

```
#!/usr/local/bin/perl
```

```
# arabjoin - a simple filter to render Arabic text
```

```
# © 1998-06-18 roman@czyborra.com
```

```
# Freeware license at http://czyborra.com/
```

```
# Latest version at http://czyborra.com/unicode/
```

```
# PostScript printout at http://czyborra.com/unicode/arabjoin.ps.gz
```

```
# This filter takes Arabic text (encoded in UTF-8 using the Unicode  
# characters from the U+0600 Arabic block in logical order) as input  
# and performs Arabic glyph joining on it and outputs a UTF-8 octet  
# stream that is no longer logically arranged but in a visual order  
# which gives readable results when formatted with a simple Unicode  
# renderer like Yudit that does not handle Arabic differently yet  
# but simply outputs all glyphs in left-to-right order.
```

```
# This little script also demonstrates that Arabic rendering is not  
# that complicated after all (it makes you wonder why some software  
# companies are still asking hundreds of dollars from poor students  
# who just want to print their Arabic texts) and that even Perl 4 can  
# handle Unicode text in UTF-8 without any nifty new add-ons.
```

```
# Usage examples:
```

```
# echo "هلاّ بالعالم!" | arabjoin
```

```
# prints أهلاً بالعالم!
```

```
# which is the Arabic version of "Hello world!"
```

```
# | recode ISO-8859-6..UTF-8 | arabjoin | uniprint -f cyberbit.ttf
```

```
# prints an Arabic mail of charset=iso-8859-6-i on your printer
```

```
# | arabjoin | xviewer yudit
```

```
# delegates an Arabic UTF-8 message to a better viewer
```

```
# ftp://sunsite.unc.edu/pub/Linux/apps/editors/X/ has uniprint in yudit-1.0
```

```
# ftp://ftp.iro.umontreal.ca/pub/contrib/pinard/pretest/ has recode-3.4g
```

```
# http://czyborra.com/unicode/ has arabjoin
```



```

for ($b=$uchar[$j=$i]; $transparent{$c=$uchar[++$j]};){};

# The following assignment is the heart of the algorithm.
# It reduces the Arabic joining algorithm described on
# pages 6-24 to 6-26 of the Arabic character block description
# in the Unicode 2.0 Standard to four lines of Perl:

$uchar[$i] = $a && $final{$c} && $medial{$b}
|| $final{$c} && $initial{$b}
|| $a && $final{$b}
|| $isolated{$b}
|| $b;

$a = $initial{$b} && $final{$c};
}

# Until the Unicode Consortium publishes its Unicode Technical
# Report #9 (Bidirectional Algorithm Reference Implementation)
# at http://www.unicode.org/unicode/reports/techreports.html
# let us oversimplify things a bit and reverse everything:

$_ = join ("", reverse @uchar);

# The following 8 obligatory LAM+ALEF ligatures are encoded in the
# U+FE70 Arabic Presentation Forms-B block in Unicode's
# compatibility zone:

s/ل/ل/g;
s/ل/ل/g;
s/ل/ل/g;
s/ل/ل/g;
s/ل/ل/g;
s/ل/ل/g;
s/ل/ل/g;
s/ل/ل/g;

# Bitstream's Cyberbit font offers 57 of the other 466 optional

```

ligatures in the U+FB50 Arabic Presentation Forms-A block:

s/تم/تَم/g;

s/في/فِي/g;

s/لج/لَج/g;

s/لح/لَح/g;

s/لخ/لَخ/g;

s/لم/لَم/g;

s/لى/لِي/g;

s/لي/لِي/g;

s/نم/نَم/g;

s/أ/أ/g;

s/أ/أ/g;

s/أ/أ/g;

s/أ/أ/g;

s/أ/أ/g;

s/بر/بِر/g;

s/بن/بِن/g;

s/بي/بِي/g;

s/تر/تِر/g;

s/تن/تِن/g;

s/تي/تِي/g;

s/ين/يِن/g;

s/ير/يِر/g;

s/ين/يِن/g;

s/بج/بَج/g;

s/بح/بَح/g;

s/بخ/بَخ/g;

s/بم/بَم/g;

s/تج/تَج/g;

s/تح/تَح/g;

s/تخ/تَخ/g;

s/تم/تَم/g;

s/ثم/ثَم/g;

s/جم/جَم/g;

s/حم/حَم/g;

s/خم/خَم/g;

إ	إإ
ئ	ئئئ
ا	اا
ب	بببب
ة	ةة
ت	تتتت
ث	ثثثث
ج	جججج
ح	حححح
خ	خخخخ
د	دد
ذ	ذذ
ر	رر
ز	زز
س	سسسس
ش	شششش
ص	صصصص
ض	ضضضض
ط	طططط
ظ	ظظظظ
ع	عععع
غ	غغغغ
—	————
ف	فففف
ق	قققق
ك	كككك
ل	لللل
م	مممم
ن	نننن
ه	هههه
و	وو
ى	ىى // □□
ي	يبيبي
□	□ // □
□	□□
□	□□
□	□