FACULTY
OF INFORMATICS
Masaryk University

Towards Universal Hyphenation Patterns

Petr Sojka (Faculty of Informatics at Masaryk University)
Ondřej Sojka ($\mathcal{CS}$TUG)


EXIT

Contents

Section 1

Introduction to Hyphenation Patterns

Patterns (in general)

> “**pattern** ORIGIN Middle English patron ‘something serving as a model’, from Old French. The change in sense is from the idea of *patron giving an example to be copied*. Metathesis in the second syllable occurred in the 16th cent. By 1700 patron ceased to be used of things, and the two forms became differentiated in sense.”
>
> — *New Oxford Dictionary of English*, *1998 edition*

Patterns (in general)

> “**pattern** ORIGIN Middle English patron ‘something serving as a model’, from Old French. The change in sense is from the idea of *patron giving an example to be copied*. Metathesis in the second syllable occurred in the 16th cent. By 1700 patron ceased to be used of things, and the two forms became differentiated in sense.”
>
> — *New Oxford Dictionary of English*, *1998 edition*

Patterns everywhere: rhythm patterns in music or poetry conveying message, patterns of behaviour, letter patterns, . . . , you name it: *hyphenation patterns*.

Patterns (of hyphenation) that compete each other

Frank Liang, DEK's student at Stanford (Ph.D., 1983), developed *the method* and algorithms for hyphenation based on the idea of competing patterns of varying length to cope with exceptions. [4].

- general, language-independent method
- pattern is a substring with a information about hyphenation between characters: `hy3ph he2n .euro7 7tex.`
- odd numbers allow hyphenation, even numbers forbid hyphenation

Patterns (of hyphenation) that compete each other

Frank Liang, DEK's student at Stanford (Ph.D., 1983), developed *the method* and algorithms for hyphenation based on the idea of competing patterns of varying length to cope with exceptions. [4].

- general, language-independent method
- pattern is a substring with a information about hyphenation between characters: `hy3ph he2n .euro7 7tex.`
- odd numbers allow hyphenation, even numbers forbid hyphenation
- patterns are as short as possible to be as general as possible (new compound words, etc)
- pattern compete each other: instead of one big set of patterns, decomposition into several layered approximations (subpatterns)
  p_1 (covering subpatterns), p_2 (inhibiting subpatterns—exceptions for p_1),
  p_3 (covering subpatterns to cover what has not been covered by "$p_1 \wedge \neg p_2$"),...

Hyphenation lookup: an instance of dictionary problem

```
    h y p h e n a t i o n
p1          1n a
p1              1t i o n
p2           n2a t
p2                2i o
p2       h e2n
p3  h y3p h
p4       h e n a4
p5       h e n5a t
    h0y3p0h0e2n5a4t2i0o0n
```

hy-phen-ation $\rightarrow$ 2 6

$\ldots \rightarrow \ldots$

$\ldots \rightarrow \ldots$

key $\rightarrow$ data

Solution to the dictionary problem:
For key part (the word) to store
the data part (its division)

Hyphenation lookup: an instance of dictionary problem

```
    h y p h e n a t i o n
p1          1n a
p1              1t i o n
p2           n2a t
p2                2i o
p2       h e2n
p3  h y3p h
p4       h e n a4
p5       h e n5a t
    h0y3p0h0e2n5a4t2i0o0n
```

hy-phen-ation $\rightarrow$ 2 6

$\ldots \rightarrow \ldots$

$\ldots \rightarrow \ldots$

key $\rightarrow$ data

Solution to the dictionary problem:
For key part (the word) to store
the data part (its division)

Given the already hyphenated word list of a language (dictionary), *how to generate the patterns*? The task was: less than 5,000 patterns, less than 30,000 bytes per language.

hyphen.tex generation by `patgen` (Liang, 1983)

level	parameters	patterns	good	bad	good	bad
1	1 2 20 (4)	458	67,604	14,156	76.6%	16.0%
2	2 1 8 (4)	509	7,407	11,942	68.2%	2.5%
3	1 4 7 (5)	985	13,198	551	83.2%	3.1%
4	3 2 1 (6)	1647	1,010	2,730	82.0%	0.0%
5	1 ∞ 4 (8)	1320	6,428	0	**89.3%**	0.0%

A total of 4,919 patterns (4,447 unique) obtained in hyphen.tex (27,860 bytes) from Webster pocket dictionary (30,000+ words only). *Suffix-compressed packed trie* occupying 5,943 locations, with 181 outputs. Patterns find 89.3% of the hyphens in the dictionary. 109 passes through the dictionary are needed. Generation required about 1 hour of CPU time on PDP-11.

hyphen.tex: [unreasonable?] power of learning from data

```
% The Plain TeX hyphenation tables [NOT TO BE CHANGED IN ANY WAY!]
% Unlimited copying and redistribution of this file are permitted as long
% as this file is not modified. Modifications are permitted, but only if
% the resulting file is not named hyphen.tex.
\patterns{ % just type <return> if you're not using INITEX
.ach4 .ad4der .af1t .al3t .am5at
...
```

hyphen.tex: [unreasonable?] power of learning from data

```
% The Plain TeX hyphenation tables [NOT TO BE CHANGED IN ANY WAY!]
% Unlimited copying and redistribution of this file are permitted as long
% as this file is not modified. Modifications are permitted, but only if
% the resulting file is not named hyphen.tex.
\patterns{ % just type <return> if you're not using INITEX
.ach4 .ad4der .af1t .al3t .am5at
...
```

"An important feature of a learning machine is that its teacher will often be very largely ignorant of quite what is going on inside, although he may still be able to some extent to predict his pupil's behaviour." — *Alan Turing*, *Mind 59:433-460*, *1950*

`patgen` program: machine learning from data

One of the very first approaches that harnessed the power of data: Liang's program `patgen` for generation of hyphenation patterns from a word list:

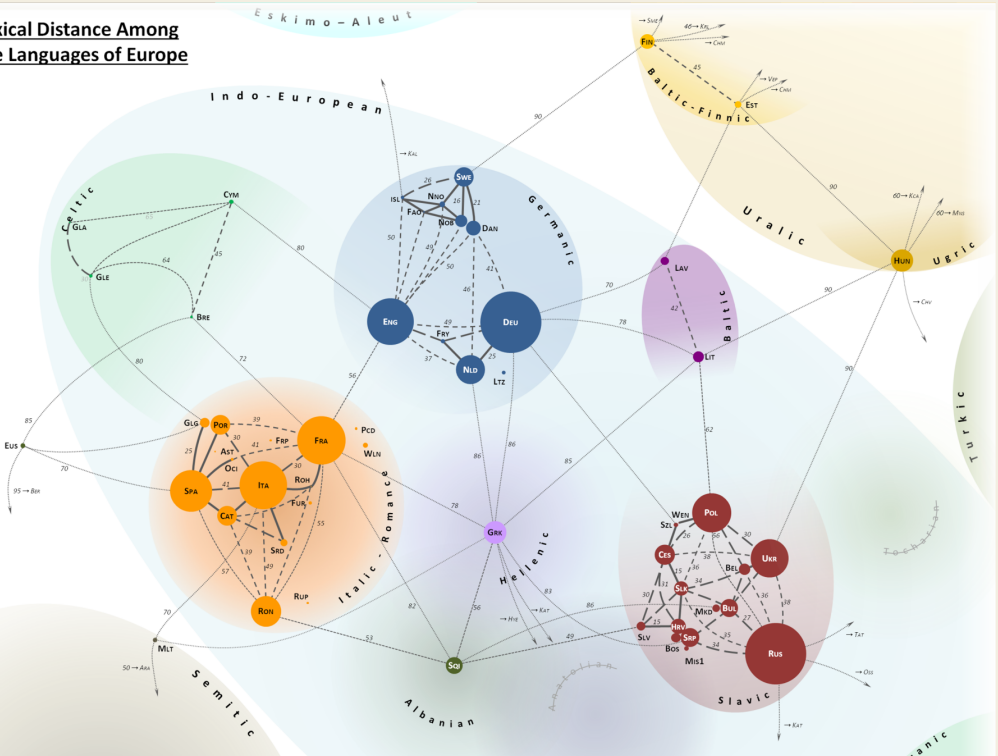
- efficient lossy or lossless *compression* of hyphenated dictionary with several orders of magnitude compression ratio.
- generated patterns have minimal length, e.g., shortest context possible, which results in their *generalization* properties.
- hyphenation of out of vocabulary words, too.

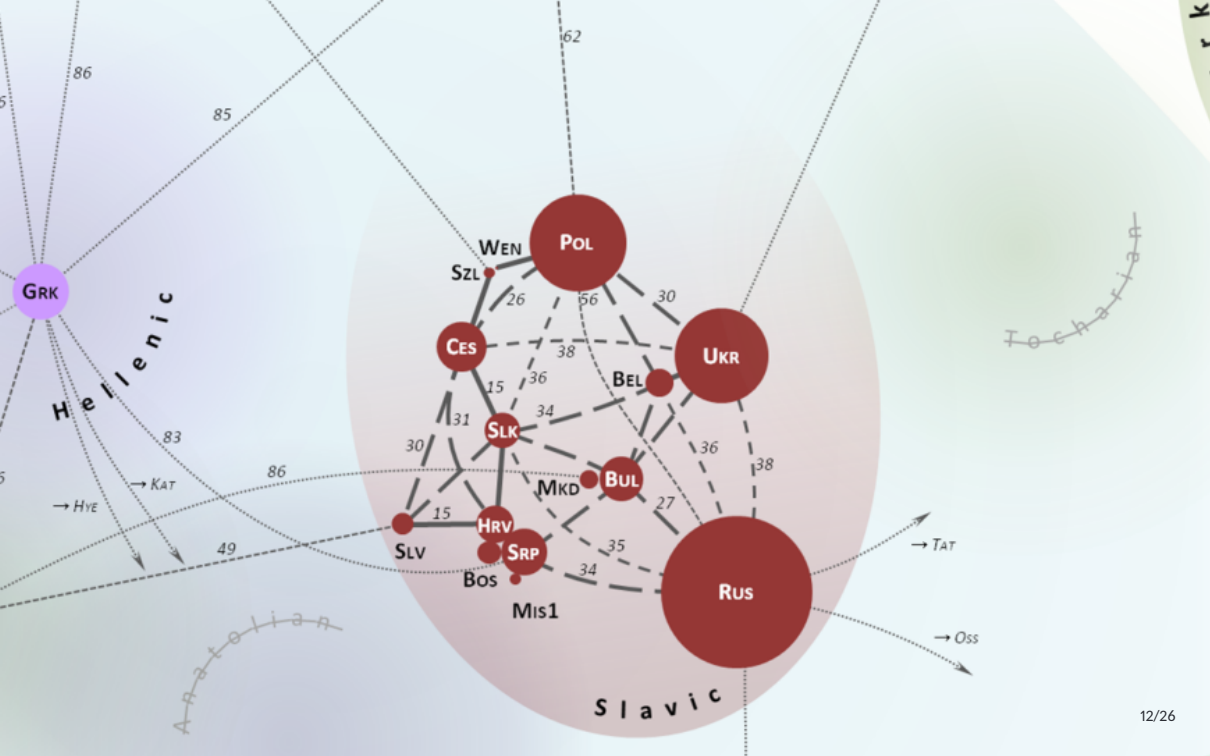
Exact lossless pattern minimization is non-polynomial by reduction to the minimum set cover problem [6]. *Exact lossless* pattern generation is feasible for Czech [7] (TUG 2019), and methodology applicable for others languages — patterns for dozens of languages are loaded at each TEX run from LATEX format file.

Section 2

The idea of Universal Hyphenation Patterns

Lexical Distance Among the Languages of Europe




62
86
85
5
Grk
Hellenic
Wen
Szl
Pol
26
56
30
Ces
38
Ukr
Bel
15
36
31
34
Slk
83
30
36
38
86
Mkd
Bul
→ Kat
→ Hye
15
27
Hrv
49
Slv
Srp
35
Bos
34
Rus
Mis1
→ Tat
→ Oss
Anatolian
Tocharian
Slavic

Hypothesis

Is there no word that has different hyphenations in the covered languages?

Then we can cover multiple languages with one set of hyphenation patterns.

Hypothesis

Is there no word that has different hyphenations in the covered languages?

Then we can cover multiple languages with one set of hyphenation patterns.

We already do this (kind of).

fa-ce-bo-ok

Hyphenation in Czech

Rules are published at [2]: https://prirucka.ujc.cas.cz/?id=135

- syllabic according to pronunciation
- morphology only secondary

Hyphenation in Slovak

Rules are published at [3] https://www.juls.savba.sk/ediela/psp2000/psp.pdf

- morphology primary according to the Ľ. Štúr Institute of Linguistics
- syllabic hyphenation secondary
- morphological boundaries are often also syllabic boundaries
- current patterns hyphenate mostly syllabically

Section 3

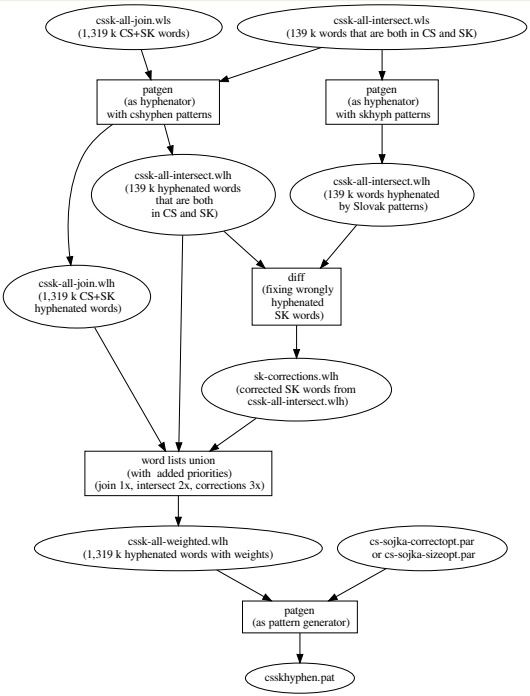
Preparation of the Czechoslovak patterns

A problem

- no hyphenated word list available for Slovak

A problem

- no hyphenated word list available for Slovak
- we have
  - Czech word list with *mostly* correct hyphenations
  - Slovak noisy word list without hyphenations
  - new Czech hyphenation patterns
  - old Slovak hyphenation patterns
- we need Czechoslovak patterns


cssk-all-join.wls
(1,319 k CS+SK words)
cssk-all-intersect.wls
(139 k words that are both in CS and SK)
patgen
(as hyphenator)
with cshyphen patterns
patgen
(as hyphenator)
with skhyph patterns
cssk-all-intersect.wlh
(139 k hyphenated words
that are both
in CS and SK)
cssk-all-intersect.wlh
(139 k words hyphenated
by Slovak patterns)
cssk-all-join.wlh
(1,319 k CS+SK
hyphenated words)
diff
(fixing wrongly
hyphenated
SK words)
sk-corrections.wlh
(corrected SK words from
cssk-all-intersect.wlh)
word lists union
(with added priorities)
(join 1x, intersect 2x, corrections 3x)
cssk-all-weighted.wlh
(1,319 k hyphenated words with weights)
cs-sojka-correctopt.par
or cs-sojka-sizeopt.par
patgen
(as pattern generator)
csskhyphen.pat

Observations

- telling apart the prefix *nej-* from *ne-* is a problem
- Slovak patterns hyphenate *syllabically* most of the time, contrary to recommendations of the Slovak language institute.
  - ne-na_u-čí
- is this correct behavior?
  - vy-ma-lo-va-ných
  - vy-maľ-o-va-ných

Results

Word list	Parameters	Good	Bad	Missed	Size	# Patterns
Czechoslovak	sizeopt	99.67%	0.00%	0.33%	32 kB	5,679
Czechoslovak	correctopt	99.96%	0.00%	0.04%	48 kB	8,199
Czech	correctopt [7]	99.76%	2.94%	0.24%	30 kB	5,593
Czech	sizeopt [7]	98.95%	2.80%	1.05%	19 kB	3,816
Slovak, `patgen`	from Table 1 of [5]	99.94%	0.01%	0.06%	56 kB	2,347
Slovak, by hand	[1]	N/A	N/A	N/A	20 kB	2,467

Section 4

Conclusions

Why is this great?

- since there are very few words that are hyphenated differently in English and Czechoslovak, universal patterns could be developed to perfectly hyphenate Czech texts full of English terms
- straightforward upgrade path from old Czech or Slovak patterns – just set language to Czechoslovak
- resource savings
  - TEX loads hyphenation patterns for all languages into memory automatically

Future work

- check hyphenations with Czech Language Institute

Future work

- check hyphenations with Czech Language Institute
- check hyphenations with Ľ. Štúr Institute of Linguistics

Future work

- check hyphenations with Czech Language Institute
- check hyphenations with Ľ. Štúr Institute of Linguistics
- generate final Czechoslovak patterns

Future work

- check hyphenations with Czech Language Institute
- check hyphenations with Ľ. Štúr Institute of Linguistics
- generate final Czechoslovak patterns
- explore joint patterns for English + Czechoslovak, and/or universal Slavic patterns

Future work

- check hyphenations with Czech Language Institute
- check hyphenations with Ľ. Štúr Institute of Linguistics
- generate final Czechoslovak patterns
- explore joint patterns for English + Czechoslovak, and/or universal Slavic patterns
- CzechoSlovak support TEXlive 2020

That's it, thanks!

- DEK for asking the right question at the right time
- Vít Suchomel of Lexical Computing for csTenTen word lists from Sketch Engine;

That's it, thanks!

- DEK for asking the right question at the right time
- Vít Suchomel of Lexical Computing for csTenTen word lists from Sketch Engine;

Questions?

Questions for you

Are there any words that should be hyphenated differently in Slovak and Czech?

Which Slavic languages could be a good fit?

[1] Janka Chlebíková.
Ako rozdeliť (slovo) Československo (How to Hyphenate (word) Czechoslovakia).
CSTUG Bulletin, 1(4):10–13, April 1991.

[2] Internetová jazyková příručka (Internet Language Reference Book).

[3] Pravidlá slovenského pravopisu.

[4] Franklin M. Liang.
Word Hy-phen-a-tion by Com-put-er.
PhD thesis, Department of Computer Science, Stanford University, August 1983.

[5] Petr Sojka.
Slovenské vzory dělení: čas pro změnu?
In *Proceedings of SLT 2004, 4th seminar on Linux and TEX*, pages 67–72, Znojmo, 2004. Konvoj.

[6] Petr Sojka.
Competing Patterns in Language Engineering and Computer Typesetting.
PhD thesis, Masaryk University, Brno, January 2005.

[7] Petr Sojka and Ondřej Sojka.

The unreasonable effectiveness of pattern generation.
TUGboat, 40(2):187–193, 2019.