Unicode HOWTO

Avtor: Bruno Haible, <*haible@clisp.cons.org*>; prevedel: Jernej Kovačič <*jkovacic@email.si*> v0.18, 4. avgust 2000, prevod 28. marec 2001

Ta spis opisuje, kako nastaviti sistem Linux, da bo kodiral besedila po standardu UTF-8. Ta spis še vedno ni v dokončni različici. Avtor bo vesel vsakega dodatnega nasveta, popravka, kazalca ali URL povezave.

Kazalo

1 Uvod

1.1 Zakaj Unicode?

Ljudje v različnih deželah uporabljajo različne črkopise za predstavitev besed v njihovih maternih jezikih. Dandanes je večina aplikacij, vključno s sistemi za elektronsko pošto in spletnimi brskalniki, čisto 8-bitnih, kar pomeni, da lahko pravilno prikažejo besedilo, če je napisano v enem izmed 8-bitnih naborov znakov, npr. ISO-8859-1 ali ISO-8859-2.

Na svetu obstaja precej več kot 256 znakov, pomislite samo na cirilico, hebrejščino, arabščino, kitajščino, japonščino, korejščino in tajščino, še vedno pa se od časa do časa uvede kakšen nov znak. Uporabnik lahko naleti na naslednje probleme:

- Nemogoče je hraniti besedilo z znaki iz različnih naborov znakov. Na primer, v publikaciji v nemščini ali francoščini (ali tudi slovenščini, op. pr.) je mogoče citirati članek v ruščini, če uporabljate TeX, xdvi in PostScript, ne morete pa tega storiti v čistem tekstu.
- Dokler je vsak dokument napisan v svojem naboru znakov in ta nabor ni prepoznan avtomatsko, so ročne nastavitve neizogibne. Če si npr. želite ogledati domačo stran distribucije XTeamLinux na naslovu *http://www.xteamlinux.com.cn/*, morate nastaviti Netscape na kodno stran GB2312.
- Uvajajo se novi simboli (npr. za evro). ISO je uvedel nov standard ISO-8859-15, ki se v veliki večini ujema z ISO-8859-1, le da so odstranjeni nekateri redko uporabljani znaki (stari znaki za valute) in je dodan znak za evro. Če uporabniki sprejmejo ta standard, imajo na disku dokumente v različnih naborih znakov in začnejo vsakodnevno razmišljati o tem. Vendar bi računalniki morali stvari poenostaviti, ne pa še bolj zakomplicirati.

Rešitev tega problema je sprejetje po vsem svetu uporabnega nabora znakov. Tak nabor znakov se imenuje Unicode (*http://www.unicode.org/*). Za več informacij o Unicode vtipkajte '`man 7 unicode`' (to je priročniška stran, vsebovana v paketu man-pages-1.20).

1.2 Kodiranja po Unicode

Unicode resda lahko odpravi probleme različnih kodnih strani, prinese pa tehnični problem: kako zapisati Unicode znake z 8-bitnimi zlogi? 8-bitni zlog je pri večini računalnikov najmanjša naslovljiva enota in tudi osnovna enota pri omrežnih povezavah preko protokola TCP/IP. Uporaba enega zloga za predstavitev enega znaka je zgodovinsko naključje, predvsem posledica dejstva, da se je razvoj računalništva pričel v Evropi in ZDA, kjer je 96 znakov zadostovalo za dolgo vrsto let.

V osnovi obstajajo štirje načini za kodiranje Unicodovih znakov v zloge:

UTF-8

128 znakov se kodira z uporabo enega zloga (znaki ASCII). 1920 znakov se kodira z uporabo dveh zlogov (rimski, grški, cirilični, koptski, armenski, hebrejski in arabski znaki). 63488 znakov se kodira z uporabo treh znakov (med drugimi kitajski in japonski znaki). Ostalih 2147418112 znakov (ki še niso povsem določeni) se lahko kodira z uporabo 4, 5 ali 6 zlogov. Za več informacij o UTF-8 vtipkajte '`man 7 utf-8`' (priročniška stran, ki je del paketa ldpman-1.20).

UCS-2

Vsak znak je predstavljen z dvema zlogoma. Ta način kodiranja lahko predstavi le prvih 65536 znakov iz Unicode.

UTF-16

To je razširitev UCS-2, ki lahko predstavi 1112064 znakov iz Unicode. Prvih 65536 znakov je predstavljenih z dvema zlogoma, ostali s štirimi.

UCS-4

Vsak znak je predstavljen s štirimi zlogi.

Prostorske zahteve za kodiranje besedil v primerjavi s trenutno uporabljanimi (8-bitni za evropske jezike, več za kitajščino / japonščino / korejščino) so razvidne iz spodnjega opisa. Pri tem gre za porabo prostora na disku in hitrost prenašanja po omrežju, če ni uporabljena nobena oblika krčenja.

UTF-8

Nobene spremembe za US ASCII, samo nekaj odstotkov več za ISO-8859-1, 50% več za kitajske, japonske oz. korejske znake, 100% več za grške in cirilične znake.

UCS-2 in UTF-16

Nobene spremembe za kitajske, japonske oz. korejske znake. 100% več za US ASCII, ISO-8859-1, ISO-8859-2, grške in cirilične znake.

UCS-4

100% več za kitajske, japonske oz. korejske znake. 300% več za US ASCII, ISO-8859-1, ISO-8859-2, grške in cirilične znake.

Ker uporaba UCS-2, UTF-16 in UCS-4 "prizadane"evropske in severnoameriške uporabnike, se ne zdi verjetno, da bi omenjeni načini kodiranja kdaj bili sprejeti za široko uporabo. Microsoftov programerski vmesnik Win32 podpira UCS-2 vsaj od leta 1995, vendar ta način kodiranja še ni bil široko sprejet za uporabo, saj npr. na Japonskem še vedno prevladuje SJIS.

Po drugi strani se zdi, da ima UTF-8 svetlejšo prihodnost pri široki uporabi, ker ne "kaznuje"evropskih in severnoameriških uporabnikov in ker precej programov za urejanje besedil sploh ne bo potrebno zamenjati zaradi podpore za UTF-8.

V nadaljevanju bomo opisali, kako si nastavite vaš sistem Linux, da bo besedila kodiral po UTF-8.

1.2.1 Opombe za programerje v C/C++

Microsoftov pristop Win32 razvijalcem olajša izdelavo različic programov, ki podpirajo Unicode. Na začetku programa je potrebno vnesti `#define UNICODE` in potem tako dolgo spreminjati '`char`' v '`TCHAR`', dokler se program ne prevede brez opozoril. Problem je, da na koncu dobite dve različici programa: ena "razume"po UCS-2 kodiran tekst, ne "razume"pa 8-bitnih kodiranj, druga pa "razumešamo stara 8-bitna kodiranja.

Še več, pri UCS-2 in UCS-4 nastopi še problem vrstnega reda zlogov. Register naborov znakov IANA na naslovu *http://www.isi.edu/in-notes/iana/assignments/character-sets* pravi o ISO-10646-UCS-2: "zlogi morajo biti podani v omrežnem vrstnem redu: standard tega namreč ne določa". Omrežni vrstni red je "big endian"(najprej zlogi z večjo utežjo, pri arhitekturah, ki temeljijo na Intelu, se uporablja ravno obraten "little endian", op.pr.). RFC 2152 je še jasnejši: "ISO/IEC 10646-1:1993(E) določa, da je oblika UCS-2 predstavljena v 8 bitnih zlogih tako, da se najprej pojavijo zlogi z večjo utežjo". Microsoft pa po drugi strani v svojih razvojnih orodjih C/C++ priporoča uporabo strojno odvisnega vrstnega reda zlogov (npr. "little endian"pri procesorjih x86) in ali oznako za vrstni red zlogov na začetku dokumenta ali neke vrste statistično hevristiko(!).

Pristop UTF-8 po drugi strani ohrani '`char*`' kot standarden tip za nize v jeziku C. Rezultat tega je, da bo vaš program obravnaval US ASCII tekst ne glede na vse spremenljivke okolja in bo prepoznal tekst kodiran tako po ISO 8859-1 ali UTF-8, če je spremenljivka okolja LANG nastavljena pravilno.

1.3 Viri dodatnih informacij

Markus Kuhn redno vzdržuje seznam virov dodatnih informacij:

- *http://www.cl.cam.ac.uk/˜mgk25/unicode.html*
- *http://www.cl.cam.ac.uk/˜mgk25/ucs-fonts.html*

Roman Czyborra ima stran s pregledom programov, ki podpirajo Unicode, UTF-8: *http://czyborra.com/utf/#UTF-8*

Primeri po UTF-8 kodiranih datotek:

- V paketu ucs-fonts Markusa Kuhna: *quickbrown.txt*, *UTF-8-test.txt*, *UTF-8-demo.txt*.
- *ftp://ftp.cs.su.oz.au/gary/x-utf8.html*
- Datoteka `iso10646` v paketu trans-1.1.1 Koste Kostisa *ftp://ftp.nid.ru/pub/os/unix/misc/trans111.tar.gz*
- *ftp://ftp.dante.de/pub/tex/info/lwc/apc/utf8.html*
- *http://www.cogsci.ed.ac.uk/˜richard/unicode-sample.html*

2 Nastavitev prikaza

Predvidevamo, da ste že prilagodili vašo konzolo in nastavili X11 na vašo tipkovnico in locale. Ta postopek je razložen v spisu Danish/International HOWTO, kot tudi v ostalih nacionalnih spisih: Finnish, French, German, Italian, Polish, Spanish, Cyrillic, Hebrew, Chinese, Thai, Esperanto in tudi Slovenian HOWTO (avtor je Primož Peterlin). Vendar vas prosimo, da ne upoštevate nasveta v Thai HOWTO, kjer se pretvarjate, da uporabljate znake iz nabora ISO-8859-1 (U0000 - U00FF), medtem ko dejansko vtipkavate tajske znake (U0E01 - U0E5B). Če boste upoštevali te napotke, boste imeli ob prehodu na Unicode le probleme.

2.1 Linux konzola

O konzoli na tem mestu ne bo veliko govora, saj avtor na tistih računalnikih, ki nimajo nameščenega xdm, v konzoli vtipka le svoje uporabniško ime, geslo in "xinit".

Vseeno, paketa kbd-0.99 (*ftp://sunsite.unc.edu/pub/Linux/system/keyboards/kbd-0.99.tar.gz*) in močno razširjena različica paketa console-tools-0.2.3 (*ftp://sunsite.unc.edu/pub/Linux/system/keyboards/console-tools-0.2.3.tar.gz*) vsebujeta v imeniku kbd-0.99/src/ (ali console-tools-0.2.3/screenfonttools/) dva programa: 'unicode_start' in 'unicode_stop'. Ko poženete 'unicode_start', se izhod konzole na zaslon interpretira, kot da je kodiran po UTF-8. Tudi vhod s tipkovnice se obravnava kot Unicode (oglejte si "man kbd_mode"). V tem načinu se znaki iz Unicode vtipkani kot Alt-x1, ..., Alt-xn, (x1 - xn so tipke na numeričnem delu tipkovnice) oddajajo kodirani po UTF-8. Če vaša tipkovnica, natančneje vaša običajna razporeditev tipk, vsebuje tipke za ne-ASCII znake (npr. nemški preglasi ali slovenski šumniki), za katere bi radi, da sledijo stanju tipke CapsLock, boste morali uporabiti popravek za jedro *linux-2.2.9-keyboard.diff* ali *linux-2.3.12-keyboard.diff*.

Verjetno boste želeli videti znake iz različnih pisav na istem zaslonu. Za to boste potrebovali konzolno pisavo Unicode. Paketa *ftp://sunsite.unc.edu/pub/Linux/system/keyboards/kbd-0.99.tar.gz* in *ftp://sunsite.unc.edu/pub/Linux/system/keyboards/console-data-1999.08.29.tar.gz* vsebujeta pisavo LatArCyrHeb-{08,14,16,19}.psf, ki pokriva latinico, cirilico, hebrejsko in arabsko pisavo. V enem kosu zajema standarde ISO 8859 s končnicami 1, 2, 3, 4, 5, 6, 8, 9 in 10. Pisavo namestite tako, da datoteko prenesete v imenik /usr/lib/kbd/consolefonts/ in izvršite "/usr/bin/setfont /usr/lib/kbd/consolefonts/LatArCyrHeb-14.psf".

Če želite, da bo "rezanje in lepljenje"(angl. cut & paste) delovalo tudi v konzoli s podporo UTF-8, boste potrebovali popravek *linux-2.3.12-console.diff*, katerega avtorja sta Edmund Thomas Grimley Evans in Stanislav Voronyi.

Aprila leta 2000 je Edmund Thomas Grimley Evans (*edmundo@rano.org*) izdelal terminalski emulator konzole UTF-8. Uporablja pisave Unicode in se zanaša na Linuxov pomnilnik za prikaz na zaslonu (angl. frame buffer).

2.2 Tuje pisave za X11

Ne oklevajte z namestitvijo ciriličnih, kitajskih, japonskih idr. pisav. Tudi če niso pisave Unicode, bodo v pomoč pri prikazu dokumentov, kodiranih po Unicode: vsaj Netscape Communicator in Java bosta uporabljala tuje pisave, kadar so na voljo.

Pri namestitvi pisav so koristni naslednji programi:

- "mkfontdir <imenik>"pripravi imenik s pisavami, da jih bo strežnik X lahko uporabljal. Ukaz je potrebno izvesti po namestitvi pisav v imenik.
- "xset fp+ <imenik>"doda imenik v trenutni seznam imenikov s pisavami za X strežnik. Če želite imenik dodati v seznam za stalno, dodajte vrstico "FontPath"v datoteko /etc/XF86Config, v razdelek "Files".
- "xset fp rehash"je potrebno izvesti po klicu ukaza mkfontdir na imeniku, ki je že vsebovan v trenutnem seznamu imenikov s pisavami X strežnika.
- "xfontsel"vam omogoča brskanje med nameščenimi pisavami z izbiranjem med različnimi lastnostmi pisav.
- "xlsfonts -fn <vzorec>"izpiše vse pisave, ki ustrezajo vzorcu za ime pisave. Prikaže tudi različne lastnosti pisav. Predvsem "xlsfonts -ll -fn <pisava>"izpiše lastnosti CHARSET_REGISTRY in CHARSET_ENCODING, ki skupaj določata kodiranje pisave.
- "xfd -fn <pisava>"prikaže pisave po straneh.

Spodaj naštete pisave so prosto dostopne (seznam ni popoln):

- Pisave, vsebovane z XFree86, včasih so zapakirane v ločenih paketih. Primer je distribucija SuSE, kjer se v osnovnem paketu 'xf86' nahajajo le pisave 75dpi. Ostale pisave se nahajajo v paketih 'xfnt100', 'xfntbig', 'xfntcyr' in 'xfntscl'.
- Mednarodne pisave za Emacs, (*ftp://ftp.gnu.org/pub/gnu/intlfonts/intlfonts-1.2.tar.gz*). Kot je bilo že omenjeno, te pisave so koristne tudi, če namesto GNU Emacsa raje uporabljate XEmacs ali Emacsa celo sploh ne uporabljate.

2.3 Pisave Unicode za X11

Programi, kjer želimo prikazati več pisav hkrati (npr. cirilica in grška pisava), lahko to dosežejo z uporabo različnih pisav za X v različnih delih besedila. To znata narediti Netscape Communicator in Java. Vendar je ta pristop bolj zapleten, ker mora programer namesto 'Font' in 'XFontStruct' uporabljati 'XFontSet' in tudi zato, ker se velikosti pri različnih pisavah lahko nekoliko razlikujejo.

- Markus Kuhn je sestavil 75 dpi pisave s fiksno širino (fixed-width), kjer so po Unicode kodirane latinica, cirilica, grška, armenska, gruzinska, hebrejska in simbolna pisava. V enem kosu so pokrite pisave po standadih ISO 8859 1, 2, 3, 4, 5, 7, 8, 9, 10, 13, 14 in 15. Ta pisava je potrebna za uporabo xterma v načinu UTF-8. Pisave se nahajajo na naslovu *http://www.cl.cam.ac.uk/˜mgk25/download/ucs-fonts.tar.gz*
- Roman Czyborra je sestavil 75 dpi pisavo velikosti 8x16 / 16x16, ki zajema ogromen del nabora Unicode. Z naslova *http://czyborra.com/unifont/* naložite unifont.hex.gz in hex2bdf. To ni pisava s fiksno širino. Evropski znaki so široki 8 pik, kitajski pa 16 pik. Sledijo navodila za namestitev:

```
$ gunzip unifont.hex.gz
$ hex2bdf < unifont.hex > unifont.bdf
$ bdftopcf -o unifont.pcf unifont.bdf
$ gzip -9 unifont.pcf
# cp unifont.pcf.gz /usr/X11R6/lib/X11/fonts/misc
# cd /usr/X11R6/lib/X11/fonts/misc
# mkfontdir
# xset fp rehash
```

- Primož Peterlin je sestavil družino pisav ETL, ki zajema latinico, cirilico, grško, armensko, gruzinsko in hebrejsko pisavo. Pisave lahko snamete z naslova *ftp://ftp.x.org/contrib/fonts/etl-unicode.tar.gz*. Za namestitev uporabite program "bdftopcf".
- Mark Leisher je sestavil proporcionalno pisavo višine 17 pik (12 točk) z imenom ClearlyU, ki zajema latinico, cirilico, grško, armensko, gruzinsko, hebrejsko, tajsko in laoško pisavo. Na voljo je na naslovu *http://crl.nmsu.edu/˜mleisher/cu.html*. Navodila za namestitev:

```
$ bdftopcf -o cu12.pcf cu12.bdf
$ gzip -9 cu12.pcf
# cp cu12.pcf.gz /usr/X11R6/lib/X11/fonts/misc
# cd /usr/X11R6/lib/X11/fonts/misc
# mkfontdir
# xset fp rehash
```

2.4 Unicode pod xtermom

Xterm je del X11R6 in XFree86, vendar ga ločeno vzdržuje Tom Dickey (*http://www.clark.net/pub/dickey/xterm/xterm.html*). Novejše različice (popravki stopnje 109 ali novejši) vsebujejo podporo za pretvorbo signalov s tipkovnice v UTF-8, še preden jih posredujejo programu, ki teče v xtermu. Podprt je tudi prikaz znakov Unicode, ki jih ta program vrne v obliki UTF-8.

Če želite pognati xterm v načinu UTF-8, morate:

- sneti datoteko xterm.tar.gz z naslova *http://www.clark.net/pub/dickey/xterm/xterm.tar.gz*,
- ga nastaviti z "./configure –enable-wide-chars ...", ga nato prevesti in namestiti.
- imeti nameščeno pisavo Unicode s fiksno širino. Za to je primeren ucs-fonts.tar.gz Markusa Kuhna (glejte zgoraj).
- pognati "xterm -u8 -fn '-misc-fixed-medium-r-semicondensed–13-120-75-75-c-60-iso10646-1'". Možnost -u8"vključi Unicode in UTF-8. Pisava, določena z dolgo možnostjo -fn"je pisava Unicode Markusa Kuhna. Brez te možnosti se uporabi privzeta pisava "fixed", to je pisava po ISO-8859-1 velikosti 6x13.
- ogledati si vzorčne datoteke iz paketa ucs-fonts Markusa Kuhna:

  ```
  $ cd .../ucs-fonts
  $ cat quickbrown.txt
  $ cat utf-8-demo.txt
  ```

  Med drugim bi morali videti grške in ruske znake.
- dodati vrstice

  ```
  XTerm*utf8:   1
  *VT100*font:  -misc-fixed-medium-r-semicondensed--13-120-75-75-c-60-iso10646-1
  ```

  v vaš $HOME/.Xdefaults (nastavitve veljajo le za vas), da bo xterm podpiral UTF-8 ob vsakem zagonu. Spreminjanja sistemskih nastavitev v /usr/X11R6/lib/X11/app-defaults/XTerm vam ne priporočamo, ker se bodo izbrisale ob naslednji nadgradnji na novo različico XFree86.

Dodaten popravek Roberta Bradyja (*rwb197@ecs.soton.ac.uk*), ki doda podporo za znake dvojne širine (večinoma gre za ideografe CJK) in kombinirane znake, je na voljo na naslovu: *http://www.zepler.org/˜rwb197/xterm/*. Temelji na popravku za xterm stopnje 140 (*http://www.clark.net/pub/dickey/xterm/xterm-140.tgz*) in ga je najbolje uporabiti z naslednjimi nastavitvami.

```
*VT100*font:     -Misc-Fixed-Medium-R-Normal--18-120-100-100-C-90-ISO10646-1
*VT100*wideFont: -Daewoo-Gothic-Medium-R-Normal--18-18-100-100-M-180-ISO10646-1
```

2.5 Pisave TrueType

Zgoraj omenjene pisave imajo fiksno širino in se jim ne da spreminjati velikosti. V nekaterih aplikacijah, še posebej v tistih za tiskanje, so nujno potrebne pisave z veliko ločljivostjo. Najpomembnejši tip pisav z nastavljivo velikostjo in veliko ločljivostjo so pisave TrueType.

Trenutno jih podpirajo

- XFree86 4.0.1; v razdelek Modules vaše datoteke XF86Config morate dodati vrstico:

  ```
  Load "freetype"
  ```

  ali

  ```
  Load "xtt"
  ```

- Gonilniki za prikaz v ostalih operacijskih sistemih.
- Urejevalnik Yudit (opisan je v nadaljevanju) in njegov gonilnik za tiskanje.

Nekatere brezplačne pisave TrueType, ki pokrivajo precejšen del Unicode, so:

Bitstream Cyberbit

Pokriva cirilico, rimsko, grško, hebrejsko, arabsko, kitajsko, korejsko, japonsko in druge pisave vključno s kombiniranimi diakritičnimi oznakami.

Naložite jo lahko z: *ftp://ftp.netscape.com/pub/communicator/extras/fonts/windows/Cyberbit.ZIP*.

Microsoft Arial

Pokriva cirilico, rimsko, grško, hebrejsko, arabsko, vietnamsko pisavo in nekatere kombinirane diakritične oznake.

Če jih želite pobrati, z internetnimi iskalniki poiščite FTP strežnike z datotekami `arial.ttf`, `ariali.ttf`, `arialbd.ttf` in `arialbi.ttf`.

Lucida Sans Unicode

Pokriva cirilico, rimsko, grško, hebrejsko pisavo in kombinirane diakritične oznake.

Vsebovani so v IBMovemu JDK 1.3.0beta za Linux, lahko pa jih neposredno poberete pod imenoma `LucidaSansRegular.ttf` in `LucidaSansOblique.ttf` z naslova *ftp://ftp.maths.tcd.ie/Linux/opt/IBMJava2-13/jre/lib/fonts/*.

Naslove za te in druge pisave TrueType lahko dobite na seznamu (avtor Christoph Singer) prosto dostopnih pisav Unicode Trutype na naslovu *http://www.ccss.de/slovo/unifonts.htm*.

Pisave Truetype lahko pretvorite v nizko ločljive, po velikosti nenastavljive pisave za X11 z uporabo pripomočka ttf2bdf (avtor Mark Leisher). Dobite ga lahko na naslovu *ftp://crl.nmsu.edu/CLR/multiling/General/ttf2bdf-2.8-LINUX.tar.gz*.

Več informacij o pisavah Truetype lahko najdete v spisu Linux TrueType HOWTO na naslovu *http://www.moisty.org/~brion/linux/TrueType-HOWTO.html*.

2.6 Razno

Programček, ki ugotovi, ali je konzola oz. xterm nastavljen na UTF-8, se nahaja na naslovu *ftp://sunsite.unc.edu/pub/Linux/system/keyboards/x-lt-1.18.tar.gz*. V paketu avtorja Ricardasa Cepasa sta datoteki testUTF-8.c in testUTF8.c. Večina programov bo delovala brez tega, zato pa naj bi preverjali vrednosti spremenljivk okolja. Glejte tudi razdelek "Spremenljivke okolja za locale".

3 Nastavitev locale

3.1 Datoteke in jedro

V imenih datotek je že mogoče uporabljati znake iz Unicode brez sprememb v jedru ali orodjih za delo z datotekami. Vzrok temu je dejstvo, da jedro za ime datoteke sprejme vse z izjemo zloga "null"in '/', ki je določen za ločevanje podimenikov. Pri kodiranju z UTF-8, se ne-ASCII znaki nikoli ne bodo pretvorili v zlog "null"ali poševnico. Zgodi se le, da ime datoteke ali imenika porabi več zlogov, kot ima znakov v imenu. Na primer ime iz petih grških znakov se bo v jedru pojavilo kot ime iz 10 zlogov. Jedro ne "ve"(in mu tega tudi ni potrebno "vedeti"), da se ti zlogi prikažejo po grško.

To je splošna teorija, dokler z datotekami upravlja samo Linux. V datotečnih sistemih, ki jih uporabljajo tudi drugi operacijski sistemi, morate poskrbeti za pretvorbo imen iz/v UTF-8.

- Datotečni sistem "vfat"ima možnost vpetja (angl. mount) "utf8". Oglejte si *file:/usr/src/linux/Documentation/filesystems/vfat.txt*. Ko možnost vpetja "iocharset"nastavite drugače, kot je privzeto (t.j. ISO-8859-1), rezultati z oz. brez "utf8"ne bodo konsistentni. Zato se možnost vpetja "iocharset"ne priporoča.
- Datotečna sistema "msdos"in "umsdos"imata ob vpetju isto možnost, ki pa nima nobenega učinka.
- Datotečni sistem "iso9660"ima možnost vpetja "utf8". Oglejte si *file:/usr/src/linux/Documentation/filesystems/isofs.txt*.
- Od Linuxovih jeder 2.2.x naprej ima tudi datotečni sistem "ntfs"možnost vpetja "utf8". Oglejte si *file:/usr/src/linux/Documentation/filesystems/ntfs.txt*.

Ostali datotečni sistemi (nfs, smbfs, ncpfs, hpfs itd.) ne pretvarjajo imen, zato podpirajo Unicode imena datotek v UTF-8 samo, če jih podpira tudi operacijski sistem na drugi strani. Da se bodo datotečni sistemi vpeli s pravilno možnostjo tudi ob ponovnih vpetjih, dodajte te možnosti v četrti stolpec ustreznih vrstic v datoteki /etc/fstab.

3.2 TTY in jedro

TTY je neke vrste dvosmerna cev med dvema programoma, ki omogoča zanimive lastnosti, npr. "odmev"pritisnjenih tipk ali urejanje v ukazni vrstici. V xtermu lahko poženete "cat"brez argumentov, vnesete in urejate lahko poljubno število vrstic, ki se "odbijejo"nazaj vrstica za vrstico. Pri jedru postopki urejanja niso pravilni, še posebej odziv na tipki Backspace in Tab ni pravilen. To odpravite na naslednji način:

- uporabite popravek jedra *linux-2.0.35-tty.diff* ali *linux-2.2.9-tty.diff* ali *linux-2.3.12-tty.diff* in ponovno prevedite jedro.
- če uporabljate glibc2, uporabite popravek *glibc211-tty.diff* in ponovno prevedite libc (oz. če niste tako pustolovski, zadostuje tudi popravek že nameščene datoteke "include": *glibc-tty.diff*),
- uporabite popravek *stty.diff* za GNU sh-utils-1.16b, ponovno zgradite program "stty"in ga preizkusite: "stty -a"in "stty iutf8".
- dodajte ukaz "stty iutf8"v skript "unicode_start"in dodajte ukaz "stty -iutf8"v skript "unicode_stop".
- Uporabite popravek *xterm.diff* za xterm-109 in ponovno prevedite "xterm", zatem ga preizkusite z zagonom "xterm -u8"/ "xterm +u8", v njem poženite "stty -a"in znotraj njega interaktivni "cat".

3.3 Splošna pretvorba podatkov

Potrebovali boste program za pretvorbo vaših lokalno (najverjetneje po ISO-8859-1 ali 2) kodiranih dokumentov v UTF-8 (druga možnost bi bila, da še naprej uporabljate različno kodirane dokumente na istem računalniku, kar pa dolgoročno ni preveč zabavno). Eden takšnih programov je ‘iconv’, ki se ga dobi poleg glibc-2.1. Enostavno vtipkajte

```
$ iconv --from-code=ISO-8859-2 --to-code=UTF-8 < stara_datoteka > nova_datoteka
```

Tukaj sta še priročna lupinska skripta "i2u"*i2u.sh* (za pretvorbo iz ISO v UTF) in "u2i"*u2i.sh* (za pretvorbo iz UTF v ISO). Prilagodite si ju v skladu z vašim trenutnim 8 bitnim naborom znakov.

Če nimate nameščenega glibc-2.1 in iconv, lahko namesto tega uporabite GNU recode 3.5. "i2u"*i2u_recode.sh* je "recode ISO-8859-1..UTF-8", in "u2i"*u2i_recode.sh* je "recode UTF-8..ISO-8859-1". Recode dobite na *ftp://ftp.iro.umontreal.ca/pub/recode/recode-3.5.tar.gz* ali *ftp://ftp.gnu.org/pub/gnu/recode/recode-3.5.tar.gz*. Opombe: potrebujete GNU recode 3.5 ali novejši. Da ga prevedete na sistemih brez glibc2 (to so skoraj vsi sistemi Linux razen najnovejših), ga morate nastaviti z možnostjo --disable-nls", sicer se ne bo povezal. Novejše različice s podporo za CJK se nahajajo na naslovu *http://www.iro.umontreal.ca/contrib/recode/*.

Namesto tega lahko uporabite tudi CLISP. Na voljo sta tudi "i2u"*i2u.lisp* in "u2i"*u2i.lisp* v lispu. Potrebna je različica z julija 1999 ali novejša. CLISP dobite na naslovu: *ftp://clisp.cons.org/pub/lisp/clisp/source/clispsrc.tar.gz*.

Ostali programi za pretvorbo podatkov, manj zmogljivi od GNU recode, so ‘trans’ (*ftp://ftp.informatik.uni-erlangen.de/pub/doc/ISO/charsets/trans113.tar.gz*), ‘tcs’ od operacijskega sistema Plan9 (*ftp://ftp.informatik.uni-erlangen.de/pub/doc/ISO/charsets/tcs.tar.gz*) in ‘utrans’ / ‘uhtrans’ / ‘hutrans’ (*ftp://ftp.cdrom.com/pub/FreeBSD/distfiles/i18ntools-1.0.tar.gz*) avtorja G. Adama Stanislava (*adam@whizkidtech.net*).

Za ponavljajočo pretvorbo datotek v UTF-8 iz različnih naborov znakov lahko uporabite polavtomatsko orodje: *to-utf8* prikaže ne-ASCII dele datoteke uporabniku, da vnese originalni nabor znakov, potem to pretvori v UTF-8.

3.4 Spremenljivke okoja za locale

Verjetno imate nastavljene naslednje spremenljivke okolja, ki se nanašajo na locale:

LANGUAGE

nadomešča LC_MESSAGES, uporablja jo samo GNU gettext

LC_ALL

nadomešča vse ostale spremenljivke LC_*

LC_CTYPE, LC_MESSAGES, LC_COLLATE, LC_NUMERIC, LC_MONETARY, LC_TIME

posamezne spremenljivke za: tip znakov in kodiranje, sporočila v naravnem jeziku, pravila za urejanje, obliko številk, obliko denarnih zneskov, prikaz datuma in časa

LANG

privzeta vrednost za vse spremenljivke LC_*

(Oglejte si ‘`man 7 locale`’ za podrobnejši opis.)

Vsaka od spremenljivk LC_* in LANG lahko vsebuje ime locale naslednje oblike.

jezik[_območje[.nabor_znakov][@modifikator]

kjer jezik pomeni kodo jezika po *ISO 639* (z malimi črkami), območje pomeni kodo države po *ISO 3166* (z velikimi črkami), nabor_znakov je samoumeven, modifikator pa označuje posebne značilnosti (npr. posebno narečje, nestandardno črkovanje).

Spremenljivka LANGUAGE lahko vsebuje več imen localov ločenih s podpičji.

Da sistemu in vsem programom poveste, da uporabljate UTF-8, morate vsem imenom locale dodati končnico UTF-8. Če ste npr. uporabljali

```
LC_CTYPE=sl_SI
```

ga boste spremenili v

```
LC_CTYPE=sl_SI.UTF-8
```

Spremenljivke okolja LANGUAGE vam *ni* potrebno spreminjati. GNU gettext ima možnost pretvoriti prevode v pravo kodiranje. Dokler ne izide glibc-2.2, je vse, kar morate storiti to, da nastavite spremenljivko okolja OUTPUT_CHARSET.

```
$ export OUTPUT_CHARSET=UTF-8
```

glibc-2.2 ne bo potreboval te spremenljivke, saj jo bo znal sam izpeljati iz spremenljivke LC_CTYPE.

3.5 Izdelava podpornih datotek za locale

Če imate nameščen glibc-2.1, glibc-2.1.1 ali glibc-2.1.2, najprej z uporabo "localdef –help"preverite, ali je sistemski imenik za razporeditev znakov enak /usr/share/i18n/charmaps. Nato popravite datoteko /usr/share/i18n/charmaps/UTF8 z ustreznim popravkom *glibc21.diff*, *glibc211.diff* ali *glibc212.diff*. Zatem naredite podporno datoteko za vsak UTF-8 locale, ki ga nameravate uporabljati, npr.:

```
$ localedef -v -c -i sl_SI -f UTF8 /usr/share/locale/sl_SI.UTF-8
```

Tukaj morate podati absolutno pot, sicer bo localedef naredil locale v imeniku "sl_SI.utf8", kar ne bo delovalo z XFree86-4.0.1.

Največkrat vam ni potrebno narediti localov "de", "fr"ali "sl"brez končnice za državo, ker te locale največkrat uporabljaja samo spremenljivka LANGUAGE, ne pa spremenljivke LC_*, LANGUAGE pa samo nadomešča LC_MESSAGES.

3.6 Dodajanje podpore v knjižnico C

Glibc-2.2 bo podpiral večzložne locale, še posebej locale UTF-8, ki smo jih naredili malo prej. Glibc-2.1.x pa tega v resnici ne podpira. Zato bo edini učinek prej omenjene izdelave datotek /usr/share/locale/sl_SI.UTF-8/* ta, da bo 'setlocale(LC_ALL,)' vrnil "sl_SI.UTF-8"v skladu z vašimi spremenljivkami okolja, namesto da bi odrezal končnico ".UTF-8".

Da dodate podporo za locale UTF-8, boste morali prevesti in namestiti naslednje tri knjižnice:

- ‘libutf8_plug.so’ iz *libutf8-0.7.3.tar.gz*,
- ‘libiconv_plug.so’, iz *libiconv-1.3.tar.gz*,
- ‘libintl_plug.so’, iz *gettext-0.10.35-iconv.tar.gz*.

Potem lahko spremenljivko okolja LD_PRELOAD nastavite, da kaže na nameščene knjižnice:

```
$ LD_PRELOAD=/usr/local/lib/libutf8_plug.so:/usr/local/lib/libiconv_plug.so:/usr/local/lib/libintl_plug.so
$ export LD_PRELOAD
```

Potem bodo v vsakem programu, ki uporablja te spremenljivke okolja, funkcije v libutf8_plug.so, libiconv_plug.so in libintl_plug.so nadomestile prvotne v /lib/libc.so.6. Za več informacij o LD_PRELOAD si oglejte "man 8 ld.so".

Vsega tega ne bo več treba počenjati, ko izide glibc2-2.

4 Posebni (specifični) programi

4.1 Programi za delo z mrežo

4.1.1 telnet

V nekaterih namestitvah telnet privzeto ni nastavljen na 8-bitni prenos. Da lahko pošiljate znake iz nabora Unicode na oddaljen računalnik, morate najprej nastaviti telnet na način "outbinary". To lahko storite na dva načina:

```
$ telnet -L <host>
```

in

```
$ telnet
telnet> set outbinary
telnet> open <host>
```

4.1.2 kermit

Komunikacijski program C-Kermit (*http://www.columbia.edu/kermit/ckermit.html*), (interaktivno orodje za nastavljanje povezave, telnet, prenos datotek, s podporo za TCP/IP in serijske povezave) v različici 7.0 ali novejši pozna kodiranji UTF-8 in UCS-2 za prenos podatkov, ravno tako pozna tudi terminalsko kodiranje po UTF-8. Zna tudi pretvarjati med obema kodiranjema in še med mnogimi drugimi kodiranji. Dokumentacijo za te značilnosti lahko najdete na *http://www.columbia.edu/kermit/ckermit2.html#x6.6*.

4.2 Brskalniki

4.2.1 Netscape

Netscape 4.05 ali novejši lahko prikaže dokumente HTML, ki so kodirani po UTF-8. V dokumentu mora med oznakama <head> in </head> stati tudi:

```
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
```

Netscape 4.05 ali novejši lahko prikaže tudi dokumente in tekstovne datoteke kodirane po UCS-2 z oznako o vrstnem redu zlogov.

Netscapova domača stran je: *http://www.netscape.com/computing/download/*.

4.2.2 Mozilla

Mozilla milestone M16 ima internacionalizacijo rešeno precej bolje kot Netscape 4. Prikaže lahko dokumente HTML, kodirane po UTF-8 s podporo za več jezikov. Zato pa obstaja manjša lepotna napaka glede pisav CJK. Nekateri simboli so lahko večji od višine vrstice, tako da prekrijejo prejšnjo oz. naslednjo vrstico.

Mozillina domača stran je: *http://www.mozilla.org/*.

4.2.3 Lynx

Lynx-2.8 ima zaslon z možnostmi (tipka 'O'), kjer se lahko nastavi nabor znakov za prikaz. Ko poganjate Lynx v xtermu ali konzoli v načinu UTF-8, nastavite to na "UNICODE UTF-8". Da ta nastavitev prične delovati v trenutnem teku brskalnika, jo morate potrditi na polju "Accept Changes"(sprejmi spremembe), da pa bo delovala vedno, morate izbrati polje "Save options to disk"(shrani možnosti na disk) in to potem potrditi s poljem "Accept Changes".

Tudi tokrat mora v dokumentu med oznakama <head> and </head> stati:

```
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
```

Pri prikazovanju tekstovnih datotek, kodiranih po UTF-8, morate v ukazni vrstici podati možnost -assume_local_charset=UTF-8"(deluje samo pri naslovih, ki se začnejo s 'file:/') ali -assume_charset=UTF-8"(deluje pri vseh vrstah naslovov). V Lynxu 2.8.2 lahko tudi na zaslonu z možnostmi (tipka 'O') spremenite predpostavljen nabor znakov v "utf-8".

Na zaslonu z možnostmi je tudi možnost, da nastavite "priljubljen nabor znakov v dokumentu"(angl. preferred document character set). Vendar to ne deluje, če se URL začne s 'file:/...' oz. če se začne s 'http://...' in se na strežniški strani nahaja Apache 1.3.0.

Problem je tudi s presledki in s prelomom vrstic. Poglejte si razdelek o ruščini v x-utf8.html ali v utf-8-demo.txt.

Če je Lynx 2.8.2 nastavljen z –enable-prettysrc, barvna shema ne deluje več pravilno, ko nabor znakov na zaslonu nastavite na "UNICODE UTF-8". To popravite s preprostim popravkom *lynx282.diff*.

Razvijalci Lynxa pravijo: "Za vsako resno uporabo z izhodom na zaslon po UTF-8 se še vedno priporoča prevajanje s knjižnico slang in -DSLANG_MBCS_HACK."

Zadnja stabilna različica: *ftp://ftp.gnu.org/pub/gnu/lynx/lynx-2.8.2.tar.gz*

Izvorno kodo Lynxa dobite na: *http://lynx.isc.org/*.

Splošna domača stran: *http://lynx.browser.org/*

Novejše različice še v razvoju: *http://lynx.isc.org/current/*, *ftp://lynx.isc.org/current/*

4.2.4 W3M

W3m avtorja Akinorija Ita (*http://ei5nazha.yz.yamagata-u.ac.jp/˜aito/w3m/eng/*) je tekstovni brskalnik po straneh HTML. Njegov prikaz tabel, naštevanj ipd. v HTML je precej lepši kot pri Lynxu. Uporaben je tudi kot zelo dober pretvornik iz HTML v tekst.

W3m ima izbire iz ukazne vrstice za tri najpomembnejša japonska kodiranja, lahko pa ga uporabite tudi za datoteke, ki so kodirane po UTF-8. Če v ukazni vrstici ne vnesete nobene izbire, morate pogosto pritiskati Ctrl-L za osveževanje prikaza, prelom vrstic v odstavkih v cirilici in v CJK pa ni dober.

To odpravite s popravkom Hironorija Sakamota (*http://www2u.biglobe.ne.jp/˜hsaka/w3m/*), ki med kodiranja za prikaz doda UTF-8.

4.2.5 Strani za preizkušanje

Nekaj strani za preizkus brskalnikov lahko najdete na spletnih straneh avtorjev Alana Wooda (*http://www.hclrss.demon.co.uk/unicode/#links*) in Jamesa Kassa (*http://home.att.net/˜jameskass/*).

4.3 Urejevalniki

4.3.1 Yudit

Yudit avtorja Gáspárja Sinaia (*http://czyborra.com/yudit/* in *http://www.yudit.org*) je prvovrsten urejevalnik besedil, kodiranih po Unicode, za sistem X Window. Podpira hkratno obdelavo pri več jezikih, načinih vnosa in pretvorbah za lokalne standarde znakov. Ima pripomočke za vnos besedila v vseh jezikih z angleško tipkovnico, kjer se uporabljajo nastivitve za razporeditve tipk.

Prevesti ga je mogoče v treh različicah: Xlib GUI, KDE GUI ali Motif GUI.

Prilagoditev je zelo lahka. Ponavadi si najprej prilagodite pisavo tako, da iz menija pisave (angl. font) izberete "Unicode". Nato izberete pisavo z velikostjo 13 (pri ukazu "xlsfonts '*-*-iso10646-1'"še vedno prihaja do nekaterih zmešnjav), ki se ujema s 13-točkovno pisavo stalne širine Markusa Kuhna.

Zatem si prilagodite način vnosa. Najbolj značilni so "Straight", "Unicode"in "SGML". Za podrobnosti o ostalih vgrajenih načinih vnosa poglejte v /usr/local/share/yudit/data/.

Da bo sprememba postala privzeta ob vseh nadaljnjih zagonih programa, vnesite želene spremembe v datoteko $HOME/.yuditrc.

Splošne zmožnosti urejevalnika so omejene na urejanje, 'rezanje & lepljenje' (angl. cut & paste) ter iskanje & nadomeščanje. Možnosti razveljavitve (angl. undo) ni.

Yudit lahko prikaže besedilo z uporabo pisave TrueType, oglejte si razdelek "Pisave TrueType". Dobre rezultate da tudi pisava Bitstream Cyberbit. Da jo Yudit lahko najde, naredite povezavo nanjo na `/usr/local/share/yudit/data/cyberbit.ttf`.

4.3.2 Vim

Vim (pri različici 6.0b) dobro podpira UTF-8: ko ga poženete v UTF-8 locale, privzame kodiranje UTF-8 za konzolo in za tekstovne datoteke, ki jih urejate. Podpira znake z dvojno širino (npr. CJK) kot tudi kombiniranje znakov in se zato odlično prilega v Xterm z omogočenim UTF-8.

Namestitev: z naslova *http://www.vim.org/* si naložite izvorno kodo. Ko odpakirate vse štiri dele, popravite datoteko src/Makefile tako, da bo vključeno tudi stikalo `-with-features=big`. Tako boste vključili značilnosti FEAT_MBYTE, FEAT_RIGHTLEFT in FEAT_LANGMAP. Zatem poženite "make"in "make install".

4.3.3 Emacs

Najprej si preberite razdelek o podpori mednarodnih naborov znakov (angl. "International Character Set Support") v priročniku za Emacs. Še posebej upoštevajte, da morate pognati Emacs z ukazom

```
$ emacs -fn fontset-standard
```

da se bo uporabila pisava s precej mednarodnimi znaki.

Na kratko povedano obstajata dva paketa za uporabo UTF-8 v Emacsu. Pri nobenem izmed njiju vam Emacsa ni potrebno še enkrat prrevajati.

- Paket emacs-utf (*http://www.cs.ust.hk/faculty/otfried/Mule/*) avtorja Otfrieda Cheonga omogoča v Emacsu kodiranje "unicode-utf8".
- Paket oc-unicode (*http://www.cs.ust.hk/faculty/otfried/Mule/*) istega avtorja je razširitev paketa Mule-UCS (*ftp://etlport.etl.go.jp/pub/mule/Mule-UCS/Mule-UCS-0.70.tar.gz*, zrcalni strežnik tudi na *http://riksun.riken.go.jp/archives/misc/mule/Mule-UCS/Mule-UCS-0.70.tar.gz*) avtorja Miyashite Hisashija in v Emacsu omogoča kodiranje "utf-8".

Uporabljate lahko kateregakoli izmed njiju, lahko tudi oba skupaj. Prednosti kodiranja po "unicode-utf8"(paket emacs-utf) sta, da se naloži hitreje in da se bolje obnese pri kombiniranju znakov (to je pomembno pri tajščini). Prednost kodiranja po "utf-8"(paket Mule-UCS / oc-unicode) pa je, da lahko tudi obdeluje medpomnilnik (kot npr. M-x shell) in ne samo nalaga in shranjuje datoteke. Prav tako bolje obravnava širine znakov (pomembno pri etiopščini). Zato pa je manj zanesljivo: po obsežnem urejanju datoteke se lahko zgodi, da se po shranjevanju datoteke nekateri znaki iz Unicode zamenjajo z U+FFFD.

Da namestite paket emacs-utf, najprej prevedite program "utf2mule"in ga namestite v enega izmed imenikov, ki so našteti v spremenljivki PATH. Nekam namestite tudi unicode.el, muleuni-1.el in unicode-char.el. Nato v datoteko $HOME/.emacs dodajte vrstice:

```
(setq load-path (cons "/home/user/somewhere/emacs" load-path))
(if (not (string-match "XEmacs" emacs-version))
  (progn
    (require 'unicode)
    ;(setq unicode-data-path "..../UnicodeData-3.0.0.txt")
    (if (eq window-system 'x)
      (progn
        (setq fontset12
          (create-fontset-from-fontset-spec
            "-misc-fixed-medium-r-normal-*-12-*-*-*-*-*-fontset-standard"))
        (setq fontset13
          (create-fontset-from-fontset-spec
            "-misc-fixed-medium-r-normal-*-13-*-*-*-*-*-fontset-standard"))
        (setq fontset14
          (create-fontset-from-fontset-spec
```

```
        "-misc-fixed-medium-r-normal-*-14-*-*-*-*-*-fontset-standard"))
    (setq fontset15
      (create-fontset-from-fontset-spec
        "-misc-fixed-medium-r-normal-*-15-*-*-*-*-*-fontset-standard"))
    (setq fontset16
      (create-fontset-from-fontset-spec
        "-misc-fixed-medium-r-normal-*-16-*-*-*-*-*-fontset-standard"))
    (setq fontset18
      (create-fontset-from-fontset-spec
        "-misc-fixed-medium-r-normal-*-18-*-*-*-*-*-fontset-standard"))
  ; (set-default-font fontset15)
   ))))
```

Katerikoli nabor pisav lahko aktivirate preko menija Mule - Set Font/Fontset ali s kombinacijo Shift - dol - miška - 1. Trenutno so v Unicode najbolje pokrite pisave z višino 15 in 13, predvsem po zaslugi pisav 9x15 in 6x13 Markusa Kuhna. Da bo nabor pisav postal začetni nabor pisav za prvo okno ob zagonu, odkomentirajte vrstico `set-default-font` v zgornjih nastavitvah Emacsa.

Za namestitev paketa oc-unicode vnesite ukaz

```
$ emacs -batch -l oc-comp.el
```

in namestite nekam tako nastale datoteko `un-define.elc` kot tudi `oc-unicode.el`, `oc-charsets.el` in `oc-tools.el`. Nato v datoteko $HOME/.emacs dodajte vrstice:

```
(setq load-path (cons "/home/user/somewhere/emacs" load-path))
(if (not (string-match "XEmacs" emacs-version))
  (progn
    (require 'oc-unicode)
    ;(setq unicode-data-path "..../UnicodeData-3.0.0.txt")
    (if (eq window-system 'x)
      (progn
        (setq fontset12
          (oc-create-fontset
            "-misc-fixed-medium-r-normal-*-12-*-*-*-*-*-fontset-standard"
            "-misc-fixed-medium-r-normal-ja-12-*-iso10646-*"))
        (setq fontset13
          (oc-create-fontset
            "-misc-fixed-medium-r-normal-*-13-*-*-*-*-*-fontset-standard"
            "-misc-fixed-medium-r-normal-ja-13-*-iso10646-*"))
        (setq fontset14
          (oc-create-fontset
            "-misc-fixed-medium-r-normal-*-14-*-*-*-*-*-fontset-standard"
            "-misc-fixed-medium-r-normal-ja-14-*-iso10646-*"))
        (setq fontset15
          (oc-create-fontset
            "-misc-fixed-medium-r-normal-*-15-*-*-*-*-*-fontset-standard"
            "-misc-fixed-medium-r-normal-ja-15-*-iso10646-*"))
        (setq fontset16
          (oc-create-fontset
            "-misc-fixed-medium-r-normal-*-16-*-*-*-*-*-fontset-standard"
```

```
            "-misc-fixed-medium-r-normal-ja-16-*-iso10646-*"))
        (setq fontset18
          (oc-create-fontset
            "-misc-fixed-medium-r-normal-*-18-*-*-*-*-*-fontset-standard"
            "-misc-fixed-medium-r-normal-ja-18-*-iso10646-*"))
      ; (set-default-font fontset15)
       ))))
```

Ustrezen nabor pisav si nastavite enako kot pri paketu emacs-utf.

Da boste lahko naložili datoteko, kodirano v UTF-8, vtipkajte

```
M-x universal-coding-system-argument unicode-utf8 RET
M-x find-file datoteka RET
```

ali

```
C-x RET c unicode-utf8 RET
C-x C-f datoteka RET
```

(oz. utf-8 namesto unicode-utf8, če imate rajši oc-unicode/Mule-UCS).

Lupinsko okno z vhodom in izhodom v UTF-8 odprete, če vtipkate:

```
M-x universal-coding-system-argument utf-8 RET
M-x shell RET
```

(To deluje samo z oc-unicode/Mule-UCS.)

Upoštevajte, da vse to deluje samo z Emacsom v okenskem načinu, ne pa v terminalskem načinu.

Richard Stallman načrtuje, da bo dolgoročno dodal v Emacs integrirano podporo za UTF-8. Podobno načrtuje tudi skupina razvijalcev XEmacsa.

4.3.4 Xemacs

(Ta razdelek je napisal Gilbert Baumann.)

XEmacs (različica 20.4 nastavljena z MULE) si lahko prilagodite na UTF-8 na naslednji način. Žal boste potrebovali izvorno kodo, da boste lahko vnesli popravke.

Najprej potrebujete naslednji datoteki, ki ju je prispeval Tomohiko Morioka:

http://turnbull.sk.tsukuba.ac.jp/Tools/XEmacs/xemacs-21.0-b55-emc-b55-ucs.diff in *http://turnbull.sk.tsukuba.ac.jp/Tools/XEmacs/xemacs-ucs-conv-0.1.tar.gz.*

Končnica .diff se nanaša na izvirnik v C-ju. V paketih tar se nahaja koda v elispu, ki omogoča precej kodnih tabel za prenose v in iz Unicode. Kot je razvidno iz imena datoteke .diff, se nanaša na XEmacs-21. Potrebno je bilo nekaj 'popravkov'. Najbolj opazna razlika v izvirniku za XEmacs 20.4 je, da se je file-coding.[ch] preimenoval v mule-coding.[ch].

Nekaj hitrih napotkov za vse, ki se podobno kot avtor ne spoznajo preveč na XEmacs-MULE:

To, kar se imenuje kodiranje (angl. encoding), se v MULE imenuje 'coding-system'. Najpomembnejša ukaza sta:

```
M-x set-file-coding-system
M-x set-buffer-process-coding-system   [comint buffers]
```

in spremenljivka ‘file-coding-system-alist’, ki ukazu ‘find-file’ pomaga uganiti uporabljen način kodiranja. Ko zadevo poženete, morate najprej storiti *tole*.

Ta koda preveri posebno vrstico z načinom, ki se prične z -*- nekje med prvimi 600 zlogi datoteke, ki jo nameravate odpreti. Če se tam pojavi polje "Encoding: xyz;"in kodiranje ’xyz’ obstaja, ga izberite. Sedaj lahko npr. storite

```
;;; -*- Mode: Lisp; Syntax: Common-Lisp; Package: CLEX; Encoding: utf-8; -*-
```

in XEmacs se preklopi v način utf.

Ko ste vse pognali, Lahko definirate \u03BB (grška lambda) kot makro:

```
(defmacro \u03BB (x) `(lambda .,x))
```

4.3.5 Nedit

4.3.6 Xedit

Če imate XFree86-4.0.1, lahko z Xeditom urejate datoteke, kodirane po UTF-8, če ste ustrezno nastavili locale (glejte zgoraj) in dodali vrstico "Xedit*international: true"v datoteko $HOME/.Xdefaults.

4.3.7 Axe

V različici 6.1.2 Axe podpira samo 8-bitne locale. Če v $HOME/.Xdefaults dodate vrstico "Axe*international: true", se bo preprosto ’sesul’.

4.3.8 Pico

4.3.9 Mined98

Mined98 (*http://www.inf.fu-berlin.de/˜wolff/mined.html*) je preprost urejevalnik avtorjev Michiela Huisjesa, Achima Müllerja in Thomasa Wolffa. V xtermu, ki podpira UTF-8 ali druga 8-bitna kodiranja, vam omogoča urejanje datotek, ki so kodirane po UTF-8 ali drugem 8-bitnem standardu. Ima tudi zelo dobre možnosti za vnos znakov iz nabora Unicode.

Mined vam omogoča urejanje 8-bitno kodiranih in po UTF-8 kodiranih datotek. Privzeto uporabi hevristično samozaznavo. Če se nočete zanašati na hevristiko, v ukazni vrstici podajte stikalo `-u`, kadar urejate datoteko po UTF-8, ali `+u`, kadar urejate 8-bitno kodirano datoteko. To lahko kadarkoli spremenite znotraj urejevalnika. V vrstici z menijem se prikaže kodiranje z "L:hža 8-bitna kodiranja in "U:hža UTF-8. Kliknite na prvega izmed teh znakov, da to spremenite.

Mined pozna znake z dvojno širino ter kombinirane znake in jih tudi pravilno prikaže.

Ima tudi lepo organizirane spuščajoče se menije, zato pa tipke, kot so Home, End ali Delete, ne delujejo.

4.4 Programi za elektronsko pošto

MIME: RFC 2279 definira UTF-8 kot nabor znakov MIME, ki se lahko prenaša pod 8-bitnimi kodiranji ter kodiranjema 'quoted-printable' in base64. Za starejši predlog MIME UTF-7 (RFC 2152) se šteje, da je v zatonu, zato se naj ga ne bi več uporabljalo.

Poštni odjemalci, ki so izšli po 1. januarju 1999, bi morali znati pošiljati in prikazovati po UTF-8 kodirana sporočila, sicer se upoštevajo kot neustrezni. Toda ta sporočila morajo nositi oznako MIME

```
Content-Type: text/plain; charset=UTF-8
Content-Transfer-Encoding: 8bit
```

Enostavno pošiljanje po UTF-8 kodirane datoteke po cevi v "mail"brez popravka oznak MIME ne be uspešno.

Programerji odjemalcev za elektronsko pošto bi si morali ogledati strani *http://www.imc.org/imc-intl/* in *http://www.imc.org/mail-i18n.html*.

Zdaj pa k posameznim poštnim odjemalcem (oz. "agentom za uporabo el. pošte"):

4.4.1 Pine

Situacija za nepopravljen Pine različice 4.10 je naslednja.

Pine ne izvaja pretvorb med nabori znakov. Omogoča pa vam, da si ogledate po UTF-8 kodirana sporočila v tekstovnem oknu, ki ima nastavljen UTF-8 (konzola za Linux ali xterm).

Ponavadi vas bo Pine opozoril o drugem naboru znakov vsakič, ko gledate po UTF-8 kodirano sporočilo. Teh opozoril se znebite, če izberete S (setup), zatem C (config), nato pa "character-set"nastavite na UTF-8. Na ta način ne boste storili nič posebnega, le opozoril ne bo več, saj Pine nima vgrajenega nobenega znanja o UTF-8.

Upoštevajte tudi, da je pri Pinu koncept znakov iz Unicode precej omejen. Prikazal bo znake v latinici in grške znake, preostalih znakov pa ne.

Popravek Roberta Bradyja (el. pošta: *rwb197@ecs.soton.ac.uk*, lokacija popravka: *http://www.ents.susu.soton.ac.uk/˜robert/pine-utf8-0.1.diff*) doda Pinu podporo za UTF-8. S tem popravkom se zaglavja in telesa dekodirajo in prikažejo pravilno. Popravek je odvisen od knjižnice libunicode (*http://cvs.gnome.org/lxr/source/libunicode/*) za GNOME.

Kljub vsemu poravnava ne deluje povsod pravilno, pri odgovarjanju se nabor znakov ne pretvori ustrezno, urejevalnik Pico pa sploh ne zna delati z večzložnimi znaki.

4.4.2 Kmail

Kmail (različica, ki se distribuira s KDE 1.0) sploh ne podpira UTF-8.

4.4.3 Netscape Communicator

Messenger (del Netscape Communicatorja) zna pošiljati in prikazati sporočila, ki so kodirana po UTF-8, vendar je potrebnega tudi nekaj malega ročnega posredovanja.

Pošiljanje po UTF-8 kodiranih sporočil: ko odprete okno za sestavljanje sporočila (Compose), morate še pred začetkom pisanja sporočila v meniju izbrati "View -> Character Set -> Unicode (UTF-8)". Potem lahko napišete sporočilo in ga odpošljete.

Ko prejmete tako kodirano sporočilo, ga Netscape žal ne bo takoj prikazal v pravem naboru znakov, niti vam ne bo tega vidno nakazal. V meniju morate ročno izbrati "View -> Character Set -> Unicode (UTF-8)". Nato izberite kategorijo pisav Unicode.

4.4.4 Emacs (Rmail, Vm)

4.4.5 Mutt

Mutt-1.0, ki ga lahko dobite na naslovu *http://www.mutt.org/*, ima zelo omejeno podporo za UTF-8. Za polno podporo je Edmund Grimley Evans naredil popravke, ki jih lahko dobite na naslovu *http://www.rano.demon.co.uk/mutt.html*.

4.4.6 Exmh

Exmh 2.1.2 s Tk 8.4a1 lahko prepozna in pravilno prikaže po UTF-8 kodirana sporočila (vendar brez znakov CJK), če v datoteko `$HOME/.Xdefaults` dodate naslednje vrstice:

```
!
! Exmh
!
exmh.mimeUCharsets:                    utf-8
exmh.mime_utf-8_registry:              iso10646
exmh.mime_utf-8_encoding:              1
exmh.mime_utf-8_plain_families:        fixed
exmh.mime_utf-8_fixed_families:        fixed
exmh.mime_utf-8_proportional_families: fixed
exmh.mime_utf-8_title_families:        fixed
```

4.5 Obdelava besedil

4.5.1 Groff

Groff 1.16, GNU izvedba tradicionalnega sistema troff/nroff za obdelavo besedil na Unixih, lahko na izhodu doda oblikovano besedilo z znaki po UTF-8. Namesto '`groff -Tlatin1`' ali '`groff -Tascii`' enostavno vtipkajte '`groff -Tutf8`'.

4.5.2 TeX

Distribucije TeTeX 0.9 in novejše vsebujejo prilagoditev Texa za Unicode, ki se imenuje Omega (*http://www.gutenberg.eu.org/omega/*, *ftp://ftp.ens.fr/pub/tex/yannis/omega*). Skupaj z datoteko unicode.tex, ki se nahaja v *utf8-tex-0.1.tar.gz* vam omogoča, da kot vhod v TeX vnesete po UTF-8 kodirane izvirnike. Trenutno je podprtih na tisoče znakov iz Unicode.

Vse, kar se spremeni, je, da poženete 'omega' (namesto 'tex') ali 'lambda' (namesto 'latex') in v glavo vašega izvirnika vrinete naslednje vrstice:

```
\ocp\TexUTF=inutf8
\InputTranslation currentfile \TexUTF

\input unicode
```

Verjetno sta tej tematiki posvečeni tudi naslednji povezavi: *http://www.dante.de/projekte/nts/NTS-FAQ.html* in *ftp://ftp.dante.de/pub/tex/language/chinese/CJK/*.

4.6 Podatkovne baze

4.6.1 PostgreSQL

PostgreSQL 6.4 ali novejši lahko zgradite z naslednjo možnostjo nastavitve: `-with-mb=UNICODE`.

4.7 Ostali programi v tekstovnem načinu

4.7.1 Less

S programom Less, ki se ga dobi na naslovu *http://www.flash.net/˜marknu/less/less-358.tar.gz*, lahko brskate po tekstovnih datotekah, ki so kodirane po UTF-8, če konzola ali xterm podpira ta način kodiranja. Prepričajte se, da spremenljivka okolja LESSCHARSET ni nastavljena (ali da je nastavljena na utf-8). Če je nastavljena tudi spremenljivka LESSKEY, se prepričajte, da datoteka, na katero kaže, ne definira spremenljivke LESSCHARSET. Če je potrebno, naredite to datoteko še enkrat z ukazom ‘lesskey’ ali pa spremenljivko LESSKEY "prekličite".

4.7.2 Lv

Lv-4.21 (*http://www.mt.cs.keio.ac.jp/person/narita/lv/*) avtorja Tomia Narite je pregledovalnik datotek z vgrajenim pretvornikom med nabori znakov. Če želite v z UTF-8 podprti konzoli pregledati po UTF-8 kodirane datoteke, vtipkajte "lv -Au8". Lahko pa ga uporabite tudi za pregledovanje po drugih kodiranjih CJK kodiranih datotek v konzoli s podporo za UTF-8.

Program ima tudi majhno napakico: utripač (kurzor) v xtermu izgine in se po koncu ne pokaže znova.

4.7.3 Expand, Wc

Priskrbite si GNU textutils-2.0 in uporabite popravek *textutils-2.0.diff*, nato si prilagodite nastavitve in v config.h dodajte vrstice "#define HAVE_MBRTOWC 1", "#define HAVE_FGETWC 1"in "#define HAVE_FPUTWC 1". V datoteki src/Makefile popravite CFLAGS in LDFLAGS tako, da vključujeta tudi imenike, kjer je nameščen libutf8. Zatem prevedite vse skupaj.

4.7.4 Col, Colcrt, Colrm, Column, Rev in Ul

Priskrbite si paket util-linux-2.9y, ga nastavite in zatem v datoteki defines.h definirajte ENABLE_WIDECHAR. V datoteki lib/widechar.h "#if 0"popravite v "if 1". V datoteki text-utils/Makefile, popravite CFLAGS in LDFLAGS tako, da vključujeta imenike, kjer je nameščen libutf8. Prevedite vse skupaj.

4.7.5 Figlet

Figlet 2.2 vsebuje stikalo za vnos po UTF-8: "figlet -C utf8"

4.7.6 Temeljni pripomočki

Seznam ukazov in pripomočkov Li18nux, ki bi moral biti povezljiv z UTF-8, je še vedno nepopoln in potrebuje še precej koristnih informacij. Avtorju to še ni uspelo :-).

Pri glibc-2.2 bodo delovali le regularni izrazi z 8-bitnimi znaki. Pri locale za UTF-8 regularni izrazi, ki vsebujejo ne-ASCII znake ali kjer naj bi primerjali posamezne večzložne znake s ".", ne bodo delovali. Posledice tega se bodo poznale v vseh ukazih in pripomočkih, ki so našteti spodaj.

alias

Na voljo še ni nobenih informacij

ar

Na voljo še ni nobenih informacij

arch

Na voljo še ni nobenih informacij

arp

Na voljo še ni nobenih informacij

asa

Na voljo še ni nobenih informacij

at

Pri at-3.1.8 sta dve uporabi isalnum v at.c neveljavni in ju je treba nadomestiti z uporabo quotearg.c ali s seznamom za izključevanje seznamov metaznakov iz lupine. Dve uporabi %8s v at.c in atd.c sta neveljavni in ju je treba popraviti na poljubno dolžino.

basename

Kot pri sh-utils-2.0i: vse je v redu.

batch

Na voljo še ni nobenih informacij

bc

Na voljo še ni nobenih informacij

bg

Na voljo še ni nobenih informacij

bunzip2

Na voljo še ni nobenih informacij

bzip2

Na voljo še ni nobenih informacij

bzip2recover

Na voljo še ni nobenih informacij

cal

Na voljo še ni nobenih informacij

cat

Na voljo še ni nobenih informacij

cd

Na voljo še ni nobenih informacij

cflow

Na voljo še ni nobenih informacij

chgrp

Kot pri fileutils-4.0u: vse v redu

chmod

Kot pri fileutils-4.0u: vse v redu

chown

Kot pri fileutils-4.0u: vse v redu

chroot

Kot pri sh-utils-2.0i: vse v redu

cksum

Kot pri textutils-2.0e: vse vredu

clear

Na voljo še ni nobenih informacij

cmp

Na voljo še ni nobenih informacij

col

Na voljo še ni nobenih informacij

comm

Na voljo še ni nobenih informacij

command

Na voljo še ni nobenih informacij

compress

Na voljo še ni nobenih informacij

cp

Kot pri fileutils-4.0u: vse v redu

cpio

Na voljo še ni nobenih informacij

csplit

Na voljo še ni nobenih informacij

ctags

Na voljo še ni nobenih informacij

crontab

Na voljo še ni nobenih informacij

cut

Na voljo še ni nobenih informacij

date

Kot pri sh-utils-2.0i: vse v redu

dd

Kot pri fileutils-4.0u: stikali conv=lcase in conv=ucase ne delujeta pravilno

depmod

Na voljo še ni nobenih informacij

df

Kot pri fileutils-4.0u: vse v redu

diff

Kot pri diffutils-2.7 (1994): diff se ne zaveda nastavitev locale, način –side-by-side zato ne izračuna pravilno širine stolpca, to velja celo za locale, ki temeljijo na ISO-8859-1

diff3

Na voljo še ni nobenih informacij

dirname

Kot pri sh-utils-2.0i: vse v redu

domainname

Na voljo še ni nobenih informacij

du

Kot pri fileutils-4.0u: vse v redu

echo

Kot pri sh-utils-2.0i: vse v redu

env

Kot pri sh-utils-2.0i: vse v redu

expand

Na voljo še ni nobenih informacij

expr

Kot pri sh-utils-2.0i: operatorji "match", "substr", "index"in "length"ne delujejo pravilno

false

Kot pri sh-utils-2.0i: vse v redu

fc

Na voljo še ni nobenih informacij

fg

Na voljo še ni nobenih informacij

file

Na voljo še ni nobenih informacij

find

Kot pri findutils-4.1.5: stikalo -ok"še ne pozna mednarodnih nastavitev, popravek je že bil posredovan vzdrževalcu. Stikalo -iregex"ne deluje pravilno, potreben je popravek v funkciji find/parser.c:insert_regex

fort77

Na voljo še ni nobenih informacij

ftp[BSD]

Na voljo še ni nobenih informacij

fuser

Na voljo še ni nobenih informacij

getconf

Na voljo še ni nobenih informacij

getopts

Na voljo še ni nobenih informacij

gunzip

Na voljo še ni nobenih informacij

gzip

Gzip je zmožen uporabljati UTF-8, vendar uporablja le angleška sporočila v naboru ASCII. Pri pravilni prilagoditvi bi bilo potrebno: uporabiti gettext, klicati setlocale. V funkciji check_ofname (datoteka gzip.c) namesto spraševanja tipa da/ne uporabite funkcijo rpmatch. Uporaba funkcije strlen v gzip.c:852 je napačna, potrebno je uporabiti funkcijo mbswidth.

hash

Na voljo še ni nobenih informacij

head

Na voljo še ni nobenih informacij

hostname

Kot pri sh-utils-2.0i: vse v redu

id

Kot pri sh-utils-2.0i: vse v redu

ifconfig

Na voljo še ni nobenih informacij

imake

Na voljo še ni nobenih informacij

insmod

Na voljo še ni nobenih informacij

ipchains

Na voljo še ni nobenih informacij

ipcrm

Na voljo še ni nobenih informacij

ipcs

Na voljo še ni nobenih informacij

ipmasqadm

Na voljo še ni nobenih informacij

jobs

Na voljo še ni nobenih informacij

join

Na voljo še ni nobenih informacij

kerneld

Na voljo še ni nobenih informacij

kill

Na voljo še ni nobenih informacij

killall

Na voljo še ni nobenih informacij

ksyms

Na voljo še ni nobenih informacij

ldd

Na voljo še ni nobenih informacij

less

Na voljo še ni popolnih informacij

lex

Na voljo še ni nobenih informacij

lilo

Na voljo še ni nobenih informacij

ln

Kot pri fileutils-4.0u: vse v redu

loadkeys

Na voljo še ni nobenih informacij

logger

Na voljo še ni nobenih informacij

Logname

Kot pri sh-utils-2.0i: vse v redu

lp

Na voljo še ni nobenih informacij

lpc[BSD]

Na voljo še ni nobenih informacij

lpr[BSD]

Na voljo še ni nobenih informacij

lprm[BSD]

Na voljo še ni nobenih informacij

lpq[BSD]

Na voljo še ni nobenih informacij

ls

Kot pri fileutils-4.oy: vse v redu

lsmod

Na voljo še ni nobenih informacij

m4

Na voljo še ni nobenih informacij

mailx

Na voljo še ni nobenih informacij

make

Na voljo še ni nobenih informacij

mesg

Na voljo še ni nobenih informacij

mkdir

Kot pri fileutils-4.0u: vse v redu

mkfifo

Kot pri fileutils-4.0u: vse v redu

mkfs

Na voljo še ni nobenih informacij

mkswap

Na voljo še ni nobenih informacij

modprobe

Na voljo še ni nobenih informacij

more

Na voljo še ni nobenih informacij

mount

Na voljo še ni nobenih informacij

mv

Kot pri fileutils-4.0u: vse v redu

netstat

Na voljo še ni nobenih informacij

newgrp

Na voljo še ni nobenih informacij

nice

Kot pri sh-utils-2.0i: vse v redu

nl

Na voljo še ni nobenih informacij

nohup

Kot pri sh-utils-2.0i: vse v redu

nslookup

Na voljo še ni nobenih informacij

nm

Na voljo še ni nobenih informacij

od

Na voljo še ni nobenih informacij

passwd[BSD]

Na voljo še ni nobenih informacij

paste

Na voljo še ni nobenih informacij

patch

Na voljo še ni nobenih informacij

pathchk

Kot pri sh-utils-2.0i: vse v redu

ping

Na voljo še ni nobenih informacij

printf

Kot pri sh-utils-2.0i: vse v redu

pr

Na voljo še ni nobenih informacij

ps

Na voljo še ni nobenih informacij

pwd

Kot pri sh-utils-2.0i: vse v redu

read

Na voljo še ni nobenih informacij

rdev

Na voljo še ni nobenih informacij

reboot

Na voljo še ni nobenih informacij

renice

Na voljo še ni nobenih informacij

rm

Kot pri fileutils-4.0u: vse vredu

rmdir

Kot pri fileutils-4.0u: vse vredu

rmmod

Na voljo še ni nobenih informacij

shar[BSD]

Na voljo še ni nobenih informacij

shutdown

Na voljo še ni nobenih informacij

sleep

Kot pri sh-utils-2.0i: vse v redu

split

Na voljo še ni nobenih informacij

strings

Na voljo še ni nobenih informacij

strip

Na voljo še ni nobenih informacij

stty

Kot pri sh-utils-2.o1: niza "<undef>še ne bi smelo prevajati, potreben je popravek v funkciji stty.c:visible.

su[BSD]

Na voljo še ni nobenih informacij

sum

Kot pri textutils-2.0e: vse v redu

tac

Na voljo še ni nobenih informacij

tail

Na voljo še ni nobenih informacij

talk

Na voljo še ni nobenih informacij

tar

Kot pri tar-1.13.17: v redu če sta imeni uporabnika in skupine vedno zapisani v ASCII

tclsh

Na voljo še ni nobenih informacij

tee

Kot pri Sh-utils-2.0i: vse v redu

telnet

Na voljo še ni nobenih informacij

test

Kot pri sh-utils-2.0i: vse v redu

time

Na voljo še ni nobenih informacij

touch

Kot pri fileutils-4.0u: vse v redu

tput

Na voljo še ni nobenih informacij

tr

Na voljo še ni nobenih informacij

true

Kot pri Sh-utils-2.0i: vse v redu

tsort

Na voljo še ni nobenih informacij

tty

Kot pri sh-utils-2.0i: vse v redu

type

Na voljo še ni nobenih informacij

ulimit

Na voljo še ni nobenih informacij

umask

Na voljo še ni nobenih informacij

umount

Na voljo še ni nobenih informacij

unalias

Na voljo še ni nobenih informacij

uname

Kot pri sh-utils-2.0i: vse v redu

uncompress

Na voljo še ni nobenih informacij

unexpand

Na voljo še ni nobenih informacij

uniq

Na voljo še ni nobenih informacij

unlink

Na voljo še ni nobenih informacij

uudecode

Na voljo še ni nobenih informacij

uuencode

Na voljo še ni nobenih informacij

wait

Na voljo še ni nobenih informacij

wc

Kot pri textutils-2.oe: wc ne more prešteti znakov, popravek je že posredovan vzdrževalcu

who

Kot pri sh-utils-2.0i: vse v redu

wish

Na voljo še ni nobenih informacij

write

Na voljo še ni nobenih informacij

xargs

Kot pri findutils-4.1.5: program uporabi strstr, popravek je že posredovan vzdrževalcu

yacc

Na voljo še ni nobenih informacij

zcat

Na voljo še ni nobenih informacij

4.8 Preostali programi za X11

Owen Taylor trenutno razvija knjižnico pango za prikaz večjezičnih besedil. Več informacij na *http://www.labs.redhat.com/˜otaylor/pango/* in *http://www.pango.org/*.

5 Tiskanje

Ker postscript sam po sebi ne podpira pisav Unicode, morajo vso odgovornost za podporo Unicode pri tiskanju prevzeti programi, ki naredijo datoteko v postscriptu, ne pa interpreter postscripta.

Obstoječe pisave postscript, ki so bile opažene do sedaj: (.pfa / .pfb / .afm / .pfm / .gsf) podpirajo samo majhno področje simbolov in niso pisave Unicode.

5.1 Tiskanje z uporabo pisav TrueType

Tako uniprint kot wprint omogočata dobro tiskanje tekstovnih datotek, ki so kodirane po Unicode. Oba zahtevata nameščene pisave TrueType. Oglejte si razdelek "Pisave TrueType". Pisava Bitstream Cyberbit da dobre rezultate.

5.1.1 Uniprint

Program "uniprint", ki je del paketa Yudit, zna tekstovno datoteko pretvoriti v postscript. Da bo uniprint lahko našel pisavo Cyberbit, naredite simbolično povezavo na `/usr/local/share/yudit/data/cyberbit.ttf`.

5.1.2 Wprint

Program "wprint"(WorldPrint) (*http://ttt.esperanto.org.uy/programoj/angle/wprint.html*) avtorja Eduarda Trapanija naknadno obdela izhod v postscriptu, ki ga iz strani HTML ali tekstovnih datotek izdelata Netscape Communicator ali Mozilla.

Rezultat je skorajda popoln, le v odstavkih s cirilico prihaja do nepravilnih prelomov vrstic, saj so vrstice široke samo polovico pričakovane širine.

5.1.3 Primerjava

Pri tekstovnih datotekah je splošna razporeditev pri uniprintu nekoliko boljša, zato pa samo wprint pravilno izpiše besedila v tajščini.

5.2 Klasični pristop

Drug način za tiskanje s pisavami TrueType je pretvorba te pisave v pisavo za postscript s pomočjo pripomočka `ttf2pt1` (*http://www.netspace.net.au/~mheath/ttf2pt1/* ali *http://quadrant.netspace.net.au/ttf2pt1/*). Podrobnosti lahko najdete v dokumentu Juliusa Chroboczeka "Printing with TrueType fonts in Unix"na naslovu *http://www.dcs.ed.ac.uk/home/jec/programs/xfsft/printing.html*.

5.2.1 TeX, Omega

Raziskati je treba še: CJK, metafont, omega, dvips, odvips, utf8-tex-0.1

5.2.2 DocBook

Raziskati je treba še: db2ps, jadetex

5.2.3 Groff -Tps

"groff -Tps"naredi izhod v postscriptu. Njegov gonilnik za postscript podpira zelo omejen nabor znakov iz Unicode (samo tisto, kar postscript podpira že sam po sebi).

5.3 Ni bilo sreče z ...

5.3.1 Tiskanje iz Netscapa

V različici 4.72 Netscape Communicator ne zna pravilno natisniti strani HTML, ki so kodirane po UTF-8. Ne preostane vam nič drugega kot wprint.

5.3.2 Tiskanje iz Mozille

V različici M16 tiskanje očitno sploh ni izvedeno.

5.3.3 Html2ps

V različici 1.0b1 pretvornik iz HTML v postscript html2ps ne podpira po UTF-8 kodiranih strani HTML in ne zna ravnati s pisavami. To pomeni, da zmeraj uporablja standardne pisave postscript.

5.3.4 A2ps

V različici 4.12 a2ps ne podpira tiskanja besedil, kodiranih po UTF-8.

5.3.5 Enscript

V različici 1.6.1 enscript ne podpira tiskanja besedil, kodiranih po UTF-8. Privzeto uporablja samo standardne pisave postscript, možno pa je vključiti tudi druge pisave.

6 Kaj narediti, da se bodo vaši programi "zavedalištandarda Unicode

6.1 C/C++

C-jevski tip ‘`char`’ je 8-bitni in bo tak tudi ostal, ker je tolikšna najmanjša naslovljiva podatkovna enota. Na voljo je več pripomočkov:

6.1.1 Za običajno ravnanje s tekstom

Standard ISO/ASCII vsebuje v skladu z dopolnili iz leta 1995 tip "širokega znaka"(angl. "wide character") ‘`wchar_t`’, nabor "dvojnikov"funkcij iz <`string.h`> in <`ctype.h`> (deklarirane so v <`wchar.h`> in <`wctype.h`>) in nabor funkcij za pretvorbo med tipoma (deklarirane so v <`stdlib.h`>).

Dobre reference za ta programerski vmesnik so

- The GNU libc-2.1 Manual, poglavji 4 "Character Handling"(ravnanje z znaki) in 6 "character Set Handling"(ravnanje z nabori znakov)
- priročniške strani *man-mbswcs.tar.gz*, zdaj se jih dobi tudi na lokaciji *ftp://ftp.win.tue.nl/pub/linux-local/manpages/man-pages-1.29.tar.gz*
- predstavitev skupine OpenGroup: *http://www.unix-systems.org/version2/whatsnew/login_mse.html*,
- Specifikacije Single Unix iste skupine: *http://www.UNIX-systems.org/online.html*,
- Standard ISO/IEC 9899:1999 (ISO C 99). Zadnji osnutek pred njegovim sprejetjem se imenuje n2794. Najdete ga na naslovu *ftp://ftp.csn.net/DMK/sc22wg14/review/* ali *http://java-tutor.com/docs/c/*.
- Predstavitev Cliva Featherja na *http://www.lysator.liu.se/c/na1.html*,
- Dinkumwarov referenčni priročnik h knjižnici C: *http://www.dinkumware.com/htm_cl/*.

Prednosti uporabe tega programerskega vmesnika:

- Od proizvajalcev neodvisen standard
- Funkcije naredijo pravo stvar, odvisno od uporabnikove nastavitve locale. V programu je potrebno le poklicati funkcijo `setlocale(LC_ALL,);`.

Pomanjkljivosti tega programerskega vmesnika:

- Nekatere funkcije niso varne v večnitnem (angl. multithread) delovanju, saj obdržijo skrito notranje stanje med klici funkcij.
- Ni prvorazrednega podatkovnega tipa za locale. Zato tega programerskega vmesnika ni pametno uporabljati povsod, kjer je istočasno potrebnih več localov ali naborov znakov.
- Pri nekaterih operacijskih sistemih ni dobre podpore za ta programerski vmesnik.

Opombe o prenosljivosti Tip ‘`wchar_t`’ je lahko ali pa tudi ne kodiran po Unicode. To je odvisno od sistema in včasih tudi od nastavitev locale. Enako velja za kodiranje večzložne sekvence sekvence ‘`char *`’ po UTF-8.

Podrobneje o tipu ’`wchar_t` pravi *Single Unix specification* tole : *Vse kode za "širokežnake v danem procesu se sestojijo iz enakega števila bitov. To je v nasprotju z znaki, ki se lahko sestojijo iz spremenljivega števila zlogov. Zlog ali zaporedje zlogov, ki predstavlja znak, je lahko predstavljen tudi kot koda "širokegažnaka. Kode "širokihžnakov tako omogočajo uniformno velikost za ravnanje s tekstovnimi podatki. Koda "širokegažnaka, kjer so vsi biti enaki 0, je znak ’null’ za "širokežnake in označuje konec niza iz "širokihžnakov. Vrednost "širokegažnaka za vsakega člana prenosnega nabora znakov* (t. j. ASCII) *bo enaka svoji vrednosti, ko se jo uporabi kot osamljen znak v celoštevilski znakovni konstanti. Kode "širokihžnakov za ostale znake so odvisne od locale in od izvedbe. Zlogi za stanje premika nimajo predstavitve med "širokimižnaki.*

Posledica tega je, da v prenosljivih programih ne smete uporabljati ne-ASCII znakov v konstantnih nizih. To pomeni, da četudi poznate Unicode kodi za dvojni narekovaj (U+201C in U+201D), v programih v jeziku C ne smete zapisati niza `L"\u201cLep pozdrav,\u201d je rekel"` ali `"\xe2\x80\x9cLep pozdrav,\xe2\x80\x9d" je rekel"`. Namesto tega uporabite GNU gettext in isto stvar napišite kot `gettext ("Lep pozdrav,’ je rekel")` in naredite podatkovno bazo sporočil sl.po, ki “Lep pozdrav,’ je rekel"prevede v "\u201cLep pozdrav,\u201d je rekel".

Sledi prikaz prenosljivosti pripomočkov za ISO/ANSI C med različnimi vrstami Unixa. GNU glibc-2.2 podpira vse izmed njih, trenutna slika pa je takšna.

GNU glibc-2.0.x, glibc-2.1.x

- <wchar.h> in <wctype.h> obstajata.
- Obstajata funkciji wcs in mbs, ne pa fgetwc, fputwc in wprintf.
- Locale za UTF-8 ne obstaja
- mbrtowc vrne EILSEQ za zloge, ki so večji ali enaki 0x80.

AIX 4.3

- <wchar.h> in <wctype.h> obstajata.
- Obstajata funkciji wcs in mbs, ravno tako fgetwc, fputwc, wprintf idr.
- Precej localov za UTF-8, po eden za vsako državo.
- Za definiranje mbstate_t je potreben -D_XOPEN_SOURCE=500.
- mbrtowc deluje.

Solaris 2.7

- <wchar.h> in <wctype.h> obstajata.
- Obstajata funkciji wcs in mbs, ravno tako fgetwc, fputwc, wprintf idr.
- Ima naslednje locale za UTF-8: en_US.UTF-8, de.UTF-8, es.UTF-8, fr.UTF-8, it.UTF-8, sv.UTF-8.
- mbrtowc vrne -1/EILSEQ (namesto -2) za zloge, ki so večji ali enaki 0x80.

OSF/1 4.0d

- <wchar.h> in <wctype.h> obstajata.
- Obstajata funkciji wcs in mbs, ravno tako fgetwc, fputwc, wprintf idr.

- Ima dodan locale universal.utf8@ucs4, oglejte si "man 5 unicode".
- mbrtowc ne pozna UTF-8.

Irix 6.5

- <wchar.h> in <wctype.h> obstajata.
- Obstajata funkciji wcs in mbs, ravno tako fgetwc in fputwc, wprintf pa ne obstaja
- Nima večzložnih localov.
- Vsebuje le deklaracijo za mbstate_t, ki bo verjetno definiran v prihodnosti.
- Ne vsebuje mbrtowc.

HP-UX 11.00

- <wchar.h> obstaja, <wctype.h> pa ne.
- Obstajata funkciji wcs in mbs, ravno tako fgetwc in fputwc, wprintf pa ne obstaja
- Vsebuje locale C.utf8.
- Ne vsebuje mbstate_t.
- Ne vsebuje mbrtowc.

Posledično se priporoča uporaba funkcij wcsr in mbsr, ki ju je mogoče ponovno pognati in sta primerni za večnitno delovanje, pozabite pa na sisteme, ki teh funkcij nimajo (Irix, HP-UX, AIX) in uporabljate vtičnico za locale, ki podpira UTF-8 - libutf8_plug.so (gl. spodaj) na sistemih, ki dovoljujejo prevajanje programov, ki uporabljajo ti funkciji wcsr in mbsr (Linux, Solaris, OSF/1).

Podobno svetuje tudi Sun v *http://www.sun.com/software/white-papers/wp-unicode/*, v razdelku "Internationalized Applications with Unicode":

Za pravilno lokalizacijo programov upoštevajte naslednje napotke:

1. *Izogibajte se neposrednemu dostopu z Unicode. To je opravilo za lokalizacijski del vašega sistema.*
2. *Uporabljajte model POSIX za vmesnik do večzložnih in "širokihžnakov.*
3. *Uporabljajte samo programerske vmesnike, ki jih pozna lokalizacijski okvir za operacije z jezikovnimi in kulturnimi posebnostmi.*
4. *Ostanite neodvisni od nabora znakov.*

Če iz kakršnega koli razloga v delu programa zares morate predpostaviti, da je 'wchar_t' Unicode (npr. če želite posebno obdelavo nekaterih znakov iz Unicode), morate ta kos programa narediti pogojno odvisen od rezultata funkcije `is_locale_utf8()`. Sicer boste pomešali obnašanje programa v različnih localih in sistemih. Funkcija `is_locale_utf8` je deklarirana v *utf8locale.h* in definirana v *utf8locale.c*.

Knjižnica libutf8 Prenosljiva izvedba programerskega vmesnika ISO/ANSI C, ki podpira 8-bitne in UTF-8 locale, se nahaja v *libutf8-0.7.3.tar.gz*.

Prednosti:

- Prenosljiva podpora za Unicode UTF-8 sedaj tudi pri operacijskih sistemih, kjer podpora za večzložne znake ne deluje ali celo sploh nimajo tovrstne podpore.
- Ista binarna koda deluje v vseh operacijskih sistemih s podporo za 8-bitne in UTF-8 locale.
- Ko proizvajalec operacijskega sistema doda ustrezno podporo za večzložne znake, lahko to izkoristite z enostavnim ponovnim prevajanjem brez uporabe stikala -DHAVE_LIBUTF8.

Rešitev pri Plan9 Operacijski sistem Plan9 (zvrst Unixa) uporablja UTF-8 za kodiranje znakov v vseh programih. Njegov tip za "širokežnake se imenuje '`Rune`' in ne '`wchar_t`'. Deli njegove knjižnice, avtor je Rob Pike, so dostopni na *ftp://ftp.cdrom.com/pub/netlib/research/9libs/9libs-1.0.tar.gz*. Podobna knjižnica avtorja Alistairja G. Crooksa se nahaja na *ftp://ftp.cdrom.com/pub/NetBSD/packages/distfiles/libutf-2.10.tar.gz*. Obe knjižnici vsebujeta iskalnik po regularnih izrazih, ki se zaveda UTF-8.

Pomanjkljivost tega programerskega vmesnika:

- UTF-8 je že vgrajen vanj, ne pa ponujen kot možnost. Programi, ki jih prevedete z njim, izgubijo podporo za 8-bitne nabore znakov, ki so še vedno pogosti v Evropi.

6.1.2 Za grafični uporabniški vmesnik

Knjižnica Qt-2.0 (*http://www.troll.no/*) vsebuje razred QString, ki pokriva celoten Unicode. Za pretvarjanje v/iz UTF-8 lahko uporabljate članski funkciji QString::utf8 in QString::fromUtf8. Funkcij QString::ascii in QString::latin1 se ne bi smelo več uporabljati.

6.1.3 Za naprednejše ravnanje s tekstom

Prej omenjene knjižnice izvedejo Unicoda zavedajoče se različice konceptov ASCII. Sedaj pa si oglejmo še knjižnice, ki imajo opravka s konceptom Unicode, npr. tretji tip črk (različen od malih in velikih črk), ločevanje med slovničnimi znamenji in simboli, kanonična razgradnja, kombiniranje razredov, kanonično urejanje ipd.

ucdata-2.4

Knjižnica ucdata (*http://crl.nmsu.edu/˜mleisher/ucdata.html*) avtorja Marka Leisherja se ukvarja z lastnostmi znakov, pretvorbo med malimi in velikim črkami, razstavljanjem in kombiniranjem razredov. Pridružen mu je tudi paket ure-0.5 (*http://crl.nmsu.edu/˜mleisher/ure-0.5.tar.gz*) za izvajanje regularnih izrazov pod Unicode.

ustring

Knižnica za C++ ustring (*http://ustring.charabia.net/*) avtorja Rodriga Reyesa se ukvarja z lastnostmi znakov, pretvorbo med velikimi in malimi črkami, razstavljanjem, kombiniranjem razredov in vključuje še izvajalnik regularnih izrazov za Unicode.

ICU

ICU pomeni International Components for Unicode (*http://oss.software.ibm.com/icu/*, glejte tudi *http://oss.software.ibm.com/icu/icuhtml/API1.5/*). IBMova zelo obsežna knjižnica za internacionalizacijo s podporo za nize po Unicode, združevanje sredstev, oblikovanje številk, časa, datuma, sporočil, povzemanje itd. Podprtih je veliko localov. Prenosljiva je na Unix in Win32, vendar se do konca (brez potrebnih naknadnih popravkov) prevede le pri libc6 na Linuxu, ne pa pri libc5.

libunicode

Gnomova knjižnica libunicode (*http://cvs.gnome.org/lxr/source/libunicode/*) avtorja Toma Tromeya in ostalih pokriva pretvorbo med nabori znakov, lastnosti znakov in razstavljanje.

6.1.4 Za pretvarjanje

Na voljo sta dve vrsti knjižnic za pretvorbo, ki podpirajo UTF-8 in precej 8-bitnih naborov znakov:

iconv Izvedba knjižnice avtorja Ulricha Drepperja se nahaja v GNU glibc-2.1.3 (*ftp://ftp.gnu.org/pub/gnu/glibc/glibc-2.1.3.tar.gz*). Priročniške strani se nahajajo na *ftp://ftp.win.tue.nl/pub/linux-local/manpages/man-pages-1.29.tar.gz*.

Prenosljiva izvedba avtorja Bruna Haibla se nahaja na *ftp://ftp.ilog.fr/pub/Users/haible/gnu/libiconv-1.3.tar.gz*.

Prenosljiva izvedba avtorja Konstantina Čugejeva (*joy@urc.ac.ru*) pa je na voljo na *ftp://ftp.urc.ac.ru/pub/local/OS/Unix/converters/iconv-0.4.tar.gz*.

Prednosti:

- Knjižnica iconv je standardizirana po POSIX, programi, ki jo uporabljajo za pretvarjanje v/iz UTF-8, bodo delovali tudi pod Solarisom. Vendar pa se imena naborov znakov nekoliko razlikujejo med različnimi operacijskimi sistemi, npr. "EUC-JP"pod glibc je "eucJP"pod HP-UX (uradno ime IANA za ta nabor znakov je "EUC-JP", tako da je to nedvomno zmota HP-UX).
- Na sistemih z glibc-2.1 ni potrebna nobena dodatna knjižnica. Na ostalih sistemih je potrebna ena izmed ostalih dveh izvedb knjižnice.

librecode Librecode avtorja Françoisa Pinarda se dobi na naslovu *ftp://ftp.gnu.org/pub/gnu/recode/recode-3.5.tar.gz*.

Prednosti:

- Podpora za transliteracijo, t.j. pretvorba ne ASCII znakov v zaporedja znakov ASCII, da se ohrani berljivost tudi, kadar ni mogoča pretvorba brez izgub.

Pomanjkljivosti:

- Nestandarden programerski vmesnik
- Počasna inicializacija.

ICU ICU pomeni International Components for Unicode in se ga dobi na *http://oss.software.ibm.com/icu/* (glejte tudi *http://oss.software.ibm.com/icu/icuhtml/API1.5/*). IBMova knjižnica za internacionalizacijo ima tudi orodja za pretvorbo, ki so deklarirana v '`ucnv.h`'.

Prednosti:

- Obsežen nabor podprtih kodiranj.

Pomanjkljivosti:

- Nestandarden programerski vmesnik

6.1.5 Ostali pristopi

libutf-8

Libutf-8 avtorja G. Adama Stanislava (*adam@whizkidtech.net*) vsebuje nove funkcije za sprotno pretvorbo v/iz po UTF-8 kodiranih tokov 'FILE*'. Dobite jo na *http://www.whizkidtech.net/i18n/libutf-8-1.0.tar.gz*.

Prednosti:

- Zelo majhna

Pomanjkljivosti:

- Nestandarden programerski vmesnik
- UTF-8 je že vgrajen vanj in ni dodan kot možnost. Programi, ki se prevedejo s to knjižnico, izgubijo podporo za 8-bitna kodiranja, ki so še vedno pogosta v Evropi.
- Namestitev ni enostavna, Makefile je potrebno ročno nastaviti, ni samodejnega nastavljanja

6.2 Java

Java ima podporo za Unicode že vgrajeno v jezik. Tip 'char' predstavlja znak po Unicode in razred 'java.lang.String' predstavlja niz, sestavljen iz znakov Unicode.

Java lahko skozi svoj okenski sistem AWT prikaže katerikoli znak iz Unicode, če: 1. ustrezno nastavite Javino sistemsko lastnost "user.language", 2. so definicije pisav /usr/lib/java/lib/font.properties.*jezik* ustrezno nastavljene, 3. so nameščene pisave, ki so določene v tej datoteki. Da lahko npr. prikažete japonske znake, morate najprej namestiti japonske pisave in pognati "java -Duser.language=ja ...". Mogoče je kombinirati nabore pisav: če želite hkrati prikazati zahodnoevropske, grške in japonske znake, morate narediti kombinacijo datotek "font.properties"(pokriva ISO-8859-1), "font.properties.el"(pokriva ISO-8859-7) in "font.properties.ja"v eno datotek. ??To ni preizkušeno??

Vmesnika java.io.DataInput in java.io.DataOutput vsebujeta metodi 'readUTF' in 'writeUTF'. Upoštevajte pa, da ne uporabljata UTF-8. Namesto tega uporabljata spremenjeno kodiranje UTF-8: znak NUL je kodiran kot dvozložno zaporedje 0xC0 0x80 namesto 0x00 in zlog 0x00 se doda na koncu. Tako kodirani nizi lahko vsebujejo znake NUL in kljub vsemu ni potrebno, da se jim na začetek doda polje length. Zato lahko z njimi manipulirajo tudi Cjevske funkcije iz <string.h>, npr. strlen() in strcpy().

6.3 Lisp

Standard Common Lisp določa dva znakovna tipa: ‘base-char’ in ‘character’. Podpora za Unicode je odvisna od izvedbe. Jezik določa tudi parameter za ključno besedo ‘:external-format’ v ‘open’ kot logično mesto za določanje nabora znakov ali kodiranja.

Med brezplačnimi izvedbami Common Lispa samo CLISP (*http://clisp.cons.org/*) podpira Unicode. Potrebujete različico CLISPA z marca 2000 ali novejšo (*ftp://clisp.cons.org/pub/lisp/clisp/source/clispsrc.tar.gz*). Tipa ‘base-char’ in ‘character’ sta ekvivalentna 16-bitnemu Unicode. Funkciji `char-width` in `string-width` omogočata programerski vmesnik primerljiv z `wcwidth()` in `wcswidth()`. Kodiranje za vhodno izhodne operacije datotek, cevi ali mrežnih priključkov se lahko določi preko parametra ‘:external-format’. Kodiranja za tty in privzeta kodiranja za datoteke, cevi oz.mrežne priključke so odvisna od locale.

Med komercialnimi izvedbami Common Lispa:

LispWorks (*http://www.xanalys.com/software_tools/products/*) podpira Unicode. Tip ‘base-char’ je ekvivalenten ISO-8859-1, tip ‘simple-char’ (podtip tipa ‘character’) vsebuje vse znake iz Unicode. Kodiranje za vhod/izhod datotečnih operacij se lahko določi preko parametra ‘:external-format’, npr. ’(:UTF-8)’. Omejitve: kodiranje se ne more uporabljati za vhodno izhodne operacije omrežnih priključkov. Z urejevalnikom ni mogoče urejati po UTF-8 kodiranih datotek.

Eclipse (*http://www.elwood.com/eclipse/eclipse.htm*) podpira Unicode. Oglejte si *http://www.elwood.com/eclipse/char.htm*. Tip ‘base-char’ je enakovreden ISO-8859-1, tip ‘character’ pa vsebuje vse znake Unicode. Kodiranje za vhodno izhodne operacije nad datotekami se lahko določi s kombinacijo parametrov ‘:element-type’ in ‘:external-format’ k ‘open’. Omejitve: funkcije za lastnosti znakov so odvisne od locale. Izvorne in prevedene izvorne datoteke ne morejo vsebovati konstantnih nizov z znaki iz Unicode.

Komercialna izvedba Common Lispa Allegro CL bo vsebovala podporo za Unicode v prihajajoči različici 6.0.

6.4 Ada95

Ada 95 je bila izdelana za podporo Unicode in standardna knjižnica Ada95 pozna posebna podatkovna tipa Wide_Character in Wide_String za ISO 10646-1, kot tudi številne pridružene funkcije. Prevajalnik GNU Ada95 (gnat-3.11 ali novejši) podpira UTF-8 kot zunaje kodiranje širokih znakov. To vam omogoča uporabo UTF-8 tako v izvorni kodi kot tudi v vhodno/izhodnih operacijah programov. Da v programu aktivirate to možnost, uporabite "WCEM=8"v nizu FORM, ko odpirate datoteko in vklopite prevajalnikovo stikalo -gnatW8", če je izvorna koda kodirana po UTF-8. Za podrobnosti si oglejte referenčna priročnika za GNAT (*ftp://cs.nyu.edu/pub/gnat/*) in Ada95 (*ftp://ftp.cnam.fr/pub/Ada/PAL/userdocs/docadalt/rm95/index.htm*).

6.5 Python

Python 2.0 (*http://starship.python.net/crew/amk/python/writing/new-python/new-python.html*) bo vseboval podporo za Unicode. Imel bo tudi podatkovni tip ‘unicode’, ki bo predstavljal niz v Unicode, modul ‘unicodedata’ za lastnosti znakov in nabor pretvornikov za najpomembnejša kodiranja. Za podrobnosti si oglejte *http://starship.python.net/crew/lemburg/unicode-proposal.txt*.

6.6 JavaScript/ECMAscript

Od Javascripta različice 1.3 naprej so nizi vedno kodirani po Unicode. Znakovnega tipa ni, lahko pa uporabljate označbo \uXXXX za znake iz Unicode znotraj nizov. Ne opravi se nobena notranja normalizacija, zato se pričakuje sprejem Unicode Normalization Form C, ki ga priporoča W3C. Za podrobnosti si oglejte *http://developer.netscape.com/docs/manuals/communicator/jsref/js13.html#Unicode* in *http://developer.netscape.com/docs/javascript/e262-pdf.pdf* za popolno specifikacijo ECMAscript.

6.7 Tcl

Tcl/Tk je pričel uoporabljati Unicode kot osnovni nabor znakov v različici 8.1. Njegova notranja predstavitev nizov je UTF-8. Podpira označevanje \uXXXX za znake iz Unicode. Oglejte si *http://dev.scriptics.com/doc/howto/i18n.html*.

6.8 Perl

Perl 5.6 notranje shranjuje nize v zapisu UTF-8, če na začetku skripta vnesete

```
use utf8;
```

`length()` vrne število znakov v nizu. Za podrobnosti si oglejte Perl-i18n FAQ na naslovu *http://rf.net/˜james/perli18n.html*.

7 Ostali viri informacij

7.1 Dopisni seznami

Širok krog občinstva lahko dosežete na naslednjih dopisnih seznamih.

Upoštevajte, da morate tam, kjer stoji 'na', dejansko vnesti '@' (zaradi preprečevanja nezaželenih sporočil)

7.1.1 linux-utf8

Naslov: `linux-utf8` na `nl.linux.org`

Ta dopisni seznam je namenjen internacionalizaciji z Unicode in zajema široko področje tem od gonilnikov za tipkovnico do pisav za X11.

Arhivi se nahajajo na naslovu *http://mail.nl.linux.org/linux-utf8/*.

Vpišete se tako, da na naslov `majordomo` na `nl.linux.org` pošljete sporočilo z vrstico "subscribe linux-utf8"v telesu sporočila.

7.1.2 li18nux

Naslov: `linux-i18n` na `sun.com`

Ta dopisni seznam se osredotoča na organiziranje dela pri internacionalizaciji Linuxa in urejanju sestankov med ljudmi.

Vpišete se tako, da na naslovu http://www.li18nux.org/ izpolnite obrazec in ga pošljete na naslov `linux-i18n-request` na `sun.com`.

7.1.3 unicode

Naslov: `unicode` na `unicode.org`

Ta dopisni seznam se osredotoča na standardizacijo, nadaljnji razvoj standarda Unicode in s tem povezane tehnologije, kot npr. Bidi in algoritmi za sortiranje.

Arhivi se nahajajo na naslovu *ftp://ftp.unicode.org/Public/MailArchive/*, vendar se ne posodabljajo redno.

Za informacije o vpisu si oglejte *http://www.unicode.org/unicode/consortium/distlist.html*.

7.1.4 Internacionalizacija X11

Naslov: `i18n` na `xfree86.org`

Ta dopisni seznam je namenjen ljudem, ki se ukvarjajo z boljšo internacionalizacijo sistema X11/XFree86.

Arhivi se nahajajo na naslovu *http://devel.xfree86.org/archives/i18n/*.

Vpišete se tako, da pošljete sporočilo prijazni osebi na naslovu `i18n-request` na `xfree86.org` z razlago vaše motivacije.

7.1.5 Pisave za X11

Naslov: `fonts` na `xfree86.org`

Ta dopisni seznam je namenjen ljudem, ki se ukvarjajo s pisavami Unicode in podsistemom pisav za sistem X11/XFree86.

Arhivi se nahajajo na *http://devel.xfree86.org/archives/fonts/*.

Vpišete se tako, da prezaposleni osebi na naslovu `fonts-request` na `xfree86.org` pošljete sporočilo z obrazložitvijo vaše motivacije.