

Multilingual
typesetting

Sylvia Blaho
Patrik Bye
Pavel Iosad

Understanding
encodings

The babel
package

The power of
XeTeX

Multilingual typesetting

L^AT_EX for Linguists

Sylvia Blaho
Patrik Bye
Pavel Iosad

Universitetet i Tromsø/CASTL

1st December 2008

Input encoding and font encoding

- An **encoding** is a set of instructions for \LaTeX to relate symbol codes stored internally with symbols input by the user and to be output on the page
- In order to handle symbols correctly, we need to know an **input encoding** (user input \Rightarrow \LaTeX) and a **font encoding** (\LaTeX \Rightarrow font in the output)
- As users, we are also confronted by the facts that not all front-ends support all encodings well.
- 👉 Thankfully for most of us, \TeX Shop has excellent encoding support

- Unfortunately encoding standards vary widely from system to system and from time to time
- There is a common standard called **Unicode**, but it does not yet have very good font coverage. However, for most of our purposes standard L^AT_EX support for Unicode is enough. (Sorry Lene & Linda.)
- In particular, Unicode obviates the need for *ad hoc* encodings where the same internal code may stand for, say, ä and ъ.
- To tell T_EXShop to always save your files in Unicode, open **Preferences** and choose **Unicode (UTF-8)** in the Encoding list. This will make T_EXShop recognize all the various symbols you input and convert them to Unicode.
- **If you send such files to someone who doesn't have a Unicode-enabled client, you're in trouble. You can try to save the file with a different encoding (say Central European), but this won't work if you have letters the encoding doesn't support, e. g. Hungarian and Russian)**

Input encodings

- Here are some of the input encodings you may want to use

System	Western European	Central European	Cyrillic
Mac OS X	applemac	macce	maccyr
Windows	ansinew	cp1250	cp1251
*nix	latin1	latin2	koi8-ru

- There are also ways to typeset CJK; there is a CJK- \LaTeX package which is compatible with Unicode (we'll come there)

Font encodings

- T_EX uses a font encoding system specific to itself. The downside is that the process for creating and installing fonts for L^AT_EX use is often convoluted. The upside is that if you make L^AT_EX understand your input encoding, the result will always come out right.
- The default font encoding we have been using so far is called OT1. It is enough for most English needs. It is, however, deprecated.

- Here are the font encodings we might want to use
- **T1**. This is the recommended encoding for European languages. Apart from all the usual suspects it also introduces new symbols:
 - `\k{}`: ogonek accent: `\k{a}` gives \mathring{a}
 - `\dh`, `\DH`: \mathring{d} , \mathring{D}
 - `\dj`, `\DJ`: \mathring{d} , \mathring{D}
 - `\th`, `\TH`: \mathring{t} , \mathring{T}
 - `\ng`, `\NG`: \mathring{n} , \mathring{N}
 - Different quote types, e. g.
`\guillemotleft a\guillemotright`
gives «a»

- **T2A**. This is the encoding for the most widespread Cyrillic languages, including Russian, Bulgarian, Serbian and Ukrainian
- **T3**. This encoding is used by `tipa`, so normally `\usepackage{tipa}` just loads it.
- **T4**. This encoding is used by the `fc` package used to typeset African languages. Thus, `FUL\m{B}E` gives FULBE
- **T5**. This encoding is needed for Vietnamese: `\ohorn` gives o

Selecting encodings

- Input encodings of course cannot be selected, since they refer to your whole file. To let \LaTeX know what you are doing, issue the following in your preamble:

```
\usepackage[<encoding>]{inputenc}
```
- If you are using Unicode (the recommended option), issue the following:

```
\usepackage{ucs}  
\usepackage[utf8x]{inputenc}
```
- 👉 **Important!** If the file is saved in a different encoding than what you told `inputenc` (e. g. you set `TeXShop` up differently), you might be in trouble.

- You can use multiple font encodings (as we did with tipa). To load them, use the following:
`\usepackage[<comma-separated list of encodings>]{fontenc}`
- Your document will start with the **last** encoding you loaded
- To change the font encoding in the middle of a document, issue the declaration
`\fontencoding{<desired encoding>}`
followed by `\selectfont`
- Small note about tipa: for extra safety, one can load T3 explicitly using
`\usepackage[T3,T1]{fontenc}`
`\usepackage[noenc]{tipa}`
- But usually we don't need to do this
- One useful gimmick is `\usepackage{cmap}`. It will make the funny letters come out right in the PDF file (e. g. you can search for them).

What is babel?

- The babel package takes care of all aspects of multilingual typesetting, such as font selection, hyphenation, automatically generated strings and so on.
- Its use is very simple:

```
\usepackage[<languages>]{babel}
```
- As with fontenc, the last language loaded will be the default
- This also takes care of font encodings, so you don't have to switch them by hand

- Language names are pretty intuitive, e. g. `hungarian`, `czech`, `polish`
- For Norwegian, there are the options `norsk` (= `bokmål`) and `nynorsk`
- Just in case, `english` is the same as `USenglish` and `american`, whereas `british` is the same as `UKenglish`
- For German, there are the options `germanb` and `ngermanb` (for the new spelling)
- Some languages define new commands. For example, the German set-up of Babel lets one input umlauted letters as `"a` instead of `\"a` and defines a command `"ck` which takes care of hyphenating `ck` as `k-k` (see the documentation)

- The package also takes care of automatically generated strings such as dates
- Try this with the output of `\today` using different babel options
- Apart from `\today`, babel also takes care of things like headings for the bibliography, table of contents etc.
- Provided you have the file with the hyphenation rules for your language (most often you do), babel also takes care of hyphenation and typographic rules (e. g. spaces on both sides of the colon for French)

- If you want to change the language in the middle of your document, there are two ways
- If it's just a short quote, the best way is
`\foreignlanguage{<language>}{your text}`
- If it's a longer chunk, use the declaration
`\selectlanguage{language}`
- Don't forget to switch back to the original language when you're finished, or enclose the whole thing in braces
- To check whether you have the hyphenation pattern for your language, check the very beginning of L^AT_EX's log. It usually says something like
LaTeX2e <2005/12/01>
Babel <v3.81> and hyphenation patterns for
english, usenglishmax, dumylang, nohyphenation,
german-x-2008-06-18, ngerman-x-2008-06-18,
ancientgreek, ibycus, arabic, basque,
bulgarian...

XeTeX and Unicode

- XeTeX (and XeLaTeX) is a pretty new extension to TeX which natively uses Unicode and lets you use fonts installed on your system rather than just the LaTeX fonts
- It also has many typographic niceties, but we won't dwell on them, see the documentation, in particular to the fontspec package
- **Important!** XeTeX only works in PDFLaTeX, so one cannot use it with Postscript-dependent packages (we will come to that later)

Using XeTeX

- Xe \LaTeX will let you use almost anything you can use with \LaTeX
- But you don't need to set up the encodings and such. A typical Xe \LaTeX preamble looks like the following:

```
\documentclass{article}  
\usepackage{xunicode}    % handles \'a and such  
\usepackage{xltextra}    % various fixes  
\usepackage{fontspec}    % see below
```

followed by the usual stuff
- Now you can type in your funny characters directly, like you'd do it in Word

The fontspec package

- The package fontspec lets you access and select fonts you have installed on your system (e. g. through the Control Panel on Windows or the Font Book on Mac)
- If you don't select a font but load fontspec you get the "normal" L^AT_EX setup
- To choose a font in the middle of a text, say
`\fontspec[font options]{font name}`
for example `\fontspec{Times New Roman}`
- Don't forget to revert to another font or include the portion you need in braces
- The font options are mostly used for OpenType fonts; if you know what they are, you can read the fontspec docs

- You can also use the following declarations in the preamble:

```
\setmainfont [Mapping=text-tex]{<font>} % serif  
\setsansfont [Mapping=text-tex]{<font>} % sans  
\setmonofont [Mapping=text-tex]{<font>} % mono
```
- See the fontspec docs for other customization options
- For example, you can define

```
\newfontfamily\familyname{definition}
```

 which can be used like `\rmfamily` and its ilk
- Say you want all of your examples to be in Charis SIL

```
\newfontfamily\examplefont{Charis SIL}  
\newcommand{\xmpl}[1]{\examplefont #1}
```

Running Xe_LA_TE_X

- In T_EXShop, just select Xe_LA_TE_X in the menu to the right of the Typeset button
- In other editors, find where the different compilation options are and create a new one (or modify one that exists) to run `xelatex` instead of `pdflatex` or something like it

Polyglossia: a babel for XeTeX

- One thing that works badly with XeLaTeX is babel, which is why we have the polyglossia package. It is invoked in the usual way

```
\usepackage{polyglossia}
```

- Once you load it, issue the command

```
\setmainlanguage{language}
```

- You can also say

```
\setotherlanguage{language}
```

as many times as you need

- See the polyglossia docs for language-specific options

- For short inserts in a foreign language, use

```
\text<language>{text}
```

e. g.

```
\textrussian{Пушкин}
```

- For longer pieces, use the environment with the language's name

```
\begin{russian}
```

```
Мой дядя самых честных правил,
```

```
Когда не в шутку занемог...
```

```
\end{russian}
```

- One exception is `\begin{Arabic}` (note the capital letter)

- Like babel, polyglossia tries to find the part of the current font that's relevant for the script. Since not all fonts supports all languages, you might get an error
- To get around the problem, use the `\newfontfamily` option of fontspec
- For example, your main document font does not support Hindi. To get around this problem, issue `\newfontfamily\hindifont[options]{font}` where the options and font are the same as you would set them up with fontspec
- Xe \LaTeX is unfinished (so bugs may appear) but under constant development. It may not yet be suitable for full-scale production tasks (such as a dissertation), but quite usable for other things. **Note that documents written for Xe \LaTeX normally won't compile in plain \LaTeX and vice versa, so it's a good idea to decide upfront what you are going to use.**