

Re: Chinese character & pinyin frequency analysis

Source: <http://sci.tech-archive.net/Archive/sci.lang/2007-10/msg00454.html>

- *From:* "Richard Wordingham" <jrw0602@xxxxxxxxxxxx>
 - *Date:* Sat, 13 Oct 2007 19:29:04 GMT
-

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

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

The real question is: why bother doing it manually? And why bother doing it at all?

Richard> 1. Older text editors aren't much good at mixing fonts.

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

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

It wasn't a problem with Chinese characters, though Windows Vista apparently includes some horrible bodes to get round the TrueType limit of 64K glyphs per font. The problem I had was with mixing Khmer and IPA.

Richard> 2. Some of the time I use these codes (jargon: character
Richard> entities) to check the interpretation of control-like
Richard> characters, such as ligature controls, or to effect a
Richard> choice of normalisation. These would not readily show up
Richard> in most editors.

Such things should be done by programs (and programmers debugging their programs). You don't use a hex-editor to create/check your

Re: Chinese character & pinyin frequency analysis

files, do you? Then, why check the unicode?

So which editor do you suggest? And sometimes I do resort to hex dumps to find out what characters are in a piece of text, though Windows 2002 and later offers an alternative method – so long as one hasn't had to resort to a hack font. And can I sure be an editor will not normalise my input?

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

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

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. :)

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

Richard> 4. There are a few characters that are best entered in
Richard> HTML text as character entities ('<', '&' and
Richard> multi-character white space immediately spring to mind),
Richard> though there are symbolic names for these.

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

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

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

What are "keyboard selections"?

Selecting keyboard layouts. For small scripts (or language systems using small subsets), one normally selects

Re: Chinese character & pinyin frequency analysis

a script- or language- specific keyboard layout. This, if I want to mix Thai, Lao, Khmer and Latin-1, I would normally switch between four different keyboard layouts. (I'm seriously considering knocking one up for IPA.) However, if a lot of keyboard layouts are enabled, switching keyboards is as tedious as switching fonts.

Richard.

.