upTEX – Unicode version of pTEX with CJK extensions

Takuji Tanaka 田中 琢爾

upTEX project

Oct 26, 2013

Outline / 概要

(1) Introduction

(2) Unicodization / Unicode 化
- Japanese / 日本語
- CJK / 中韓 / 中・日・한
- with European languages / 欧文との親和性
- world languages / 世界の言語

(3) Impremention / 実装
- Unicodization / Unicode 化
- \kcatcode
- set3

(4) upTEX vs. Ω, XƎTEX, ...

(5) Present & future / 現在と今後

Part I

Introduction

ASCII pTEX/pLATEX

It's **great**:

- **High quality** Japanese typesetting
  - incl. vertical writing, Japanese hyphenation, ...
- Japanese **standard** TEX/LATEX
- **Strong support** by environment
  - —DVIware, packages, macros, softwares, books, ...

but has **weakness**:

- **Japanese local**
  - — 8bit Latin/Chinese/Korean are not available
- **Limited character set**
  - by legacy encodings (Shift_JIS, EUC-JP)

Motivation

Support **wider character set** of Japanese
by Unicode

Support **babel**
by switching Latin–CJK tokens

Support **Chinese/Korean**

Keep quality & environment of pTEX

Feature of upTEX/upLATEX

(1) **High quality** CJK typesetting
based on pTEX/pLATEX

(2) **Compatible** with pTEX/pLATEX

(3) **Unicode** / **UTF-8**

(4) Switching Latin (12bit) / **CJK (29bit) tokens**

(5) **CJK with Babel** (Latin/Cyrillic/Greek. . .)

(6) **Over BMP** — incl. SIP (U+2xxxx)

Part II

Unicodization / Unicode化

Unicodization / Unicode 化

Strategies of Unicodization

(1) Unicodize only IO
    Ex: \usepackage[utf8]{inputenc}

(2) Imprement Unicode functions
    Ex: XeTeX

(3) Comromise
    upTeX: Intenal: Unicodize only CJK,
    IO: Fully Unicodize

Partial Unicodization / 折衷的Unicode化

		TEX	pTEX	upTEX
Latin	7bit Latin	azAZ	azAZ	azAZ
	8bit Latin	æœÆŒ		æœÆŒ
	inputenc	гдГД		гдГД
Japanese	JIS X 0208		あア亜	あア亜
	Unicode			①Ⅳ髙
CK	Unicode			**汉字** 漢字 한글

pTEX, upTEXconsists of two parts
(1) As same as original TEX
(2) pTeX–JIS X 0208, upTeX–Unicode

New JIS : JIS X 0213

upTEX treats new JIS X 0213 (over JIS X 0208)

崍〽崏崏♬♩♤♠♢♦♡♥♧♣崝崝〠☎☀☁☂☃
♨ゔかけヷヸヹヺ⅓⅔柰柬柰崮峑㈱㈲
①②③❶❷❸嵛崚崚ⅰⅱⅲⅠⅡⅢⓐⓑⓒ㋐㋑㋒
鄧小平 李承燁 里見弴 草彅剛 朴璐美 森鷗外 森雞二
王銘琬 宮﨑あおい 蔣介石 你好 深圳 東日本旅客鉃道株式会社 尾骶骨 生酛仕込 凮月堂 㐂寿 仐寿
圓壔函數
啞然 火焰 嚙む 任俠 長身瘦軀 石鹼 屢〻 刺繡 醬油
蟬時雨 隔靴搔痒 奧飛驒 簞笥 摑む 充塡 顚末 祈禱
瀆職 土囊 潑剌 醱酵 頰紅 素麵 麴町 蓬萊 蠟燭 攢竹

Characters out of JIS / JIS 外字

over JIS X 0213 (new JIS)

髙島屋、内田百閒、
杮落とし、安全才一、𠮷野家

source

髙島屋、内田百閒、杮落
とし、安全才一、𠮷野家

output

Platform dependent characters are now in Unicode

①②③④⑤⑥⑦⑧⑨⑩⑪⑫⑬⑭⑮⑯⑰⑱⑲⑳ⅠⅡⅢ
ⅣⅤⅥⅦⅧⅨⅩ㍉㌔㌢㍍㌘㌧㌃㌶㍑㍗㌍㌦㌣㌫㍊㌻
㎜㎝㎞㎎㎏㏄㎡㍻〝〟№㏍℡㊤㊥㊦㊧㊨㈱㈲㈹㍾㍽㍼
≡∫∮√⊥∠∟⊿∵∩∪髙閒塚德豐﨑彅弴燁珉鄧

Chinese/Japanese/Korean
中・日・한

```
\schrm 简体中文: 你好

\tchrm 繁體中文: 早晨

\jpnrm 日本語: こんにちは

\korrm 한국어: 안녕하세요
```

source

简体中文: 你好
繁體中文: 早晨
日本語: こんにちは
한국어: 안녕하세요

output

Difference of glyphs among CJK / CJKのグリフの違い

Simplified Chinese	**骨練,**	**平直。**	**神祀,**	**才次.**
Traditional Chinese	骨練,	平直。	神祀,	才次.
Japanese	骨練,	平直。	神祀,	才次.
Korean	骨練,	平直。	神祀,	才次.

end-of-line

```
Please give↓
me beer.

请给我↓
啤酒。

ビールを私に↓
下さい。

맥주를 나에게↓
주세요.
```

Please give me beer.

(treated as space)

请给我啤酒。

(ignored)

ビールを私に下さい。

(ignored)

맥주를 나에게 주세요.

(treated as space)

Control word by CJK characters

```
\def\오늘{%
 \number\year 연%
 \number\month 월%
 \number\day 일%
}
Today: 《\오늘》
```

Today: 《2013 연 10 월 26 일》

Japanese-OTF package

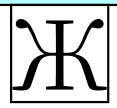
```
\usepackage[uplatex,...]{otf}
...
Adobe-Korea1-1:\\
\CIDK{8322}\CIDK{8588}
...
Adobe-Japan1-5:\\
\●問\◇答\ajRecycle{10}%
\ajLig{学校法人}%
\ajPICT{野球}\\
\ajMaru{1}...
```

Adobe-Korea1-1:
鞣☯譱颱颼饅

Adobe-Japan1-5:
問答稔呉
①❷3 4(5)6.(七)㈧IX

Japanese-OTF package also supports CK.

Unification / 統合

	standard	full-width
Cyrillic	 U+0416	 U+0416
Latin	 U+0057	 U+FF37

No “full-width” code in Greek, Cyrillic in Unicode.
It is a barrier to Unicodize Japanese softs.
upTEX can treat full-width Greek, Cyrillic by markup.

inputenc & UTF-8

```
\usepackage[utf8]{inputenc}
\usepackage[T1]{fontenc}
\kcatcode`ç=15
...
“¿But aren’t Kafka’s
Schloß and Æsop’s
Œuvres often naïve
vis-à-vis the dæmonic
phœnix’s official
rôle in fluffy soufflés?”
```

“¿But aren’t Kafka’s Schloß and Æsop’s Œuvres often naïve vis-à-vis the dæmonic phœnix’s official rôle in fluffy soufflés?”

Babel

```
\usepackage[french,...]%
{babel}
...
\selectlanguage{english}
English ... \today
...
\selectlanguage{russian}
Русский ... \today

\selectlanguage{japanese}
日本語 ... \today
```

English
October 26, 2013
Français
26 octobre 2013
Deutsch
26. Oktober 2013
Czech
26. října 2013
Русский
26 октября 2013 г.
日本語
2013 年 10 月 26 日

It's a small world

upTEX can treat CJK, Latin, Cyrillic and Greek.
upTEX cannot directly treat Arabic, Brahmic, . . .

Part III

Impremention / 実装

Unicodization / Unicode 化

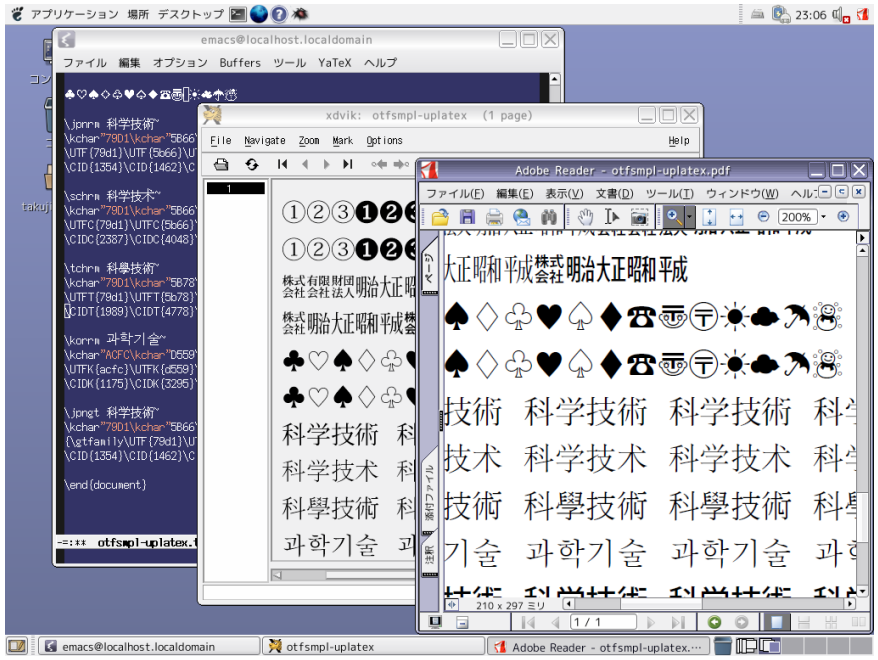
(1) IO: EUC/SJIS in pTEX → UTF8 in upTEX
    (ptexenc library)

(2) Internal buffer: 16bit in pTEX → 29bit in upTEX
    (Ref. Omega)

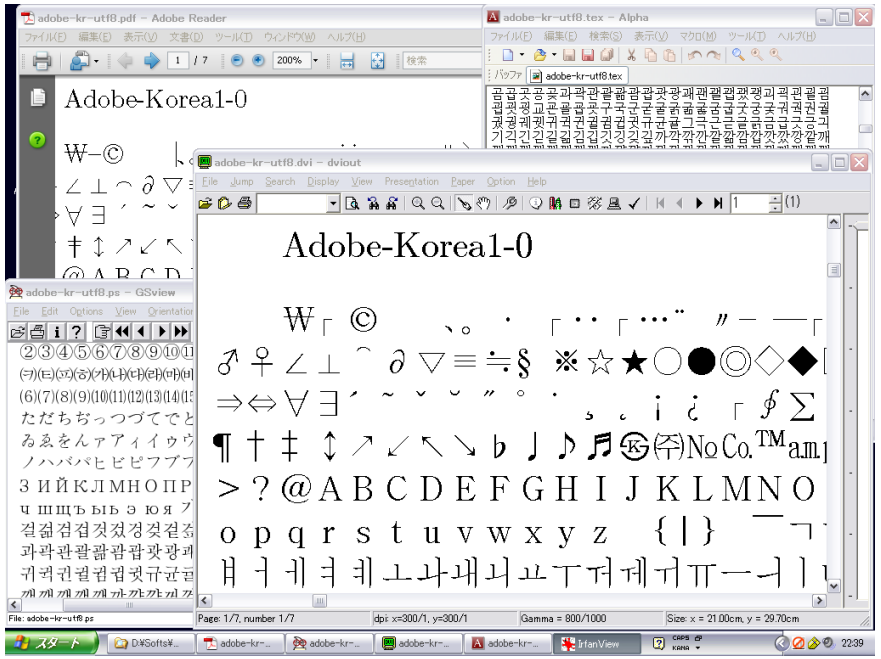
(3) Unicodize standard macros, libraries

(4) upTEX support of DVIWARE

DVIware



ptetex3+ / Linux



W32TeX / Windows

dvipdfmx, dvips, xdvi, dvi2tty &
DVIOUT are available

\kcatcode

kcat code	cat code	kind	e.g.	control word	end of line
	...	...			
	10	space	␣		
15	11	char	azAZ	yes	as space
	12	other char	(.!?	no	as space
	...	...			
16		**Kanji**	汉漢	yes	ignore
17		**Kana**	かナ	yes	ignore
18		**CJK symbol**	《・。』	no	ignore
19		**Hangul**	한글	yes	as space

If \kcatcode is 15, the character is treat as Latin and upTEX works as same as original TEX.

set3 & over BMP

弥丈弨弩人弱弽彁彂彄彄彅彉八彋彎彐当彔录彗彘彣彫彯彰彴彶径土径待
很徍徎幌徒従徕徜徹健循徬徭徳徵幫忉忒志幝志忘忘忙忚応忞忞幡忟忨快
忭念恁忻怍怖怜怞怤怦怦佛怭怮怰恃恲恲恲恈恢恣恣恟恧恧恨恩恮息息恲
悅悅庁悋悔悕悖梋庇悜庇悡患您您悥悰悵悽泰悫惀惎惏惐惩惵惷惸惾愀愐
惕座感愯惰愷愻㤥 熊慌慎𢡄慖慘慙慜慝慝憬慠慥慵慨廎傷𢦏慷慷廂慘𪫧憄
憄憍憑憒廇憝憝憝憟憠憥憣憤憦憩憩憫憬憯憰憵憸憹憻憼憿懀懁懃懄懁臼
懋懐懦懨懩懪廈懾攤戚戊或戟戠戠戯戭截房所扃扃扆扆扊手扌扏扐扔扖扡
扪扳扵扶抵承技𢪸 扢抉延抒抓抛折抛抨抪披抲抴拆拆拉抾拎拔拕拙拚拜拜
拡拢拧拨拲拳拷拷拹拺拽拫持挃指挏挐挚挤𢰝 挫挭振挺捆捕捉捃捃捜捤捧
捩捶捸捺捼捼掅掊掌掍掕掖掘掘吡

(JIS2004 includes a lot of CJK Ideograph Extension B)

upTEX supports SIP (Supplementary Ideograph Plane) U+2xxxx
by using DVI command set3.
How visionary Knuth is!!

Part IV

upTEX vs. Ω, XƎTEX, . . .

upTEX vs. Ω, XƎTEX, ...

		TEX	pTEX	upTEX	Ω	XƎTEX
Compatibility	Latin	◎	○	◎	○	△
	Japanese	—	◎	◎	×	×
Advancedness		×	×	×	×	◎
Multilingual	Latin	◎	○	◎	◎	◎
	Japanese	—	○	◎	△	△
	CK	—	—	◎	△	△
	others	—	—	—	△	◎
Integrity	(Japanese)	◎	◎	◎	△	△
Popularity	Japan	◎	◎	○	△	△
	World	◎	△	△	△	○

◎ > ○ > △ > ×

Part V

Present & Future / 現在と今後

History

Year	
1995	ASCII pTeX ver.2, pLaTeX2e
2007	upTEX first release, alpha version
2007	upTEX is in W32TeX
2008	e-upTEX by Kitagawa-san
2012	upTEX 1.00
2012	upTEX is in TeX Live
2013	upTEX presentation in TUG2013

Future / 今後

Currently, upTEX has capability of multilingual (CJK, Latin, Cyrillic, Greek) typesetting.
Possible items in the future are:

(1) **Document classes** for Chinese/Korean
    (Any volunteer?)

(2) **Babel options** for Chinese/Korean
    (It will be useful in ko.TeX etc. Any volunteer?)

(3) Does upTEX have a potential
    to be **a useful CJK TEX?**

Part VI

Appendix / おまけ

Latin/CJK tokens

			TEX	pTEX	upTEX
Latin	I/O		8bit (multibytes)†	7bit 1byte	8bit (multibytes)†
	token	charcode	8bit	8bit	8bit
		catcode	4bit	4bit	4bit
CJK	I/O		—	EUC etc. 8bit 2bytes	**UTF-8** 8bit 2–4bytes
	token	charcode	—	16bit	24bit
		kcatcode	—	—	5bit
Latin/CJK **classification**			—	fixed	**customizable**
inputenc			OK	NG	**OK**
Babel			full	partial	**full**

†: with inputenc

Character encoding in upTEX

	Latin	**CJK**		
	TEX compatible	upTEX extended		
	<256	BMP	over BMP	comment
.tex / .aux I/O buffer		**UTF8**		
	1byte	2–3bytes	4bytes	
token	12bit	29bit		with (k)catcode
.dvi / .vf	set1 **T1 etc.** 8bit	set2 **UCS2** 16bit	set3 **UTF32** 24bit	
.tfm	**T1 etc.** 8bit	**UCS2** 16bit	—†	†treated as Kanji 'jfm' for CJK
.ps / CMap	**T1 etc.** 8bit	**UCS2** 16bit	**UTF16** 2×16bit	

kcatcode

kcat code	cat code	kind	e.g.	control word	end of line
15	...	...			
	10	space	␣		
	11	char	azAZ	yes	as space
	12	other char	(.!?	no	as space
	...	...			
16		**Kanji**	汉漢	yes	ignore
17		**Kana**	かナ	yes	ignore
18		**CJK symbol**	《・。』	no	ignore
19		**Hangul**	한글	yes	as space