

hyph-utf8

Mojca Miklavec (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Arthur Reutenauer (no longer active on hyph-utf8)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 $\mathrm{T\!E\!X}$, converted to UTF-8. They can be used directly by UTF-8-aware $\mathrm{T\!E\!X}$ engines such as $\mathrm{LuaT\!E\!X}$ and $\mathrm{XeT\!E\!X}$, and there is a mechanism to convert the patterns to some 8-bit encoding when used with $\mathrm{pdfT\!E\!X}$ or Knuth's $\mathrm{T\!E\!X}$.

List of supported languages

English			
-	english	usenglish, USenglish, american	
en-us	usenglishmax		
en-gb	ukenglish	british, UKenglish	
Afrikaans			Esperanto
af	afrikaans		eo
			esperanto
Ancientgreek			Estonian
grc	ancientgreek		et
grc-x-ibycus	ibycus		estonian
Arabic			Ethiopic
ar	arabic		mul-ethi
			ethiopic
Armenian			
hy	armenian		amharic, geez
Assamese			Farsi
as	assamese		fa
			farsi
Basque			
eu	basque		persian
Bengali			Finnish
bn	bengali		fi
			finnish
Bulgarian			French
bg	bulgarian		fr
			french
Catalan			
ca	catalan		patois, francais
Chinese			Friulan
zh-latn-pinyin	pinyin		fur
			friulan
Coptic			Galician
cop	coptic		gl
			galician
Croatian			German
hr	croatian		de-1901
			german
Czech			de-1996
cs	czech		ngerman
			de-ch-1901
Danish			swissgerman
da	danish		Greek
Dutch			el-monoton
nl	dutch		monogreek
			el-polyton
			greek
			polygreek
			Gujarati
			gu
			gujarati
			Hindi
			hi
			hindi
			Hungarian
			hu
			hungarian
			Icelandic
			is
			icelandic
			Indonesian
			id
			indonesian

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

```
\usepackage[language]{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}
Deutŕcher Text ... (set in Fraktur, with traditional hyphenation).
\end{german}
```

```
\end{document}
```