

Transitioning to Unicode: Tricks of the Trade

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

1

Outline

- Structure of Unicode
- Typing Unicode Text
- On the Web
- Gotchas!

2

What are your Target Languages?

- Every language has its own quirk

3

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

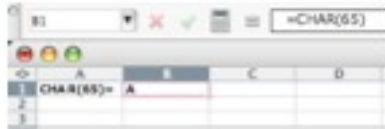
4

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 49=1 58=2
97=a 98=b 99=c 39=' 38=&
48=(41=) 91=[32=___

5

Excel Char Function



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

6

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



7

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

8

Other Alphabets

Char 0-127 = ASCII

Char 128-255 = other script

Script	Win	Mac	ISO/etc.
Russian	Win-1251	MacCyrillic	KOI-8
Greek	Win-1253	MacGreek	ISO-8859-7
Hebrew	Win-1255	MacHebrew	ISO-8859-8

9

“Exotic” Roman Alphabets

- 256 characters not enough for Polish (ą, ğ)
Czech, Croatian (š, č)
- So “Latin-2” developed
- But still missing ð, ò (Hawaiian), ş (Turkish), ŵ, ŷ (Welsh)...
- **Where would it end?**

10

East Asian Encodings

- 65,000+ (2¹⁶) Characters
- 0-127:ASCII
- Cyrillic, Greek included
- Chinese Characters plus "local" characters
- **This is the Future**

11

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>

12

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://Wlits.psu.edu/suggestions/international/keyboard/>
- A Mac OS X Demo

13

GOTCHA 1: Use the right text editor

- **Win:** Notepad, Unipad, Global Writer, Open Office (NOT EditPlus 2)
- **Mac:** TextEdit, BBEdit (recent), Mellel, Nisus, Neo Office J
- **What about Word?** Results not guaranteed although cross-platform compatibility improved.

14

GOTCHA 2: Save As UTF-8

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



15

Accent Codes

- Accent codes for Western European languages and punctuation can be used
- <http://lils.psu.edu/suggestions/international/accents/>
- Eg. Copyright Sign =
ALT+0169 (Win)
Option+G (Mac)
- Mac Extended Roman Keyboard
Includes codes for macrons (ā), caron (č),...
<http://lils.psu.edu/suggestions/international/accents/codessoc.html#osx>

16

Insert a Character

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



17

Unicode On the Web

- **Step 1:** Activate and use keyboard
You must switch between English and other
+ Notepad, Unipad, BBEdit
+ Dreamweaver, FrontPage (check code)
+ Mozilla, Composer
- **Step 2:** Declare UTF-8 encoding in header

```
<!DOCTYPE  
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">  
</head>
```

18

GOTCHA 3: No UTF-8 Meta Tag

French (Fréqûte)

You do not have access to this page in English or other languages. You may have access to it in French.

Русский (Русский)

Вы не можете увидеть эту страницу на английском языке. Вы можете увидеть эту страницу на русском языке.

Without the utf-8 tag
Need to manually
switch encoding in
View menu of browser

With the utf-8 tag

Русский (Русский)

Вы можете увидеть эту страницу на английском языке. Вы можете увидеть эту страницу на русском языке.

19

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.travelphrases.info/fonts.html>
 - <http://www.alswood.net/unicode/fonts.html>

20

GOTCHA 5: Unicode on ANGEL

BECAUSE you can't control encoding tag:

- You should use entity codes like © for © <http://ht.it.sjtu.edu/suggestions/internationalweb/csdetom.html>
- You should use Mozilla which converts Unicode to four-digit escape codes.

21

Cyrillic on ANGEL



22

RSS and Unicode

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



23

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



24

Flavors of Unicode

- Unicode comes in a variety of flavors depending on how bytes are delivered
- **UTF-16 Big Endian** (4 bytes)
L = 00.4C
- **UTF-16 Little Endian** (bytes reversed)
L = 4C.00
- **UTF-32** (6 bytes)
L = 00.00.4C
- **UTF-8** (bytes compacted into 8-bit chunks)
L = 4C (just like ASCII)

25

Is it hopeless? NO

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

26

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://it.its.psu.edu/suggestions/international/>

27
