Getting Reasonable Chinese Characters in LaTeX Documents

Paul E. Johnson

January 22, 2009

From the LyX wiki (http://wiki.lyx.org/LyX/XeTeX) , I learned that the program xelatex from the XeTeX addon for LaTeX can replace the latex program. The xelatex program can generate pdf documents from TeX files that include Chinese and other international characters that are in the utf-8 font format. It is not just for Chinese, but that was my major purpose when I started to look into this. XeTeX is supposed to accept all unicode characters.

XeTeX may accept all characters, but that does not mean they show up in the output. I was very discouraged after I experimented with the file “xetex.lyx” on the LyX wiki. While the five Chinese characters in the demonstration do work, a user quickly realizes that additional characters are almost all missing from output. I installed the SCIM input system and typed in 20 or 30 characters and about 3/4 of them were missing from the pdf output. The characters show on the screen if there is an X11 screen font for them, but the fonts selected in that example file are less comprehensive. How discouraging. And the worst part is that the missing characters show in pdf output as “blanks”, not as errors. So the author of a long document may never realize that a character is missing. That is simply unacceptable, as far as I’m concerned.

Finding XeTeX discouraging, I started researching an alternative approach, the so-called CJK-LaTeX approach. That uses some LaTeX packages to provide Chinese-Japanese-Korean fonts in TeX documents. I have found the CJK-LaTeX strategy to be very difficult. Not only is there a lot of LaTeX syntax to learn, but there is also a lot of system configuration required. The Chinese fonts for latex are not widely available in high quality sets. The addition of Chinese fonts to allow use of the latex processor is a very difficult project. In addition, while editing a document that includes Chinese-Japanese-Korean characters, one faces the problem of deciding on an “encoding” for the characters, such as big5 or gb (or many other possibilities that make my head spin).

The strong point in favor of XeTeX is that it does not require the complicated font management scheme or management of encodings that latex requires in order to handle Chinese. Rather, we can use utf-8 encodings and nice documents come out.

After frustrating myself with CJK-LaTeX, I looked again at XeTeX and discovered that the missing characters were not caused byXeTeXitself, but rather by the particular font used in the XeTeX example on the LyX wiki. I've learned that there are more complete unicode fonts that appear to offer much broader converge of international characters. This TeX file represents my effort to find a working system for Chinese authors.

Make LyX work

In the LyX preferences, the following can be inserted to create a menu item “PDF (xetex)” under the LyX pulldown.

```
#FORMAT SECTION
\format "pdf4" "pdf" "PDF(xelatex)" "" "xdg-open" "" "document,vector"
#CONVERTER SECTION
\converter "pdflatex" "pdf4" "xelatex $$i " "latex"
```

If you don't have a system with “xdg-open,” insert your favorite pdf viewer in place of that command.

You may just want to use this document as a template for new LyX documents that will be processed with XeTeX, but if you start from scratch, the only vital changes are the following.

1. In the Lyx Document/Settings/Language menu, change the encoding to “utf8”. Don't choose any language-specific encoding. not even CJK-utf8 or whatever other variant looks nice. We are using Unicode UTF-8 because it includes just about every character you need.

2. Put these things in your preamble under LyX Document/Settings/Preamble:

   This is the bare minimum

```
\usepackage{fontspec}
\usepackage{xunicode}
\usepackage{xltxtra}
```

That is not quite sufficient to get a nice looking document. The double quotation mark will not be rendered in the pretty LaTeX style unless this is inserted:

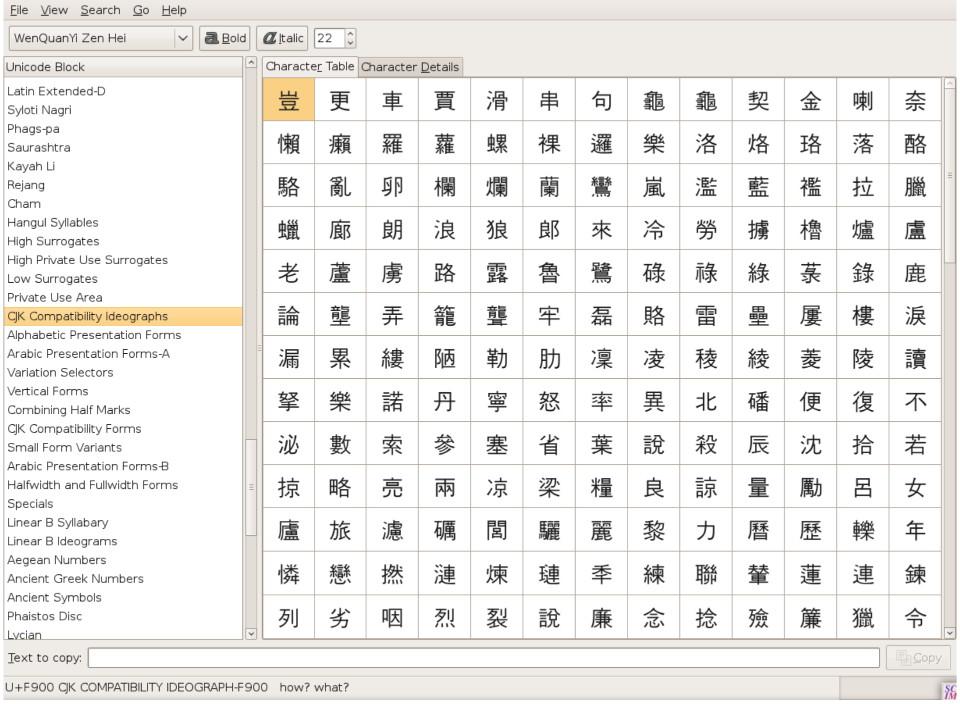
```
\defaultfontfeatures{Mapping=tex-text}
```

Without that, then “this” appears as ‘ ‘ this ’ ’. The preamble can be used to specify one's preferred fonts for the roman, sans serif, and typewriter fonts as follows. Note that the Mapping feature is included where desired.

```
\setromanfont[Mapping=tex-text]{LMRoman12}
\setsansfont[Mapping=tex-text]{LMSans12}
\setmonofont{LMTypewriter12}
```

At the current time, I am experiencing a bug in either LyX or XeTeX that causes xelatex to ignore the Mapping option as applied for all fonts, and so a workaround is to include it separately for both the roman and sans fonts.

Figure 1: Gnome Character Map



WenQuanYi Zen Hei

22

Unicode Block

Latin Extended-D
Syloti Nagri
Phags-pa
Saurashtra
Kayah Li
Rejang
Cham
Hangul Syllables
High Surrogates
High Private Use Surrogates
Low Surrogates
Private Use Area
CJK Compatibility Ideographs
Alphabetic Presentation Forms
Arabic Presentation Forms-A
Variation Selectors
Vertical Forms
Combining Half Marks
CJK Compatibility Forms
Small Form Variants
Arabic Presentation Forms-B
Halfwidth and Fullwidth Forms
Specials
Linear B Syllabary
Linear B Ideograms
Aegean Numbers
Ancient Greek Numbers
Ancient Symbols
Phaistos Disc
Lycian

豈	更	車	賈	滑	串	句	龜	龜	契	金	喇	奈
懶	癩	羅	蘿	螺	裸	邏	樂	洛	烙	珞	落	酪
駱	亂	卵	欄	爛	蘭	鸞	嵐	濫	藍	襤	拉	臘
蠟	廊	朗	浪	狼	郎	來	冷	勞	擄	櫓	爐	盧
老	蘆	虜	路	露	魯	鷺	碌	祿	綠	菉	錄	鹿
論	壟	弄	籠	聾	牢	磊	賂	雷	壘	屢	樓	淚
漏	累	縷	陋	勒	肋	凜	凌	稜	綾	菱	陵	讀
拏	樂	諾	丹	寧	怒	率	異	北	磻	便	復	不
泌	數	索	參	塞	省	葉	說	殺	辰	沈	拾	若
掠	略	亮	兩	凉	梁	糧	良	諒	量	勵	呂	女
廬	旅	濾	礪	閭	驪	麗	黎	力	曆	歷	轢	年
憐	戀	撚	漣	煉	璉	秊	練	聯	輦	蓮	連	鍊
列	劣	咽	烈	裂	說	廉	念	捻	殮	簾	獵	令

U+F900 CJK COMPATIBILITY IDEOGRAPH-F900 how? what?

Choosing fonts

The XƎTEXframework uses the OpenType Latin Modern fonts by default. Its manual wryly observes that if Latin Modern fonts are not installed, one would be well advised to obtain them.

The fontspec package is not actually required, but if you don't use it, then you will not be able to choose fonts for yourself. After including the fontspec package, the following sort of command will be used

\fontspec{LMRoman12} to indicate that the following text will be in the Latin Modern font size 12.

The fontspec option accepts the so-called "display name" of the font. The display name can contain spaces.

I had a little trouble finding the available fonts and their legal names. On a Gnome Linux system, the "display names" of the fonts can be obtained from the Gnome Character Map program. Its interface is illustrated in Figure 1. On the top left, observe that I've selected a font with a display name "WenQuanYi Zen Hei" and below that is a list of Unicode "plates", the many different blocks of symbols that are available.

According to the fontenc manual, italics can be obtained by inserting the option \textit{something} to italicize *something*. The ERT is not needed in LyX. xelatex also accepts LyX's italic menu item *E* to operate correctly. One can also obtain **bold** by LyX's character style setting

Demo

Within the tex file, fonts may be changed with the fontspec macro like this:

```
\fontspec{Times New Roman} or \fontspec{LMRoman12}
```

Again, any valid display font name is allowed. Also, note that this declaration remains in effect until it is changed.

I've found 3 sets of fonts that provide a more-or-less comprehensive set of Chinese characters that xelatex can used when generating pdf files.

When I was trying to use CJK-LaTeX, I went through a rather tedious process to install the Bitstream Cyberbit true type font and then go through a long fontconfig exercise to generate the tfm files that LaTeX needs. Many Chinese-for-latex sites seem to say we should Cyberbit, but I'm forming the conclusion that they are offering out-dated or bad advice. That font is commercial and somewhat illegitimate in heritage because the Bitstream company no longer offers it and it only survives by people trading it among themselves. Noncommercial use of Cyberbit is a violation of its licensing. After installing Cyberbit, I learned of the GNU unifont packages, which are available for Linux, and the UNIbit project, which is also a very complete version of unicode with rigorous coverage of Chinese characters. So, if I can get good results with them, I'd rather not use Bitstream Cyberbit.

Anyway, here's how you let the document know which font to use.

```
\fontspec{Bitstream Cyberbit} or {unifont} or {WenQuanYi Zen Hei}
```

I'm compiling this document with the xelatex program that is shipped with TeXLive-2007. TeXLive-2008 is out, but not packaged for Linux distributions as far as I know. TeXLive-2008 is supposed to have an even better version of xelatex. We'll wait and see.

I don't speak/write Chinese, so I have to ask my Chinese friends about these results. Do you care to advise me? Is one set of fonts noticeably better in the PDF output?

Demonstration of Chinese from WenQuanYi Zen Hei, the unibit font:

客独美国

本常问问答集

是从一些经常被问到的问题及其适当的解答中，

以方便的形式摘要而出的。跟上一版不同的是，

其编排结构已彻底改变。有关新结构的细节，

可参考「如何阅读本问答集及了解其编排结构」该 项中的说明。

Demo of Chinese from the "GNU unifont":

客独美国

本常问问答集

是从一些经常被问到的问题及其适当的解答中，

以方便的形式摘要而出的。跟上一版不同的是，

其编排结构已彻底改变。有关新结构的细节，

可参考「如何阅读本问答集及了解其编排结构」该 项中的说明。

Here is the same from Bitstream Cyberbit:

客独美国

本常问问答集

是从一些经常被问到的问题及其适当的解答中,

以方便的形式摘要而出的。跟上一版不同的是,

其编排结构已彻底改变。有关新结构的细节,

可参考「如何阅读本问答集及了解其编排结构」该 项中的说明。

The Firefly font was the new big thing in 2005. It's described here: `http://unifont.org/fontguide`. The Font is called "AR PL New Sung"

客独美国

本常问问答集

是从一些经常被问到的问题及其适当的解答中，

以方便的形式摘要而出的。跟上一版不同的是，

其编排结构已彻底改变。有关新结构的细节，

可参考「如何阅读本问答集及了解其编排结构」该 项中的说明。

For purposes of comparison, here is the same from a font called AR PS KaitiM Big5 (which is recommended on this website: http://www.cantonese.sheik.co.uk/fonts.htm). It looks very incomplete to me:

客　美

本常　　答集

是　一些　常被　到的　　及其适　的解答中，

以方便的形式摘要而出的。跟上一版不同的是，

其　排　构已　底改　。有　新　构的　　，

可　考「如何　　本　答集及了解其　排　构」　　　中的　明。

Maybe AR PL KaitiM Big5 is not intended for this kind of work. As I said, I don't speak Chinese.

How do you like "AR PL UMing CN" ?

客独美国

本常问问答集

是从一些经常被问到的问题及其适当的解答中,

以方便的形式摘要而出的。跟上一版不同的是,

其编排结构已彻底改变。有关新结构的细节,

可参考「如何阅读本问答集及了解其编排结构」该 项中的说明。

How do you like AR PL UKai CN ?

客独美国

本常问问答集

是从一些经常被问到的问题及其适当的解答中，

以方便的形式摘要而出的。跟上一版不同的是，

其编排结构已彻底改变。有关新结构的细节，

可参考「如何阅读本问答集及了解其编排结构」该 项中的说明。

With the Microsoft Internet Explorer 5.5, Microsoft distributed a true type fonts “MS Song” and “MS Hei”. I’m trying to find out what the rules are about that font.

This is MS Song:

客独美国

本常问问答集

是从一些经常被问到的问题及其适当的解答中，

以方便的形式摘要而出的。跟上一版不同的是，

其编排结构已彻底改变。有关新结构的细节，

可参考「如何阅读本问答集及了解其编排结构」该 项中的说明。

This is MS Hei:

客独美国

本常问问答集

是从一些经常被问到的问题及其适当的解答中，

以方便的形式摘要而出的。跟上一版不同的是，

其编排结构已彻底改变。有关新结构的细节，

可参考「如何阅读本问答集及了解其编排结构」该 项中的说明。

How do you like Times New Roman? This surprised me a little bit, because the Gnome Character Map seems to indicate that “Times New Roman” has some Chinese characters. What do you see?

□□□□

□□□□□□

□□□□□□□□□□□□□□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□

□□□□□□□□□□□□□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□ □□□□□□

I see nothing there!

I do not know why the Cyberbit output is lighter than the others. I find that I can specify bold fonts in this TEX document, but the output pdf does not have have a bold font.

Oh, well, that's enough Linux LATEX frustration for a weekend.