingua::KO::Romanize::Hangul*.

The *Unicode::Tussle* CPAN module includes many programs to help with working with Unicode, including these programs to fully or partly replace standard utilities: *tcgrep* instead of *egrep*, *uniquote* instead of *cat -v* or *hexdump*, *uniwc* instead of *wc*, *unilook* instead of *look*, *unifmt* instead of *fmt*, and *ucsort* instead of *sort*. For exploring Unicode character names and character properties, see its *uniprops*, *unichars*, and *uninames* programs. It also supplies these programs, all of which are general filters that do Unicode-y things: *unititle* and *unicaps*; *uniwide* and *uninarrow*, *unisupers* and *unisubs*; *nfd*, *nfc*, *nfkd*, and *nfkc*; and *uc*, *lc*, and *tc*.

Finally, see the published Unicode Standard (page numbers are from version 6.0.0), including these specific annexes and technical reports:

Â§3.13 Default Case Algorithms, page 113; Â§4.2 Case, pages 120â€“122; Case Mappings, page 166â€“172, especially Caseless Matching starting on page 170.

UAX #44: Unicode Character Database

UTS #18: Unicode Regular Expressions

UAX #15: Unicode Normalization Forms

UTS #10: Unicode Collation Algorithm

UAX #29: Unicode Text Segmentation

UAX #14: Unicode Line Breaking Algorithm

UAX #11: East Asian Width

AUTHOR

Tom Christiansen <tchrist@perl.com> wrote this, with occasional kibbitzing from Larry Wall and Jeffrey Friedl in the background.

COPYRIGHT AND LICENCE

Copyright Â© 2012 Tom Christiansen.

This program is free software; you may redistribute it and/or modify it under the same terms as Perl itself.

Most of these examples taken from the current edition of the “Camel Book”; that is, from the 4th Edition of *Programming Perl*, Copyright Â© 2012 Tom Christiansen <et al.>, 2012-02-13 by Oâ€™Reilly Media. The code itself is freely redistributable, and you are encouraged to transplant, fold, spindle, and mutilate any of the examples in this manpage however you please for inclusion into your own programs without any encumbrance whatsoever. Acknowledgement via code comment is polite but not required.

REVISION HISTORY

v1.0.0 “first public release, 2012-02-27