The State of Babel Polish🇵🇱

Marcin Woliński

BachoTEX 2023

Motivation

- The package Babel is the key LaTeX infrastructure for typesetting in various languages.
- Around 2021 some bugs surfaced related to the Polish module of Babel. This module
  - uses internally the symbol `\lll`, which conflicts with AMS-LaTeX's name for the symbol ⋘,
  - redefines `\selectfont`, which leads to conflicts with some standard LaTeX packages. 😱🤯
- At that time, version 1.2l of the Polish module was current, dated 2005/03/31.

What did the offending code do?

- It was created (for LaTeX 2.07) when no TeX fonts supported typesetting Polish.
- Polish uses some non-ASCII letters:

ąćęłńóśźżĄĆĘŁŃÓŚŹŻ

What did the offending code do?

- It was created (for LaTeX 2.07) when no TeX fonts supported typesetting Polish.
- Polish uses some non-ASCII letters: ąćęłńóśźżĄĆĘŁŃÓŚŹŻ
- Most problematic:

ąĄęĘłŁ

What did the offending code do?

- It was created (for LaTeX 2.07) when no TeX fonts supported typesetting Polish.
- Polish uses some non-ASCII letters: ąćęłńóśźżĄĆĘŁŃÓŚŹŻ
- Most problematic: ąĄęĘłŁ
- Someone (at the University of Warsaw? Leszek Holenderski?) spotted the character '54 in cmmi10:

$$A \hookrightarrow B$$

What did the offending code do?

- It was created (for LaTeX 2.07) when no TeX fonts supported typesetting Polish.
- Polish uses some non-ASCII letters: ąćęłńóśźżĄĆĘŁŃÓŚŹŻ
- Most problematic: ąĄęĘłŁ
- Someone (at the University of Warsaw? Leszek Holenderski?) spotted the character '54 in cmmi10: $A \hookrightarrow B$
- This character was being combined with letters:

ąę (vs. ąę)

What did the offending code do?

- It was created (for LaTeX 2.07) when no TeX fonts supported typesetting Polish.
- Polish uses some non-ASCII letters: ąćęłńóśźżĄĆĘŁŃÓŚŹŻ
- Most problematic: ąĄęĘłŁ
- Someone (at the University of Warsaw? Leszek Holenderski?) spotted the character '54 in cmmi10: $A \hookrightarrow B$
- This character was being combined with letters: ąę (vs. ąę)
- Babel shorthands were defined:

  `"L"od"z g"abk"e` → Łódź gąbkę

Non-lethal problems of the old code

There are known problems with this code:

- The "-shorthands for Polish characters do not use LICR, so even with modern fonts the precomposed characters from the font are not used (in the case of *ą*, *ę*, *ł*, *ż*).
- The substitute for ogonek accent defined for OT1 is not accessible with standard `\k` (besides being hideous).
- Wrong shape of *ę* is used in the word *Część* generated by the command `\part`.

How to correct this?

How to correct this?

Delete the old code! 

How to correct this?

Delete the old code! 

In fact, the old code has been moved to a module named `polish-compat`:

```
\usepackage[polish-compat]{babel}
```

The new Babel Polish

Version 1.3 of Babel Polish is a pretty typical language module:

- The code uses the new syntax provided by Babel to declare language specific elements.
- The module provides Polish translation for strings such as “section”, “chapter”, “appendix”, “table of contents”, month names, etc.
- It ensures Polish hyphenation patterns are used.
- Activates `\frenchspacing`, so that spaces after punctuation marks are equal to typical inter-word spaces.
- New documentation was provided, covering also the use of Unicode engines.

How to typeset in Polish

Case 1: Unicode TEX engines (XeLATEX, luaLATEX)

```
\documentclass{mwart}
\usepackage[polish]{babel}

\begin{document}
Zażółć gęślą jaźń w~wiaderku.
\end{document}
```

- Assumed input encoding: UTF-8.
- Default font: Latin Modern with an ample repertoire of accented Latin letters.

How to typeset in Polish

Case 2: classic TEX (pdfLATEX)

```
\documentclass{mwart}
\usepackage{lmodern}
\usepackage[T1]{fontenc}
\usepackage[polish]{babel}

\begin{document}
Zażółć gęślą jaźń w~wiaderku.
\end{document}
```

- Assumed input encoding: UTF-8 (from 2018).
- Default font: Computer Modern without Polish letters.
- Font encoding needs switching to the European enc. T1.
- Default T1 font is bitmap-based, so switch to Latin Modern.

A side note on T1 fonts

- Many font switching packages work with T1: Latin Modern, the TEX Gyre collection (`tgtermes`, `tgbonum`, `tgschola`, etc.) and others. With Unicode engines you can also use any OpenType font present on your system.
- Beware the fonts from the PSNFSS set, in particular `times.sty`:

TEX Gyre Termes	gąbkę	*gąbkę*
PSNFF Times	gąbkę	*gąbkę*

A remaining design decision

Which shorthands should be provided?

"-	soft hyphen
"=	explicit hyphen repeated after the break
"|	disable ligature at this position.
"`	Polish left double quote: „
"'	Polish right double quote: ”
"<	left guillemet: « (used in Polish as second level quotes)
">	right guillemet: »

Two problematic Polish typographic rules

- If a hyphen used in compound words (e.g., *biało-czerwony*) occurs at a line break, it should be repeated after the break.
- A line break should not occur after a one-letter word.

Two problematic Polish typographic rules

- If a hyphen used in compound words (e.g., *biało-czerwony*) occurs at a line break, it should be repeated after the break.
- A line break should not occur after a one-letter word.

```
Zażółć biało"=czerwoną gęślą jaźń w~wiaderku.
```

Babel solution

With luaLaTeX, current version of Babel can install pre-hyphenation filters which implement these rules:

```
\usepackage{babel}
\babelprovide[main,
  transforms={hyphen.repeat,oneletter.nobreak}
  ]{polish}
```

…

```
Zażółć biało-czerwoną gęślą jaźń w wiaderku.
```

Summary

- Polish module of Babel became completely predictable and boring.
- Some interesting things get developed in Babel itself.