

The `hyph-utf8` package and hyphenation with \TeX

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

- *Mojca Miklavc*, *Arthur Reutenauer* (core, patterns)
- *Manuel Pégourié-Gonnard*, *Khaled Hosny* (Lua \TeX support)
- *Élie Roux* (no longer active)

Date:

2014-06-11

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 Lua \TeX and Xe \LaTeX , and there is a mechanism to convert the patterns to the appropriate 8-bit encoding when used with p \TeX , pdf \TeX or Knuth's \TeX .

Table of Contents:

1	USING HYPHENATION PATTERNS	2
1.1	Plain \TeX	2
1.2	L ^A \TeX	2
	<i>L^A\TeX with Babel</i>	2
	<i>L^A\TeX with Polyglossia</i>	2
	<i>Low-level commands</i>	2
1.3	Con \TeX t	3
	<i>Con\TeXt MkII</i>	3
1.4	Some advanced examples	4
	<i>Example for Polyglossia</i>	4
2	LIST OF SUPPORTED LANGUAGES	4

1 Using hyphenation patterns

1.1 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.def` (see Section 2).

1.2 L^AT_EX

1.2.1 L^AT_EX with Babel

You can switch the language in L^AT_EX 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 T_EX engine, however it has never been properly adapted to work well with Unicode engines. If you are using X_YT_EX it is advisable to use Polyglossia instead.

1.2.2 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.

1.2.3 Low-level commands

Since Babel's `hyphen.cfg` is built into the L^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.

1.3 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]

% 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 ConT_EXt MkII

When using ConT_EXt 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]
```

ConT_EXt 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}
Deutlicher 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	
----	--------	--

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	
Irish		
ga	irish	
Italian		
it	italian	
Kannada		
kn	kannada	
Kurmanji		
kmr	kurmanji	
Latin		
la	latin	
la-x-classic	classicalatin	
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	
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 L^AT_EX):

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