Unicode and org mode

William Henney

03 novembre 2007

Notes on using unicode characters in org mode

How to enter the unicode characters

Use either the SGML or TeX input method.

Using the TeX input method

- Type `C-u C-\ tex` to activate
- Type things like α or x^2 and they will be translated into the unicode glyph. Use tab for completion help.
- Pro: "Intuitive" to use.
- Con: Gets in the way of typing a "real" backslash

Using the SGML input method

- Type `C-u C-\ sgml` to activate
- Type things like `α` or `°` to get α and °.
- Pro: Access to more glyphs than with TeX it seems
- Con: No access to sub/superscripts

Punctuation

We can use the em and en dashes—this clause is bounded by em dashes—directly in the org file. However, they aren't very easily distinguishable in some fonts, especially fixed width ones at small sizes. Here is a range of numbers separated by an en dash: 223–999. In this sentence – following British typographic convention – the en dash is used like the em dash is used in American typography. Here are some minus signs:— binary $(223 - 999)$ and unary (-0.2). Finally, here is a hyphen for comparison: a-b. They look good in proportional fonts, such as Times, Futura and Optima. Baskerville

is the font where they look most like their Computer Modern versions. In fact, Baskerville looks quite a lot like CMR in other ways too... Oh, and that was an ellipsis.

```
Test in fixed-width font:-- range 666-999
```

symbol	examples
hyphen	1-2 a-b
en dash	1–2 a–b
em dash	1—2 a—b
minus	1−2 a−b

It seems that the glyphs for the non-ascii characters are always taken from those of the font family of the `default` face, even where the font-lock face is specifically set to another font family.

Dealing with pre-formatted text

```
This uses the org-code face, so we can easily make
it fixed-width
```

Even if we are using a proportional font family for the `default` face, by customizing the `org-code` face, we can use a fixed-width font (such as Monaco) for pre-formatted material (lines starting with ":" and words delimited with "="). We can do the same with the `org-table` face, so that the alignment of table lines still works. In the case of the pairing of Monaco and Times, it is also necessary to set the height of the fixed-width faces to 0.85, so that the character sizes match up.

Bugs

1. Table alignment still won't be quite right if there are unicode characters in the table cells, since the glyphs for these have variable widths, even in a *supposedly* fixed-width font like Monaco.

2. It doesn't work for sections with the QUOTE keyword, since these do not use any special face.

Other typographical symbols (e.g., §)

Diamond character: ◇

Greek letters and math symbols: $\alpha = x^2 - y^2$

Examples: $\frac{1}{2} \int \Xi_0 \, \mathrm{d}z = \hbar c / \lambda \Rightarrow$

Variations between fonts (Mac OS X 10.4/Aquamacs 1.2)

As far as I can see, only a few fonts have their own set of glyphs for the Greek letters. Times has a nice set of glyphs, although it does have the problem that italic nu and italic v look *very* similar. Spot the difference: νv ! Most font families use a common set of glyphs that have a Sans Serif feel to them, as though they were designed to go with Helvetica (although Helvetica actually uses a slightly different set). These glyphs have the problem that the "gamma" looks too much like a "y" and the "tau" looks like a "t". When used with Monaco, they look too small.

Super- and sub-scripts

These don't exist for all letters.

Example alphabets

$\alpha\beta\gamma\delta\varepsilon\zeta\eta\vartheta\iota\kappa\lambda\mu\nu\xi o\pi\varrho\sigma\tau\upsilon\varphi\chi\psi\omega$
$/\alpha\beta\gamma\delta\varepsilon\zeta\eta\vartheta\iota\kappa\lambda\mu\nu\xi o\pi\varrho\sigma\tau\upsilon\varphi\chi\psi\omega/$
abcdefghijklmnopqrstuvwxyz
abcdefghijklmnopqrstuvwxyz
$AB\Gamma\Delta EZH\Theta IK\Lambda MN\Xi O\Pi P\Sigma T\Upsilon\Phi X\Psi\Omega$
$AB\Gamma\Delta EZH\Theta IK\Lambda MN\Xi O\Pi P\Sigma T\Upsilon\Phi X\Psi\Omega$
ABCDEFGHIJKLMNOPQRSTUVWXYZ
ABCDEFGHIJKLMNOPQRSTUVWXYZ

$\alpha\beta\gamma\delta\varepsilon\zeta\eta\vartheta\iota\kappa\lambda\mu\nu\xi o\pi\varrho\sigma\tau\upsilon\varphi\chi\psi\omega$
$/\alpha\beta\gamma\delta\varepsilon\zeta\eta\vartheta\iota\kappa\lambda\mu\nu\xi o\pi\varrho\sigma\tau\upsilon\varphi\chi\psi\omega/$
abcdefghijklmnopqrstuvwxyz
/abcdefghijklmnopqrstuvwxyz/
$AB\Gamma\Delta EZH\Theta IK\Lambda MN\Xi O\Pi P\Sigma T\Upsilon\Phi X\Psi\Omega$
$/AB\Gamma\Delta EZH\Theta IK\Lambda MN\Xi O\Pi P\Sigma T\Upsilon\Phi X\Psi\Omega/$
ABCDEFGHIJKLMNOPQRSTUVWXYZ
/ABCDEFGHIJKLMNOPQRSTUVWXYZ/

Export to HTML

This should work since the charset is declared as utf-8. However, support in browsers is variable.

- Safari and Opera work the best—everything looks pretty nice in both.

- Firefox does OK, but the minus signs come out as hyphen. The bold math looks funny too with greek letters being **very** bold.

Export to LaTeX

Presumably, this won't work out of the box. I haven't tried it yet. However, see this blog post by Graham Lee for a possible solution:

```
\usepackage{ucs} % Unicode support
\usepackage[utf8x]{inputenc} % UCS' UTF-8 driver is better than the \LaTeX{} kernel'
\usepackage[T1]{fontenc} % The default font encoding only contains Latin characters
\usepackage{ae,aecompl} % Almost European fonts/hyphenation do a better job than Com
```

Update [**2007-11-02 Fri**]

Best to use the option [`mathletters`], since otherwise it tries to use commands like `textalpha` and I have no idea where these are defined (and Google wasn't much help). With `mathletters` it uses the standard math symbol greek alphabet, whether you are in math mode or not. I guess a better solution would be to use `ifmmode` to test if we are in math mode and use `upalpha` if we are not.

Problems encountered with `org-export-latex`
- Backslashes in quoted text are not properly escaped.

Integration with calc

Calc does not understand unicode as afar as I can see (e.g., it doesn't recognise 2.3 ± 0.4 as an error form). Presumably, this could be fixed rather easily since calc already has the concept of display styles.