The `hyph-utf8` package and hyphenation with TeX

Maintainers of the `hyph-utf8` package and collectors of patterns:

- *Mojca Miklavec, Arthur Reutenauer*
- With contributions by *Khaled Hosny, Manuel Pégourié-Gonnard, Élie Roux*

Main description:
June 2011

Latest editorial change:
18 May 2017

Abstract:

In 2008 all the existing hyphenation patterns from TeX distributions have been collected in a single package `hyph-utf8`, converted into UTF-8 encoding and adapted for use in different TeX engines. The patterns can be used directly by Unicode-aware engines such as LuaTeX and XƎTeX, and there is a mechanism to convert the patterns to the appropriate 8-bit encoding when used with pTeX, pdfTeX or Knuth's TeX.

Table of Contents:

1 Using hyphenation patterns

1.1 Plain TeX

In engines that support ε-TeX you can select the desired hyphenation patterns with:

```
\uselanguage{langname}
```

where `langname` is the string identifying a particular hyphenation file in `language.def` (see Section 2).

1.2 LaTeX

1.2.1 LaTeX with Babel

You can switch the language in LaTeX with:

```
\usepackage[languagename]{babel}
```

In 8-bit engines you also need to make sure that you load the proper font encoding which supports all the characters used in the language of your choice, for example:

```
\usepackage[T1]{fontenc}
```

N.B.: You can use Babel with any TeX engine, however it has never been properly adapted to work well with Unicode engines. If you are using XeTeX it is advisable to use Polyglossia instead.

1.2.2 LaTeX with Polyglossia

Polyglossia should be the preferred choice when using XeLaTeX. It doesn't support LuaLaTeX yet, but it is planned to extend it in future.

```
\usepackage{polyglossia}
\setmainlanguage[optional settings]{langname}
\setotherlanguages{otherlangname}

\begin[optional settings]{otherlangname} ... \end{otherlangname}
```

See Polyglossia manual for extensive list of options.

1.2.3 Low-level commands

Since Babel's `hyphen.cfg` is built into the LaTeX format, hyphenation patterns can be used without even loading Babel or Polyglossia. At the low-level this simply corresponds to defining

```
\language=\l@<langname>
```

The user command is supposed to be

```
\hyphenrules{langname}
```

or

```
\begin{hyphenrules}{langname} ... \end{hyphenrules}.
```

and *should* work with any flavour of LATEX, however we couldn't make it work.

1.3 ConTEXt

ConTEXt doesn't load patterns for all the language that `hyph-utf8` provides. If you miss any language, please contact the mailing list. The general syntax for supported languages is the following:

```
% language of the main document
\mainlanguage[language]

% to switch to another language locally
{\language[otherlanguage] language of some short fragment}
```

You can use full language name or the two-letter language code.

1.3.1 ConTEXt MkII

When using ConTEXt MkII the EC/T1 font encoding is used by default already, but you might need to change the encoding when using Polish, languages written in Cyrillic scripts, etc. For example:

```
\usetypescript[iwona][qx]
\setupbodyfont[iwona]
\mainlanguage[polish]
```

ConTEXt loads hyphenation patterns in several encodings. The Czech or Slovak patterns can be used with both EC and IL2 font encoding for example. The right hyphenation patterns will be chosen based on current font encoding.

1.4 Some advanced examples

1.4.1 Example for Polyglossia

```
\usepackage{polyglossia}
% the language used for main document
\setmainlanguage{asturian}
% American English with extended hyphenation patterns
\setotherlanguage[variant=usmax]{english}
% German with experimental patterns "ngerman-x-latest"
\setotherlanguage[spelling=new,latesthyphen=true]{german}
\setotherlanguages{spanish,catalan,french}

\begin{document}

Long Asturian text ... (Hyphenation for Asturian is not available,
but polyglossia automatically falls back on Catalan for now,
which seems to be a reasonable choice.)

\begin{german}
Deutscher Text ... (with the hyphenation patterns selected above:
"ngerman-x-latest")
\end{german}

\begin[script=fraktur,spelling=old]{german}
Deutſcher Text ... (set in Fraktur, with traditional hyphenation).
\end{german}

\end{document}
```

2 List of supported languages

English

-	**english**	usenglish, USenglish, american
en-us	usenglishmax	
en-gb	ukenglish	british, UKenglish

Afrikaans

af	afrikaans

Ancientgreek

grc	ancientgreek
grc-x-ibycus	ibycus

Arabic

ar	arabic

Armenian

hy	armenian

Assamese

as	assamese

Basque

eu	basque

Belarusian

be	belarusian

Bengali

bn	bengali

Bulgarian

bg	bulgarian

Catalan

ca	catalan

Chinese

zh-latn-pinyin	pinyin

Church Slavonic

cu	churchslavonic

Coptic

cop	coptic

Croatian

hr	croatian

Czech

cs	czech

Danish

da	danish

Dutch

nl	dutch

Esperanto

eo	esperanto

Estonian

et	estonian

Ethiopic

mul-ethi	ethiopic	amharic, geez

Farsi

fa	farsi	persian

Finnish

fi	finnish

French

fr	french	patois, francais

Friulan

fur	friulan

Galician

gl	galician

Georgian

ka	georgian

German

de-1901	german
de-1996	ngerman
de-ch-1901	swissgerman

Greek

el-monoton	monogreek	
el-polyton	greek	polygreek

Gujarati

gu	gujarati

Hindi

hi	hindi

Hungarian

hu	hungarian

Icelandic

is	icelandic

Indonesian

id	indonesian

Interlingua

ia	interlingua

Irish

ga	irish

Italian

it	italian

Kannada

kn	kannada

Kurmanji

kmr	kurmanji

Latin

la	latin
la-x-classic	classiclatin
la-x-liturgic	liturgicallatin

Latvian

lv	latvian

Lithuanian

lt	lithuanian

Malayalam

ml	malayalam

Marathi

  mr   marathi

Mongolian

  mn-cyrl   mongolian

  mn-cyrl-x-lmc mongolianlmc

Norwegian

  nb   bokmal   norwegian, norsk

  nn   nynorsk

Occitan

  oc   occitan

Oriya

  or   oriya

Panjabi

  pa   panjabi

Polish

  pl   polish

Piedmontese

  pms   piedmontese

Portuguese

  pt   portuguese   portuges

Romanian

  ro   romanian

Romansh

  rm   romansh

Russian

  ru   russian

Sanskrit

  sa   sanskrit

Serbian

  sr-latn   serbian

  sr-cyrl   serbianc

Slovak

  sk   slovak

Slovenian

  sl   slovenian   slovene

Spanish

  es   spanish   espanol

Swedish

  sv   swedish

Tamil

  ta   tamil

Telugu

  te   telugu

Thai

  th   thai

Turkish

  tr   turkish

Turkmen

  tk   turkmen

Ukrainian

  uk   ukrainian

Uppersorbian

  hsb   uppersorbian

Welsh

  cy   welsh

Babel defines a few more synonyms (which consequently only work in LaTeX):

english	canadian
british	australian, newzealand
german	austrian
ngerman	naustrian
portuguese	brazilian, brazil