Implementing PATGEN in Python

Ryszard Kubiak

BachoTEX 2019

Odkaszlnąć

```
.o d k a s z l n ą ć.
.o2d2    s4z      8ć.
.o0d3k2           ą1
  2d1k      2l1n
        2s0z1l
        2s0z0l0n
 o1    a1
.o2d3k2a2s4z2l1n0ą8ć.
```

od-kaszl-nąć

The authors of patterns

Petr Sojka, Pavel Ševeček
Hyphenation in TEX – Quo Vadis?,
TUGboat, Vol. 16 (1995)

Hanna Kołodziejska
Dzielenie wyrazów polskich w systemie TEX
Sprawozdania Instytutu Informatyki UW, 1987

Bogusław Jackowski, Marek Ryćko
Tam gdzie minus oznacza dzielenie
Biuletyn GUST, Zeszyt 2, 1993

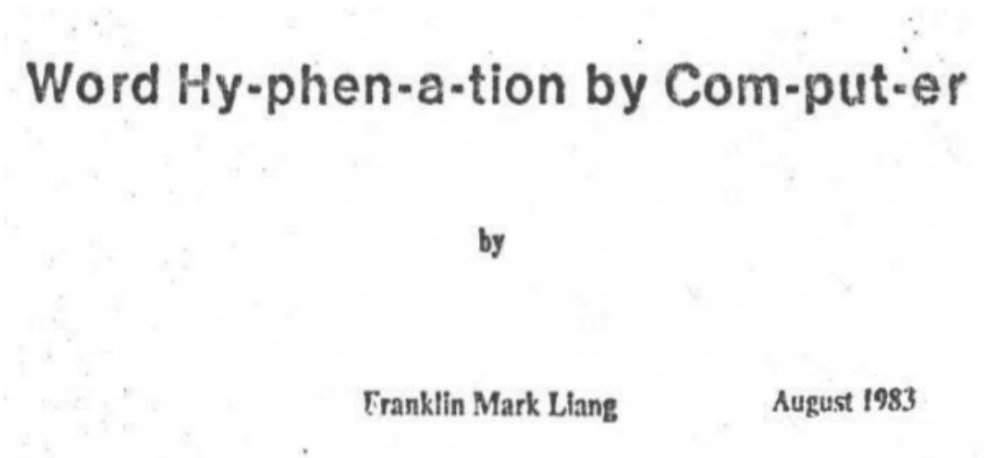
Creating patterns by hand

Group of patterns	What they mean
a1 ą1 e1 ę1 i1	Generally hyphenation allowed after a vowel
i2u i2ą i2e i2ę	But: firbidden between 'i' and a vowel
a2u e2u	Generally do not hyphenate combinations 'au', 'eu'
e3u2sz .na3u .nie3u .prze3u .za3u .za4uto	But: there are exceptions
c4h c4z d4ż d4ź r4z s4z	Generally do not hyphenate pairs that mean a single sound
ma2r5zn	But: there are exceptions

4033 of Polish patterns – hyph-pl.tex

.żv8	.dk8	4n3n	sze2z1long
.żw8	.dl8	4p3p	sze4ść
.żx8	.dm8	4r3r	szto2k1holm
.żz8	.dn8	4t3t	szyn2k1was
.a2b2s3t	.do3ć2	4w3w	to3y2o3t
.a2d3	.do3ł2	4z3z	turboo2d3rzut
.ad4a	.do3ś2	8ć.	tygo3d2ni
.ad4e	.do3ź2	8ćć.	u1

Frank Liang's Thesis



Word Hy-phen-a-tion by Com-put-er

by

Franklin Mark Liang August 1983

The idea of hyphenation patterns

PATGEN – automatic patterns generator

Use of 'trie' data structure for optimization

Various implementations

Version	Author	Year	Language	Features
patgen	F. Liang	1983	Pascal	original
patgen2	P. Breitenlohner	1991	C	consistent
opatgen	D. Antoš	2001	C++	UTF-8
hydra	A. Reutenauer	2016	Ruby	consistent
orthos	M. Nater	2016	Javascript	inconsistent
pypatgen	M. Kroutikov	2016	Python	inconsistent

Motivation

To understand

> Liang, p. 30: "In each pass we take into account only the effects of patterns chosen in previous passes."
>
> Can the 2 in s2z be replaced with 3?

Fun

Playing with the concepts of *clean code*

Python

Complex and atomic patterns

2cz1k

Dot	Pattern
cz·k	cz1k
c·zk	c0zk
czk·	czk0
·czk	2czk

PATGEN looks for atomic patterns!

Level by level. Length by Length. In organ-pipe order.

Coefficients for patterns selection

$$w_{good} * good - w_{bad} * bad \geqslant threshold$$

w_{good}	w_{bad}	$threshold$	
1	1	1	More good hits than bad ones
1	8	1	Errors highly deprecated

make-full-pattern.sh

Shell script by Norbert Schwartz, Bernd Raichle

Level	w_{good}	w_{bad}	*threshold*
1	1	1	1
2	1	2	1
3	1	1	1
4	1	4	1
5	1	1	1
6	1	6	1
7	1	4	1
8	1	8	1

Sample source dictionary

a-ba-żur
a-ba-żu-rem
a-ba-żu-ry
a-ba-żu-rze
a-bi-tu-rient
a-bi-tu-rien-tów
a-bra-ka-da-brze
abs-tra-hu-jąc
abs-tra-hu-ję
abs-trak-cjo-ni-stów
ab-surd
ab-sur-dal-ny

Level 1, after dot 1

a-ba-żur	a1d	a-ba*żur
a-ba-żu-rem	o1n	a-ba*żu-rem
a-ba-żu-ry	a1h	a-ba*żu-ry
a-ba-żu-rze	k1c	a-ba*żu-rze
a-bi-tu-rient	s1t	a-bi*tu-rient
a-bi-tu-rien-tów	a1ż	a-bi*tu-rien-tów
a-bra-ka-da-brze	i1t	a-bra-ka*da-brze
abs-tra-hu-jąc	i1s	abs*tra*hu*jąc
abs-tra-hu-ję	u1j	abs*tra*hu*ję
abs-trak-cjo-ni-stów	l1n	abs*trak*cjo*ni*s.tów
ab-surd		ab-surd
ab-sur-dal-ny		ab-sur-dal*ny

Level 1, after dots 1, 0

a-ba*żur	a1d 1br	a*ba*żur
a-ba*żu-rem	o1n 1ka	a*ba*żu*rem
a-ba*żu-ry	a1h 1da	a*ba*żu*ry
a-ba*żu-rze	k1c 1tó	a*ba*żu-rze
a-bi*tu-rient	s1t 1ba	a*bi*tu*rient
a-bi*tu-rien-tów	a1ż 1ry	a*bi*tu*rien*tów
a-bra-ka*da-brze	i1t 1bi	a*bra*ka*da*brze
abs*tra*hu*jąc	i1s 1su	abs*tra*hu*jąc
abs*tra*hu*ję	u1j 1re	abs*tra*hu*ję
abs*trak*cjo*ni*s.tów	l1n 1ri	abs*trak*cjo*ni*s.tów
ab-surd		ab*surd
ab-sur-dal*ny		ab*sur*dal*ny

Level 1, after dots 1, 0, 2

a*ba*żur
a*ba*żu*rem
a*ba*żu*ry
a*ba*żu-rze
a*bi*tu*rient
a*bi*tu*rien*tów
a*bra*ka*da*brze
abs*tra*hu*jąc
abs*tra*hu*ję
abs*trak*cjo*ni*s.tów
ab*surd
ab*sur*dal*ny

a1d 1br
o1n 1ka
a1h 1da
k1c 1tó
s1t 1ba
a1ż 1ry
i1t 1bi
i1s 1su
u1j 1re
l1n 1ri

a*ba*żur
a*ba*żu*rem
a*ba*żu*ry
a*ba*żu-rze
a*bi*tu*rient
a*bi*tu*rien*tów
a*bra*ka*da*brze
abs*tra*hu*jąc
abs*tra*hu*ję
abs*trak*cjo*ni*s.tów
ab*surd
ab*sur*dal*ny

Level 2, after dots 1, 0, 2

a*ba*żur
a*ba*żu*rem
a*ba*żu*ry
a*ba*żu-rze
a*bi*tu*rient
a*bi*tu*rien*tów
a*bra*ka*da*brze
abs*tra*hu*jąc
abs*tra*hu*ję
abs*trak*cjo*ni*s.tów
ab*surd
ab*sur*dal*ny

a1d
o1n
a1h
k1c
s1t
a1ż
i1t
i1s2
u1j
l1n

1br
1ka
1da
1tó
1ba
1ry
1bi
1su
1re
1ri

a*ba*żur
a*ba*żu*rem
a*ba*żu*ry
a*ba*żu-rze
a*bi*tu*rient
a*bi*tu*rien*tów
a*bra*ka*da*brze
abs*tra*hu*jąc
abs*tra*hu*ję
abs*trak*cjo*ni*stów
ab*surd
ab*sur*dal*ny

Dictionaries – instead of Tries

TEX pattern	Score
2sz1l	{0:2, 2:1}
	{0:2, 1:0, 2:1, 3:0}
.o2d2	{1:2, 2:2}

Scored patterns

```
{
'szl': {0:2, 2:1},
'.od': {1:2, 2:2}
}
```

Data of the Algorithm

`Dict` – dictionary of hyphenated words

$\texttt{Level}_{\texttt{min}}$, $\texttt{Level}_{\texttt{max}}$ – levels range

$\texttt{Length}_{\texttt{min}}$, $\texttt{Length}_{\texttt{max}}$ – lengths range

`Patt` – patterns from previous steps

`Margins` – `\lefthyphenmin`, `\righthyphenmin`

$\texttt{W}_{\texttt{good}}$, $\texttt{W}_{\texttt{bad}}$, `T` – selection coefficients

Main Loop

```
patterns := Patt
for level in Level_min..Level_max do
    for length in Length_min..Length_max do
        for dot in organ_pipe_order_range(length) do
            new_patterns := finder(patterns, level, length, dot)
            patterns.extend(new_patterns)
```

Body, given: patterns, level, length, dot

```
missed, bad := patterns.verify_against_dict(Dict, level)
candidates, c_good, c_bad := {}, {}, {}
for w in Dict do
    for p, d in pinned_subwords(w, length, dot, Margins) do
        if missed.contains(p, d, w) then
            candidates.add(p)
        adjust(p, d, w, Dict, level, missed, bad, c_good, c_bad)
return candidates.select(c_good, c_bad, W_good, W_bad, T)
```

Features

Python version 3.5, static typing

ca. 700 lines of code

Modules for: hyphenating, verifying, generating

Slow (minutes compared to patgen's seconds)

TODOs

Publish

- still work on cleaning the code
- more unit tests
- main routines
- documentation (literate style using Sphinx)
- join efforts with Mike Kroutikov (pypatgen)

Reimplementation using tries

GUI?

Various encodings?

Reimplementation in Lua