LaTeX for NLP Researchers

Tristan Miller
Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt

25 May 2017

Presented at:
School of Data Analysis and Artificial Intelligence
National Research University – Higher School of Economics

Overview

The basics

Matters of style

Packages for (computational) linguistics

Collaborative editing and source control

Converting to/from other formats

Bibliographies and citations

Online resources

The basics

Your TeX installation

- Like GNU/Linux, (La)TeX is a huge, decentralized project
- New features and bug fixes all the time
- Distributions ease burden of installation and maintenance
- Keep your distribution reasonably up to date!
- On *nix and Windows, consider manually installing TeX Live:
  - installable as single user or system-wide
  - includes all CTAN packages and fonts, and many scripts
  - automatic live updates
  - built-in documentation for all packages – just type `texdoc packagename`
  - community-supported by TUG (the TeX Users Group)

Essential packages

`\usepackage{microtype}`

Better line breaking. Can save you *a lot* of space!

`\usepackage[pdfusetitle]{hyperref}`

Sets PDF title and author metadata from `\title` and `\author`. Makes your preprints more discoverable through search engines.

`\RequirePackage[l2tabu, orthodox]{nag}`

Warns against use of obsolete/deprecated commands and packages. (Add this *before* your `\documentclass`.)

Avoid obsolete packages and commands

`l2tabu` documents obsolete/deprecated commands and packages, as well as their replacements. Examples:

- **Don't use `\usepackage{times}`!**
  It fails to set matching math and sans-serif fonts.
  Use `\usepackage{newtxtext,newtxmath}` instead.
- **Don't use `\bf`, `\it`, etc.!**
  They don't combine or kern correctly.
  (e.g., `{\it \bf foo}` produces **foo** instead of *foo*.)
  Use `\textbf{…}`, `\emph{…}`, etc. instead.

Matters of style

Typesetting math

- In formulas, multi-letter variables should be enclosed in `\mathit{…}`. Otherwise the letters will be typeset as if they are products of single-letter variables.

  `$f(myvar)$` → $f(m\,y\,v\,a\,r)$

  `$f(\mathit{myvar})$` → $f(myvar)$

- Don't set operator names as variables.

  `$max(x)$` → $m\,a\,x(x)$

  `$\max(x)$` → $\max(x)$

- If LaTeX doesn't provide the operator, define it yourself.

  `\DeclareMathOperator*{\argmax}{arg\,max}`

Tabular material

- Don’t force positioning of tables and other floats (for example, using `\begin{table}[H]`). In general, LaTeX knows better than you how to best place floats.

Tabular material: booktabs

Avoid using vertical rules in tables. They make reading difficult. Use the `booktabs` package for “proper” rules.

ferret	colour	mass (g)	bites
Þóra	self	670	no
Þjófur	self	770	yes
Pafnuty	sable	810	no

Tabular material: booktabs

Avoid using vertical rules in tables. They make reading difficult. Use the `booktabs` package for “proper” rules.

ferret	colour	mass (g)	bites
Þóra	self	670	no
Þjófur	self	770	yes
Pafnuty	sable	810	no

Tabular material: siunitx

Always align numerical columns at the decimal point. The `siunitx` package is great for special cases (varying levels of precision, bolded cells, units, etc.)

algorithm	precision
our method	67.89
Smith et al. (2016)	66.7*
Ivanov & Ivanov (2014)	68.33*
naïve baseline	9.00

* Evaluated on a slightly different data set

Tabular material: siunitx

Always align numerical columns at the decimal point. The `siunitx` package is great for special cases (varying levels of precision, bolded cells, units, etc.)

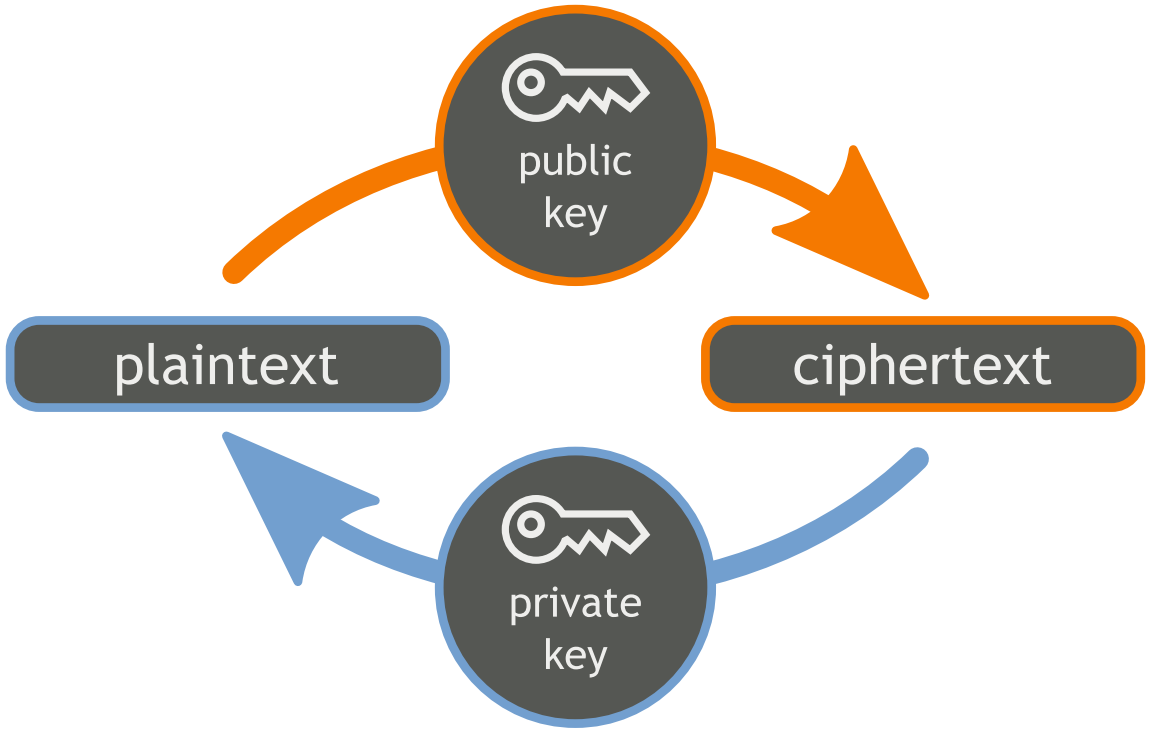
algorithm	precision
our method	67.89
Smith et al. (2016)	66.7 *
Ivanov & Ivanov (2014)	68.33*
naïve baseline	9.00

* Evaluated on a slightly different data set

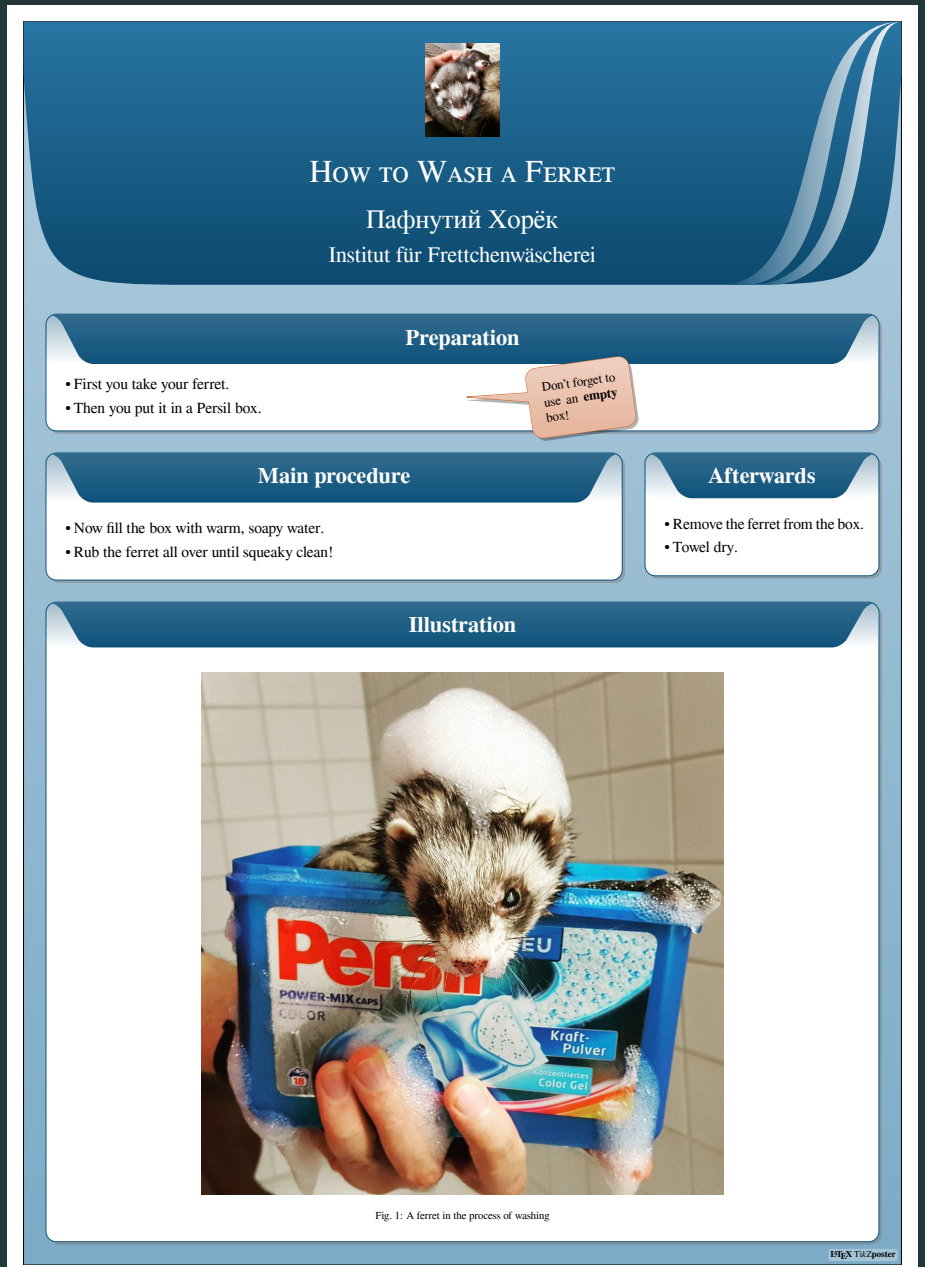
Slides and posters

- Advantages of using LaTeX for slides and posters:
  - Easily reuse tables, figures, citations, etc. from your papers
  - Compatible with source control systems
  - Generated PDFs are faithfully reproduced
- Disadvantages:
  - No GUI for manually positioning elements
  - Harder to do complex animations
- Popular slide packages: `powerdot`, `beamer`
- Popular poster packages: `tikzposter`, `beamerposter`

Public-key ciphers



tikzposter


How to Wash a Ferret
Пафнутий Хорёк
Institut für Frettchenwäscherei
Preparation
• First you take your ferret.
• Then you put it in a Persil box.
Don't forget to use an empty box!
Main procedure
• Now fill the box with warm, soapy water.
• Rub the ferret all over until squeaky clean!
Afterwards
• Remove the ferret from the box.
• Towel dry.
Illustration
Persil
POWER-MIX CAPS
COLOR
Kraft-Pulver
Konzentriertes Color Gel
Fig. 1: A ferret in the process of washing
LaTeX TikZposter

Assignments and exams

- Advantages of using LaTeX for assignments and exams:
  - Easily reuse formulas, diagrams, etc. from your lecture slides
  - Automatically track and sum point values of questions
  - Randomly select questions from a database
  - Easily format long-answer, fill-in-the-blank, and multiple-choice questions
  - Automatically print grading tables
  - Selective printing of solutions (for the answer key)
- Assignment packages: `probsoln`, `exsheets`, `answers`
- Exam packages: `exam`, `exsheets`

Packages for (computational) linguistics

Setting text in foreign languages (LaTeX)

- The `babel` package sets text in non-English languages.
- It will ensure that all your foreign text is correctly typeset and hyphenated.

```
\usepackage[ngerman,english]{babel}

You are \foreignlanguage{ngerman}{wunderbar}!

% Use environments for longer spans.
\begin{ngerman}
  Das ist toll!
\end{ngerman}
```

Setting text in foreign languages (XeLaTeX)

- XeLaTeX has better support for fonts and encodings
- Multilingual typesetting is available from the `polyglossia` package, which has better support than `babel` for non-Latin scripts

```
\usepackage{polyglossia}
\setotherlanguage{russian}
\newfontfamily\cyrillicfont{Times New Roman}

What's English for \textrussian{злорадство}?

\begin{russian}
  Эх, чужак, общий съём цен шляп (юфть) – вдрызг!
\end{russian}
```

Enumerated examples

The `gb4e` and `lingmacros` packages typeset examples in standard linguistics style:

```
\begin{exe}
  \ex This is an example sentence.
  \ex[*]{This example ungrammatical is.}
\end{exe}
```

(1) This is an example sentence.

(2) * This example ungrammatical is.

Glosses

`gb4e` and `lingmacros` can also typeset glosses:

```
\begin{exe}
\ex
\gll Хорёк ест мороженое.\\
ferret.NOM eat.3.SG.PRS ice-cream.ACC\\
\trans `The ferret eats the ice cream.'
\end{exe}
```

(3) Хорёк ест мороженое.
ferret.NOM eat.3.SG.PRS ice-cream.ACC
'The ferret eats the ice cream.'

International Phonetic Alphabet

If using `pdflatex`, then the `tipa` package can be used to enter IPA symbols using an ASCII code:

`[\textipa{f@'nEtIk}]` → [fə'nɛtɪk]

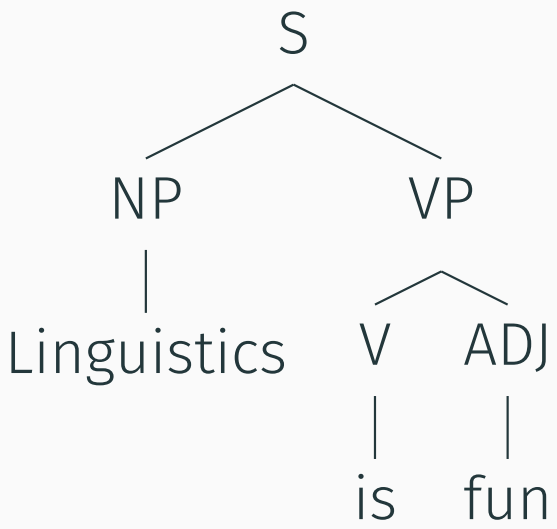
This can be inconvenient for large amounts of IPA. Using XeLaTeX, you can enter IPA symbols directly in the source:

`[fə'nɛtɪk]` → [fə'nɛtɪk]

Syntactic trees

The `qtree` package can draw syntactic trees.

```
\Tree [.S [.NP Linguistics ] [.VP [.V is ] [.ADJ fun ] ] ]
```



Collaborative editing and source control

Producing diffs

(`color`)`diff` visualizes line-by-line changes in your LaTeX source:

```
$ colordiff file_1.tex file_2.tex
3c3
< Four score and seven years ago our four fathers brought forth, on this
< continent, a new nation, conceived in Liberty, and dedicated to the
< proposition that all men are created equal. Now we are engaged in a great
< civil war, testing whether that nation, or any nation so conceived, can
< long endure. We are met on a great battlefield of that war.
---
> Four score and seven years ago our forefathers brought forth, on this
> continent, a new nation, conceived in \emph{Liberty}, and dedicated to the
> proposition that all persons are created equal. Now we are engaged in a
> great civil war, testing whether that nation, or any nation so conceived
> and so dedicated, can long endure. We are met on a great battlefield of
> that war.
```

Producing diffs

`wdiff` visualizes word-by-word changes in your LaTeX source:

```
$ alias cwdiff="wdiff -n -w $'\033[1;31m' -x $'\033[m' \
                         -y $'\033[1;34m' -z $'\033[m'"
$ cwdiff file_1.tex file_2.tex

\documentclass[preview=true,12pt]{standalone}
\begin{document}\LARGE
Four score and seven years ago our four fathers forefathers
brought forth, on this continent, a new nation, conceived in
Liberty, \emph{Liberty}, and dedicated to the proposition that all
men persons are created equal. Now we are engaged in a great civil
war, testing whether that nation, or any nation so conceived,
conceived and so dedicated, can long endure. We are met on a great
battlefield of that war.
\end{document}
```

`gitwdiff` automates this for Git and SVN.

Producing diffs

`latexdiff` visualizes changes in the typeset document:

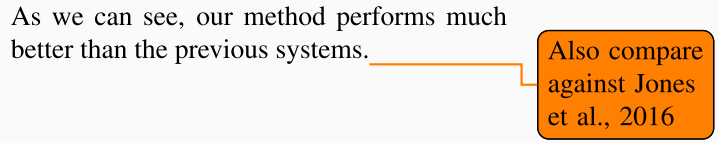
```
$ latexdiff file_1.tex file_2.tex > file_diff.tex
$ pdflatex file_diff
```

Four score and seven years ago our ~~four fathers~~ forefathers brought forth, on this continent, a new nation, conceived in ~~Liberty~~*Liberty*, and dedicated to the proposition that all ~~men~~ persons are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war.

“To-do” and “fix-me” annotations

- Comments help keep track of things to be fixed
- LaTeX comments (`%`) and inline comments (e.g., via `\textbf{TODO: …}`) are too ad-hoc and inconspicuous
- Several packages provide “to-do” macros: `easy-todo`, `fixme`, `fixmetodonotes`, `todo`, `todonotes`
- Some packages can generate lists of to-dos

```
\documentclass{article}
\usepackage{todonotes}
\begin{document}
As we can see, our method performs
much better than the previous
systems.\todo{Also compare against
Jones et al., 2016}
\end{document}
```



Converting to/from other formats

Converting documents to Microsoft Word

- Most non–computer science publication venues expect submissions in Microsoft Word
- Adapting an existing LaTeX paper is never trivial – budget a lot of time for this!
- Converters aren't perfect, but can be faster than (re)typing from scratch
- `latex2rtf` can do most of the work
- Other possibilities:
  - convert first to HTML with Hevea, LaTeX2HTML, TeX4ht, TtH, LaTeXML, pdf2htmlEX, or plasTeX
  - convert first to plain text with `detex` or `tex2mail`

Exporting diagrams, etc. to images

- The `standalone` class can:
  - crop your document so the page size fits the content
  - output to PDF (for inclusion in other LaTeX documents)
  - output to EPS (for inclusion in Microsoft Office documents)
  - output to PNG or other bitmap formats (for use on web pages)

```
\documentclass{standalone}
\begin{document}
$\displaystyle\int_0^\infty
\sum_{i=0}^\infty
\frac{f_i(x)}{2\pi}$
\end{document}
```

→ $\displaystyle\int_0^\infty \sum_{i=0}^\infty \frac{f_i(x)}{2\pi}$

Exporting diagrams from PowerPoint 2010

1. Right-click on the graphic. Select *Size and position…* → *Size*. Note the height and width. Press Close.
2. Copy the graphic (Ctrl+C).
3. Create a new presentation (Ctrl+N).
4. *Design* → *Page Setup*. Change the height and width to the values you noted previously. (You may need to add a few millimetres to each dimension.) Press OK.
5. Right-click on the slide and select Paste Options: Picture.
6. *File* → *Save & Send* → *Create PDF/XPS Document* → *Create PDF/XPS*.
7. Set “Save as type” to “PDF (*.pdf)”. Enter a filename and press the “Publish” button.

Bibliographies and citations

Bibliography processors: BibTeX vs. Biber

BibTeX

- stable and widely used
- poor support for non-ASCII characters
- arcane, Forth-based style language

Biber

- Unicode-capable
- easily extensible
- works only with `biblatex`

Citation tools: `natbib` vs. `biblatex`

`natbib`

- works only with BibTeX
- designed for author–year and numeric citation styles

`biblatex`

- works with BibTeX or Biber
- easily extensible
- not widely used by publication venues

Online resources

Online resources

- The Comprehensive TeX Archive Network (CTAN)
- UKTUG FAQ
- TeX – LaTeX Stack Exchange (in particular questions tagged with “thesis” and “linguistics”)
- Academia Stack Exchange (in particular questions tagged with “thesis” and “latex”)
- The LaTeX for Linguists Home Page
- LaTeX on *Wikibooks* (includes sections on typesetting technical and linguistics texts)

Thank you!