Benchmark Dataset of Hyphenated Words for Generating Hyphenation Patterns

Ondřej Metelka, 514098@mail.muni.cz
Petr Sojka, sojka@fi.muni.cz

Faculty of Informatics, Masaryk University, Brno, Czech Republic

December 6, 2025

Contents

Introduction

The hyph-bench dataset

Evaluation and Results

Conclusions and Future Steps

Discusion

Bibliography

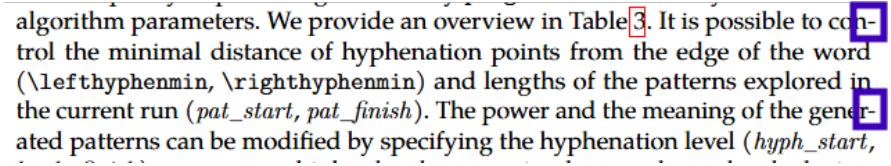
Section 1

Introduction

Hyphenation patterns - Why bother?

Every typesetting system needs a hyphenation mechanism.

Example 1: Excerpt from our article


algorithm parameters. We provide an overview in Table 3. It is possible to con-
trol the minimal distance of hyphenation points from the edge of the word
(`\lefthyphenmin`, `\righthyphenmin`) and lengths of the patterns explored in
the current run (*pat_start*, *pat_finish*). The power and the meaning of the gener-
ated patterns can be modified by specifying the hyphenation level (*hyph_start*,

Example 2: Very bad hyphenation

`ther-apist` vs. `the-rapist`

Tools: `patgen` and the others

Several paradigms for hyphenation generation:

- **pattern based**: the oldest, baseline algorithm [5],
  - highly optimized, *packed trie* structure for storing data
- **finite state methods**: modelling with deterministic automata [1],
- **neural approaches**: e.g., Bi-LSTM [3, 6], or Transformer [2] architecture.

Section 2

The hyph-bench dataset

Quick peek

- **18** datasets comprising **15** languages
  - Slavic: **Czech**, **Czecho-Slovak***, **Russian**, **Ukrainian**
  - Germanic: **German**, **Icelandic**, **Dutch**
  - Romance: **Spanish**, **Italian**, **Portuguese**
  - Hellenic: **Modern Greek**
  - non-European: **Malay**, **Thai**, **Turkish**
- over **5 million** lines with marked hyphenation points in total
- languages with **non-Latin scripts** (Modern Greek, Russian, Thai, Ukrainian) included

Language/Name	Size	#Lines	$\overline{\#\text{Lines}}$	$\overline{\#\text{Hyphs}}$	#Chars
cs/cshyphen_cstenten	8,202.6	597,170	11.83	2.79	48
cs/cshyphen_ujc	1,377.1	104,277	11.48	2.59	48
cs/wiktionary	632.0	53,962	9.95	2.16	48
cssk/cshyphen	10,646.4	835,145	11.10	2.54	50
de/wiktionary	12,821.6	807,813	15.01	2.96	86
de/wortliste	10,035.4	604,970	15.78	2.99	47
el/wiktionary	603.9	28,208	11.90	2.87	52
es/wiktionary	11,753.4	816,723	13.36	3.27	50
is/hyphenation-is	3,037.8	218,308	12.12	1.91	45
it/wiktionary	981.1	77,351	11.96	2.90	39
ms/wiktionary	13.6	1,247	8.52	1.81	69
nl/wiktionary	8,124.5	528,598	14.73	2.82	68
pl/wiktionary	1,713.6	141,916	10.75	2.20	52
pt/wiktionary	541.6	44,321	11.17	2.74	40
ru/wiktionary	3,795.6	152,608	13.10	2.74	51
th/orchid	3,567.1	34,868	40.69	6.02	95
tr/wiktionary	187.2	15,619	10.41	2.37	42
uk/wiktionary	351.6	15,701	11.79	2.65	41

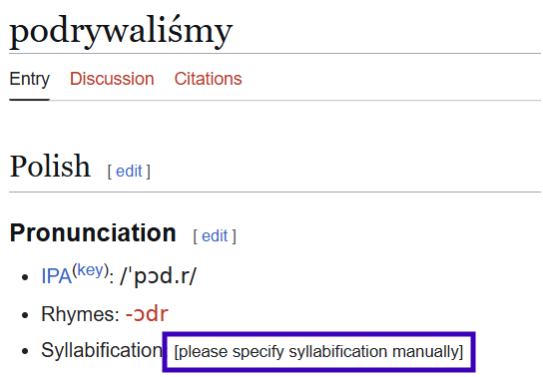
Data collection

We aggregated data from various sources:

- **parsed entries from Wiktionary [12, 13]**: Czech, German, Modern Greek, Spanish, Italian, Malay, Dutch, Polish, Portuguese, Russian, Turkish,
- **freely accessible word lists**: German [4], Icelandic [7],
- **text corpora with hyphenation points**: Thai [11],
- **outputs of previous work in the field**: Czech [9, 10], Czecho-Slovak [9], Ukrainian [8].

The Curse of Community Data

Wiktionary datasets: large, but no review assured



podrywaliśmy

Entry Discussion Citations

Polish [edit]

Pronunciation [edit]

- IPA(key): /ˈpɔd.r/
- Rhymes: -ɔdr
- Syllabification: [please specify syllabification manually]

Pronunciation [edit]

- (*Brazil*) IPA(key): /fi.lo.neˈis.te/
- (*Portugal*) IPA(key): /fi.lu.neˈiʃ.te/
- Hyphenation: mi·so·ne·ís·ta

Noun [edit]

filoneísta *m or f* (*plural* **filoneístas**)

1. philoneist [synonym, antonym ▲]

Synonym: neófilo

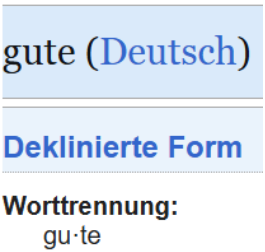
Antonym: misoneísta

The Curse of Community Data I

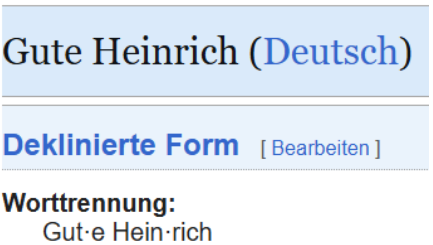
Several ambiguous word hyphenations in the data

- `de/wiktionary`: **741** ambiguous hyphenations, e.g., `gu-te` vs. `gut-e`
- `de/wortliste`: **0** ambiguous hyphenations, only `gu-te` present here

Example 1: `gute` [= *CS dobří, EN good*]



Example 2: `Gute Heinrich` [= *CS merlík všedobr, EN Good King Henry*]



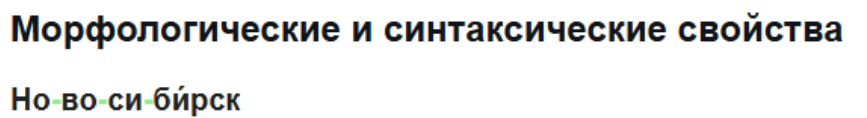
The Curse of Community Data II

Non-hyphenated words are often stored in the *hyphenations* field in the dump file

- problem particularly within `ru/wiktionary` dataset

Example 1: мышонок [= *CS myška, EN little mouse*]

Example 2: Новосибирск [= Novosibirsk, Russian city]

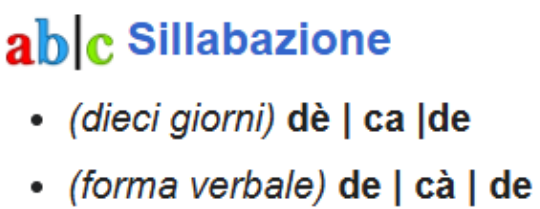


The Curse of Community Data III

Inconsistent markup in dump files

- different hyphenation marks, extra characters, parts of the word missing
- linguistic notes (e.g., *'traditional'*)
- pronunciation accents added in some languages, e.g., Italian

Example: The accentuation in the Italian word `decade`



Preprocessing

Data cleaning and disambiguation: noisy input

- recall previous slides

Evaluation preparation: limitations posed by `patgen`

- lines longer than **50 characters**
- characters with **special meaning** (numeric literals, asterisk $*$, full stop .)

Example: Too-long German word

Do·nau·dampf·schiff·fahrts·elek·t·ri·zi·tä·ten·haupt·be·triebs·werk·bau·un·ter·be·am·ten·ge·sell·schaft, *kein Plural*

Section 3

Evaluation and Results

Methodology

Each dataset ran through `patgen` with a predefined set of parameters (we call these **parameter profiles**)

- profiles taken from prior work [4, 9], or crafted by ourselves
- 10-fold cross-validation used for generalization assessment

Example: `base` parameter profile

Profile	Level	Lengths	Params
base	1	2 4	1 1 1
	2	3 5	1 2 1
	3	4 7	1 3 1
	4	5 9	1 4 1

Metrics

For each dataset/parameter profile combination, we measured

- $F_{1/7}$**-score**: weighted harmonic mean of **precision** and **recall**, assessing the correctness of the generated patterns,
  - more importance is put on precision since wrong hyphenation is way worse than missing one
- **pattern trie size**: the number of nodes in the pattern storing trie, assessing the space efficiency,

and reported their mean across all 10 folds.

Language/Name	base		cshyphen		wortliste		wortliste8	
	$F_{1/7}$	trie size	$F_{1/7}$	trie size	$F_{1/7}$	trie size	$F_{1/7}$	trie size
cs/cshyphen_cstenten	0.9266	14630	0.9444	12052	0.9243	12121	0.9243	13008
cs/cshyphen_ujc	0.9803	19339	0.9830	17046	0.9803	14736	0.9794	20761
cs/wiktionary	0.9712	11275	0.9759	10643	0.9744	8330	0.9739	9960
cssk/cshyphen	0.9981	27653	0.9986	25302	0.9981	21685	0.9982	23690
de/wiktionary	0.9982	37751	0.9983	32394	0.9977	29449	0.9983	37648
de/wortliste	0.9000	54035	0.9775	47544	0.8985	43214	0.9007	55424
el/wiktionary	0.9846	4232	0.9875	3602	0.9866	3223	0.9864	3722
es/wiktionary	0.9997	3299	0.9998	2023	0.9998	1784	0.9998	2129
is/hyphenation-is	0.9804	30244	0.9839	32284	0.9802	28664	0.9797	31229
it/wiktionary	0.9963	4026	0.9970	3593	0.9968	2875	0.9964	4114
ms/wiktionary	0.9039	991	0.9296	796	0.9244	721	0.9236	733
nl/wiktionary	0.9967	33889	0.9972	30859	0.9963	27514	0.9967	33954
pl/wiktionary	0.9904	10058	0.9913	8564	0.9904	7262	0.9901	9845
pt/wiktionary	0.9955	2449	0.9969	1761	0.9970	1422	0.9969	1768
ru/wiktionary	0.9192	68380	0.9270	53820	0.9212	42164	0.9173	92188
th/orchid	0.9569	32666	0.9622	28795	0.9488	24258	0.9551	38587
tr/wiktionary	0.9870	2217	0.9920	1406	0.9916	1273	0.9911	1407
uk/wiktionary	0.9517	5458	0.9602	4741	0.9564	4272	0.9552	4880

Section 4

Conclusions and Future Steps

What did we do and find?

We...

- compiled a first version of the benchmark dataset of hyphenated words,
- proposed a framework for hyphenation evaluation,
- evaluated a baseline approach against the framework.

You can find all the data and scripts at
`https://github.com/tondach01/hyph-bench`.

What next?

Follow-up research may focus on...

- **hyph-bench v2.0**: more languages, better parsing and disambiguation, combination of the datasets into a mega-benchmark,
- **more evaluation**: compare baseline algorithm with other hyphenation methods (FSM, neural, etc.),
- **dynamic profiles**: parameters for `patgen` determined on-the-fly via optimization methods, e.g., Bayesian optimization, allowing for smaller and more accurate patterns for each language,
- **universal patterns**: training a set of common patterns for a group of (e.g., Slavic) languages.

Section 5

Discusion

Section 6

Bibliography

Bibliography I

[1] Gosse Bouma. “Finite State Methods for Hyphenation.” In: *Natural Language Engineering* 9.1 (2003), pp. 5–20. DOI: 10.1017/S1351324903003073.

[2] Jan Franěk. “Transformer-Based Word Hyphenation.” Master’s thesis. Masaryk University, Brno, Faculty of Informatics (advisor: Petr Sojka), 2025. URL: https://is.muni.cz/th/wbgkw/.

[3] Jacob Krantz, Maxwell Dulin, and Paul De Palma. “Language-Agnostic Syllabification with Neural Sequence Labeling.” In: *18th IEEE International Conference On Machine Learning And Applications, ICMLA 2019, Boca Raton, FL, USA, December 16–19, 2019*. Ed. by M. Arif Wani et al. IEEE, 2019, pp. 804–810. DOI: 10.1109/ICMLA.2019.00141.

Bibliography II

[4] Werner Lemberg. *A database of German words with hyphenation information*. German. 2023. URL: https://repo.or.cz/wortliste.git (visited on 11/22/2025).

[5] Franklin M. Liang. “Word Hy-phen-a-tion by Com-put-er.” PhD thesis. Stanford University, Aug. 1983, p. 44.

[6] Xuezhe Ma and Eduard Hovy. “End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF.” In: *Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers)*. ACL, 2016, pp. 1064–1074. DOI: 10.18653/v1/P16-1101.

[7] Kristján Rúnarsson, Baldur Jónsson, and Magnús Gíslason. *Icelandic Hyphenation Dictionary (2.0)*. CLARIN-IS. 2020. URL: http://hdl.handle.net/20.500.12537/86.

Bibliography III

[8] Ondřej Sojka. “Transfer Learning of Slavic Syllabification for Hyphenation Patterns.” Bachelor Thesis, Masaryk University, Brno, Faculty of Informatics (advisor: Pavel Šmerk, https://is.muni.cz/th/s17d6/?lang=en. Masaryk University, Brno, Feb. 2025.

[9] Petr Sojka and Ondřej Sojka. “New Czechoslovak Hyphenation Patterns, Word Lists, and Workflow.” eng. In: *TUGboat* 42.2 (2021). ISSN: 0896-3207. URL: https://doi.org/10.47397/tb/42-2/tb131sojka-czech.

[10] Petr Sojka and Ondřej Sojka. “The Unreasonable Effectiveness of Pattern Generation.” In: *TUGboat* 40.2 (2019), pp. 187–193. URL: https://tug.org/TUGboat/tb40-2/tb125sojka-patgen.pdf.

Bibliography IV

[11] Virach Sornlertlamvanich, Thatsanee Charoenporn, and Hitoshi Isahara. *ORCHID: Thai Part-Of-Speech Tagged Corpus*. Tech. rep. TR-NECTEC-1997-001. http://www.links.nectec.or.th/. Thai National Electronics and Computer Technology Center, Thailand, Dec. 1997, pp. 5–19.

[12] Wiktionary contributors. *Wiktionary*. URL: https://www.wiktionary.org/ (visited on 11/22/2025).

[13] Tatu Ylonen. "Wiktextract: Wiktionary as Machine-Readable Structured Data." In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 1317–1325. URL: https://aclanthology.org/2022.lrec-1.140/.