

# Re: Chinese character & pinyin frequency analysis

---

*Source:* <http://sci.tech-archive.net/Archive/sci.lang/2007-10/msg00439.html>

---

- *From:* LEE Sau Dan <[danlee@xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx](mailto:danlee@xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx)>
  - *Date:* Mon, 15 Oct 2007 08:03:11 +0800
- 

"Richard" == Richard  
Wordingham  
<[jrw0602@xxxxxxxxxxx](mailto:jrw0602@xxxxxxxxxxx)>  
writes:

>> Bad design. Font and characters are 2 separate things. A  
>> text-editor should only be concerned with text characters, not  
>> fonts. Mixing the element of "font" into a text-search is  
>> absurd.

Richard> You misunderstand me. The issue is how the editor is to  
Richard> display the text one is editing. The obvious technique  
Richard> is to take the characters from a font, perhaps with a set  
Richard> of customisation options. The problem arises when the  
Richard> characters being used are not covered by a single font.

Usually, it is hard to find a single font that covers the whole  
Unicode character set. So, software developers have come up with the  
"font set" abstraction: a collection of fonts. The fonts in a font  
set may cover overlapping subsets of the Unicode character set. But  
if you have enough fonts to cover the whole set, or the subset of  
Unicode that you're using, things will display well.

If your editor cannot do that, then, that's not a good one.

Richard> The solution with a simple user interface (i.e. no user  
Richard> control) is for the editor (or one of its agents) to  
Richard> search the installed fonts when the selected font does  
Richard> not handle the specific character. Of course, this can  
Richard> work badly when diacritics and base characters do not  
Richard> come from the same font. (It seems that Emacs provides  
Richard> user control over font mixing.)

Emacs has been supporting fontsets for a decade.

## Re: Chinese character & pinyin frequency analysis

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'.

What are ZWJ and ZWNJ?

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?

Richard> I have, very occasionally, resorted to fixing Word files  
Richard> by editing the RTF files as plain text.

>> Why do you need to do that? Because Word sucks?

Richard> I've had difficulties lining up cells in tables and with  
Richard> controlling the incorporation of pictures.

So, use a more decent tool instead of Word.

Richard> This had nothing to do with Word. There was a bug in  
Richard> VAX/VMS Run-Time Library and we couldn't move up to the  
Richard> next version of the operating system because we were  
Richard> dependent on a cheap bolt-on that didn't work on the next  
Richard> version of the operating system.

Sigh... legacy systems. They're like plague. The sooner you get rid  
of them, the better. Legacy systems only drags progress.

>> Would you choose to buy a badly designed car and then be  
>> obliged to fix a few screws everyday, or buy a decent car that  
>> have all the screws working well out of the factory?

Richard> Quite a few people choose to run old bangers, some even  
Richard> resorting to specially tuning the engines themselves to  
Richard> get them through emission tests and then retuning them to  
Richard> get adequate power. They reckon they are saving money.

If they know what they're doing and they possess the necessary skills,  
then that's fine.

## Re: Chinese character & pinyin frequency analysis

But in your case, do you have such skills? Do you have the skills to modify or fix your broken tools to do a better job?

Richard> So if I am to use Emacs, am I to create an appropriate  
Richard> encoding?

You can. You can even automate many things by written some E-Lisp code. It's a Turing-complete language. So, what cannot be done?

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 of 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?

Richard> 5. It's a lot quicker to type '&#331;' for eng than to  
Richard> fiddle about with keyboard selections.

>> You mean different Input Methods?

Richard> Strictly yes, but to me 'input method' suggests something  
Richard> more elaborate than a software layout.

Yes. "software keyboard layout" is just a very simple special case of input methods. :)

Richard> Incidentally, I've been trying out Emacs 22.1.1 (build  
Richard> environment i386-mingw-nt5.1.2600) on Windows XP. I'm  
Richard> having severe problems with Lao and Khmer. If I type Lao  
Richard> or Khmer in, it seems as though the input is not  
Richard> understood. If I paste them in, I still can't yet get  
Richard> them to display - I hope simply because, not really  
Richard> grokking Unix font specifications, I haven't worked out  
Richard> how to specify an appropriate font

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.

Re: Chinese character & pinyin frequency analysis

You may need to consult the manual, though, to find out the details.

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.

Richard> I've a lot of study to do in setting up Emacs 22.1.1.  
Richard> I've yet to work out why my customisation script for  
Richard> MS-DOS Emacs 20.5 made Emacs 22.1.1 sick.

I can't help you here. I've given up Windows/DOS for over a decade. One thing is sure: Emacs 22 and Emacs 20 are 2 major revisions apart. There are many differences.

—  
Lee Sau Dan N^f ~{ @nJX6X~ }

E-mail: danlee@xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx  
Home page: <http://www.informatik.uni-freiburg.de/~danlee>