

# Re: Chinese character & pinyin frequency analysis

---

*Source:* <http://sci.tech-archive.net/Archive/sci.lang/2007-10/msg00432.html>

---

- *From:* "Richard Wordingham" <jrw0602@xxxxxxxxxxxx>
  - *Date:* Mon, 15 Oct 2007 02:29:03 GMT
- 

"LEE Sau Dan" <danlee@xxxxxxxxxxxxxxxxxxxxxxxxxxxx> wrote:

"Richard"  
== Richard  
Wordingham  
<jrw0602@xxxxxxxxxxxx>  
writes:

Richard> How does Emacs help with my telling where ZWJ and ZWNJ  
Richard> have been placed? I did an experiment with 'Ca esar'  
Richard> (with ZWJ between 'a' and 'e') and found 'a ZWNJ e'  
Richard> displayed as 'a e'.

Correction: 'a ZWJ e' displayed as 'a e'.

What are ZWJ and ZWNJ?

Basically characters respectively mandating (ZWJ = Zero Width Joiner) and prohibiting (ZWNJ = Zero Width Non-Joiner) ligation of the characters either side. There are some complications in their semantics for cursive scripts (prototypically Arabic) and Indic scripts (prototypically Devanagari). I understand they'd be rather important for a proper fraktur font.

Richard> In Code2000 it displays the same as 'æ'. I see I've  
Richard> still got some work to do to get Emacs to use the  
Richard> Code2000 font, assuming that it's possible.

What's Code2000?

The usual example of a font that attempts to cover the entire Basic Multilingual Plan (less the Private Usage Area).

## Re: Chinese character & pinyin frequency analysis

Richard> Using a legacy encoding for compacting will also result  
Richard> in one's having a pair of source and derived files, and  
Richard> possible problems if the editor is not clever enough to  
Richard> convert the character encoding declaration.

It's a file or workflow management issue. We always have lots of files and copies (real ones or virtual) around. Having them isn't a problem. Not managing them properly is.

Nobody has ever complained that after writing a C program, one has to compile it and that generates a second file---the executable. (Intermediate-level programs may need to use a third, intermediate file---object code, too.) Why?

Because there is little temptation for maintenance to be done on the object or executable files. On the other hand, when one uses automatic code generators, (e.g. Statemate or Simulink to C) there is a severe risk of maintenance being done on the derived files rather than, in these cases, the 'pictorial' sources. In the case I have in mind, the temptation is extremely high.

If you're using the native Win32 port of Emacs, then Unix font specs. are irrelevant. The Win32 port should be using native Windows fonts. You may need to consult the manual, though, to find out the details.

Does one exist for the Win32 port? I'm looking through a FAQ, and it chiefly refers to the Unix-style names. I've tried some of the example commands. They tell me that Arial (for example) is available, but it doesn't appear on the GUI font menu.

Richard> – but they do get saved when I save the file in UTF-8 or  
Richard> UTF-16. It looks as though there is more work to do on  
Richard> Emacs for Unicode support.

Yes. Unicode is not native for Emacs. But given that many people use Emacs to edit UTF-8 and UTF-16 files, and they have been doing that for years, I don't think there would be a very big problem. You may want to search the web for some additional Emacs packages that can facilitate your work.

Possibly. That is getting distinctly obscure.

## Re: Chinese character & pinyin frequency analysis

There also definitely seems to be a problem with proportional fonts. It would seem that one once had to use fixed width Thai fonts (i.e. 0 or 1 cells per character) for MULE, and I suspect that may still be the case for serious editing. A freshly pasted line is legible, but moving a cursor through it or typing it afresh causes serious problems. The word I3 is unrecognisable after typing it, though repasting the line does cause it to display properly.

.