Chinese TeX Using the CJK LaTeX Package, Unicode TrueType Fonts and pdfTeX under Windows

I studied English, French and German in school in Norway, and Chinese for two years at UC Berkeley. (Unfortunately, my pronunciation is horrible in all languages!)

Because of my web page on the Chinese calendar, I need to put Chinese characters and pinyin on the web and in my TeX files.

The goal of this article is very modest. If you are an expert on CJK (Chinese, Japanese, Korean) typesetting, you will probably not learn anything new. But if you need to include some pieces of CJK text in your TeX document, then I hope that this article will prove useful.

There are two issues involved in using CJK in TeX. First you need to input the CJK text in your text editor, and then you need to compile it under TeX.

The CJK LaTeX package by Werner Lemberg is a wonderful tool, but it can be a bit hard to install, especially if you're not using Linux. Fortunately, it is now fairly easy to install under Windows using either TeXLive or MiKTeX. Using Unicode TrueType fonts and PDFTeX involves some additional steps, so I wanted to share my solution with other people.

In this article you will learn about

Inputting Chinese Characters

First of all, you need to have installed Chinese, both simplified and tradional if necessary, as an additional input language on Windows. You can do that when you first install Windows, using the language bar, or from "Regional and Language Options" in the Control Panel.

There are many ways to input Chinese text, but if you just need to write a few words, a simple solution is to use MS Word and the MS IME. Here are some links about the MS IME.

- Microsoft Global Input Method Editors (IMEs)
- What is an IME (Input Method Editor) and how do I use it?

- IME Tutorial

Save as plain text and choose the appropriate encoding, either UTF-8 (simplified/traditional), GB2312 (simplified) or Big5 (traditional). My emphasis will be on UTF-8. Make sure that no characters appear in red! Then open the text files in your TeX editor and copy over to your TeX document. I use WinEdt, and I cannot see the Chinese characters there. If I open the text files in Notepad, the Chinese appear in the UTF-8 case (assuming that you have installed Chinese, both simplified and tradional if necessary, as an input language), but not in the other cases. However, if I try to copy the UTF from Notepad to WinEdt, it does not come out right, I must open the text file in WinEdt.

Here are three sample files, `test-UTF8.tex`, `test-GB2312.tex` and `test-Big5.tex`. If you want to test all three, you probably need to have both simplified and traditional input support installed.

If you're looking for a good Windows Unicode editor, you may want to check out EmEditor. You can use the MS IME with it! They give out academic licenses for free!

Some Unicode editors may add the "BOM" (Byte-order Mark) or UTF-8 signature at the very beginning of the file, even if the output file encoding is set to UTF-8. The BOM under UTF-8 is the byte sequence `0xEF 0xBB 0xBF`. If the output file starts with those three bytes, they should be removed, or you may get strange warnings, log entries, or even errors while processing with LaTeX.

Winedt is a very good program, but not so good at dealing with UTF-8 CJK. If you get a message about "Some Characters were lost during the UTF-8 to ANSI conversion", you have probably saved your file with the UTF-8 signature.

Installing CJK under Windows

This is no problem with either MiKTeX or TeXLive. Just select the packages the usual way.

Compiling the Example Files

Under MiKTeX, the GB and Big5 example files work fine, but the UTF8 example does not work. There are several ways of resolving this. One method is to convert TrueType fonts to PostScript Type 1 format by creating tfm, pfb and map files. The MiKTeX version of the CJK package has done this for the gbsnlp and bsmilp fonts used in the GB and Big5 examples, but not for the cyberbit font used in the UTF8 example. However, PDFTeX can use TrueType fonts directly, by creating tfm, enc and map files. I will describe how to use this method for the cyberbit font.

- Download `Cyberbit.ZIP` from ftp://ftp.netscape.com/pub/communicator/extras/fonts/windows/Cyberbit.ZIP. Rename `Cyberbit.ttf` to `cyberbit.ttf` and put it in `localtexmf\fonts\truetype\bitstream`.
- Run

```
ttf2tfm cyberbit.ttf -w cyberb@Unicode.sfd@ > cyberbit.log
```

This creates 165 tfm and 165 enc files. The -w option is important! That is how you get the enc files.

- Put all the tfm files in `localtexmf\fonts\tfm\bitstream\cyberb`.
- Put all the enc files in `localtexmf\fonts\enc\pdftex\cyberb`.
- Check if `texmf\ttf2tfm\base\ttfonts.map` contains the line

  ```
  cyberb@Unicode@ cyberbit.ttf
  ```

  If not, create the file `localtexmf\ttf2tfm\base\ttfonts.map` with that line in it.
- Download delloye.free.fr/cyberbit.map, rename it `cyberb.map`, replace `cyberbit` with `cyberb`, except for `cyberbit.ttf`, throughout the file, and place it in `localtexmf\fonts\map\pdftex\cyberb`.
- Create the file `localtexmf\web2c\updmap.cfg` with the line

  ```
  Map cyberb.map #localtexmf\fonts\map\pdftex\cyberb\cyberb.map
  ```

  in it.
- Run `mkfntmap` (or `updmap` or `initexmf --mkmaps`).
- Refresh the file name database.

The cyberb name is a remnant of old file name restrictions. If you want to write `cyberbit` instead of `cyberb`, you'll have to do the following.

- Run

  ```
  ttf2tfm cyberbit.ttf -w cyberbit@Unicode.sfd@ > cyberbit.log
  ```

- Put `cyberbit@Unicode@ cyberbit.ttf` in `localtexmf\ttf2tfm\base\ttfonts.map`.
- Use delloye.free.fr/cyberbit.map, as it is.
- Put the enc file in `localtexmf\fonts\enc\pdftex\cyberbit`.
- Put the map files in `localtexmf\fonts\map\pdftex\cyberbit`.
- Add the line

  ```
  Map cyberbit.map #localtexmf\fonts\map\pdftex\cyberbit\cyberbit.map
  ```

  to localtexmf\web2c\updmap.cfg.
- Download delloye.free.fr/c70cyberbit.fd into `localtexmf\tex\latex\cyberbit`, or copy `texmf\tex\latex\CJK\UTF8\c70song.fd` into that directory, rename it `c70cyberbit.fd` and replace cyberb with cyberbit.

This should work on most modern TeX systems. If your system insists on making pk files, you need to fix your `updmap`.

With TeXLive, it is enough to create the tfm files (no need to use the `-w` option) and to edit `ttfonts.map`. However, this will only generate pk fonts.

If you want to run the `CJKbabel.tex` file, you can download the `t5.sty` file from CTAN, put it somewhere in localtexmf and refresh the file name database. Another alternative is to use the vntex package.

Some useful links.

- [Using Truetype fonts and Unicode in Pdflatex]() by Otfried Cheong.
- [How to make LaTeX (teTeX) handle unicode and CJK in MacOSX]() by Pai H. Chou.
- [CJK Support]()

Thanks to Werner Lemberg (the author of the CJK package), Danai Sae-Han, and Harald Hanche-Olsen for patient e-mails, to Olivier Delloye for his useful posting, and to Pai H. Chou for his helpful web page!

Combining Simplified and Traditional Characters

A simple solution is of course to use Unicode, but it can also easily be achieved by using the `\CJKencoding` command. You may want to look at sample-utf8.tex and sample-gb-big5.tex.

Character Encoding Conversion

Sometimes your Chinese text doesn't come out right, because it uses the wrong character encoding. This used to be a big problem, but now you can do character encoding conversion with Chinese Encoding Converter at Erik E. Peterson's On-line Chinese Tools.

Finding Unicode Codes

I often need to know the Unicode code for Chinese characters, either for TeX or HTML. You can input the characters in MS Words and copy them into Chinese Character Dictionary - Unicode Version at Erik E. Peterson's On-line Chinese Tools. You have to select the box for showing Unicode Value in the results and select UTF-8, and not Unicode, for the input. The other version, Chinese Character Dictionary, will not work, since it does not have the UTF-8 option. To convert to octal, you can use Conversion Table - Decimal, Hexadecimal, Octal, Binary.

You can also use Convert characters to Unicode at pinyin.info.

Creating CJK Bookmarks with Hyperref and Beamer

To get anything unusual, like for instance pinyin tone marks, into the bookmarks when using hyperref, you have to use the `\texorpdfstring{}{}` command, for example ``\section{\texorpdfstring{\Jie2 \Qi4}{Ji\'e Q\`i}}``. This example was easy, because the marks for the second or fourth tones can be obtained by standard accented characters. If you want to get the first tone, you have to first get the Unicode code from either Reading and Writing Chinese Characters and Pinyin on the Web Using Unicode, Code Charts (PDF Version) at the Unicode Home Page or from the file puenc.def in the hyperref package. You then write for example ``\section{\texorpdfstring{\Tang1 \Ruo4 \Wang4}{T{\001\001}ng Ru\`o W\`ang}}``. If you want to get the Chinese characters, you can use Chinese Character Dictionary - Unicode Version or Convert characters to Unicode, and write for example `\section{\texorpdfstring{湯若望}{\156\157 \202\345 \147\033}}`. Please note that the pdfstring use octal numbers for the hex Unicode strings. To convert to octal, you can use Conversion Table - Decimal, Hexadecimal, Octal, Binary.

For more info on pinyin tone marks you may look at my page on Reading and Writing Chinese Characters and Pinyin on the Web Using Unicode.

Helmer Aslaksen
Department of Mathematics
National University of Singapore
helmer.aslaksen@gmail.com

I use the W3C MarkUp Validation Service and the W3C Link Checker.