Guide to p\TeX \text{ for developers unfamiliar with Japanese}

Japanese \TeX \text{ Development Community*}

version p3.8.2, November 26, 2019

p\TeX \text{ and its variants, up\TeX, } \varepsilon-\text{p\TeX} \text{ and } \varepsilon-\text{up\TeX}, \text{ are all } \TeX \text{ engines with native Japanese support. Its output is always a DVI file, which can be processed by several DVI drivers with Japanese support including } dvips \text{ and } dvipdfmx. \text{ Formats based on } \LaTeX \text{ is called p\LaTeX \text{ when running on p\TeX/ } \varepsilon-\text{p\TeX}, \text{ and called up\LaTeX \text{ when running on up\TeX/ } \varepsilon-\text{up\TeX}.}

\textbf{Purpose of this document}

This document is written for developers of \TeX/\LaTeX, who aim to support p\TeX/\pLaTeX \text{ and its variants up\TeX/} up\LaTeX. Knowledge of the followings are assumed:

- Basic knowledge of Western \TeX \text{ (Knuthian } \TeX, \varepsilon-\TeX \text{ and pdf\TeX)},
- ... and its programming conventions.

Any knowledge of Japanese (characters, encodings, typesetting conventions etc.) is not assumed; some explanations are provided in this document when needed. We hope that this document helps authors of packages or classes to proceed with supporting \p\TeX \text{ family smoothly.}

Note: This English edition (ptex-guide-en.pdf) is \textit{not} meant to be a complete translation of Japanese edition (ptex-manual.pdf). For example, this document does not cover the following aspects of p\TeX:  

- Typesetting conventions of Japanese characters  
- Details of vertical writing

For beginners of writing Japanese texts, please refer to the Japanese edition.
Contents

1 Brief introduction 4
2 pT\TeX and its variants 4
3 Compatibility with Western T\TeX 4
4 \LaTeX on pT\TeX/\up\LaTeX — p\LaTeX/\up\LaTeX 4
5 Eminent characteristics of pT\TeX/\up\TeX 5

II Details 6

5 Output format — DVI 6
5.1 Extensions of DVI format in pT\TeX family 7
5.2 DVI drivers with Japanese support 8
5.2.1 Using dvipdfmx 8
5.2.2 Using dvips 8

6 Programming on pT\TeX family 8
6.1 Number of registers and marks 8
6.2 Number of math families 9
6.3 Additional primitives and keywords 9
6.3.1 Sync\TeX additions (available in p\TeX, up\TeX, \varepsilon-p\TeX, \varepsilon-up\TeX) 9
6.3.2 p\TeX additions (available in p\TeX, up\TeX, \varepsilon-p\TeX, \varepsilon-up\TeX) 9
6.3.3 up\TeX additions (available in up\TeX, \varepsilon-up\TeX) 11
6.3.4 \varepsilon-p\TeX additions (available in \varepsilon-p\TeX, \varepsilon-up\TeX) 11
6.3.5 \varepsilon-up\TeX additions (available in \varepsilon-up\TeX) 13
6.4 Omitted primitives and unsupported features 13
6.5 Behavior of Western \TeX primitives 13
6.5.1 Primitives with limitations in handling Japanese 13
6.5.2 Primitives capable of handling Japanese 14
6.6 Case study 14
6.6.1 Detecting p\TeX 14
6.6.2 Detecting up\TeX 15
6.6.3 Defining large integer constants 15
6.6.4 Creating a Japanese character token with a specified code 16
6.7 Difference from pdf\TeX in DVI mode 18
6.8 Recommendation for file encoding
6.9 Input handling
6.10 Japanese tokens

7 Basic introduction to Japanese typesetting
7.1 Automatic insertion of glue and penalties
7.2 Japanese fonts

8 Other strange beasts
8.1 Internal kanji encodings
Part I

Brief introduction

1 p\TeX\ and its variants

There is no advantage to choose p\TeX/\up\TeX\ over \varepsilon\-p\TeX/\varepsilon\-up\TeX, so we focus mainly on \varepsilon\-p\TeX/\varepsilon\-up\TeX.

2 Compatibility with Western \TeX

p\TeX/\up\TeX are almost upward compatible with Knuthian \TeX, however, they do not pass the TRIP test. In p\TeX/\up\TeX, input handling is different from Knuthian \TeX; if a pair of two or more 8-bit codes matches Japanese character code, it is regarded as one Japanese character. There is no difference in handling 8-bit TFM font.

\varepsilon\-p\TeX/\varepsilon\-up\TeX are almost upward compatible with \varepsilon\-\TeX, however, input handling is similar to p\TeX/\up\TeX. It does not pass the e-TRIP test. That said, please note that “raw \varepsilon\-\TeX” is unavailable anymore in \TeX Live and derived distributions; they provide a command \texttt{etex} only as “DVI mode of pdf\TeX.” We should note that \varepsilon\-p\TeX/\varepsilon\-up\TeX are not upward compatible with DVI mode of pdf\TeX, which will be discussed later in section 6.7.

3 \LaTeX\ on p\TeX/\up\TeX — \LaTeX/\up\LaTeX

Formats based on \LaTeX\ is called \LaTeX\ when running on p\TeX/\varepsilon\-p\TeX, and called up\LaTeX\ when running on up\TeX/\varepsilon\-up\TeX. In recent versions (around 2011) of \TeX\ Live and its derivatives, the default engines of \LaTeX\ and up\LaTeX\ are \varepsilon\-p\TeX\ and \varepsilon\-up\TeX. That is, the command \texttt{platex} starts \varepsilon\-p\TeX\ (not p\TeX) with preloaded format \texttt{platex.fmt}.

In the kernel level (\texttt{platex.ltx} and \texttt{uplatex.ltx}), \LaTeX\ and up\LaTeX\ adds some additional commands related to the followings:

- Selection of Japanese fonts
- Crop marks (called “tombow”) for printings
- Adjustment for mixing horizontal and vertical texts

For authors, \LaTeX/\up\LaTeX\ are almost upward compatible with original \LaTeX, except for the followings:

- Order of float objects; in \LaTeX/\up\LaTeX, \texttt{⟨bottom float⟩} is placed above \texttt{⟨footnote⟩}. That is, the complete order is \texttt{⟨top float⟩ → ⟨body text⟩ → ⟨bottom float⟩ → ⟨footnote⟩}. 
For developers, additional care may be needed, for changes in the kernel macros and/or absence of pdf\TeX \ features.

4 Eminent characteristics of p\TeX /up\TeX

The most important characteristics of p\TeX /up\TeX can be summarized as follows:

- Japanese characters are interpreted and handled completely apart from Western characters.

- Texts can be aligned vertically, called “tate-gumi” (縦組). The horizontal alignment of texts is called “yoko-gumi” (横組), and both “tate-gumi” and “yoko-gumi” can be mixed even within a single document.
Part II
Details

5 Output format — DVI

The output of \LaTeX{} family is always a DVI file. This is in contrast to the mainstream of pdf\LaTeX{} in the Western \TeX{} world.

In case you are not familiar with DVI output processing, first we give some general notice on how to get a "correct" output using \LaTeX{} in DVI mode.

- The DVI format is, as its name suggests, inherently driver-independent. However, some \LaTeX{} packages (graphicx, color, hyperref etc.) embed some \texttt{\special} commands into the DVI, which can be interpreted later by some specific DVI driver. Such a DVI is no longer driver-independent, thus those are called driver-dependent packages.

- In almost all major \TeX{} distributions (of course including \TeX{} Live), the default DVI driver is set to \texttt{dvips}. When you choose to process the resulting DVI file with a driver other than dvips (e.g. dvipdfmx) after running \LaTeX{}, you need to pass a proper driver option (e.g. \texttt{[dvipdfmx]}) to all driver-dependent packages.

Now, let's move on to the situation in Japan, which is slightly complicated due to historical reasons but may also apply to other countries:

- There are two major conventions to pass a proper driver option to all driver-dependent packages:
  
  1. To give a driver option to each driver-dependent package:

     \begin{verbatim}
     \documentclass{article}
     \usepackage[dvipdfmx]{graphicx}
     \usepackage[dvipdfmx]{color}
     \end{verbatim}

  2. To have a driver option as global:

     \begin{verbatim}
     \documentclass[dvipdfmx]{article}
     \usepackage{graphicx}
     \usepackage{color}
     \end{verbatim}

The former convention has been used for many years since 1990s when the number of driver-dependent packages was limited. But in recent years (around 2010–), there are much more driver-dependent packages available. Thus we (Japanese \TeX{} experts) advise
a global driver option rather than individual package options for simplicity, but not yet fully widespread.\footnote{The fact that there had been a mismatch in option names ([dvipdfm] vs. [dvipdfmx]) between packages may also have been part of it; geometry did not understand [dvipdfmx] option until 2018!}

- Many people still see driver options as “optional”; they do without driver options unless really needed. For example, the convention of having a global driver option does no harm even when no driver-dependent package is used, but some users choose to omit a driver option to avoid a warning:

\texttt{LaTeX Warning: Unused global option(s): [dvipdfmx].}

\subsection*{5.1 Extensions of DVI format in p\TeX family}

The DVI format output by p\TeX family is fully compatible with Knuthian \TeX, as long as the following conditions are met:

- No Japanese characters are typeset.
- There is no portion of vertical text alignment.

However, some additional DVI commands, which are defined in the standard \cite{He2007} but never used in \TeX82, can come out.

- \texttt{set2} (129), \texttt{put2} (134): Appears in both p\TeX and up\TeX DVI. Used to typeset a Japanese character with 2-byte code.
- \texttt{set3} (130), \texttt{put3} (135): Appears in only up\TeX DVI. Used to typeset a Japanese character with 3-byte code.

When p\TeX is going to typeset a Japanese character into DVI, it is encoded in JIS, which is always a 2-byte code. For this purpose, \texttt{set2} or \texttt{put2} are used. When up\TeX is going to output a Japanese character into DVI, it is encoded in UTF-32. If the code is equal to or less than \texttt{U+FFFF}, the lower 16-bit is used with \texttt{set2} or \texttt{put2}. If the code is equal to or greater than \texttt{U+10000}, the lower 24-bit is used with \texttt{set3} or \texttt{put3}.

In addition, p\TeX\up/Up\TeX defines one additional DVI command.

- \texttt{dir} (255): Used to change directions of text alignment.

The DVI format in the preamble is always set to 2, as with \TeX82. On the other hand, the DVI ID in the postamble can be special. Normally it is set to 2, as with \TeX82; however, when \texttt{dir} (255) appears at least once in a single p\TeX\up/Up\TeX DVI, the post\_post table of postamble contains \texttt{ID = 3}. 
5.2 DVI drivers with Japanese support

There are some DVI drivers with Japanese support. The most eminent drivers are dvips and dvipdfmx. Nowadays most of casual Japanese users are using dvipdfmx as a DVI driver. On the other hand, users of dvips are unignorable, especially those working in publishing industry. In recent years, most of major driver-dependent packages support both two drivers.

5.2.1 Using dvipdfmx

A DVI file which is output by p\TeX can be converted directly to a PDF file using dvipdfmx.

5.2.2 Using dvips

A DVI file which is output by p\TeX can be converted to a PostScript file using dvips.

The resulting PostScript file can then be converted to a PDF file using Ghostscript (ps2pdf) or Adobe Distiller. When using Ghostscript, a proper setup of Japanese font must be done before converting PostScript into PDF. An easy solution for the setup is a script cjk-gs-integrate developed by Japanese \TeX Development Community.

6 Programming on p\TeX family

We focus on programming aspects of p\TeX and its variants.

6.1 Number of registers and marks

p\TeX and up\TeX have exactly the same number (= 256) of registers (count, dimen, skip, muskip, box, and token) as Knuthian \TeX. e-p\TeX and e-up\TeX in extended mode have more registers; there are 65536, which is twice as many as 32768 of e-\TeX. Similarly e-p\TeX and e-up\TeX have 65536 mark classes, which is twice as many as 32768 of e-\TeX.

The following code presents an example of detecting the number of registers and mark classes available:

```latex
\ifx\eTeXversion\undefined
  % Knuthian TeX, pTeX, upTeX:
  % 256 registers, 1 mark
\else
  \ifx\omathchar\undefined
    % e-TeX, pdfTeX (in extended mode):
    % 32768 registers, 32768 mark classes
  \else
    % e-pTeX, e-upTeX (in extended mode):
```

8
Here a primitive \omathchar, which is derived from Ω, is used as a marker of a change file fam256.ch.\footnote{There is another p\TeX-\N derived engine named p\TeX-\ng (or Asiatic p\TeX) \url{https://github.com/clerkma/ptex-ng}, it is based on \$\varepsilon$-\TeX and up\TeX, but currently does not adopt fam256.ch so it has the same number of registers and mark classes as \$\varepsilon$-\TeX.}

6.2 Number of math families

In \p\TeX and up\TeX, the number of math fonts is restricted to 16, each of which can contain 256 characters (same as Knuthian \TeX). In ε-\p\TeX and ε-up\TeX, a change file fam256.ch, which is derived from Ω, extends the upper limit to 256. As a consequence, ε-\p\TeX and ε-up\TeX allows 256 math fonts, each of which can contain 256 characters.\footnote{Ω allows 256 math fonts, each of which can contain 65536 characters.}

For \p\TeX and \up\TeX users to use more than 16 math fonts, it is necessary to use macros which exploit Ω-derived primitives such as \omathchar. Recent (u)p\TeX (since 2016/11/29) partially supports this, and the maximum number of math alphabets that can be defined by ™\DeclareMathAlphabet is extended to 256 (™\e@mathgroup@top) without needing any extension package. However, symbol fonts are restricted to 16 as ™\DeclareMathSymbol etc still use the standard ™\mathchar etc. A simple solution to use more symbol fonts as well as math alphabets is to load a package mathfam256\footnote{\url{https://www.ctan.org/pkg/mathfam256}} though it’s still preliminary.

6.3 Additional primitives and keywords

Here we provide only complete lists of additional primitives of \p\TeX family in alphabetical order. The features of each primitive can be found in Japanese edition.

6.3.1 Sync\TeX additions (available in \p\TeX, up\TeX, ε-\p\TeX, ε-up\TeX)

In the standard build of \TeX Live, Sync\TeX extension is unavailable in Knuthian \TeX; however, it is enabled in \p\TeX family.

▷ \synctex

6.3.2 p\TeX additions (available in \p\TeX, up\TeX, ε-\p\TeX, ε-up\TeX)

▷ \autospacing
\autoxspacing
\disinhibitglue — New primitive since p3.8.2 (\TeX\ Live 2019)
\dtou
\euc
\ifdbox — New primitive since p3.2 (\TeX\ Live 2011)
\ifddir — New primitive since p3.2 (\TeX\ Live 2011)
\ifmbox — New primitive since p3.7.1 (\TeX\ Live 2017)
\ifmdir
\iftbox
\iftdir
\ifybox
\ifydir
\inhibitglue
\inhibitxspcode
\jcharwidowpenalty
\jfam
\jfont
\jis
\kanjiskip
\kansuji
\kansujichar
\kcatcode
\kuten
\noautospacing
\noautoxspacing
\postbreakpenalty
\prebreakpenalty
\ptexminorversion — New primitive since p3.8.0 (\TeX\ Live 2018)
\ptexrevision — New primitive since p3.8.0 (\TeX\ Live 2018)
\ptexversion — New primitive since p3.8.0 (\TeX\ Live 2018)
- \scriptbaselineshiftfactor — New primitive since p3.7 (\TeX Live 2016)
- \scriptscriptbaselineshiftfactor — New primitive since p3.7 (\TeX Live 2016)
- \showmode
- \sjis
- \tate
- \tbaselineshift
- \textbaselineshiftfactor — New primitive since p3.7 (\TeX Live 2016)
- \tfont
- \xkanjiskip
- \xspcode
- \ybaselineshift
- \yoko
- \H
- \Q
- \zh
- \zw

6.3.3 \texttt{up\TeX} additions (available in \texttt{up\TeX}, \texttt{\varepsilon\-up\TeX})
- \disablecjktoken
- \enablecjktoken
- \forcecjktoken
- \kchar
- \kchardef
- \ucs
- \uptexrevision — New primitive since u1.23 (\TeX Live 2018)
- \uptexversion — New primitive since u1.23 (\TeX Live 2018)

6.3.4 \texttt{\varepsilon\-p\TeX} additions (available in \texttt{\varepsilon\-p\TeX}, \texttt{\varepsilon\-up\TeX})
- \currentspacingmode — New primitive since 191112 (\TeX Live 2020)
- \currentxspacingmode — New primitive since 191112 (\TeX Live 2020)
- \epTeXinputencoding — New primitive since 160201 (\TeX Live 2016)
\epTeXversion — New primitive since 180121 (\TeX\ Live 2018)
\expanded — New primitive since 180518 (\TeX\ Live 2019)
\hfi
\ifincsname — New primitive since 190709 (\TeX\ Live 2020)
\ifpdfprimitive — New primitive since 150805 (\TeX\ Live 2016)
\lastnodechar — New primitive since 141108 (\TeX\ Live 2015)
\lastnodesubtype — New primitive since 180226 (\TeX\ Live 2018)
\odelcode
\odelimiter
\omathaccent
\omathchar
\omathchardef
\omathcode
\oradical
\pagefistretch
\pdffcreationdate — New primitive since 130605 (\TeX\ Live 2014)
\pdffelapsedtime — New primitive since 161114 (\TeX\ Live 2017)
\pdffiledump — New primitive since 140506 (\TeX\ Live 2015)
\pdffilemoddate — New primitive since 130605 (\TeX\ Live 2014)
\pdffilesz — New primitive since 130605 (\TeX\ Live 2014)
\pdflastxpos
\pdflastypos
\pdfmdfivesum — New primitive since 150702 (\TeX\ Live 2016)
\pdfnormaldeviate — New primitive since 161114 (\TeX\ Live 2017)
\pdfpageheight
\pdfpagewidth
\pdfprimitive — New primitive since 150805 (\TeX\ Live 2016)
\pdfrandomseed — New primitive since 161114 (\TeX\ Live 2017)
\pdfresettimer — New primitive since 161114 (\TeX\ Live 2017)
\pdfsavepos
\pdfsetrandomseed — New primitive since 161114 (\TeX\ Live 2017)
\pdfshellescape — New primitive since 141108 (\TeX\ Live 2015)
\pdfstrcmp
\pdfuniformdeviate — New primitive since 161114 (\TeX\ Live 2017)
\readpapersizespecial — New primitive since 180901 (\TeX\ Live 2019)
\Uchar — New primitive since 191112 (\TeX\ Live 2020)
\Ucharcat — New primitive since 191112 (\TeX\ Live 2020)
\vfi
\fi

6.3.5 \TeX\ additions (available in \TeX\)
\currentcjktoken — New primitive since 191112 (\TeX\ Live 2020)

6.4 Omitted primitives and unsupported features

Compared to Knuthian \TeX\ and \TeX, some primitives are omitted due to conflict with Japanese handling. One is enc\TeX\ extension, such as \texttt{\mubyte}. The ML\TeX\ extension, such as \texttt{\charsubdef}, is included but not well-tested.

6.5 Behavior of Western \TeX\ primitives

Here we provide some notes on behavior of Knuthian \TeX\ and \TeX\ primitives when used within \p\TeX\ family.

6.5.1 Primitives with limitations in handling Japanese

Each of the following primitives allows only character codes 0–255; other codes will give an error “! Bad character code.”

\texttt{\catcode, \scode, \mathcode, \dcode, \lcode, \ucode.}

Each of the following primitives has \texttt{\ldots}char in its name, however, the effective values are restricted to 0–255.

\texttt{\endlinechar, \newlinechar, \escapechar, \defaulthyphenchar, \defaultskewchar.}
6.5.2 Primitives capable of handling Japanese

The following primitives are extended to support Japanese characters:

- \texttt{\char \langle character code \rangle}, \texttt{\chardef \langle control sequence \rangle = \langle character code \rangle}

  In addition to 0–255, internal codes of Japanese characters are allowed. For putting Japanese characters, a Japanese font is chosen. More information can be found in 6.6.3.

- \texttt{\font}, \texttt{\fontname}, \texttt{\fontdimen}

- \texttt{\accent \langle character code \rangle = \langle character \rangle}

- \texttt{\if \langle token_1 \rangle \langle token_2 \rangle, \ifcat \langle token_1 \rangle \langle token_2 \rangle}

  Japanese character token is also allowed. In that case,
  - \texttt{\if} tests the internal character code of the Japanese character.
  - \texttt{\ifcat} tests the \texttt{\kcatcode} of the Japanese character.

  \TeXbook{} describes the behavior of \texttt{\if} and \texttt{\ifcat} as follows;

  If either token is a control sequence, \TeX{} considers it to have character code 256 and category code 16, unless the current equivalent of that control sequence has been \texttt{\let} equal to a non-active character token.

  However, this includes a lie; in the real implementation of \texttt{tex.web}, a control sequence is considered to have a category code 0.

6.6 Case study

Here we provide some code examples which may be useful for package developers.

6.6.1 Detecting p\TeX{}

Since the primitive \texttt{\ptexversion} is rather new (added in 2018), the safer solution for detecting p\TeX{} is to test if a primitive \texttt{\kanjiskip} is defined.

\begin{verbatim}
\newif\if\kanjiskip
\ifx\kanjiskip\undefined
  \else
    \% \TeX{} / up\TeX{} / e-\TeX{} / e-up\TeX{}
  \fi
\end{verbatim}
6.6.2 Detecting up\TeX X

up\TeX X/\epsilon-up\TeX X are almost upward compatible with p\TeX X/\epsilon-p\TeX X respectively, however, there are two major differences:

1. Improvements in the \texttt{\textbackslash{}kcatcode} business, mainly for better handling of Latin-1 characters and CJK tokens.

2. Unicode as the default internal kanji encoding, for direct use of its huge character set.

The first difference can be detected by checking if \texttt{\textbackslash{}enablecjktoken} primitive is defined.

\begin{verbatim}
\ifx\enablecjktoken\undefined
  \else
    \% up\TeX X/e-up\TeX X
  \fi
\end{verbatim}

The second difference can be detected by checking if the character 0x2121 (fullwidth space in JIS encoding) is stored as "3000 internally.

\begin{verbatim}
\ifx\kanjiskip\undefined
  \else
    \ifnum\jis"2121="3000
      \% up\TeX X/e-up\TeX X with internal Unicode
    \else
      \% p\TeX X/e-p\TeX X
      \% or, up\TeX X/e-up\TeX X with internal EUC-JP or Shift-JIS
    \fi
  \fi
\end{verbatim}

Please note that the format-build setting of \texttt{-kanji-internal=(sjis|euc)} with up\TeX X makes it effectively p\TeX X regarding the character set, which means that only JIS X 0208 character set is supported.

6.6.3 Defining large integer constants

According to \cite{2} (Section 3.3),

A control sequence that has been defined with a \texttt{\textbackslash{}chardef} command can also be used as a \texttt{(number)}.

Tokens defined with \texttt{\textbackslash{}mathchardef} can also be used this way.

Here is the list of primitives which can be used for this purpose in p\TeX X family:
\chardef \langle control sequence\rangle = \langle character code\rangle
Defines a control sequence to be a synonym for \char \langle character code\rangle.

\kchardef \langle control sequence\rangle = \langle character code\rangle (for up\TeX/\varepsilon-up\TeX)
Defines a control sequence to be a synonym for \kchar \langle character code\rangle.

\mathchardef \langle control sequence\rangle = \langle 15-bit number\rangle
Defines a control sequence to be a synonym for \mathchar \langle 15-bit number\rangle.

\omathchardef \langle control sequence\rangle = \langle 27-bit number\rangle (for \varepsilon-p\TeX/\varepsilon-up\TeX)
Defines a control sequence to be a synonym for \omathchar \langle 27-bit number\rangle.

The first two (\chardef and \kchardef) are usable only when the integer being defined is in the range of valid character codes, which is not necessarily continuous (see §8.1). The most efficient and convenient way of defining integer constants is as follows:

- 0–255: \chardef
- 256–32767: \mathchardef
- 32768–134217727: \omathchardef (only for \varepsilon-p\TeX/\varepsilon-up\TeX)
- (optional) 256–2147483647: \chardef (only for up\TeX/\varepsilon-up\TeX)

6.6.4 Creating a Japanese character token with a specified code

Short version:

- With \varepsilon-p\TeX 191112 or later (\TeX Live 2020), you can use expandable primitives \Uchar and \Ucharcat
- Otherwise, use the “\kansuji trick”.

The “\kansuji trick”
This is a modified version of the “\lowercase trick” available in p\TeX family.

Short note on the “\lowercase trick”: to create a character token with a specified code value between 0–255 with Knuthian \TeX, the “\lowercase trick” can be used; for example,

\begingroup
\lccode`?=%mycount
\lowercase\endgroup \def\X{?}
defines \texttt{X} which expands to a character number \texttt{mycount} while the \texttt{catcode} of \texttt{?} (12) is preserved. However, the trick cannot be applied to Japanese characters, since πT\TeX family does not support \texttt{lccode} outside 0–255.

\texttt{kansuji} is an expandable primitive like \texttt{number} or \texttt{romannumeral}, and it converts an integer into its corresponding kanji notation called “kansuji” (漢数字). The important point here is that the number-kanji mapping can be altered by \texttt{kansujichar}.

Example 1: equivalent to \texttt{\def\X{\jis 2422}} (JIS code \texttt{0x2422} is “ぁ”):

\begin{verbatim}
\begingroup
  \kansujichar1=\jis 2422 \xdef\X{\kansuji1}
\endgroup
\end{verbatim}

Example 2: equivalent to \texttt{\def\jis {Japan}}.

\begin{verbatim}
\begingroup
  \kansujichar5=\jis 467C \relax
  \kansujichar6=\jis 4B5C \relax
  \expandafter\gdef\csname\kansuji56\endcsname{Japan}
\endgroup
\end{verbatim}

Since \texttt{kansujichar} accepts only Japanese character code, the “\texttt{kansuji trick}” and the \texttt{"\lowercase trick"} should be used complementarily.

\textbf{\texttt{\Uchar}, \texttt{\Ucharcat}}

The “\texttt{kansuji trick}” above include an assignment of \texttt{kansujichar} which is unexpandable. ε-π\TeX 191112 or later (π\TeX Live 2020) provides expandable primitives \texttt{\Uchar} and \texttt{\Ucharcat}, which are derived from \texttt{Xₜₜₑₓ}.

\begin{itemize}
  \item \texttt{\Uchar} \texttt{(character code)}
    
    Expands to a character token with specified slot \texttt{(character code)}.
    
    \begin{itemize}
      \item When an 8-bit number (0–255) is given, it expands to a Latin character token with category code 12, except for a space character (32) which has category code 10.
      \item When a Japanese character code greater than 255 is given, it expands to a Japanese character token with its current category code; 16–18 for ε-π\TeX, 16–19 for ε-up\TeX.
    \end{itemize}

  \item \texttt{\Ucharcat} \texttt{(character code)} \texttt{(category code)}
    
    Expands to a character token with slot \texttt{(character code)} and \texttt{(category code)} specified.
    
    \begin{itemize}
      \item With π\TeX:
        
        \begin{itemize}
          \item Only 8-bit number (0–255) are allowed for \texttt{(character code)}; that is, only Latin characters can be generated.
        \end{itemize}
    \end{itemize}
\end{itemize}
– The values allowed for \textit{category code} are 1–4, 6–8, 10–13.

• With \texttt{\textasciitilde}\texttt{up}\texttt{T}\texttt{E}\texttt{X}:
  – When \textit{character code} is between 0–127, only Latin characters can be generated. Thus, the values allowed for \textit{category code} are 1–4, 6–8, 10–13.
  – When \textit{character code} is between 128–255, both Latin and Japanese characters can be generated depending on the specified \textit{category code}; 1–4, 6–8, 10–13: Latin character, 16–19: Japanese character.
  – When \textit{character code} is greater than 255, only Japanese characters can be generated. Thus, the values allowed for \textit{category code} are 16–19.

6.7 Difference from pdf\texttt{T}\texttt{E}\texttt{X} in DVI mode

As stated above, \texttt{\textasciitilde}\texttt{p}\texttt{T}\texttt{E}\texttt{X}/\texttt{\textasciitilde}\texttt{up}\texttt{T}\texttt{E}\texttt{X} are not upward compatible with DVI mode of pdf\texttt{T}\texttt{E}\texttt{X}.

6.8 Recommendation for file encoding

6.9 Input handling

For simplicity, first we introduce of input handling of \texttt{\textasciitilde}\texttt{up}\texttt{T}\texttt{E}\texttt{X}.

6.10 Japanese tokens

7 Basic introduction to Japanese typesetting

This section does not aim to explain Japanese typesetting completely; here we provide a minimum requirement for “getting away” with Japanese.

7.1 Automatic insertion of glue and penalties

Sometimes \texttt{p}\texttt{T}\texttt{E}\texttt{X} family automatically inserts glue and penalties between characters.

7.2 Japanese fonts

\texttt{p}\texttt{T}\texttt{E}\texttt{X} family can have 3 different “current” fonts at the same time; a Latin font, a Japanese font for horizontal writing (“yoko-gumi”), and a Japanese font for vertical writing (“tate-gumi”). The first one is the same as in the Knuthian \texttt{T}\texttt{E}\texttt{X}, which is defined in a standard TFM format. The latter two are specific to \texttt{p}\texttt{T}\texttt{E}\texttt{X} family, which are defined in a JFM (Japanese \texttt{T}\texttt{E}\texttt{X} font metric) format.\footnote{A JFM is a modified version of the standard TFM. It can be created by \texttt{(u)p}\texttt{LtoTF}, and decoded by \texttt{(u)pTFtoPL}. Please also refer to the man pages of these programs (\texttt{ppltotf.man1.pdf} and \texttt{ptftopl.man1.pdf}).}

18
While typesetting, p\(\TeX\) family automatically switches between these 3 fonts, depending on the character code and the writing direction:

- For typesetting Latin characters, the current Latin font shown by \texttt{\the\font} is selected.
- For typesetting Japanese characters, the current Japanese font suitable for the current writing direction is selected. It is shown by \texttt{\the\jfont} for horizontal writing and \texttt{\the\tfont} for vertical writing.

In Knuthian \(\TeX\), the primitive \texttt{\nullfont} refers to an “empty font” in which all characters are undefined. However in p\(\TeX\) family, this is regarded as a Latin font and there is no equivalent to “Japanese \texttt{\nullfont}” by design. To elaborate, it is possible only when no Japanese font is set globally, i.e. in ini\(\TeX\) mode. Once a valid Japanese font is selected, there is no way of selecting “Japanese \texttt{\nullfont}” to discard all characters.

Moreover, p\(\TeX\) and friends assume that each Japanese font (except “Japanese \texttt{\nullfont}” in ini\(\TeX\) mode) contains all valid Japanese character code. In other words, all Japanese fonts share the same character set corresponding to the whole valid Japanese character code range.

\section{Other strange beasts}

\subsection{Internal kanji encodings}

The \texttt{\langle character code\rangle} is a union of the following two:

- Range of numbers between 0–255, and
- Numbers allowed for internal code of Japanese characters.

The former is the same as Knuthian \(\TeX\), but the latter is a problem. In up\(\TeX\)/\texttt{-kanji-internal=uptex} (default on), the range is very simple:

\[ c \geq 0 \]

However in p\(\TeX\)/\texttt{-kanji-internal}, only legacy encodings (EUC-JP as \texttt{euc}, or Shift-JIS as \texttt{sjis}) are available for \texttt{-kanji-internal}. In this case, the range can be represented as follows:

\[ c = 256c_1 + c_2 \ (c_i \in C_i) \]

where

\[
\begin{align*}
C_1 &= C_2 = \{"a1, \ldots, "fe\} \quad \text{(euc)} \\
C_1 &= \{"81, \ldots, "9f\} \cup \{"e0, \ldots, "fc\}, C_2 = \{"40, \ldots, "7e\} \cup \{"80, \ldots, "fc\} \quad \text{(sjis)}.
\end{align*}
\]
Therefore, the overall range of \textit{character code} is not continuous.

To check whether an integer is a valid Japanese character code or not, you can use \texttt{\iffontchar} with \texttt{-p\TeX{} 190709} or later (\TeX{} Live 2020). Suppose a count register \texttt{\mycount} stores an integer, you can do it as follows:

\begin{verbatim}
\iffontchar\jfont\mycount
  \% \mycount is a valid Japanese character code
\fi
\end{verbatim}

Here the primitive \texttt{\jfont} is used merely as a representative non-empty\footnote{This assumption is always safe after one of the standard \TeX{} formats (e.g. plain \TeX{}, \LaTeX{}) is loaded.} Japanese font containing all valid Japanese character code (see \hyperref[sec:font]{\ref*{sec:font}}).
References


Index

Symbols

\accent
\autosize
\autosize
\autoxspace
\char
\charedf
\currentcjktoken
\currentxspacingmode
\disablecjktoken
\epTeXinputencoding
\epTeXversion
\euc
\expanded
\font
\fontdimen
\forcecjktoken
\if
\ifcat
\ifddir
\ifdefbox
\ifdir
\iffcsname
\ififstream
\ifdir
\ifdefpdfprimitive
\ifdefbox
\ifdefdir
\ifdef
\ifxbox
\ifdef
\ifpdfprimitive
\ifinhibitglue
\inhibitxspacecode
\jcharwidowpenalty

\jfam
\jfont
\jis
\kanjiskip
\kansuji
\kansujichar
\kcatcode
\kchar
\kchardef
\kuten
\lastnodechar
\lastnodesubtype
\mathchardef
\mathchar
\mathaccent
\omathaccent
\omathchar
\omathchardef
\omathcode
\pagefistretch
\pdfcreationdate
\pdfelapsestime
\pdfcomputedtime
\pdfdefaulttime
\pdfdump
\pdffiledump
\pdffilemodate
\pdffilesize
\pdflastxpathos
\pdflastxpos
\pdflastypos
\pdfmdfivesum
\pdfmedialim
\pdfnormaldeviate
\pdfpagelayout
\pdfpagewidth
\pdfpageheight
\pdfpage
\pdfbbox
\pdfpageinfo
\pdfpageinfo
\pdfpagerecur
\pdfpagelastpage
\pdfpagecount
\pdfpagecount
\pdfpagecount
\pdfpagelastpage
\pdfpagelastpage
\pdfpagecount
\pdfpagelastpage
\pdfpagecount
\pdfpagelastpage
\pdfpagecount
\pdfpagelastpage
\pdfpagecount
\pdfpagelastpage
\pdfpagecount
\pdfpagelastpage
\pdfpagecount
\pdfpagelastpage
\pdfpagecount
\pdfpagelastpage

22