hyph-utf8

Mojca Miklavec (current maintainer) Arthur Reutenauer (no longer active on hyph-utf8)

April 9, 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 LuaTeX and XeTeX, and there is a mechanism to convert the patterns to some 8-bit encoding when used with pdfTeX or Knuth's TeX.

List of supported languages

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

Interlingua normanian portugues portugues Interlingua ro romanian romanian Irish ro romansh	Indonesian			Portuguese		
ia interlingua	id	indonesian		pt	portuguese	portuges
Prish	Interlingua			Romanian		
Idalian Italian Russian it italian ru russian Kannada sa sanskrit km kannada sa sanskrit Kurmanji serbian serbian Lao sr-latn serbian lo lao Slovak Latin sk slovak la latin Slovenian slovenian la latin Spanish spanish spanish ly latvian sy sy spanish spanish lt lithuanian sy	ia	interlingua		ro	romanian	
Italian Russian it italian ru russian Kannada sanskrit kanskrit kn kannada sanskrit kmranji serbian kmracilam kmr kurmanji serbian Lao serbian serbian lo lao Slovak Latin sk slovak la latin sk slovenian	Irish			Romansh		
ru russian Kannada Sanskrit km kannada sa sanskrit Kurmanji Serbian serbian kmr kurmanji sr-cyrl serbian Lao sr-cyrl serbianc serbianc lo la Slovak serbianc Latin sk slovak slovak Latin sk slovak slovenian spanish espanolo respanolo spanish respanolo spanish respanolo suedish mill mill mill mill suedish mill mill <th< td=""><td>ga</td><td>irish</td><td></td><td>rm</td><td>romansh</td><td></td></th<>	ga	irish		rm	romansh	
Kannada Sanskrit kn kannada sa sanskrit Kurmanji Serbian serbian kmr kurmanji sr-latn serbian Lao sr-cyrl serbianc serbianc lo lao Slovak Latin sk slovak Latin sk slovak slovenian	Italian			Russian		
kn kannada sa sanskrit Kurmanji kmr kurmanji sr-latn serbian Lao sr-cyrl serbianc lo lao Slovak Latin sk slovak la latin sk slovak Latvian slovenian slovenian slovenian lv latvian spanish espanol lt lithuanian sv spanish espanol lt lithuanian sv swedish suedish ml malayalam sv swedish suedish mr marathi ta tamil se sanish sepanol Mongolian te telugu se	it	italian		ru	russian	
Kurmanji Serbian kmr kurmanji sr-latn serbian Lao sr-cyrl serbianc lo lao Slovak Latin sk slovak la latin sk slovenian Latvian slovenian slovenian slovenian lx latvian spanish espanol lt lithuanian sw swedish ml malayalam sv swedish ml malayalam Tamil Marathi ta tamil mr marathi Telugu Mongolian te telugu mn-cyrl mongolianlmc tr turkish mr-cyrl-x-lmc mongolianlmc tr turkish Norwegian Turkmen nn nynorsk tk turkmen Oriya Ukrainian typersorbian or oriya Uppersorbian pa panjabi	Kannada			Sanskrit		
Rum	kn	kannada		sa	sanskrit	
Lao sr-cyrl serbianc lo lao Slovak Latin sk slovak la latin Slovenian Latvian sl slovenian lv latvian Spanish espanols Lithuanian syanish espanols It lithuanian sv swedish Malayalam syanish espanol ml malayalam Tamil tamil mr marathi ta tamil mr marathi te telugu Mongolian te telugu mn-cyrl mongolianlmc tr turkish Norwegian tr turkish Norwegian tyrkinian Ukrainian Oriya uk ukrainian or oriya Uppersorbian Panjabi hsb uppersorbian Polish te te	Kurmanji			Serbian		
Io Iao Slovak Latin sk slovak Ia latin Slovenian slovenian	kmr	kurmanji		sr-latn		
Latin Isk slovenian Latvian Isl slovenian slovenian Iv latvian Spanish spanish espanol Lithuanian es spanish espanol It lithuanian Swedish swedish Malayalam sv swedish ml malayalam Tamil tamil Marathi ta tamil tamil mr marathi te telugu mn-cyrl mongolian tr turkish tr Morougilan nongolian mongolianlmc tr turkish turkish Norwegian turkmen Turkmen turkmen Norwegian Ukrainian turkmen turkmen Oriya uk ukrainian turkmen Panjabi hsb uppersorbian turkmen Polish Velsh turkmen	Lao				serbianc	
Ial Ialin Slovenian spanish espanish espanish spanish espanish spanish spanish	lo	lao				
Latvian Spanish slovenian slovenian Lithuanian es spanish espanol It lithuanian Swedish swedish Malayalam sv swedish swedish Marathi ta tamil tamil Mr marathi Telugu telugu telugu Mongolian Turkish tr turkish tr Norwegian tr turkish tr tr Norwegian Turkmen Ukrainian tr turkmen tr Oriya uk ukrainian tr non cyrl-x-lmc nongolian norsk tk tk turkmen tr nor tr					slovak	
Iv latvian Spanish ses spanish espanol Lithuanian es spanish espanol It lithuanian Swedish swedish Malayalam sy swedish ml malayalam Tamil Marathi ta tamil mr marathi Telugu Mongolian te telugu mn-cyrl mongolianlmc tr turkish Norwegian Turkmen Turkmen nb bokmal norwegian, norsk tk turkmen Oriya Ukrainian Ukrainian Oriya Uppersorbian hsb uppersorbian panjabi hsb uppersorbian panjabi Welsh Welsh	la	latin		Slovenian		
LithuanianesspanishespanolItlithuanianSwedishsvswedishMalayalamsvswedishmlmalayalamTamiltatamilMarathitatamilmrmarathiTeluguteteluguMongoliantrturkishmn-cyrlmongolianlmctrturkishNorwegianTurkmentrturkmennbbokmalnorwegian, norsktkturkmennnnynorskUkrainianUkrainianOriyaukukrainianororiyaUppersorbianuppersorbianPanjabihsbuppersorbianpapanjabiWelshPolishcywelsh					slovenian	slovene
It		latvian		Spanish		
MalayalamsvswedishmlmalayalamTamilMarathitatamilmrmarathiTeluguMongoliantetelugumn-cyrlmongolianTurkishmn-cyrl-x-lmcmongolianlmctrturkishNorwegianTurkmennbbokmalnorwegian, norsktkturkmennnnynorskUkrainianOriyaukukrainianororiyaUppersorbianPanjabihsbuppersorbianpapanjabiWelshPolishtywelsh					spanish	espanol
Marathi ta tamil mr marathi Telugu Mongolian te telugu mn-cyrl mongolianlmc Turkish mn-cyrl-x-lmc mongolianlmc tr turkish Norwegian Turkmen Turkmen nb bokmal norwegian, norsk tk turkmen nn nynorsk Ukrainian Oriya uk ukrainian or oriya Uppersorbian Panjabi hsb uppersorbian pa panjabi Welsh Polish cy welsh	lt	lithuanian		Swedish		
Marathi ta tamil mr marathi Telugu Mongolian te telugu mn-cyrl mongolianlmc Turkish nn-cyrl-x-lmc mongolianlmc tr turkish Norwegian Turkmen nb bokmal norwegian, norsk tk turkmen nn nynorsk Ukrainian Oriya uk ukrainian or oriya Uppersorbian Panjabi hsb uppersorbian pa panjabi Welsh Polish ty welsh	•				swedish	
mr marathi Telugu Mongolian te telugu mn-cyrl mongolianlmc Turkish nn-cyrl-x-lmc mongolianlmc tr turkish Norwegian Turkmen nb bokmal norwegian, norsk tk turkmen nn nynorsk Ukrainian Oriya uk ukrainian or oriya Uppersorbian Panjabi hsb uppersorbian pa panjabi Welsh Polish ty welsh		malayalam		Tamil		
Mongoliante delugumn-cyrlmongolianTurkishmn-cyrl-x-lmcmongolianlmctrturkishNorwegianTurkmennbbokmalnorwegian, norsktkturkmennnnynorskUkrainianOriyaukukrainianororiyaUppersorbianPanjabihsbuppersorbianpapanjabiWelshPolishtywelsh	Marathi				tamil	
mn-cyrl mongolian Turkish nn-cyrl-x-lmc mongolianlmc tr turkish Norwegian Turkmen nb bokmal norwegian, norsk tk turkmen nn nynorsk Ukrainian Oriya uk ukrainian or oriya Uppersorbian Panjabi hsb uppersorbian pa panjabi Welsh Polish ty welsh		marathi		Telugu		
mn-cyrl-x-lmc mongolianlmc tr Turkmen Norwegian nb bokmal norwegian, norsk tk turkmen nn nynorsk Ukrainian Oriya uk ukrainian or oriya Uppersorbian Panjabi hsb uppersorbian pa panjabi Welsh Polish					telugu	
Norwegian nb bokmal norwegian, norsk tk turkmen nn nynorsk Ukrainian Oriya uk ukrainian or oriya Uppersorbian Panjabi hsb uppersorbian pa panjabi Welsh Polish turkmen Ukrainian Ukrainian Uppersorbian ksb uppersorbian Welsh Ory welsh	Ü	0		Turkish		
nb bokmal norwegian, norsk tk turkmen nn nynorsk Ukrainian Oriya uk ukrainian or oriya Uppersorbian Panjabi hsb uppersorbian pa panjabi Welsh Polish turkmen Ukrainian Ukrainian Uppersorbian	-	mongolianlmc			turkish	
nn nynorsk Ukrainian Oriya uk ukrainian or oriya Uppersorbian Panjabi hsb uppersorbian pa panjabi Welsh Polish cy welsh	•					
Oriya ukrainian or oriya Uppersorbian Panjabi hsb uppersorbian pa panjabi Welsh Polish cy welsh	nb	bokmal	norwegian, norsk		turkmen	
or oriya Uppersorbian Panjabi hsb uppersorbian pa panjabi Welsh Polish cy welsh	nn	nynorsk		Ukrainian		
PanjabihsbuppersorbianpapanjabiWelshPolishcywelsh	Oriya			uk	ukrainian	
pa panjabi Welsh Polish cy welsh		oriya				
Polish cy welsh	Panjabi				uppersorbian	
θ		panjabi		Welsh		
* *. *				су	welsh	
pl polish	pl	polish				

Using hyphenation patterns

Plain T_FX

In engines that support ε-T_FX 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.

LATEX

Since Babel's hyphen.cfg is built in the Xel^ATeXformat, 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.

LATEX with Babel

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

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

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.

ConT_FXt

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

You can use full language name or language code. When using ConTeXt 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]
```

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