upTEX — Unicode version of pTEX with CJK extensions

Takuji Tanaka

Abstract

upTEX is a Unicode extension of ASCII's pTEX (a Japanese-localized TEX). It not only improves Japanese support, but also handles Chinese and Korean characters, i.e., Kanji (Hanzi, Hanja), Kana, CJK symbols, and Hangul with Unicode. Moreover, it can process multilingual typesetting of original LATEX with inputenc and Babel (Latin, Cyrillic, Greek, etc.) by switching its \kcatcode tables. This paper describes the main features of upTEX.

1 Introduction

1.1 Motivation

The objective of upTEX [1] is the "Unicodization" of ASCII Corp.'s pTEX [2]. p(LA)TEX is the most popular TEX system in Japan and is widely used as a typesetter for commercial printing. pTEX achieves professional quality Japanese typesetting [3] including Japanese hyphenation and vertical writing.

However, p(LA)TEX has some limitations due to the legacy encodings (double byte Japanese encodings, EUC-JP or Shift_JIS). up(LA)TEX tries to improve this, while keeping the benefits of pTEX, as follows.

1.2 Enhancement of Japanese character set

The Japanese government has standardized several character set versions as JIS (Japanese Industrial Standard). In pTEX, the Japanese character set is basically limited to JIS C 6226 and JIS X 0208, namely JIS level-1 and -2 (6879 characters in 1990). In 2000, a new standard for character set JIS X 0213 was standardized, with an additional character set JIS level-3 and -4, 11233 characters in 2004. Moreover, additional characters which are useful in Japanese are defined in the Unicode standard. upTEX supports both sets, using the UTF-8 encoding.

1.3 Support of Chinese, Korean

pTEX basically supports Japanese and English (7-bit Latin). upTEX adds support of Chinese and Korean using procedures similar to Japanese typesetting. So now upTEX supports CJK (Chinese, Japanese and Korean).

1.4 Cooperation with Babel

pTEX has some difficulty handling 8-bit Latin due to the double byte legacy encodings. upTEX more easily treats 8-bit Latin, compared to pTEX. The inputenc (option utf8) and Babel (Latin, Cyrillic, Greek, etc.) packages are available in upLATEX.

1.5 Compatibility with pTEX

Since upTEX keeps the typesetting procedures of pTEX, most macros for p(LA)TEX are expected to work in up(LA)TEX with minimal or no modifications.

1.6 Limitation

Current upTEX has several limitations.

In the area of multilingual support, upTEX handle CJK, Latin, Cyrillic and Greek. However, upTEX cannot directly treat complex characters, e.g., Arabic, Hindi.

upTEX includes the ε-TEX extensions, thanks to Kitagawa [4]. On the other hand, upTEX does not have pdfTEX extensions yet.

The (u)pTEX engine remains based on original TEX and written using WEB. It includes two special extensions: (1) (u)pTEX treats Japanese (CJK) glyph metrics with an extended format of tfm, called jfm; (2) (u)pTEX defines a special DVI command 255 for vertical writing.

Thus, (u)pTEX requires that related DVIware support jfm and this DVI command. Advanced features of OpenType fonts are hardly touched by pTEX and upTEX.

2 Implementation

This section describes the implementation of upTEX.

2.1 Unicodization

Table 1 compares the internal structure of original TEX, pTEX and upTEX. pTEX uses internally a 16-bit token for Japanese, which is not enough to cover the wide range of the Unicode character set. upTEX expands that to an internal 29-bit token for CJK, where the Ω (Omega) implementation was used for reference. upTEX internally "unicodizes" only CJK characters, treating 8-bit Latin characters the same as in original TEX.

Table 2 summarizes encoding in upTEX. upTEX I/O accepts UTF-8 with a variable length of one–four octets. Originally, upTEX was not so refined. The routine was cleaned up via the ptexenc library written by Tsuchimura [5]. upTEX assumes that characters beyond Unicode's Basic Multilingual Plane (BMP) have a fixed font metric to support Kanji on the Supplementary Ideograph Plane (SIP).

2.2 Extension of \kcatcode

Table 3 shows the classifications of the \kcatcode primitive in upTEX. pTEX defined \kcatcode values of 16 (Kanji), 17 (Kana), and 18 (Japanese symbol) for Japanese typesetting. upTEX extends the

`\kcatcode` of 15 (not CJK), defines 19 (Hangul), and 18 is redefined as CJK symbol.

When `\kcatcode` is set to 15 (not CJK), the character is treated like Latin characters in original TeX. This feature provides the improved Babel support in upTeX compared to pTeX. For example, users can switch `\kcatcode` to select whether Cyrillic and Greek characters are treated as proportional glyphs in their language, or treated as full-width glyphs in conventional Japanese fonts.

When `\kcatcode` is set to 19 (Hangul), end-of-line is treated as space.

2.3 Use of the DVI command `set3`

upTeX uses the DVI command `set3` and supports Kanji of SIP (U+2xxxx), where some Kanji characters used daily in Japanese are included. The upTeX project prepared patches for some DVI software (`dvips`, `dvipdfmx`, `dviout`, `xdvi` and `dvi2tty`) to support SIP. In TeX Live, the patches for all but `xdvi` are (or will be) applied.

3 upTeX vs. other Unicode TeX

Table 4 compares features of original TeX, upTeX and other TeX families with Unicode support (Ω, XeTeX [6] and LuaTeX-ja [7]).

The recommended practice for using upTeX is as a "better pTeX" with better Japanese support and/or better multilingualization with CJK and Babel support.

4 History and future

A very brief timeline of upTeX-related events:

1995	ASCII pTeX ver.2, pLaTeX 2_ε [2]
2007	upTeX first release, alpha version
2007	upTeX in W32TeX [8]
2008	ε-upTeX by Kitagawa-san [4]
2012	upTeX 1.00
2012	upTeX in TeX Live

Currently, I believe upTeX has the capability of multilingual (CJK, Latin, Cyrillic, Greek) typesetting. Possible enhancements for the future:

- Support IVS (Ideographic Variation Sequence).
- Document classes for Chinese/Korean.
- Babel options for Chinese/Korean.
- Add pdfTeX extensions.
- Merge LuaTeX [9] or XeTeX [6].

Acknowledgements

I deeply appreciate the work of ASCII Corporation (currently ASCII Media Works, Inc.). upTeX could not exist without the achievement of pTeX.

I thank Tsuchimura-san (土村展之氏), ZR-san (八登崇之氏) [10], Inoue-san (井上浩一氏), Kakuto-san (角藤亮氏), Okumura-san (奥村晴彦氏), Yasuda-san (安田功氏), Kuriyama-san (栗山雅俊氏), Kitagawa-san (北川弘典氏), Dora-san, Norbert Preining-san and the many people who discussed upTeX at the Japanese website TeX Q&A [11] and other places for their fruitful observations.

References

[1] Tanaka, Takuji. *upTeX, upLaTeX - Unicode version of pTeX, pLaTeX*, `http://homepage3.nifty.com/ttk/comp/tex/uptex_en.html`

[2] ASCII Corporation (currently ASCII Media Works, Inc.). *ASCII Japanese TeX (pTeX)* (in Japanese), `http://ascii.asciimw.jp/pb/ptex/`

[3] Okumura, Haruhiko. *pTeX and Japanese typesetting*, `http://oku.edu.mie-u.ac.jp/~okumura/texfaq/japanese/ptex.html`

[4] Hironori, Kitagawa, *e-pTeX Wiki* (in Japanese), `http://sourceforge.jp/projects/eptex/wiki/FrontPage`

[5] *ptetex Wiki*, "UTF-8 対応 (4)" (in Japanese), `http://tutimura.ath.cx/ptetex/?UTF-8%C2%D0%B1%FE%284%29`

[6] SIL International. *The XeTeX typesetting system*, `http://scripts.sil.org/xetex`

[7] *LuaTeX-ja*, `http://sourceforge.jp/projects/luatex-ja/wiki/FrontPage%28en%29`, `http://ctan.org/pkg/luatexja`

[8] Kakuto, Akira. *W32TeX*, `http://w32tex.org/index.html`

[9] The LuaTeX team. *LuaTeX*, `http://www.luatex.org/`

[10] Yato, Takayuki. *En toi Pythmeni tes TeXnopoleos* 電脳世界の奥底にて (in Japanese), `http://zrbabbler.sp.land.to/`, *upLaTeX* を使おう (in Japanese), `http://zrbabbler.sp.land.to/uplatex.html`

[11] *TeX Q&A* (in Japanese), `http://oku.edu.mie-u.ac.jp/~okumura/texwiki/`

◇ Takuji Tanaka
Kokubunji
Tokyo
Japan
KXD02663 (at) nifty dot ne dot jp
http://homepage3.nifty.com/ttk/comp/tex/uptex_en.html

Table 1: Comparison of structure among TEX, pTEX and upTEX.
† denotes that it works with the `inputenc` package.

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

Table 2: Encoding in upTEX.

	Latin TEX compatible <256	**CJK** upTEX extended BMP	 over BMP	 comment
`.tex` / `.aux` I/O buffer	UTF-8 1 byte	2–3 bytes	4 bytes	
token	12 bit	29 bit		with (k)catcode
`.dvi` / `.vf`	`set1` T1 etc. 8 bit	`set2` UCS-2 16 bit	`set3` UTF-32 24 bit	
`.tfm`	T1 etc. 8 bit	UCS-2 16 bit	—‡	‡ treated as Kanji `jfm` for CJK
`.ps` / CMap	T1 etc. 8 bit	UCS-2 16 bit	UTF-16 2×16 bit	

Table 3: \kcatcode in pTEX and upTEX. * denotes upTEX extension.

\kcatcode	\catcode	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

Table 4: Comparison of features among upTEX and other TEX families.
Symbols express the following: Better . . . ◎ > ○ > △ > ▽ . . . worse.

		TEX	pTEX	upTEX	Ω	XƎTEX	LuaTEX-ja
Compatibility	Latin	◎	○	◎	○	△	◎
	Japanese	—	◎	◎	▽	▽	△
Multilingual	Latin	◎	○	◎	◎	◎	◎
	Japanese	—	○	◎	△	△	○
	CK	—	—	◎	△	△	?
	others	—	—	—	△	◎	○
Integrity	(Japanese)	◎	◎	◎	△	△	○
Popularity	Japan	◎	◎	○	△	△	△
	World	◎	△	△	△	○	△