Towards Universal Hyphenation Patterns

Petr Sojka and Ondřej Sojka

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
sojka@fi.muni.cz, ondrej.sojka@gmail.com

Abstract. Hyphenation is at the core of every document preparation system, being that typesetting system such as TEX or modern web browser. For every language, there have to be algorithms, rules, or patterns hyphenating according to that. We are proposing the development of generic hyphenation patterns for a set of languages sharing the same principles, e.g., for all syllable-based languages. We have tested this idea by the development of Czechoslovak hyphenation patterns. At the minimal price of a tiny increase in the size of hyphenation patterns, we have shown that further development of universal syllabic hyphenation patterns is feasible.

Keywords: hyphenation, hyphenation patterns, patgen, syllabification, Unicode, TEX, syllabic hyphenation, Czech, Slovak

"Any respectable word processing package includes a hyphenation facility. Those based on an algorithm, also called logic systems, often break words incorrectly."
Major Keary in [6]

1 Introduction

Hyphenation is at the core of every document preparation system, be it TEX or any modern web browser.[1] There are about 5,000 languages supported by Unicode Consortium that are still in use today. In a digital typography system supporting Unicode and its languages in full, there should be support: algorithms, rules, or language hyphenation patterns. Recently, there were attempts to tackle the word segmentation problem in different languages by Shao et al. [9], primarily for speech recognition or language representation tasks, where the algorithm is error-prone – small number of errors is tolerated. On the contrary, in a typesetting system like TEX, errors in hyphenation are not tolerated at all – all exceptions have to be covered by the algorithm.

Current typesetting support in the TEXlive distribution contains [8] hyphenation patterns for about 80 different languages. All these patterns have to be loaded into TEX's memory at the start of every compilation, which slows down compilation significantly.

There are essentially two quite different approaches to hyphenation:

[1] Cascading style sheet version 3 hyphenation implementation is supported in Firefox and Safari since 2011.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019, pp. 63–68, 2019.

etymology-based Rule is to cut word on the border of a compound word or the border of the stem and ending or prefix or negation. A typical example is British hyphenation rules created by the Oxford University Press [1].

phonology-based Hyphenation based on the pronunciation of syllables allows reading text with hyphenated lines similarly or the same as if the hyphenation were not there. This pragmatic approach is preferred by the American publishers [4] and the Chicago Manual of Style [2].

In this paper, we evaluate the feasibility of the development of universal phonology-based (syllabic) hyphenation patterns. We describe the development from word lists of Czech [11,15,12] and Slovak [14] used on the web pages. We describe the reproducible approach, and document the reproducible workflow and resources in the public repositories as a language resource and methods to be followed.

"Hyphenation does not lend itself to any set of unequivocal rules. Indeed, the many exceptions and disagreements suggest it is all something dreamed up at an anarchists' convention." Major Keary in [6]

2 Methods

The core idea is to develop common hyphenation patterns for phonology-based languages. In the case these languages share the pronunciation rules, homographs from different languages typically do not cause problems, as they are hyphenated the same. The rare cases that hyphenation is dictated by the seam of compound word contrary to phonology (`ro-zum` vs. `roz-um`) could be solved by not allowing the hyphenation.

Recently, we have shown that the approach to generate hyphenation patterns from word list by program `patgen` is unreasonably effective [16]. One can set the parameters of the generation process so that the patterns cover 100% of hyphenation points, and the size of the patterns remains reasonably small. For the Czech language, hyphenation points from 3,000,000 hyphenated words are squeezed into 30,000 bytes of patterns, stored as in the compressed trie data structure. That means achieving a compression ratio of several orders of magnitude with 100% coverage and nearly zero errors. [16] For a similar language such as Slovak, the pronunciation is very similar, syllable-forming principles are the same, and also compositional rules and prefixes are pretty close, if not identical.

We have decided to verify the approach by developing hyphenation patterns that will hyphenate both Czech and Slovak words without errors, with only a few missed hyphens. That means that only words like `oblít` will not be hyphenated, because the typesetting system cannot decide in which meaning the word is used: `o-blít` or `ob-lít`.

To generate these hyphenation patterns, we needed to create lists of correctly hyphenated Czech and Slovak words.

Data Preparation

For our work, word lists with frequencies for Czech and Slovak were donated from the TenTen family of corpora [5,7]. Only words that occurred more than ten times were used in further processing.

Czech word list was cleaned up and extended as described by us in [16], using Czech morphological analyzer `majka`. Final word list `cs-all-cstenten.wls` has 606,494 words.

For Slovak, we have got 1,048,860 Slovak words with frequency higher than 10 from SkTenTen corpora from 2011 [5]. Filtering only words containing ISO-Latin2 characters we obtained file `sk-all-latin2.wls` with 991,552 words.

Together, we have 1,319,334 Czech and Slovak words in `cssk-all-join.wls`, of which 139,356 were contained in both word lists: `cssk-all-intersect.wls`.

Pattern Development

The workflow of Czechoslovak pattern development is illustrated in Figure 1 on the following page. We have used recent accurate Czech patterns [16] for hyphenation of the joint Czech and Slovak word list. We had to manually fix bad hyphenation typically near the prefix and stem of words when phoneme-based hyphenation was one character close to the seam of the prefix or compound word: `neja-traktivnější`, `neja-teističtější`, `neje-kologičtější`.

We have then hyphenated words used in both languages also by current Slovak patterns. There were only a few word hyphenations that needed to be corrected – we created the `sk-corrections.wlh` that contained the fixed hyphenated words. Finally, we used them with a higher weight to generate final Czechoslovak hyphenated patterns.

Results

We have tried several parameters to generate Czechoslovak patterns, namely those developed in previous research [16], e.g. parameters for size- and correct-optimized set of patterns. After short fine tuning we have generated patterns that are both correct (Table 2 on page 67) and small (Table 1 on page 67). That means that `patgen` was able to generalize hyphenation rules common for both languages with a negligible enlargement of the generated patterns.

We have made all results and workflow reproducible by putting all files and necessary software (scripts, Makefiles) publicly available in our repository [10].

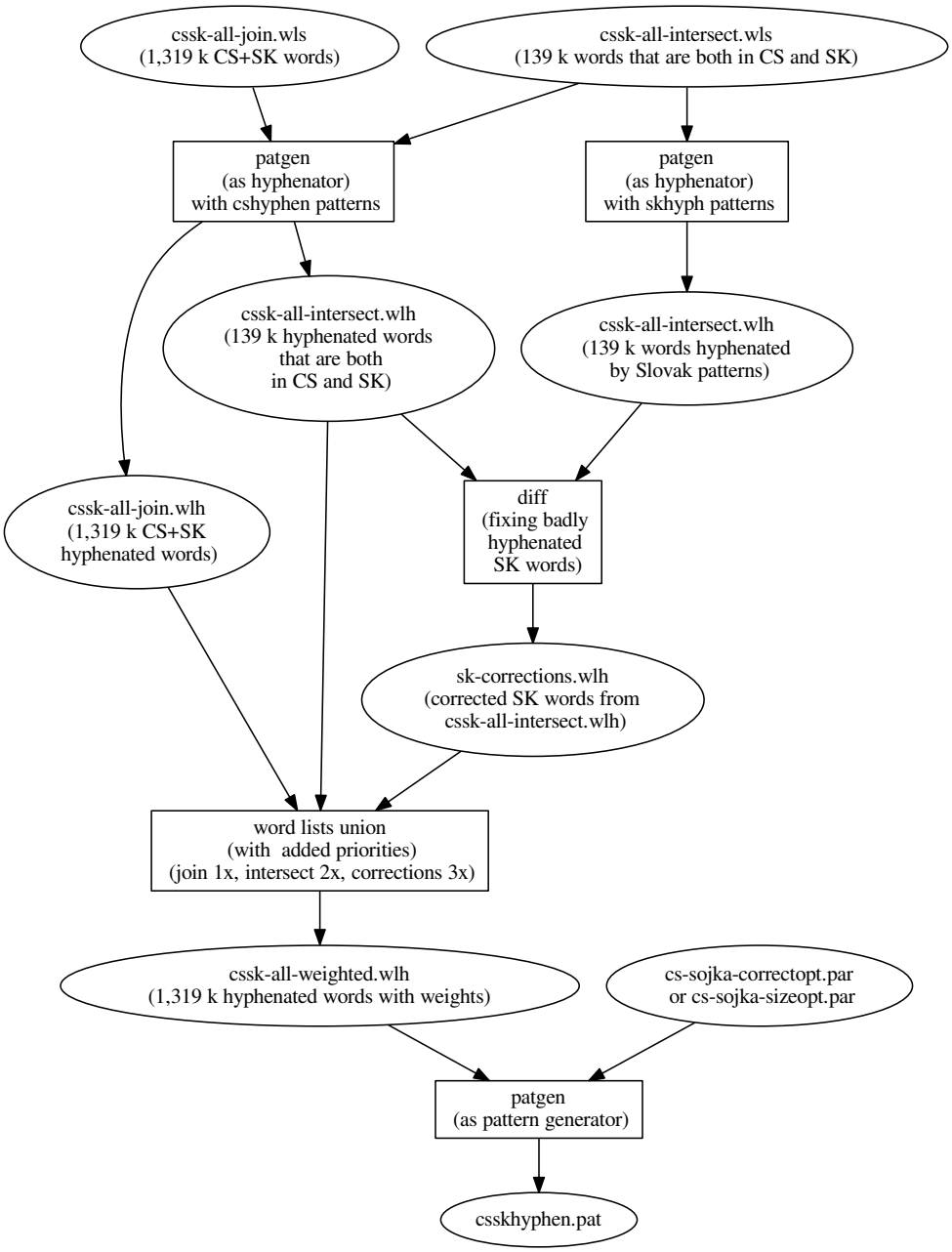

Fig. 1. The development process of the new Czechoslovak patterns. Bootstrapping with Czech patterns, checking and fixing with higher weight Slovak words that are common with Czech ones.

Table 1. Statistics from the generation of Czechoslovak hyphenation patterns with size optimized parameters.

Level	Patterns	Good	Bad	Missed	Lengths	Params
1	491	4,126,566	917,233	37,871	1 3	1 2 20
2	1,651	3,610,957	1,244	553,480	2 4	2 1 8
3	4,031	4,153,820	21,816	10,617	3 5	1 4 7
4	2,647	4,150,588	0	13,849	4 7	3 2 1

Table 2. Statistics from the generation of Czechoslovak hyphenation patterns with correct optimized parameters.

Level	Patterns	Good	Bad	Missed	Lengths	Params
1	2,221	4,082,523	345,325	81,914	1 3	1 5 1
2	2,237	4,071,117	9,416	93,320	1 3	1 5 1
3	4,337	4,164,049	11,001	395	2 6	1 3 1
4	2,647	4,162,859	0	1,625	2 7	1 3 1

Table 3. Comparison of the efficiency of different approaches to hyphenating Czech and Slovak. Note that the Czechoslovak patterns are comparable in size and quality to single-language ones – there is only a negligible difference compared e.g., to purely Czech patterns.

Word list	Patterns	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 [16]	99.76%	2.94%	0.24%	30 kB	5,593
Czech	sizeopt [16]	98.95%	2.80%	1.05%	19 kB	3,816
Slovak	[13, Table 1, `patgen`]	99.94%	0.01%	0.06%	56 kB	2,347
Slovak	[3, by hand]	N/A	N/A	N/A	20 kB	2,467

> "Esoteric Nonsense? Hyphenation is neither anarchy nor the sole province of pedants and pedagogues.... If the author wants to attract and hold an audience, then hyphenation needs just as careful attention as any other aspect of presentation."
>
> Major Keary in [6]

3 Conclusion and Future Works

We have shown that the development of common hyphenation patterns for languages with similar pronunciation is feasible. The resulting Czechoslovak patterns are only slightly bigger than single-language patterns and hyphenate the source word list without a single error.

Development is expected to continue in the repository [10]. Final word lists and versions of hyphenation patterns will be deposited into the LINDAT-Clarin archive.

We will double-check the hyphenated word lists with members of Czechoslovak TEX Users Group CSTUG. We will finally offer the new patterns for "Czechoslovak language" to the TEXlive distribution, creating the first language support package to be shared by multiple languages.

Acknowledgement This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin infrastructure LM2015071. We are indebted to Don Knuth for questioning common properties of Czech and Slovak hyphenation during our presentation of [16] at TUG 2019 that lead us into this direction.

References

1. Allen, R.: The Oxford Spelling Dictionary, The Oxford Library of English Usage, vol. II. Oxford University Press (1990)
2. Anonymous: The Chicago Manual of Style. University of Chicago Press, Chicago, 17 edn. (Sep 2017)
3. Chlebíková, J.: Ako rozdeliť (slovo) Československo (How to Hyphenate (word) Czechoslovakia). CSTUG Bulletin **1**(4), 10–13 (Apr 1991)
4. Gove, P.B., Webster, M.: Webster's Third New International Dictionary of the English language Unabridged. Merriam-Webster Inc., Springfield, Massachusetts, U.S.A (Jan 2002)
5. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family. In: Proc. of 7th International Corpus Linguistics Conference (CL). pp. 125–127. Lancaster (Jul 2013)
6. Keary, M.: On hyphenation – anarchy of pedantry. PC Update, The magazine of the Melbourne PC User Group (2005), https://web.archive.org/web/20050310054738/http://www.melbpc.org.au/pcupdate/9100/9112article4.htm
7. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. In: Proceedings of the Eleventh EURALEX International Congress. pp. 105–116. Lorient, France (2004)
8. Reutenauer, A., Miklavec, M.: TEX hyphenation patterns, https://tug.org/texhyphen/, accessed 2019-11-24
9. Shao, Y., Hardmeier, C., Nivre, J.: Universal Word Segmentation: Implementation and Interpretation. Transactions of the Association for Computational Linguistics **6**, 421–435 (2018). https://doi.org/10.1162/tacl_a_00033, https://www.aclweb.org/anthology/Q18-1033
10. Sojka, O., Sojka, P.: cshyphen repository, https://github.com/tensojka/cshyphen
11. Sojka, P.: Notes on Compound Word Hyphenation in TEX. TUGboat **16**(3), 290–297 (1995)
12. Sojka, P.: Hyphenation on Demand. TUGboat **20**(3), 241–247 (1999)
13. Sojka, P.: Slovenské vzory dělení: čas pro změnu? In: Proceedings of SLT 2004, 4th seminar on Linux and TEX. pp. 67–72. Konvoj, Znojmo (2004)
14. Sojka, P.: Slovenské vzory dělení: čas pro změnu? (Slovak Hyphenation Patterns: A Time for Change?). CSTUG Bulletin **14**(3–4), 183–189 (2004)
15. Sojka, P., Ševeček, P.: Hyphenation in TEX – Quo Vadis? TUGboat **16**(3), 280–289 (1995)
16. Sojka, P., Sojka, O.: The unreasonable effectiveness of pattern generation. TUGboat **40**(2), 187–193 (2019), https://tug.org/TUGboat/tb40-2/tb125sojka-patgen.pdf