Towards Universal Hyphenation Patterns

Petr Sojka (Faculty of Informatics at Masaryk University)
Ondřej Sojka (ČS TUG)
Contents

1. Introduction to Hyphenation Patterns

2. The idea of Universal Hyphenation Patterns

3. Preparation of the Czechoslovak patterns

4. Conclusions
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.”

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.”


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 “$p_1 \land \neg p_2$”), …
Hyphenation lookup: an instance of dictionary problem

\[ \text{hyphenation} \]

\[
\begin{align*}
\text{p1} & \quad \text{ln a} \\
\text{p1} & \quad \text{t i o n} \\
\text{p2} & \quad \text{n2a t} \\
\text{p2} & \quad \text{2i o} \\
\text{p2} & \quad \text{h e2n} \\
\text{p3} & \quad \text{y3p h} \\
\text{p4} & \quad \text{e n a4} \\
\text{p5} & \quad \text{e n5a t} \\
\text{h0y3p0h0e2n5a4t2i0o0n} \end{align*}
\]

\[ \text{hy-phen-at} \rightarrow 2 \ 6 \\
\ldots \rightarrow \ldots \\
\ldots \rightarrow \ldots \\
\text{key} \rightarrow \text{data} \\
\]

Solution to the dictionary problem:

For key part (the word) to store the data part (its division)
Hyphenation lookup: an instance of dictionary problem

\[
\text{hyphenation} \\
p_1 \quad \text{ln} \ a \\
p_1 \quad \text{lt} \ i \ o \ n \\
p_2 \quad \text{n2a} \ t \\
p_2 \quad 2i \ o \\
p_2 \quad \text{h e2n} \\
p_3 \quad \text{h y3p} \ h \\
p_4 \quad \text{h e n a4} \\
p_5 \quad \text{h e n5a} \ t
\]

hy-phen-ation $\rightarrow$ 26

\[
\ldots \rightarrow \ldots \\
\ldots \rightarrow \ldots \\
\text{key} \rightarrow \text{data}
\]

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)

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

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.
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,
"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 \texttt{\LaTeX} run from \texttt{\LaTeX} format file.
Section 2

The idea of Universal Hyphenation Patterns
Lexical Distance Among the Languages of Europe
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


- 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
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

<table>
<thead>
<tr>
<th>Word list</th>
<th>Parameters</th>
<th>Good</th>
<th>Bad</th>
<th>Missed</th>
<th>Size</th>
<th># Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czechoslovak</td>
<td>sizeopt</td>
<td>99.67%</td>
<td>0.00%</td>
<td>0.33%</td>
<td>32 kB</td>
<td>5,679</td>
</tr>
<tr>
<td>Czechoslovak</td>
<td>correctopt</td>
<td>99.96%</td>
<td>0.00%</td>
<td>0.04%</td>
<td>48 kB</td>
<td>8,199</td>
</tr>
<tr>
<td>Czech</td>
<td>correctopt [7]</td>
<td>99.76%</td>
<td>2.94%</td>
<td>0.24%</td>
<td>30 kB</td>
<td>5,593</td>
</tr>
<tr>
<td>Czech</td>
<td>sizeopt [7]</td>
<td>98.95%</td>
<td>2.80%</td>
<td>1.05%</td>
<td>19 kB</td>
<td>3,816</td>
</tr>
<tr>
<td>Slovak, patgen</td>
<td>from Table 1 of [5]</td>
<td>99.94%</td>
<td>0.01%</td>
<td>0.06%</td>
<td>56 kB</td>
<td>2,347</td>
</tr>
<tr>
<td>Slovak, by hand</td>
<td>[1]</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>20 kB</td>
<td>2,467</td>
</tr>
</tbody>
</table>
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 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?
   Ako rozdeliť (slovo) Československo (How to Hyphenate (word) Czechoslovakia).

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

3. Pravidlá slovenského pravopisu.

   *Word Hyphen-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?

6. Petr Sojka.
   *Competing Patterns in Language Engineering and Computer Typesetting.*

7. Petr Sojka and Ondřej Sojka.
The unreasonable effectiveness of pattern generation.