The ever-regenerating hydra: hyphenation patterns in Unicode, and beyond

Figure 1 The queen & the hero: a primitive hyphenation technique

Mojca Miklavec & Arthur Reutenauer

4th ConTEXt meeting and TEXperience, 2010, Mlýn Brejlov, 17.09.2010

Why do we need hyphenation?

Sometimes we come across words like "Rinderkennzeichnungs- und Rindfleischetikettierunsüberwachungsaufgabenübertragungsgesetz".

Rinderkennzeichnungs- und Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz

The hyph-utf8 started in 2008. Its original goal was to convert all the existing hyphenation patterns to UTF-8, and to provide automatic conversion to 8-bit encodings for "older" TEX engines.

This was completed rather quickly and we then moved to extend the existing hyphenation set.

The code

```
% Test whether we received one or two arguments

\def\testengine#1#2!{\def\secondarg{#2}}

% That's Tau (as in Taco or TEX, Tau-Epsilon-Chi),
% a 2-byte UTF-8 character

\testengine T!\relax

% Unicode-aware engines (such as XeTeX or LuaTeX)
% only see a single (2-byte) argument

\ifx\secondarg\empty
    \message{UTF-8 Slovenian Hyphenation Patterns}
\else
    \message{EC Slovenian Hyphenation Patterns}
    \input conv-utf8-ec.tex
\fi
\input hyph-sl.tex
```

The languages

ar	Arabic
fa	Persian
eu	Basque
bg	Bulgarian
cop	Coptic
hr	Croatian
cz	Czech
da	Danish
nl	Dutch
eo	Esperanto
et	Estonian
fi	Finnish
fr	French
de-1901	German, "old" spelling
de-1996	German, "new" spelling
el-monoton	Modern Greek, monotonic spelling
el-polyton	Modern Greek, polytonic spelling
grc	Ancient Greek
grc-x-ibycus	Modern Greek in the Ibycus encoding
hu	Hungarian
is	Icelandic
id	Indonesian
ia	Interlingua
ga	Irish
it	Italien
la	Latin
mn-cyrl	Mongolian, Cyrillic script
mn-cyrl-x-2a	Mongolian, Cyrillic script (new patterns)
no	Norwegian
nb	Norwegian Bokmål
nn	Norwegian Nynorsk
zh-latn	Chinese Pinyin
pl	Polish
pt	Portuguese
ro	Romanian
ru	Russian
sr-latn	Serbian, Latin script
sr-cyrl	Serbian, Cyrillic script
sh-latn	Serbo-Croatian, Latin script
sh-cyrl	Serbo-Croatian, Cyrillic script
sk	Slovak
sl	Slovenian
es	Spanish
sv	Swedish
tr	Turkish
en-gb	British English
en-us	American English
uk	Ukrainian
hsb	Upper Sorbian
cy	Welsh

New languages added: Lithuanian and Latvian, Sanskrit, modern Indic languages.

Collaboration: OpenOffice, `hyphenator.js`, Apache FOP.

Future plans

We always have future plans!

Extend the existing patterns: the ones that have been created with `patgen` usually don't work when `\lefthyphenmin` or `\righthyphenmin` is 1.

New hyphenation patterns (German).

An example

ubrouskem

Where to get help (or not)

Mailing-list: `tex-hyphen@tug.org`

Web page: `http://tug.org/tex-hyphen/`

People: Mojca & Arthur