Transitioning to Unicode: Tricks of the Trade

Elizabeth J. Pyatt,
Teaching and Learning with Technology
ejp10@jmu.edu

Outline
• Structure of Unicode
• Typing Unicode Text
• On the Web
• Gotchas!

What are your Target Languages?
• Every language has its own quirk

Optimal Language Support (Like English)
• Text can be sent as is in e-mails, Word files, in HTML
• Can be read on Mac, Windows, Linux
• Minimal text translation when opening a file
• You don’t have to activate or install custom fonts, or utilities

Remember ASCII?
• ASCII - American Standard Code for Information Exchange
• Assigns a Number to each letter/symbol/character/number
• Examples
  65=A  66=B  67=C  68=D  69=E  70=F
  97=a  98=b  99=c  100=d  101=e  102=f
  48=0  49=1  50=2  51=3  52=4  53=5

Excel Char Function

The Char function in Excel converts numbers to the equivalent character in ASCII

Limitations of ASCII

- 128 Characters only, so advanced punctuations, technical symbols accented letters, other scripts not supported.
- Vendors added more characters, but encodings not consistent

8-Bit Encodings (256)

- Vendors and other standards bodies devised encodings of 256 (2^8) characters
- Characters 0-127 = ASCII
- Characters 128-255 = Something else
- Western Europe (Spanish/French/German...)
  - MacRoman
  - Windows-1252
  - ISO-8859-1 (Latin-1) for the Web

Other Alphabets

Char 0-127 = ASCII
Char 128-255 = other script

<table>
<thead>
<tr>
<th>Script</th>
<th>Win</th>
<th>Mac</th>
<th>ISO/etc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russian</td>
<td>Win-1251</td>
<td>MacCyrillic</td>
<td>KODI-8</td>
</tr>
<tr>
<td>Greek</td>
<td>Win-1253</td>
<td>MacGreek</td>
<td>ISO-8859-7</td>
</tr>
<tr>
<td>Hebrew</td>
<td>Win-1255</td>
<td>MacHebrew</td>
<td>ISO-8859-8</td>
</tr>
</tbody>
</table>

“Exotic” Roman Alphabets

- 256 characters not enough for Polish (i, o)
- Czech, Croatian (i, c)
- No “Latin-2” developed
- But still missing Ǯ (Hawaiian), Ҫ (Turkish), Ǯ (Welsh)...
- Where would it end?
East Asian Encodings

- 65,000+ (2^16) Characters
- 0-127-ASCII
- Cyrillic, Greek included
- Chinese Characters plus "local" characters
- This is the Future

Enter Unicode

- Encoding scheme for all scripts
- 0-127-ASCII
  128-255: Latin 1
  Scripts by block
- BUT 65,536 characters won't be enough.
  (Maybe 1 million plus will be)
- http://www.unicode.org

Create A Unicode Doc

- **Activate Keyboard**
  Win: Input Locale in Regional Control Panel
  Mac: International System Preferences
  UNIX/Linux: Install
- Make sure document is saved as Unicode
- http://ktics.psu.edu/suggestions/international/ keybords/
- A Mac OS X Demo

GOTCHA 1: Use the right text editor

- **Win:** NotePad, Uni-Not, Global Writer.
  Open Office (NOT EditPlus 2)
- **Mac:** TextEdit, BBEdit (recent), Mutt, Nau, Neo Office 
- **What about Word?** Results not guaranteed although cross-platform compatibility improved.

GOTCHA 2: Save As UTF-8

- Make Plain Text
- Select Encoding in Save-As dialogue
Accent Codes

- Accent codes for Western European languages and punctuation can be used
  - http://eufb.psu.edu/suggestions/international/accents/
  - E.g. Copyright Sign = ALT+0169 (Win)
  - Option+G (Mac)
  - Mac Extended Roman Keyboard
    - Includes codes for macrons (˘), caron (ˇ),
    - http://eufb.psu.edu/suggestions/international/accents/codewsc.html#office

Insert a Character

- Windows: CharMap
  - Allows you to enter a single character
- Mac: Character Palette

Unicode On the Web

- **Step 1**: Activate and use keyboards
  - You must switch between English and other
    - Notepad, Unipad, BBEd
    - Dreamweaver, FrontPage (check code)
    - Mozilla Composer
- **Step 2**: Declare UTF-8 encoding in header
  ```html
  <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
  </head>
  ```

GOTCHA 3: No UTF-8 Meta Tag

French (Français)

Russian (Русский)

Without UTF-8 Tag

Need to manually switch encoding in
View menu of browser

GOTCHA 4: User Missing Font

- The more obscure the script, the more likely you will need to point a user to the correct font.
- Fortunately, lots of freeware fonts exist
  - http://www.travelprokes.info/fonts.html
  - http://www.alanwood.net/unicode/fonts.html
GOTCHA 5: Unicode on ANGEL

BECAUSE you can’t control encoding tag:
- You should use entity codes like &copy; for ©
  http://ricty.ics.s.ucla.edu/suggestions/international/web/
cdata.htm
- You should use Mozilla which converts Unicode
to four-digit escape codes.

Cyrillic on ANGEL

Cyrillic Message Board in ANGEL, Top in Mozilla

RSS and Unicode

- NO escape codes allowed (it’s an XML document)
- Should be raw Unicode

GOTCHA 6: Invisible Unicode Characters

- These can cause older applications to do
  some “Weird” things
  + Mysterious Japanese characters appear
    in cut and paste operations
  + CSS formats don’t apply
  + Plain text not parsable

Flavors of Unicode

- Unicode comes in a variety of flavors
depending on how bytes are delivered
- UTF-16 Big Endian (4 bytes)
  L = 0.4C
- UTF-16 Little Endian (bytes reversed)
  L = 4C.00
- UTF-32 (6 bytes)
  L = 00.00.00.4C
- UTF-8 (bytes compacted into 8-bit chunks)
  L = 4C (just like ASCII)
Is it hopeless? NO

• Substantial improvement with each year
• In a few years, you may be able to forget some gobbledygook
• Newer applications (e.g. Open Office) able to incorporate Unicode
• Strong Unicode support in language communities. Lots of freeware out there!

Can we Improve? YES

• Why not choose UTF-8 route over ISO-8859-1?
• Don’t use old fonts in new documents
• Help users make the transition
• Test Spanish text in Web tools. Maybe even some Russian.
• Go to http://lib.it.psu.edu/suggestions/international!