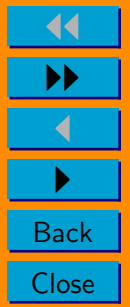
Bachotek 2001

Języki dalekowschodnie i pakiet CJK

Michał Piskorski

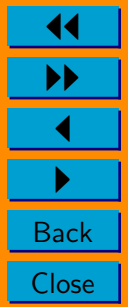
Uniwersytet Warszawski, Instytut Orientalistyczny
Zakład Japonistyki i Koreanistyki,Sekcja Koreanistyki
Zakład Zastosowań Informatycznych
email: micpis@mercury.ci.uw.edu.pl



Pakiet CJK

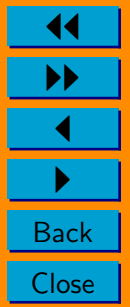
C – chiński
J – japoński
K – koreański

Autorem pakietu jest Werner Lemberg.



Cechy pakietu

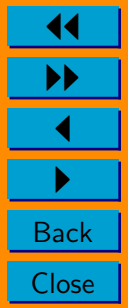
- wspomaga skład tekstów w językach orientalnych przy użyciu LaTeX 2_ε
- akceptuje większość metod kodowania powszechnie stosowanych na Dalekim Wschodzie
- udostępnia mechanizmy typowe dla konwencji typograficznych w krajach kręgu CJK
- definiuje możliwość uzyskania odmiany grubej fontu przy braku rzeczywistego fontu CJK
- umożliwia osadzanie znaków innych niż znaki CJK wewnątrz środowiska CJK
- współpracuje z pakietami polonizacyjnymi



Przetwarzanie dokumentów CJK

W zależności od rodzaju kodowania przetwarzanie tekstów z reguły przeprowadza się w jeden z podanych sposobów:

- w trybie z przetwarzaniem wstępnym
- w trybie bezpośrednim (bez procesu konwersji)



Dokumenty, które nie wymagają przetwarzania wstępnego

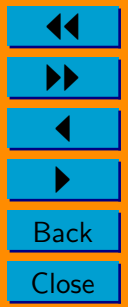
– kodowane zgodne z EUC (*Extended UNIX Code*, 8-bitowe kodowanie zgodne ze standardem ISO) dla języków CJK:

  * `EUC-GB` (również `GBt`), `EUC-TW` (`CNS1` - `CNS7` z tzw. *single shift*) – dla języka chińskiego
  * `EUC-JP` (również z *single shift*) – dla języka japońskiego
  * `EUC-KR` – dla języka koreańskiego

– `UTF-8` – format transformacji kodowania Unicode

Dokumenty, które wymagają przetwarzania wstępnego

- `Big5`, `Big5+`, `GBK` – dla języka chińskiego
- `SJIS` – dla języka japońskiego
- `emacs-mule` – format wielojęzyczny
- `CEF` – *Chinese Encoding Framework*



CEF to metoda kodowania która:

- została opracowana przez Christiana Witterna (`cwitter@gwdg.de`) dla potrzeb Research Institute for Zen Buddism (IRIZ) w Kyoto
- wykorzystuje makra zbudowane w standardzie SGML
- służy do osadzania znaków CJK w tekście zapisanym w dowolnym kodowaniu
- może służyć do osadzania rzadko spotykanych znaków CJK – obsługuje znaki w kodowaniu zdefiniowanym przez użytkownika, tzw. kodowaniu prywatnym (*private encoding*)
- współpracuje z *KanjiBase for Windows*, który umożliwia edycję w tym formacie (rozprowadzany przez IRIZ, `http://www.iijnet.or.jp/iriz/irizhtml/irizhome.htm`)





Back
Close

Mechanizmy wspomagające konwencje typograficzne w krajach kręgu CJK

– definiowanie środowiska, które w określony sposób traktują odstęp:

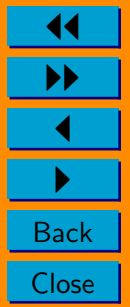
- `CJK*` – dla języka japońskiego i chińskiego: usuwanie wszekich odstępów (oprócz tzw. *shibuaki*) oraz znaków nowej linii, jakie wystąpią po znaku CJK
- `CJK` – dla języka koreańskiego: typowy (europejski) sposób traktowania odstępów w LaTeXu

– elementy typowe dla składu japońskiego:

- *shibuaki* – spacja o szerokości 1/4 ideogramu chińskiego do oddzielania osadzonego tekstu innego niż CJK
- *furigana* – zapis czytania ideogramu chińskiego ponad opisywanym znakiem

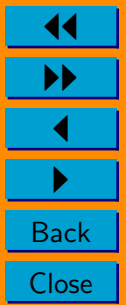
Pakiet CJK wykorzystuje sposób wyboru fontów zgodny z NFSS

- umożliwia to ładowanie fontów tylko w razie potrzeby (uwaga: w przypadku korzystania z fontów `pmC` załadowane zostaną wszystkie fonty składowe)
- możliwość jednoczesnego korzystania z 256 fontów (od wersji web2c 7.2 do 1000 fontów)
- pliki definiujące fonty `.fd` zawierają deklaracje typowe dla NFSS
- możliwość korzystania z nowych poleceń określających wysokość fontu: `CJK`, `sCJK`, `CJKfixed`, `sCJKfixed`, `CJKsub`, `CJKssub`
- definicja tzw. *poor man's boldface*, czyli sztucznie uzyskanej odmiany grubej, przy braku rzeczywistego fontu: `CJKb`, `sCJKb`, `CJKfixedb`, `sCJKfixedb`, `DNPb`, `DNPgenb`



Definicja kodu fontu zgodnie z NFSS

Kodowanie fontu zgodnie z NFSS dla fontów CJK oznaczone jest przez literę *C* i następującą po niej dwucyfrową liczbę. Pierwsza z cyfr określa kodowanie (i standardowy zestaw znaków), np. `Big5`, `GB`, `JIS`, druga kodowanie fontu (układ znaków) np. ' ' (pusty), `pmC`, `dnp`, `HL`. Nie wszystkie pozycje są praktycznie wykorzystane.



Przykłady definicji kodu fontu

kodowanie fontu NFSS – kodowanie (fontu)

C00 – Bg5

C09 – Bg5+

C11 – GB w układzie `pmC`

C40 – JIS

C42 – JIS w układzie `dnp`

C60 – KS dla fontów zawierających ideogramy chińskie (*hanja*)

C61 – KS dla fontów zawierających litery i sylaby alfabetu *hangŭl*

C63 – KS dla fontów z alfabetem *hangŭl* w układzie *HLaTeXa*

C70 – Unicode

Pliki definicji fontów

Nazwa plików .fd składa sięz kodu NFSS i nazwy kroju fontu np. `c00fs.fd`, co oznacza, że definiowany jest fontu C00 z krojem *fangsong* (skrót *fs* został wymuszony przez ograniczenie nazwy kroju do 5 znaków).
Definicje fontu mogą również być podane w preambule.

Przykład definicji fontu w pliku c70arial.fd:

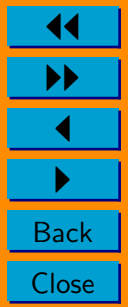
```
\DeclareFontFamily{C70}{arial}{}
\DeclareFontShape{C70}{arial}{m}{n}{<-> CJK * aruni}{}
\DeclareFontShape{C70}{arial}{bx}{n}{<-> CJKb * aruni}{\CJKbold}
```

Przetwarzanie wstępne

– W pakiecie znajdują się preprocesory (konwertery) do przetwarzania wstępnego wymaganego dla niektórych metod kodowania.

– na płycie TEX Live 5 dostępne są preprocesory dla systemów operacyjnych Linux i Windows, wykorzystywane w trybie wsadowym

– przetwarzanie wstępne dla tekstów edytowanych w Emacsie i kodowanych w `emacs-mule` dokonywane jest wewnątrz edytora



Preprocesory dla trybu wsadowego

* `bg5conv` dla kodowania Big5
* `sjisconv` dla kodowania SJIS
* `extconv` dla kodowania Big5+ oraz GBK
* `cefconv`, `cef5conv`, `cefsconv` dla kodowania CEF

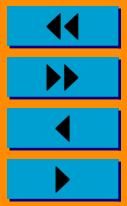
Przetwarzanie wstępne w Emacsie

- kodowanie `emacs-mule` jest kodowaniem, które umożliwia zapis tekstów wielojęzycznych, m.in. pisanych w językach dalekowschodnich i języku polskim
- do przetwarzania wstępnego konieczny jest plik `cjk-enc.el`, który należy do dystrybucji pakietu CJK
- komendy Emacsa uruchamiające przetwarzanie wstępne:

  `M-x cjk-write-file`

  `M-X cjk-write-all-files`

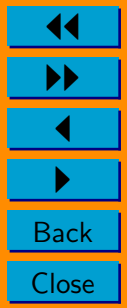
- plik wynikowy: `.tex` $\longrightarrow$ `.cjk`



Przełączanie między trybami

– Od wersji 4.4.0 będzie istniała możliwość przełączania między trybem z przetwarzaniem wstępnym i trybem bezpośrednim

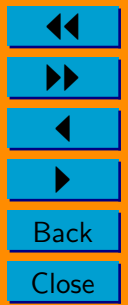
– Praktyczne wykorzystanie zdarza się stosunkowo rzadko. Może ono zaistnieć, gdy dołączamy do dokumentu teksty w innym kodowaniu,a których skład wymaga drugiego z trybów.



Przykład przełączania między trybami

Tekst referatu wygłoszonego na konferencji koreanistycznej

- kodowanie EUC nie pozwala na zapis wszystkich sylab afabetu *hangŭl*, z których część w przeszłości znajdowała się w praktycznym użyciu
- zarówno kodowanie kodowanie EUC jak również żadne inne powszechnie używane kodowanie nie pozwala na zapis sylab, które zawierają archaiczne litery alfabetu *hangŭl*
- rozwiązanie problemu — wykorzystanie przełączania między trybem z prztwarzaniem wstępnym tekstu kodowanego w `emacs-mule` oraz trybem bez przetwarzania wstępnego tekstu zapisanego w standardzie Unicode (kodowanie `utf-8`)



Podział dokumentu przetwarzanie wstępne/przetwarzanie bezpośrednie

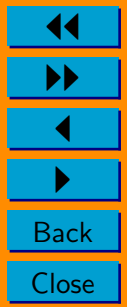
– do zapisu tekstu podstawowego w języku polskim z drobnymi odwołaniami do oryginału (nazwiska, tytuły utworów) zostało użyte kodowanie `emacs-mule` i wykorzystane były fonty typu 1 z płyty TeX Live 5

– do zapisu cytowanych utworów w oryginale zostało użyte kodowanie `utf-8` oraz odpowiednio przygotowany zestaw fontów typu 1, powstały ze źródłowego fontu ttf. Wykorzystany font zawierał glify wszystkich generycznych sylab alfabetu *hangŭl* zgodnie ze standardem Unicode 2.0 oraz glify sylab z archaicznymi literami tego alfabetu w zakresie znaków użytkownika (*private use area*)

Konwersja fontów ttf → type1

sposób przygotowania fontów:

- tfm: ttf2tfm z wykorzystaniem pliku unicode.sfd do podziału na fonty składowe
- pfb: ttf2pfb, t1asm
- utworzenie odpowiedniego pliku .fd
- aktualizacja plików konfiguracyjnych dvipsa i pdflatexa



Konwersja fontów ttf → type1

ograniczenia:

- konieczność wywołania dvipsa z opcją całkowitego ładowania fontu
- problemy z utworzeniem plików pdf z użyciem pdflatexa i dvipspdfm

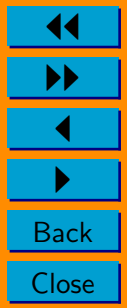
Dlaczego nie użyłem tylko Unicode'u?

- dostępny font dalekowschodni posiadał niedopracowane glify dla języka polskiego
- pakiet CJK nie jest przystosowany do wsparcia składu tekstów ze znakami innymi niż CJK:
  * nie uwzględnia podcięć kernowych i ewentualnych ligatur między fontami składowymi
  * jeżeli font nie posiada odmiany grubej, dla znaków ascii nie zostanie utworzony jej substytut (poor man's boldface)

Współpraca z pakietami do składu w języku polskim

Pakiet współpracuje z pakietami Babel i Polski. Jeżeli korzystamy z pakietu Polski dostępne są:

- instrukcja `\dywiz`
- instrukcje wstawiania polskich symboli znaków matematycznych
- notacja ciachowa



Ograniczenia w korzystaniu z pakietu Polski

– dla kodowania emacs-mule:

  * polecenie `\prefixing` musi być użyte po `\begin{document}`

– dla kodowania `latin-2` lub `cp1250`:

  * polecenie `\prefixing` musi być użyte po `\begin{document}`
  * przed wystąpieniem środowiska CJK musi nastąpić `\nonprefixing`

W kodowaniu emacs-mule instrukcję `\dywiz`, polskie symbole znaków matematycznych oraz notację ciachową można swobodnie wykorzystywać wewnątrz środowiska CJK.

Cechy pakietu

- wspomaga skład tekstów w językach orientalnych przy użyciu LaTeX 2_ε
- akceptuje większość metod kodowania powszechnie stosowanych na Dalekim Wschodzie
- udostępnia mechanizmy typowe dla konwencji typograficznych w krajach kręgu CJK
- definiuje możliwość uzyskania odmiany grubej fontu przy braku rzeczywistego fontu CJK
- umożliwia osadzanie znaków innych niż znaki CJK wewnątrz środowiska CJK
- współpracuje z pakietami polonizacyjnymi