

Re: Chinese character & pinyin frequency analysis

Source: <http://sci.tech-archive.net/Archive/sci.lang/2007-10/msg00446.html>

- *From:* "Richard Wordingham" <jrw0602@xxxxxxxxxxxx>
 - *Date:* Sun, 14 Oct 2007 20:18:41 GMT
-

"LEE Sau Dan" <danlee@xxxxxxxxxxxxxxxxxxxxxxxxxxxx> wrote in message
<news:87sl4el7dq.fsf@xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx>

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

Then, upgrade your editor. If you're serious enough to use Chinese characters, you should be using one that does it properly.

Richard> Microsoft recently upgraded Notepad, at least for Windows
Richard> XP users. It now searches one's fonts for characters not
Richard> in the font you are using, so that problem has *now*
Richard> largely gone away.

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.

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

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

Re: Chinese character & pinyin frequency analysis

Richard> So which editor do you suggest?

I use Emacs.

How does Emacs help with my telling where ZWJ and ZWNJ have been placed? I did an experiment with 'Ca esar' (with ZWJ between 'a' and 'e') and found 'a ZWNJ e' displayed as 'a e'. In Code2000 it displays the same as 'æ'. I see I've still got some work to do to get Emacs to use the Code2000 font, assuming that it's possible.

Richard> And can I sure be an editor will not normalise my input?

A text editor should NOT fiddle with your input unless you explicitly instruct it to. That's What You Get Is What You Want.

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?

I've had difficulties lining up cells in tables and with controlling the incorporation of pictures. Viewing RTF has also been the most straightforward way of ensuring that pictures have been embedded into what started as a programmatically generated RTF file. The smarter approach would have been to embed the pictures when generating the file, but I don't know how to do that, so I had to convert references to embeddings manually.

Richard> And, yes, I do resort to editing binary files when the
Richard> need arises – the worst case was having to edit a VAX
Richard> object file to initialise an additional register.

You should be using something that's better than Word in that aspect.

This had nothing to do with Word. There was a bug in VAX/VMS Run-Time Library and we couldn't move up to the next version of the operating system

Re: Chinese character & pinyin frequency analysis

because we were dependent on a cheap bolt-on that didn't work on the next version of the operating system.

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?

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

Richard> 3. It can be tempting to compact text by using a legacy
Richard> encoding. There are also message boards where characters
Richard> will get misinterpreted – I have had to enter accented
Richard> letters as character entities to avoid them being
Richard> misinterpreted according to a legacy code.

Use an editor that can do that automatically. :)

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

Richard> If you have one to hand.

Emacs. And if you editor can't convert between encodings, use a tool to do that (e.g. GNU iconv). It isn't that difficult to write a simple Perl script to do custom conversions, either.

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.

Write a Perl script and the process can be automated. Write it once and use it millions of times.

Re: Chinese character & pinyin frequency analysis

Again, these ought to be delegated to a decent editor.

Richard> So which cheap editor do you suggest for HTML
Richard> incorporating ECMA-script ('javascript')?

I use Emacs.

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

You mean different Input Methods?

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

Incidentally, I've been trying out Emacs 22.1.1 (build environment i386-mingw-nt5.1.2600) on Windows XP. I'm having severe problems with Lao and Khmer. If I type Lao or Khmer in, it seems as though the input is not understood. If I paste them in, I still can't yet get them to display – I hope simply because, not really grokking Unix font specifications, I haven't worked out how to specify an appropriate font – but they do get saved when I save the file in UTF-8 or UTF-16. It looks as though there is more work to do on Emacs for Unicode support.

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

Richard.

.