

ConvPL 2.3

Uniwersalny konwerter polskich znaków

ConvPL jest programem freeware, rozprowadzany wraz z kodem źródłowym. To co go wyróżnia od innych tego typu programów, to liczba obsługiwanych standardów, dowolność kierunków konwersji i maksymalnie uproszczona - choć dająca wiele możliwości - linia poleceń.

SKŁADNIA

```
cpl <nn><konwersja> [<wejscie.txt>] [<wyjscie.txt>]
```

Parametr <nn> to dwie cyfry/litery, określające standard źródłowy (pierwsza) i docelowy (druga) przeprowadzanej konwersji.

Parametr <konwersja> określa rodzaj opcjonalnie stosowanej konwersji zakończenia linii pomiędzy DOSem a Unixem i Amigą.

Parametry „wejscie.txt” i „wyjscie.txt” to nazwy plików odpowiednio źródłowego i docelowego.

Podanie tylko jednej nazwy - pliku wejściowego - spowoduje zapisanie tekstu po konwersji pod tą samą nazwą.



OBSŁUGIWANE STANDARDY

Dostępne standardy (podajemy dwie cyfry/litery):

- ✓ 0: bez polskich liter (ASCII,CP473)
- ✓ 1: Mazovia
- ✓ 2: FidoMazovia
- ✓ 3: IBM-Latin/2 (CP852)
- ✓ 4: ISO-Latin/2 (ISO-8859/2)
- ✓ 5: DHN
- ✓ 6: CSK
- ✓ 7: Cyfromat
- ✓ 8: IEA
- ✓ 9: LOGIC
- ✓ a: Microvex
- ✓ b: <Amiga> Amiga PL
- ✓ c: CP/J (Elwro 800 Junior)
- ✓ d: <Amiga> FAT
- ✓ e: Windows 3.0 (CP1250)
- ✓ f: <Amiga> XJP
- ✓ g: Ventura

- ✓ h: Macintosh
- ✓ i: TeX PL
- ✓ j: Corel 2.0

Skróty

Zamiast najczęściej stosowanych konwersji można stosować skróty. Podajemy wówczas tylko jedną literę:

- ✓ l: Mazovia na IBM-Latin/2
- ✓ m: IBM-Latin/2 na Mazovię
- ✓ x: Mazovia na ASCII
- ✓ y: IBM-Latin/2 na ASCII
- ✓ z: Fido-Mazovia na Mazovię

Dowolność konwersji jest całkowita, z jednym wyjątkiem. Standard ASCII nie może być standardem źródłowym konwersji (przyczyna jest oczywista - żeby nie pokaszaniĆ całego tekstu - ale jest od tej reguły wyjątek! patrz UWAGI).

Konwersja zakończeń linii

ConvPL umożliwia równoległą konwersję pomiędzy tekstem spod DOSa, a tekstem spod Unixa czy Amigi. Rodzaj docelowej konwersji określają parametry:

- ✓ -: konwersja do tekstu Unixa (Amigi) -> NL
- ✓ =: konwersja do tekstu DOSa -> CR+LF

Zmieniłem występujące w poprzednich wersjach oznaczenia tych parametrów (/,\), ponieważ bekslesz jest znakiem specjalnym pod U*xem. Nowe ustawienia można równie łatwo zapamiętać: pod U*xem koniec linii to jeden znak (NL), więc jedna kreska: „-“; pod DOSem są to dwa znaki (CR+LF), stąd dwie kreski: „=” . ;)

Patrz również rozdział Uwagi, akapit Konwersja Unix<->DOS.

PRACA W TRYBIE FILTRU

ConvPL może pracować również w trybie filtru. Brak nazw plików we/wy spowoduje pobranie tekstu z standardowego urządzenia wejścia - stdin (klawiatura) i zapisanie go do standardowego wyjścia - stdout (monitor, terminal, drukarka). Przykłady:

```
cpl l
```

każda linia wpisywana z klawiatury będzie pojawiała się poniżej po wciśnięciu <Enter>, znaki z kodu Mazovia będą zamieniane na IBM Latin 2

```
cpl l <joke.txt >prn
```

konwersja pliku joke.txt i wysłanie wyników bezpośrednio na drukarkę (DOS);

```
cpl | <joke.txt >joke2.txt
```

równoznaczne poleceniu „cpl | joke.txt”

Jeśli nie chcemy zaśmiecać sobie dysku, wydajemy następujące polecenia:

```
cpl | <joke.txt >nul (MSDOS) lub >/dev/null (Unix)
```

Tryb filtru umożliwia również przetwarzanie tekstu w potoku (pipe).

Przykład:

```
type joke1.txt | cpl 14- >joke3.txt (MSDOS)
```

```
cat joke.txt1 joke2.txt | cpl14- >joke3.txt (Unix)
```

Przyjmując, że pliki joke.txt i joke2.txt są napisane oryginalnie w Mazovii i pod MSDOSem, takie polecenie spowoduje połączenie obu plików, konwersję z Mazovii na ISO-Latin-2, i przy okazji z CR+LF -> NL, a następnie zapis do pliku joke3.txt.

Uruchamiając ConvPL w tym trybie należy uważać na pomyłki takie jak ta:

```
cpl | joke.txt >prn
```

Tekst zostanie przekonwertowany zgodnie z regułą 13 (skrót „l”), ale na drukarkę nie zostanie wysłany, gdyż ConvPL będzie go zapisywał do pliku joke.txt a nie na standardowe wyjście. W rzeczywistości powyższe (błędne) polecenie oznacza: konwertuj plik joke.txt, zapisując wynik pod tą samą nazwą, zaś na drukarkę wysyłaj wyjście przez program normalnie wyświetlane na ekranie, czyli komunikaty.

Polecenie poniższe spowoduje najprawdopodobniej uszkodzenie oryginalnego pliku, więc należy się go wystrzeżać :

```
cpl | <joke.txt >joke.txt (obie nazwy takie same!)
```

Po prostu, system straci głowę próbując równocześnie czytać i pisać do pliku, a ucierpi na tym tenże plik.

Pamiętajmy: aby uruchomić ConvPL w trybie filtru, potok danych musi przychodzić ze stdin i wychodzić na stdout (a nie częściowo z pliku, częściowo z potoku). Jeśli nie wiesz o co chodzi, to poczytaj o potokach, filtrach oraz roli symboli <, |, > itp. w poleceniach DOSa i Unixa.

UWAGI

Komentarz do standardów

Za definicję Amiga PL dziękuję Marcinowi Kądziołce <2:484/15.16@fidonet>.

Definicje dodane w wersji 1.75 udostępnił mi Edwin Wierszelis <2:482/16@fidonet>, autor konwertera KPN.

„Standard” ASCII to oczywiście zwykłe literki ASCII, zastępujące odpowiednie polskie znaki. Taki tekst zwie się zwykle „polskawym”, a standard określa jako CP473 (podstawowa, amerykańska strona kodowa).

Fido-Mazovia jest półformalnym standardem, przyjętym w środowisku sieci Fido. Jedyna różnica w stosunku do klasycznej Mazovii to zamiana znaku c' (141) na c-cedilla (135), wynikająca ze względów technicznych.

Standard IBM-Latin/2 (CP852), rozpowszechniony głównie w środowisku DOSa, jest lansowany m.in. przez MicroSoft.

ISO-8859/2, znany też jako ISO-Latin/2 jest standardem rozpowszechnionym zwłaszcza w systemach unixowych, pracujących pod X-windowsami. Jest również standardem zalecanym przez Polskie Normy oraz uznanym za obowiązujący w polskojęzycznym WWW.

Standard FAT jest jednym z amigowskich dialektów, znalezionym w dyskmagu FAT przez Edwina Wierszelisa.

Standard Windows 3.0 jest uproszczoną, 8-bitową wersją UniCode, zdefiniowaną jako strona kodowa 1250 (CP1250). Znany też jako Windows-EE, czyli strona kodowa dla Europy Wschodniej.

Standard DHN jest rozpowszechniony wraz z polskim ChiWriterem.

Konwersja tekstu Unix<->DOS

Pod Unixem każda linijka tekstu kończy się znakiem NL (new line) o kodzie 0xa, równoznaczny DOSowemu LF (line feed). Pod DOSem zakończenie linii stanowi para znaków CR+LF, czyli 0x0d,0x0a. Parametr „-” wymusza zakończenie linii znakiem NL, zaś „=” - parą CR+LF. Brak określonego kierunku konwersji nie spowoduje żadnej zmiany końców linii w pliku wynikowym. Parametry konwersji zakończeń linii dodaje się „na trzeciego” do parametrów konwersji standardów.

UWAGA: od wersji 2.01 istnieje możliwość wykonywania konwersji zakończeń linii bez zmiany standardu polskich znaków. Wystarczy jako standard źródłowy podać ASCII - „0” (do tej pory było to niedopuszczalne), zaś jako drugi parametr opcję konwersji zakończeń linii „-” lub „=”. Drugi parametr, podany zgodnie z konwencją i określający standard docelowy jest w tym wypadku ignorowany. A zatem, każdy z poniższych przykładów ma takie samo działanie:

```
cpl 0- joke.txt
```

```
cpl 00- joke.txt
```

```
cpl 02- joke.txt    („2” jest tu ignorowane)
```

czyli konwersję zakończeń linii do pojedynczych NL, bez naruszania polskich liter.

Kod źródłowy

Program w wersji 2.02b był z powodzeniem kompilowany Borlandem 3.1, gcc 2.7.0 pod Linuxem (a.out) oraz DJGPP pod DOSem (gcc 2.6.3). W wersji 2.3 kompiluje się poprawnie pod Linuxem oraz Visual Studio 2005.

Licznik linii

Typowy „wodotrysk”. Zwalnia pracę o ok. 30%. Standardowo wyłączony, w celu włączenia trzeba przekompilować źródła bez opcji NO_LINE_COUNT. Licznik jest zawsze wyłączony podczas pracy w trybie stdin/stdout (czyli zazwyczaj jako filtr, CGI itp.).

CGI

Począwszy od wersji 2.02 można CPL skompilować jako CGI, czyli program współpracujący z serwerem WWW. W tej wersji CPL traci niepotrzebne opcje oraz modyfikacji ulega sposób wywoływania. Składnia CPL-CGI jest następująca:

```
cpl <nn><T | H><nazwa.pliku> (bez odstępów!)
```

Gdzie nn to standard źródłowy i docelowy (bez zmian) a T/H określa typ tekstu wg standardu MIME:

H: *text/html* (tekst źródłowy HTML)

T: *text/plain* (każdy inny tekst)

Przykład:

cpl 4eHplik.html = konwersja ISO-Latin do Windows, plik typu HTML

cpl 41Tplik.txt = ISO-Latin do Mazovii, zwykły plik tekstowy

Wywołanie z dokumentów HTML (w formacie URLa) powinno wyglądać np. tak:

<http://host.domena.pl/cgi-bin/cpl?3eHplik.html>

Inne opcje

- ✓ -?, -h,/? , --help: wyświetlają listę wszystkich dostępnych parametrów, składnię linii komend itp.
- ✓ -V, --version: wyświetla (na stdout) numer wersji CPL -L, --licencja: wyświetlaną licencję użytkownika programu.
- ✓ -S, --standard: wyświetlają tabelkę wszystkich obsługiwanych standardów polskich znaków.

HISTORIA

- 1.0 pierwsza wersja (quick fix dla pakietu PGP-pl);
- 1.2 definicje standardów przeniesione do polish.c; parę nowych definicji;
- 1.3 Amiga PL; nowy układ polish.c;
- 1.5 nieco przyspieszona praca; nowy kod źródłowy, makefile;
- 1.6 skróty dla najczęściej używanych konwersji; tryb stdin/stdout (filtr)
- 1.7 konwersja między tekstem unixowym (amigowym) i DOSowym
- 1.72 licznik linii

- 1.75 nowe definicje (Win 3.0, FAT, XJP, CP/J);
- 1.76 nowe definicje (Ventura, Macintosh, TeX PL i Corel);
- 1.90 dodana konwersja wielu plików, pare innych zmian;
- 2.00 napisany w 90% od nowa, usunięte stare pluskwy, wprowadzone nowe ;)
dodane rozpoznawanie standardow
- 2.01 kilka dodatków, zmiany w interpretacji niektórych poleceń (patrz Uwagi)
- 2.02 możliwość skompilowania jako CGI; umożliwia to wywoływanie CPL przez
serwer WWW i konwersję tekstów dokonywaną on-line (patrz Uwagi)
_ta wersja nie była przeze mnie jeszcze dostatecznie przetestowana,
używaj raczej 2.01 jeśli nie potrzebujesz nowych funkcji_
- 2.3 wersja odkurzona po 12 latach

QueryPL 2.3

Narzędzie do rozpoznawania standardów polskich znaków

QueryPL jest programem freeware, rozprowadzany wraz z kodem źródłowym. Służy do rozpoznawania standardu kodowania polskich znaków w plikach tekstowych, jeśli standard ten jest nieznan lub niepewny.

SKŁADNIA

qpl <plik> [-s | -v | -b]

Jako pierwszy argument podaje się zawsze plik który ma być sprawdzony.

Po nazwie pliku mogą następować dwa opcjonalne argumenty:

- ✓ -s użycie metody statystycznej, generalnie pewniejszej i koniecznej w niektórych wypadkach; patrz rozdział ALGORYTM i UWAGI nr 3.
- ✓ -v wyświetlenie szczegółowych informacji o pracy programu, tabeli zgodności itp.
- ✓ -b podanie wyniku analizy w postaci parametru dla programu ConvPL - cyfry lub litery; opcji tej można użyć do automatycznego konwertowania plików o nieznanym standardzie, podstawiając uzyskany znak do linii poleceń CPL jako

standard źródłowy; patrz dokumentacja ConvPL; opcja -b automatycznie wyłącza opcję -v

ALGORYTM

Zasada działania jest następująca: zliczane są wszystkie wystąpienia wszystkich polskich znaków narodowych. Ponieważ znaki powtarzają się w różnych standardach, każde wystąpienie jest zliczane oddzielnie dla każdego ze standardów. Następnie wystąpienia wszystkich znaków są sumowane w obrębie standardów. Wygrywa standard, w którym zmieściło się więcej znaków znalezionych w tekście.

Może się zdarzyć że w tekście będą występować znaki nie mające funkcji znaków diakrytycznych (np. jako elementy tabelki), ale należące do któregoś z licznych standardów. W pewnych przypadkach mogą one zmienić wynik analizy zwykłych zliczeń, prowadząc do złego wyboru standardu. Dla zapobieżenia temu QPL może dodatkowo wykorzystać wiedzę o tym, jak często każdy z znaków diakrytycznych występuje w normalnym, polskim tekście. Metodę statystyczną, wykorzystującą tę funkcję, włącza się opcją -s.

W tym wypadku, przed zsumowaniem zliczeń każde z nich zostanie pomnożone przez odpowiedni współczynnik, wskazujący z jaką częstotliwością dana litera zwykle występuje. Tabelę zliczeń znaków dla obu metod wyświetla opcja -v.

Opcji -s należy używać do tekstów, w których poza normalnymi polskimi znakami występują inne znaki, mogące być znakami diakrytycznymi w jakimś innym standardzie. Warto też jej spróbować, gdy metoda standardowa daje w wyniku jakiś rzadki i niespotykany standard.

UWAGI

- ✓ Winiетка programu i wszystkie informacje są wyświetlane na stderr, dzięki czemu są „niewidzialne” przy zrzucaniu wyjścia do pliku, filtrowaniu itp. Tabela zgodności (-v) i wynik - parametr dla ConvPL (-b) są natomiast wyświetlane na stdout (nigdy razem, bo jedno wyklucza drugie).
- ✓ QPL jest rozprowadzany razem z konwerterem ConvPL. Zdecydowałem się na rozprowadzanie go w postaci odrębnego programu z powodów praktycznych. Jako taki, może on mieć większe możliwości, więcej zastosowań i bogatszą linię poleceń. Współpracę między obydwoma programami - w celu np. automatyzacji konwersji można dość łatwo zapewnić sobie za pomocą skryptów (U*x) lub baczów (MSDOS/4DOS).
- ✓ Rozróżnienie standardów DHN i Logic jest trudne, bo składają się one z tych samych znaków, a różnica między nimi polega na tym, że cztery są zamienione kolejnością. QPL zazwyczaj trafnie rozpoznaje różnicę przy użyciu metody statystycznej (-s), jednak w wyniku podaje na wszelki wypadek obie możliwości. Druga z nich jest mniej prawdopodobna, ale teoretycznie jest możliwa.

Historia

1.0 pierwsza wersja

QueryPL jest rozprowadzany na tych samych warunkach co ConvPL.

Są one wyświetlane po wydaniu polecenia: `cpl—licencja`.

Autor

Pawel Krawczyk <pawel.krawczyk@hush.com>

Najnowsza wersja : <http://ipsec.pl/cpl>