

hyph-utf8

Mojca Miklavc (current maintainer)
Arthur Reutenauer (no longer active on hyph-utf8)

May 23, 2012

Abstract

The `hyph-utf8` package gathers all the existing hyphenation patterns for \TeX , converted to UTF-8. They can be used directly by UTF-8-aware \TeX engines such as $\text{Lua}\TeX$ and $\text{Xe}\TeX$, and there is a mechanism to convert the patterns to some 8-bit encoding when used with $\text{pdf}\TeX$ or Knuth's \TeX .

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**

Bengali

bn **bengali**

Bulgarian

bg **bulgarian**

Catalan

ca **catalan**

Chinese

zh-latn-pinyin pinyin

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**

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	Romanian	ro	romanian
Irish	ga	irish	Romansh	rm	romansh
Italian	it	italian	Russian	ru	russian
Kannada	kn	kannada	Sanskrit	sa	sanskrit
Kurmanji	kmr	kurmanji	Serbian	sr-latn	serbian
Latin	la	latin		sr-cyrl	serbianc
Latvian	lv	latvian	Slovak	sk	slovak
Lithuanian	lt	lithuanian	Slovenian	sl	slovenian slovene
Malayalam	ml	malayalam	Spanish	es	spanish espanol
Marathi	mr	marathi	Swedish	sv	swedish
Mongolian	mn-cyrl	mongolian	Tamil	ta	tamil
	mn-cyrl-x-lmc	mongolianlmc	Telugu	te	telugu
Norwegian	nb	bokmal	Turkish	tr	turkish
	nn	nynorsk	Turkmen	tk	turkmen
Oriya	or	oriya	Ukrainian	uk	ukrainian
Panjabi	pa	panjabi	Uppersorbian	hsb	uppersorbian
Polish	pl	polish	Welsh	cy	welsh
Portuguese	pt	portuguese portuges			

Babel defines a few more synonyms (which consequently only work in L^AT_EX):

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

Using hyphenation patterns

Plain T_EX

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

```
\uselanguage{langname}
```

where `langname` is the string identifying a particular hyphenation file in `language.dat` and can be taken from table on the first two pages.

L^AT_EX

Since Babel's `hyphen.cfg` is built in the XeL^AT_EX 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 L^AT_EX, however we couldn't make it work.

L^AT_EX with Babel

You can use Babel with any T_EX engine, however it is currently unmaintained and has never been adapted to work well with Unicode engines. If you are using X_YL^AT_EX please use Polyglossia instead.

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

L^AT_EX with Polyglossia

Polyglossia should be the preferred choice when using XeL^AT_EX. It doesn't support LuaL^AT_EX 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.

ConT_EXt

ConT_EXt 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]
```

```
{\language[otherlanguage] language of some short fragment}
```

You can use full language name or language code. When using ConT_EXt MKII you might need to select the appropriate font encoding for Cyrillic scripts, Polish and some other languages:

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

ConT_EXt loads hyphenation patterns in several encodings, so that you can for example use Czech patterns with either ec or il2 font encodings. The right hyphenation patterns will be chosen based on current font encoding.

More examples

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}
Deutfcher Text ... (set in Fraktur, with traditional hyphenation).
\end{german}
```

```
\end{document}
```