Guide to p\TeX\ for developers unfamiliar with Japanese

Japanese \TeX\ Development Community*

version p3.9.0, August 4, 2021

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

Purpose of this document

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

- Basic knowledge of Western \TeX\ (Knuthian \TeX, $\varepsilon$-\TeX\ and \texttt{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 family smoothly.

Note: This English guide (ptex-guide-en.pdf) is not meant to be a complete translation of Japanese manual (ptex-manual.pdf). For example, this document does not cover issues regarding Japanese typesettings. If you are interested in typesetting conventions of Japanese text, please also refer to \texttt{japanese.pdf} distributed with babel-japanese.

This document is maintained at: \url{https://github.com/texjporg/ptex-manual/}

*\url{https://texjp.org}, e-mail: issue(at)texjp.org
Contents

I Brief introduction 4

1 pT EX and its variants 4

2 Eminent characteristics of pT EX family 4

3 Compatibility with Western T EX 5

4 L ATEX on pT EX — pL ATEX 5

II Details 6

5 Output format — DVI 6

5.1 Extensions of DVI format in pT EX 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 EX 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 pT EX additions (available in pT EX, upT EX, 𝜀-pT EX, 𝜀-upT EX) 9

6.3.2 upT EX additions (available in upT EX, 𝜀-upT EX) 11

6.3.3 𝜀-pT EX additions (available in 𝜀-pT EX, 𝜀-upT EX) 12

6.3.4 𝜀-upT EX additions (available in 𝜀-upT EX) 13

6.3.5 Other cross-engine additions 13

6.4 Omitted primitives and unsupported features 13

6.5 Behavior of Western T EX primitives 14

6.5.1 Primitives with limitations in handling Japanese 14

6.5.2 Primitives capable of handling Japanese 14

6.6 Case study 15

6.6.1 Detecting pT EX 15

6.6.2 Detecting upT EX 15

6.6.3 Defining large integer constants 16

6.6.4 Creating a Japanese character token with a specified code 16

6.7 Difference from pdfT EX 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

The figure below shows the relationship between engines.

\[
\begin{array}{c}
pdf\TeX \\
\varepsilon-\TeX \\
\varepsilon-p\TeX \\
p\TeX \\
\varepsilon-\up\TeX \\
up\TeX \\
\end{array}
\]

\(p\TeX\) is an old Japanese-specific extension of T\(\TeX\)82, which aims to support proper type-setting of Japanese text but only supports a limited character set, JIS X 0208 (6879 characters).

\(up\TeX\) is developed as an extension of \(p\TeX\) to support full Unicode characters. It also includes extensions to overcome the difficulties of \(p\TeX\) in processing 8-bit Latin characters due to conflicts with legacy multibyte Japanese encodings.

\(\varepsilon-p\TeX\) and \(\varepsilon-up\TeX\) are \(\varepsilon\)-\(\TeX\) extensions of \(p\TeX\) and \(up\TeX\) respectively. In the current release, some extensions derived from pdf\(\TeX\) and \(\Omega\) are also available.

2 Eminent characteristics of \(p\TeX\) family

The most important characteristics of \(p\TeX\) family can be summarized as follows:

- Japanese characters are interpreted and handled completely apart from Western characters. If a pair of two or more 8-bit codes in the input matches the pattern of Japanese character codes, it is regarded as one Japanese character and given a different \texttt{\catcode} value.

- There are two text directions; horizontal (yoko-gumi; 横組) and vertical (tate-gumi; 縦組). Two directions can be mixed even within a single document.
3 Compatibility with Western \TeX

\pTeX/\upTeX/ are almost upward compatible with Knuthian \TeX/, however, they do not pass the TRIP test. The most important difference lies in the handling of 8-bit code inputs; some 8-bit Latin characters may be subject to the encoding conversion. There is no difference in handling 8-bit TFM font.

\varepsilon-\pTeX/\varepsilon-\upTeX/ are almost upward compatible with \varepsilon-\TeX/, however, input handling is similar to \pTeX/\upTeX/. 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/.” You should note that \varepsilon-\pTeX/\varepsilon-\upTeX/ are not upward compatible with DVI mode of pdf\TeX/, which will be discussed later in Section 6.7.

There is no advantage to choose \pTeX/\upTeX/ over \varepsilon-\pTeX/\varepsilon-\upTeX/, so we focus mainly on \varepsilon-\pTeX/\varepsilon-\upTeX/.

4 \LaTeX/ on \pTeX/\upTeX/ — \pl\LaTeX/\upl\LaTeX/

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

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

- Selection of Japanese fonts,
- Crop marks (called \texttt{tombow}; トンボ) for printings,
- Adjustment for mixing horizontal and vertical texts.

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

- Order of float objects; in \pl\LaTeX/\upl\LaTeX/, (bottom float) is placed above (footnote). That is, the complete order is (top float) → (body text) → (bottom float) → (footnote).

For developers, additional care may be needed, for changes in the kernel macros and/or the absence of pdf\TeX/ features.
Part II
Details

5 Output format — DVI

The output of p\TeX\ family is always a DVI file. This is in contrast to the mainstream of pdf\TeX\ 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 L\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 \texttt{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\footnote{Since LaTeX2ε 2020-02-02, this warning is effectively gone. This is due to preloading of expl3 into the format, and the driver-dependent code of expl3 interprets the global driver option.}:

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

### 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{pTeX} but never used in \TeX82, can come out.

- `set2` (129): Used to typeset a Japanese character with 2-byte code (both p\TeX and up\TeX).
- `set3` (130): Used to typeset a Japanese character with 3-byte code (up\TeX only).

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, `set2` or `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 `set2` or `put2`. If the code is equal to or greater than \texttt{U+10000}, the lower 24-bit is used with `set3` or `put3`.

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

- `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 `dir` (255) appears at least once in a single p\TeX/\upTeX DVI, the post_post table of postamble contains 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 *pTeX* can be converted directly to a PDF file using *dvipdfmx*. For Japanese fonts to be used in the output PDF, *dvipdfmx* refers to *kanjix.map* generated by the command *updmap*. You can use the script *kanji-config-updmap* to change font settings; please refer to its help message or documentation.

5.2.2 Using *dvips*

A DVI file which is output by *pTeX* can be converted to a PostScript file using *dvips*. For Japanese fonts to be used in the output PostScript, *dvips* refers to *psfonts.map* generated by the command *updmap*. You can use the script *kanji-config-updmap* to change font settings; please refer to its help message or documentation.

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 to run a script *cjk-gs-integrate* developed by Japanese *TeX* Development Community.

6 Programming on *pTeX* family

We focus on programming aspects of *pTeX* and its variants.

6.1 Number of registers and marks

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

```
\if\eTeXversion\undefined
  % Knuthian TeX, pTeX, upTeX:
  % 256 registers, 1 mark
```
Here a primitive \omathchar, which is derived from \Omega, is used as a marker of a change file fam256.ch.\(^3\)

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 \(\varepsilon\)-p\(\TeX\) and \(\varepsilon\)-up\(\TeX\), a change file fam256.ch, which is derived from \Omega, extends the upper limit to 256. As a consequence, \(\varepsilon\)-p\(\TeX\) and \(\varepsilon\)-up\(\TeX\) allows 256 math fonts, each of which can contain 256 characters.\(^4\)

For p\(\LaTeX\)/up\(\LaTeX\) users to use more than 16 math fonts, it is necessary to use macros which exploit \Omega-derived primitives such as \omathchar. Recent (u)p\LaTeX\ (since 2016/11/29) partially supports this, and the maximum number of math alphabets that can be defined by \\texttt{\DeclareMathAlphabet} is extended to 256 (\texttt{\e@mathgroup@top}) without needing any extension package. However, symbol fonts are restricted to 16 as \texttt{\DeclareMathSymbol} etc still use the standard \texttt{\mathchar} etc. A simple solution to use more symbol fonts as well as math alphabets is to load a package mathfam256\(^5\) 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 p\(\TeX\) additions (available in p\(\TeX\), up\(\TeX\), \(\varepsilon\)-p\(\TeX\), \(\varepsilon\)-up\(\TeX\))

- \\texttt{\autospacing}
- \\texttt{\autoxspacing}

\(^3\)There is another p\(\TeX\)-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 \(\varepsilon\)-\(\TeX\).

\(^4\)\Omega allows 256 math fonts, each of which can contain 65536 characters.

\(^5\)\url{https://www.ctan.org/pkg/mathfam256}
\texttt{\textbackslash disinhibitglue} — New primitive since p3.8.2 (\TeX{} Live 2019)
\texttt{\textbackslash dtou}
\texttt{\textbackslash euc}
\texttt{\textbackslash ifdbox} — New primitive since p3.2 (\TeX{} Live 2011)
\texttt{\textbackslash ifddir} — New primitive since p3.2 (\TeX{} Live 2011)
\texttt{\textbackslash ifjfont} — New primitive since p3.8.3 (\TeX{} Live 2020)
\texttt{\textbackslash ifmbox} — New primitive since p3.7.1 (\TeX{} Live 2017)
\texttt{\textbackslash ifmdir}
\texttt{\textbackslash iftbox}
\texttt{\textbackslash iftdir}
\texttt{\textbackslash ifftfont} — New primitive since p3.8.3 (\TeX{} Live 2020)
\texttt{\textbackslash ifybox}
\texttt{\textbackslash ifydir}
\texttt{\textbackslash inhibitglue}
\texttt{\textbackslash inhibitxspcode}
\texttt{\textbackslash jcharwidowpenalty}
\texttt{\textbackslash jfam}
\texttt{\textbackslash jfont}
\texttt{\textbackslash jis}
\texttt{\textbackslash kanjiskip}
\texttt{\textbackslash kansuji}
\texttt{\textbackslash kansujichar}
\texttt{\textbackslash kcatcode}
\texttt{\textbackslash kuten}
\texttt{\textbackslash noautospacing}
\texttt{\textbackslash noautoxspacing}
\texttt{\textbackslash postbreakpenalty}
\texttt{\textbackslash prebreakpenalty}
\texttt{\textbackslash ptexminorversion} — New primitive since p3.8.0 (\TeX{} Live 2018)
\texttt{\textbackslash ptexrevision} — New primitive since p3.8.0 (\TeX{} Live 2018)

10
\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
\toucs — New primitive since p3.10.0 (\TeX\ Live 2022)
\ucs — Imported from up\TeX, since p3.10.0 (\TeX\ Live 2022)\(^6\)
\xkanjiskip
\xspcode
\ybaselineshift
\yoko
\H
\Q
\zh
\zw

6.3.2 up\TeX\ additions (available in up\TeX, \v\TeX-\up\TeX)
\disablecjktoken
\enablecjktoken
\forcecjktoken
\kchar
\kchardef
\uptexrevision — New primitive since u1.23 (\TeX\ Live 2018)
\uptexversion — New primitive since u1.23 (\TeX\ Live 2018)

\(^6\)The primitive \ucs was part of “up\TeX\ additions” until \TeX\ Live 2021.
6.3.3 $\varepsilon$-p\TeX{} additions (available in $\varepsilon$-p\TeX{}, $\varepsilon$-up\TeX{}):

- \texttt{\currentspacingmode} — New primitive since 191112 (\TeX{} Live 2020)
- \texttt{\currentxspacingmode} — New primitive since 191112 (\TeX{} Live 2020)
- \texttt{\epTeXinputencoding} — New primitive since 160201 (\TeX{} Live 2016)
- \texttt{\epTeXversion} — New primitive since 180121 (\TeX{} Live 2018)
- \texttt{\expanded} — New primitive since 180518 (\TeX{} Live 2019)
- \texttt{\hfi}
- \texttt{\ifincsname} — New primitive since 190709 (\TeX{} Live 2020)
- \texttt{\ifpdfprimitive} — New primitive since 150805 (\TeX{} Live 2016)
- \texttt{\lastnodechar} — New primitive since 141108 (\TeX{} Live 2015)
- \texttt{\lastnodesubtype} — New primitive since 180226 (\TeX{} Live 2018)
- \texttt{\odelcode}
- \texttt{\odelimiter}
- \texttt{\omathaccent}
- \texttt{\omathchar}
- \texttt{\omathchardef}
- \texttt{\omathcode}
- \texttt{\oradical}
- \texttt{\pagefistretch}
- \texttt{\pdfcreationdate} — New primitive since 130605 (\TeX{} Live 2014)
- \texttt{\pdfelapsedtime} — New primitive since 161114 (\TeX{} Live 2017)
- \texttt{\pdffiledump} — New primitive since 140506 (\TeX{} Live 2015)
- \texttt{\pdffilemoddate} — New primitive since 130605 (\TeX{} Live 2014)
- \texttt{\pdffilesize} — New primitive since 130605 (\TeX{} Live 2014)
- \texttt{\pdflastxpos}
- \texttt{\pdflastypos}
- \texttt{\pdfmdfivesum} — New primitive since 150702 (\TeX{} Live 2016)
- \texttt{\pdfnormaldeviate} — New primitive since 161114 (\TeX{} Live 2017)
- \texttt{\pdfpageheight}
- \texttt{\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)
\pdfrandomseed
\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)
\vadjust pre — New primitive since 210701 (\TeX\ Live 2022)
\vfi
\fi

6.3.4 $\varepsilon$-up$\TeX$ additions (available in $\varepsilon$-up$\TeX$)
    \currentcjktoken — New primitive since 191112 (\TeX\ Live 2020)

6.3.5 Other cross-engine additions
Sync$\TeX$ extension (available in p$\TeX$, up$\TeX$, $\varepsilon$-p$\TeX$, $\varepsilon$-up$\TeX$):
    \synctex
\TeX\ Live additions (available in p$\TeX$, up$\TeX$, $\varepsilon$-p$\TeX$, $\varepsilon$-up$\TeX$):
    \partokencontext — New primitive since $\TeX$ Live 2022
    \partokenname — New primitive since $\TeX$ Live 2022
    \tracingstacklevels — New primitive since $\TeX$ Live 2021

6.4 Omitted primitives and unsupported features
Compared to Knuthian $\TeX$ and $\varepsilon$-$\TeX$, some primitives and extensions are omitted due to conflict with Japanese handling.

- The enc$\TeX$ extension, including the primitives \mubyte etc., is unavailable.
- The ML$\TeX$ extension, such as \charsubdef, is not enabled by default. It becomes available with the command-line option \mltex, but not well-tested.
6.5 Behavior of Western $\TeX$ primitives

Here we provide some notes on behavior of Knuthian $\TeX$ and $\varepsilon$-$\TeX$ 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.”

\catcode, \sfcode, \mathcode, \delcode, \lccode, \uccode. 

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

\endlinechar, \newlinechar, \escapechar, \defaultthyphenchar, \defaultskeewchar.

6.5.2 Primitives capable of handling Japanese

The following primitives are extended to support Japanese characters:

▶ \char ⟨character code⟩, \chardef ⟨control sequence⟩=⟨character code⟩

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.

▶ \font, \fontname, \fontdimen

▶ \accent ⟨character code⟩=⟨character⟩

▶ \if ⟨token1⟩ ⟨token2⟩, \ifcat ⟨token1⟩ ⟨token2⟩

Japanese character token is also allowed. In that case,

• \if tests the internal character code of the Japanese character.
• \ifcat tests the \kcatcode of the Japanese character.

$\TeX$book describes the behavior of \if and \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 \let equal to a non-active character token.

However, this includes a lie; in the real implementation of 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 \ptexversion is rather new (added in 2018), the safer solution for detecting p\TeX is to test if a primitive \kanjiskip is defined.

\begin{verbatim}
\ifdef\kanjiskip\undefined
\else
  % p\TeX / up\TeX / e-p\TeX / e-up\TeX
\fi
\end{verbatim}

6.6.2 Detecting up\TeX

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

1. Improvements in the \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 \...cjktoken primitive is defined.

\begin{verbatim}
\ifdef\enablecjktoken\undefined
\else
  % up\TeX/e-up\TeX
\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}
\ifdef\kanjiskip\undefined
\else
  \ifnum\jis"2121="3000
    % up\TeX/e-up\TeX with internal Unicode
  \else
    % p\TeX/e-p\TeX
    % or, up\TeX/e-up\TeX with internal EUC-JP or Shift-JIS
  \fi
\fi
\end{verbatim}
Please note that the format-build setting of `-kanji-internal=(sjis|euc)` with upTeX makes it effectively pTeX regarding the character set, which means that only JIS X 0208 character set is supported.

6.6.3 Defining large integer constants

According to [2] (Section 3.3),

A control sequence that has been defined with a `\chardef` command can also be used as a `⟨number⟩`. This fact is used in allocation commands such as `\newbox`. Tokens defined with `\mathchardef` can also be used this way.

Here is the list of primitives which can be used for this purpose in pTeX family:

- `\chardef ⟨control sequence⟩=⟨character code⟩`
  Defines a control sequence to be a synonym for `\char ⟨character code⟩`.
- `\kchardef ⟨control sequence⟩=⟨character code⟩` (for upTeX/ε-upTeX)
  Defines a control sequence to be a synonym for `\kchar ⟨character code⟩`.
- `\mathchardef ⟨control sequence⟩=⟨15-bit number⟩`
  Defines a control sequence to be a synonym for `\mathchar ⟨15-bit number⟩`.
- `\omathchardef ⟨control sequence⟩=⟨27-bit number⟩` (for ε-pTeX/ε-upTeX)
  Defines a control sequence to be a synonym for `\omathchar ⟨27-bit number⟩`.

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 §10). The most efficient and convenient way of defining integer constants is as follows:

- 0–255: `\chardef`
- 256–32767: `\mathchardef`
- 32768–134217727: `\omathchardef` (only for ε-pTeX/ε-upTeX)
- (optional) 256–2147483647: `\chardef` (only for upTeX/ε-upTeX)

6.6.4 Creating a Japanese character token with a specified code

Short version:

- With ε-pTeX 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,

\begin{verbatim}
\begingroup
\lccode`?=\mycount
\lowercase{\endgroup \def\X{?}}
\end{verbatim}

defines \X which expands to a character number \mycount while the \catcode of ? (12) is preserved. However, the trick cannot be applied to Japanese characters, since p\TeX family does not support \lccode outside 0-255.

\kansuji is an expandable primitive like \number or \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 \kansujichar.

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

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

**Example 2:** equivalent to \def\日本{Japan}.

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

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

**\Uchar, \Ucharcat**

The “\kansuji trick” above includes an assignment of \kansujichar which is unexpandable. e-\TeX 191112 or later (\TeX Live 2020) provides expandable primitives \Uchar and \Ucharcat, which are derived from \Xe\TeX. Regardless of their names, and unlike \Xe\TeX or Lua\TeX, these primitives do not necessarily take a Unicode value as an argument. These primitives in e-\TeX and e-up\TeX take a valid character code (see \texttt{8.1}) based on the internal kanji encoding.
\texttt{\Uchar \langle character code \rangle}

Expands to a character token with specified slot \langle character code \rangle.

- 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.
- 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 \texttt{\epsilon-p\TeX}, 16–19 for \texttt{\epsilon-up\TeX}.

\texttt{\Ucharcat \langle character code \rangle \langle category code \rangle}

Expands to a character token with slot \langle character code \rangle and \langle category code \rangle specified.

- With \texttt{\epsilon-p\TeX}:
  - Only 8-bit number (0–255) are allowed for \langle character code \rangle; that is, only Latin characters can be generated.
  - The values allowed for \langle category code \rangle are 1–4, 6–8, 10–13.
- With \texttt{\epsilon-up\TeX}:
  - When \langle character code \rangle is between 0–127, only Latin characters can be generated. Thus, the values allowed for \langle category code \rangle are 1–4, 6–8, 10–13.
  - When \langle character code \rangle is between 128–255, both Latin and Japanese characters can be generated depending on the specified \langle category code \rangle; 1–4, 6–8, 10–13: Latin character, 16–19: Japanese character.
  - When \langle character code \rangle is greater than 255, only Japanese characters can be generated. Thus, the values allowed for \langle category code \rangle are 16–19.

6.7 Difference from pdf\TeX\ in DVI mode

As stated in Section 3, \texttt{\epsilon-p\TeX/\epsilon-up\TeX} are not upward compatible with DVI mode of pdf\TeX, which is available as the \texttt{etex} command in \TeX\ Live. Here we list some important differences: First, some pdf\TeX-specific primitives are absent. Examples:

- All primitives specific to PDF output: \texttt{\pdfoutput}, \texttt{\pdfinfo}, \texttt{\pdfobj} etc.\footnote{\texttt{\epsilon-p\TeX/\epsilon-up\TeX} has primitives \texttt{\pdfpagewidth} and \texttt{\pdfpageheight}; this is just because they were convenient for implementing \texttt{\pdfsavepos}, and their behavior is somewhat different from that of pdf\TeX. Also note that \texttt{\epsilon-p\TeX/\epsilon-up\TeX} does not have \texttt{\pdfhorigin} and \texttt{\pdfvorigin}.}
- All primitives related to micro-typography: \texttt{\pdffontexpand}, \texttt{\pdfprotrudechars}, etc.
- Some primitives related to handling of strings: \texttt{\pdfescapestring}, \texttt{\pdfescapehex} etc.
6.8 Recommendation for file encoding

6.9 Input handling

For simplicity, first we introduce of input handling of \textsc{e}-up\TeX{.}

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 \textsc{p}\TeX{ family automatically inserts glue and penalties between characters.

7.2 Japanese fonts

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

While typesetting, \textsc{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 \textsc{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 \textsc{ini}\TeX{} mode. Once a valid Japanese font is selected, there is no way of selecting “Japanese \texttt{\nullfont}” to discard all characters.

Moreover, \textsc{p}\TeX{} and friends assume that each Japanese font (except “Japanese \texttt{\nullfont}” in \textsc{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.
8 Other strange beasts

8.1 Internal kanji encodings

The (character code) 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/\up\TeX with \-kanji-internal=uptex (default on), the range is very simple:

\[ c \geq 0 \]

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

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

where

\[
\begin{align*}
C_1 &= C_2 = \{"a1,\ldots,"fe\} \quad \text{(euc)}, \\
C_1 &= \{"81,\ldots,"9f\}\cup\{"e0,\ldots,"fc\}, \quad C_2 = \{"40,\ldots,"7e\}\cup\{"80,\ldots,"fc\} \quad \text{(jis)}. \\
\end{align*}
\]

Therefore, the overall range of (character code) is not continuous.

To check whether an integer is a valid Japanese character code or not, you can use \iffontchar with \p\TeX 190709 or later (\TeX Live 2020). Suppose a count register \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 \jfont is used merely as a representative non-empty\(^9\) Japanese font containing all valid Japanese character code (see \textsection 7.2).

\footnote{\textsuperscript{9}This assumption is always safe after one of the standard p\TeX formats (e.g. plain p\TeX, pL\TeX) is loaded.}
References

https://ctan.org/pkg/dvistd

https://www.eijkhout.net/texbytopic/texbytopic.html
Index

Symbols

\accent ........................................ \inhibitxspcode ................. 10
\autospacing .............................. \jcharwidowpenalty .......... 10
\autooxspacing .......................... \jfam ................................. 10
\char ......................................... \jis ................................ 10
\chardef .................................. \kanjiskip ....................... 10
\currentcjktoken ...................... \kansuii ........................... 10
\currentxspacingmode ................\kansujichar....................... 10
\disablecjktoken .................. \kcatcode .......................... 10
\disinhibitglue ..................... \kchar ................................ 10
\dtou ........................................ \kchardef ......................... 10
\enablecjktoken ................... \kuts ................................ 10
\epTeXinputencoding ............. \lastnodesubtype ................ 10
\epTeXversion ......................... \lastnodeclasses ................ 10
\euc ........................................ \mathchardef ..................... 10
\expanded ................................ \noautospec ....................... 10
\font ...................................... \noautoxspacing ............ 10
\fontdimen ................................ \odelcode ................................ 10
\fontname ................................ \odelimiter .......................... 10
\forcecjktoken .................. \omathaccent ........................ 10
\fmsbox ................................ \omathchar ........................ 10
\if ........................................ \omathchardef ................ 10
\if ........................................ \omathcode ........................ 10
\ifcat ................................ \oradical ................................ 10
\ifdbx ................................ \pagefistretch ................ 10
\ifdef ................................ \partokencontext ............. 10
\ifincsname ......................... \partokenname ..................... 10
\ifjfont ................................ \pdfcreationdate ............. 10
\ifmbox ................................ \pdelfelapsedtime .......... 10
\ifmbox ................................ \pdfcreationtime .......... 10
\ifpdfdir ................................ \pdfcreationtime .......... 10
\iftbox ................................ \pdfcreationtime .......... 10
\iftdir ................................ \pdfcreationtime .......... 10
\iftfont ................................ \pdfcreationtime .......... 10
\ifybox ................................ \pdfcreationtime .......... 10
\ifydir ................................ \pdfcreationtime .......... 10
\inhibitglue ......................... \pdfcreationtime .......... 10

22