# A Lightweight Approach to a Giga-Corpus of Historical Periodicals: The Story of a Slovenian Historical Newspaper Collection

Filip Dobranić[1], Bojan Evkoski[2], Nikola Ljubešić[3,1]

[1]Institute of Contemporary History
[2]Department of Network and Data Science, Central European University
[3]Department of Knowledge Technologies, Jožef Stefan Institute

May 2024

# The Most Uninteresting Thing In the World


Figure 1: Yesterdays' newspapers.

# Not That Uninteresting

**Meteorologično poročilo.**

| Dan | Čas opazovanja | Stanje barometra v mm. | Temperatura | Vetrovi | Nebo | Mokrina v mm. |
|---|---|---|---|---|---|---|
| 5. okt. | 7. zjutraj | 731·7 mm. | 11·4° C | sl. zah. | obl. | 6·10 mm dežja. |
| | 2. popol. | 733·0 mm. | 18·0° C | sl. jzh. | obl. | |
| | 9. zvečer | 734·9 mm. | 15·6° C | sl. zah. | d. jas. | |

Srednja temperatura 15·0°, za 2·0° nad normalom.

**Dr. IVAN MILAN HRIBAR**

vljudno naznanja, da je

**otvoril**

svojo **odvetniško pisarno**

**v Ljubljani**

Sodnijske ulice štev. 2.

3242—2

**Radenska**

**kisla voda po natriju in litiju najbogatejša**

preskušeno zdravilno sredstvo proti **mehurnim boleznim, kamenu v mehurji, protinu**, dalje proti **boleznim v želodci, mokril, dolgotrajnemu kataru dihal, zlatej žili** in **zlatenici.**

Poskusi doktorjev: Garrod, Biswanger, Cantani, Ure so dokazali, da ima **ogljenokisli litij** največjo raztopilno moč pri sečnokislih usednih izločkih, iz česar se sklepa na najugodnejše učinke Radenske kisle vode.

Kot okrepčujoča pijača z vinom ali sadnimi soki in sladkorjem pomešana, je Radenska voda v obče priljubljena.

**Kopališče Radensko**

Kopanje v slatini in v jeklenici upliva posebno pri: protinu, bolečinah mokril, malokrvnosti, ženskih boleznih, slabostih i. t. d.

**Stanovališče po ceni.**

**Prospekt** (brezplačno) od **ravnateljstva: Kopališče Radence** blizu **Ljutomera.**

**Radenska kisla voda,** ob vznožji Slovenskih goric, **ne zamenjati z Radgonsko,** to je **Radkersburger.**

Čeravno Radenska kisla voda stoji več kot jeden dan v odprti posodi, vender se peni prav močno, ako se z vinom pomeša, ker ima v sebi spojene ogljikove kisline. To svojstvo pa jo odlikuje pred mnogimi drugimi kislimi vodami, katere imajo navadno le prosto ali manj trdno spojeno ogljikovo kislino, ki se že pogubi in razkadi, kakor hitro se steklenica odmaši. (264—5)

**Prvega petelina v letošnji sezoni je v soboto 1. aprila ustrelil na Koloveu ljubljanski zobozdravnik g. dr. Bretl.**

**Poslano.***)

Res je, da sem svojo ženo A. R. pretepel zaradi njenega hudega jezika in ker namesto mene ljubi „repetničarje“. Če bi ji bil jaz rebro zlomil, ne bi bila v štirih dneh zdrava.

**Anton Rajk,**

**Vavpčavas, Dobernče, Dolenjsko.**

*) Za vsebino tega spisa je uredništvo odgovorno le toliko, kolikor določa zakon. (2579)

Figure 2: Some examples of interesting things in yesterdays' newspapers.

# Humble beginnings

- 250k historical "digitised" newspaper issues. [3]
- Varied and "questionable" quality.
- PDF + TXT files.
- No annotations.
- Disappointing search.

# How to build a corpus in 7 easy steps

1. Remove suspiciously short documents.
2. Filter by language.
3. Filter by "quality".
4. Correct split words.
5. Align pages.
6. Correct (some) OCR errors.
7. Lingustically annotate the corpus. [6, 8]

# Removing suspiciously short documents

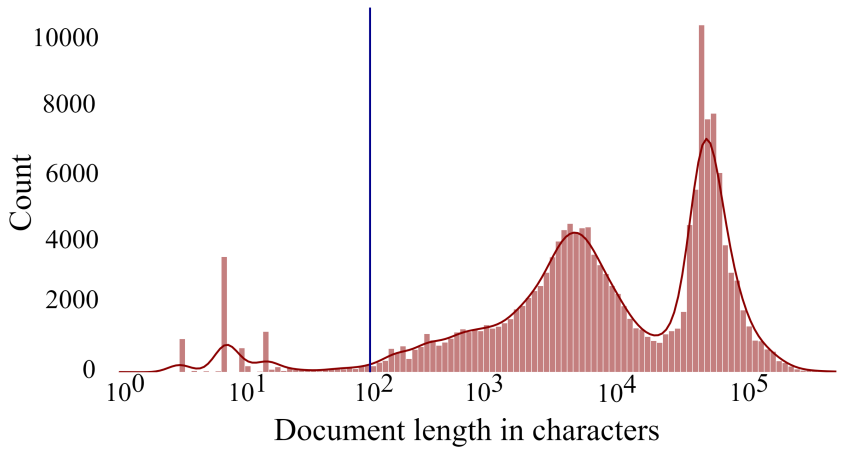


Figure 3: Documents' character-size distribution with the cut-off point of excluding document shorter than 100 characters.

# Filtering by language

- Librarians did most of the work for us ...
- ... the rest was done by FastText. [1]

# Filtering by quality

- Intuition: the more it reads like contemporary news, the more it is like contemporary news.
- Use KenLM [5] + SentiNews [2] to determine contemporary-news-like-ness.
- Quality is in the eye of the beholder, so talk to those beholden.

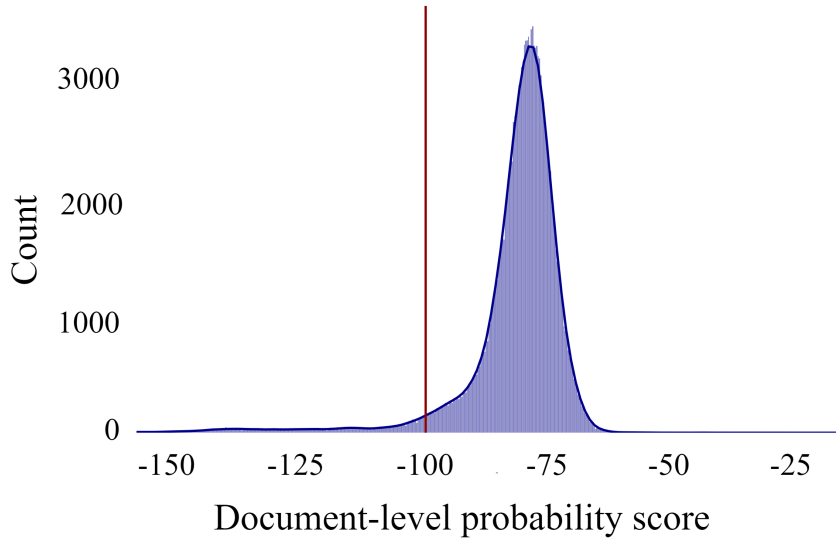


Figure 4: Documents' probability scores as reported by the language model.

# Fixing things

- Correct split words with regular expressions and a bit of stats.
- Align pages between PDF and TXT files.
- Correct OCR errors with cSMTiser [7].
  - Text normalisation model: 300 random paragraphs of training data.
  - Statistical language model: all paragraphs with at least 100 characters & ParlaMint. [4]
  - WER: 5.4% $\rightarrow$ 4.4%
  - CER 1.2% $\rightarrow$ 1%

# Post-OCR error correction

| OCR-ed trigram | Corrected trigram | Suitability |
|---|---|---|
| ki ae je | ki se je | suitable |
| vcčkrat pa 12 | večkrat pa 12 | suitable |
| bii bi na | bil bi na | suitable |
| Izdajatelj In odgovorni | Izdajatelj in odgovorni | suitable |
| nočejo oirok pustiti | nočejo otrok pustiti | suitable |
| Štev. 8. | štev. 8. | ambiguous |
| Članov in obresti | članov in obresti | ambiguous |
| Številke po 4 | številke po 4 | ambiguous |
| Aadaiko - \fomkm | Aadaiko - fomkm | unsuccessful |
| Ui aa\|-dovrženeHa kar | Ui aaj-dovrženela kar | unsuccessful |
| Družtvo sv. Jožefa | Društvo sv. Jožefa | historical difference |

# Corpus totals

| | |
|---|---|
| Number of tokens | 928,540,876 |
| Number of sentences | 52,604,613 |
| Number of documents | 157,669 |
| Number of named entities | 39,705,149 |
| Person | 21,052,963 |
| Location | 14,085,060 |
| Organization | 2,387,221 |
| Miscellaneous | 2,179,905 |

# Corpus persons


Figure 5: Most common person named entities.

# Corpus places

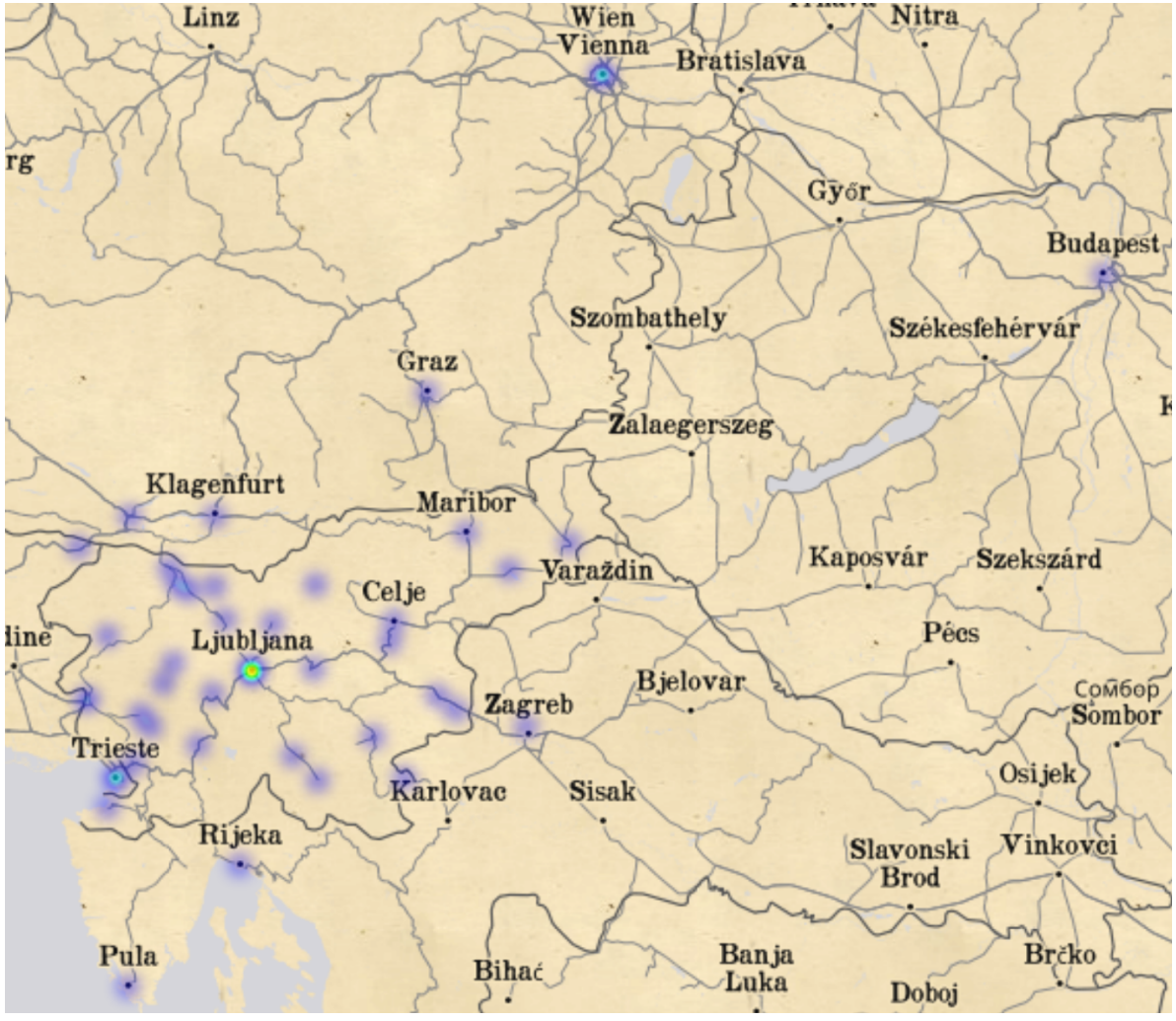


Figure 6: Geographic heat map of Slovenia and neighboring regions. It marks the most common locations detected in the corpus.

# Finally

- We built a useful giga-level historical corpus without breaking the bank and so can you!
- Corpus links:
  - Download: `https://djnd.si/ejao`
  - Explore: `https://djnd.si/ejap`
  - Replicate: `https://djnd.si/ejaq`

# References I


- Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the association for computational linguistics **5**, 135–146 (2017)
- Bučar, J.: Manually sentiment annotated slovenian news corpus SentiNews 1.0 (2017), `http://hdl.handle.net/11356/1110`, slovenian language resource repository CLARIN.SI
- dLib, D.k.S.: dLib.si - periodika, `http://dlib.si/Publications.aspx`
- Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Grigorova, V., Rudolf, M., Pančur, A., Kopp, M., Barkarson, S., Steingrímsson, S., van der Pol, H., Depoorter, G., de Does, J., Jongejan, B., Haltrup Hansen, D., Navarretta, C., Calzada Pérez, M., de Macedo, L.D., van Heusden, R., Marx, M., Çöltekin, Ç., Coole, M., Agnoloni, T., Frontini, F.,

# References II


Montemagni, S., Quochi, V., Venturi, G., Ruisi, M., Marchetti, C., Battistoni, R., Sebők, M., Ring, O., Dargis, R., Utka, A., Petkevičius, M., Briedienė, M., Krilavičius, T., Morkevičius, V., Diwersy, S., Luxardo, G., Rayson, P.: Multilingual comparable corpora of parliamentary debates ParlaMint 2.1 (2021), http://hdl.handle.net/11356/1432, slovenian language resource repository CLARIN.SI

Heafield, K., Pouzyrevsky, I., Clark, J.H., Koehn, P.: Scalable modified Kneser-Ney language model estimation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 690–696. Association for Computational Linguistics, Sofia, Bulgaria (Aug 2013), https://www.aclweb.org/anthology/P13-2121

# References III


- Ljubešić, N., Dobrovoljc, K.: What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In: Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing. pp. 29–34. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-3704, https://www.aclweb.org/anthology/W19-3704
- Ljubešić, N., Zupan, K., Fišer, D., Erjavec, T.: Normalising Slovene data: historical texts vs. user-generated content. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016). pp. 146–155 (2016)
- Terčon, L., Ljubešić, N.: CLASSLA-Stanza: The next step for linguistic processing of south slavic languages (2023)