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, Privoz 11, SI-1000 Ljubljana
[2]Department of Network and Data Science, Central European University, Quellenstraße 51, AT-1100 Wien
[3]Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana
filip.dobranic@inz.si, evkoski_bojan@phd.ceu.edu, nikola.ljubesic@ijs.si

Abstract

Preparing historical newspaper collections is a complicated endeavour, consisting of multiple steps that have to be carefully adapted to the specific content in question, including imaging, layout prediction, optical character recognition, and linguistic annotation. To address the high costs associated with the process, we present a lightweight approach to producing high-quality corpora and apply it to a massive collection of Slovenian historical newspapers from the 18th, 19th and 20th century resulting in a billion-word giga-corpus. We start with noisy OCR-ed data produced by different technologies in varying periods by the National and University Library of Slovenia. To address the inherent variability in the quality of textual data, a challenge commonly encountered in digital libraries globally, we perform a targeted post-digitisation correction procedure, coupled with a robust curation mechanism for noisy texts via language model inference. Subsequently, we subject the corrected and filtered output to comprehensive linguistic annotation, enriching the corpus with part-of-speech tags, lemmas, and named entity labels. Finally, we perform an analysis through topic modeling at the noun lemma level, along with a frequency analysis of the named entities, to confirm the viability of our corpus preparation method.

Keywords: Historical giga-corpus, Slovenian, periodicals, Post-OCR correction

1. Introduction

In recent years, the application of Optical Character Recognition (OCR) technology has made it possible to transform centuries-old texts into machine-readable formats (Hamad and Mehmet, 2016; Chaudhuri et al., 2017). This development has paved the way for extensive computational analyses and linguistic studies of historical texts, enabling scholars and researchers to delve deeper into the intricate nuances of the past (Singh et al., 2012; Marjanen et al., 2020). These endeavours are supported by digital corpora of historical print records, the creation of which usually takes the form of multi-person-year projects of interdisciplinary teams (Ahnert et al.; Ehrmann et al., 2020; Cordell et al., 2017; Cordell and Smith, 2022). We present the creation of a curated corpus of Slovenian historical periodicals from the 18th, 19th and 20th centuries on a shoestring budget, made possible by previous upstream digitisation efforts made by archivists at the National and University Library of Slovenia (NUK).

Similar digitised source collections exist across GLAM institutions all over the world (e.g. BnF, retrieved 2023; Austrian National Library, retrieved 2023; Deutsche Digitale Bibliothek, retrieved 2023), but so far their unknown and often unreliable quality prevented them from being fully utilised. While bare access to digitised sources is a very important achievement, such collections are often noisy and the sources present are of varying quality. They often lack deep search indices and linguistic annotations, as well as searchable metadata, e.g. Nacionalna i sveučilišna knjižnica u Zagrebu, retrieved 2023. Navigating through vast amounts of material of unknown quality poses challenges for researchers, including difficulties in locating specific information and conducting computational analysis. Moving from repositories of scans to curated concordancers is a crucial step in supporting comprehensive historical research (Pfanzelter et al., 2021).

In this paper, we outline the methodology employed in creating and curating the Slovenian historical newspapers corpus and provide a broad overview of some of its content which encompasses a wide range of historical texts, including manuscripts, letters, diaries, newspapers, and other written records. The task of curating a dataset from noisy historical data is a complex endeavor which is bound to introduce additional biases (Beelen et al., 2022). By embracing this biased reality of corpora creation, we advocate for allowing researchers to build corpora tailored to their specific research interests. Achieving this requires efficient, easily applicable methods that significantly reduce effort without compromising the quality necessary for downstream research.

By outlining our curation, correction, and enrichment process, we aim to empower researchers and small teams to create their own corpora, especially for underserved languages and subject matters, both in terms of funding and digitised materials

LREC-COLING 2024, pages 695–703
20-25 May, 2024.

available. Such collections will provide researchers, historians, linguists, and enthusiasts with valuable resources and means to explore the cultural, political, and linguistic landscape of our past. We are releasing the source code of our computational methods along with this report.[1]

2. Compilation of the corpus

2.1. Origin

The collection we present is a selection of the expansive digital archive, dLib[2], maintained by The National and University Library of Slovenia. Among other digitised sources, it currently contains over 800,000 freely accessible periodical issues of which we initially obtained about 250,000 periodicals published between 1771 and 1914.

The documents published by dLib are a diverse set in terms of layout, printing technique, paper quality, age, as well as the quality of digital reproduction. Each digitised document is presented as a PDF file with images of individual pages and text laid over them. Alongside the PDF, a text-only file is available, containing the output produced by OCR software. Original documents were scanned and processed with OCR software at different times, using varied technologies, information about which is not preserved and available. As such, a portion of dLib's resources are of a quality unsuitable for distant reading (Moretti, 2013) or corpus linguistics.

While the PDF files encode more data (i.e. page alignment and precise character coordinates) compared to accompanying text files, we noticed several inconsistencies when extracting text directly from PDF files. Character spacing is often ambiguous, and words split over lines are never joined (as they often are in raw text produced by OCR software). Consequently, text extracted from PDF files differs based on the software used to read the file itself. In light of this, we chose the text files to be our source of linguistic truth and used the PDF files to align the text to individual pages. The page alignment allowed us to link images of individual pages to the tokens in the corpus to facilitate close reading and manual verification of digitised text.

2.2. Corpus curation

To ensure the reliability and quality of the collection, we perform several steps to curate the source collection. First, we remove documents that are evidently of a quality too low to work with or predominantly in a language we are not interested in.

[1] https://dihur.si/muki/nuknec/.

[2] The Digital Library of Slovenia https://dlib.si

2.2.1. Removal of suspiciously short documents

After examining the character-size distribution of the original set of documents presented in Figure 1 and close reading of documents in several size bands, we excluded documents with fewer than 100 alphanumerical characters in total. Such documents often lack the necessary context and substantive information required for meaningful analysis. Additionally, very short documents are often based on unsuccessful optical character recognition and documents with little consecutive textual material, such as sheet music or advertisements.

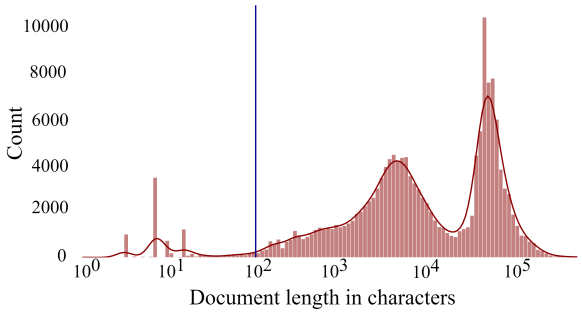

Figure 1: Documents' character-size distribution with the cut-off point of excluding document shorter than 100 characters.

2.2.2. Language filtering

Using metadata supplied by dLib which contained information on the language of documents, we filtered to select only Slovenian texts. Noticing various documents in languages other than Slovenian after this step, we further employed automatic language detection using FastText (Bojanowski et al., 2017) to remove residual documents not in Slovenian.

2.3. Filtering by quality

In order to assess the quality of the OCR-ed documents we evaluated them with a statistical language model. We trained a statistical word-level 5-gram language model (Heafield et al., 2013) on the SentiNews dataset (Bučar, 2017), and used the model to assess the quality of our documents based on model-assigned probability scores. The intuition here is that contemporary news data are generic enough to be successfully used in identifying texts that are malformed.

After training the model, we sampled 25 character sequences of 100 characters from every document. Probability scores were assigned to each sample and then the mean and standard deviation of the scores were logged for each document.

After manually inspecting random documents, we concluded that the majority of sequences with the lowest probability scores are low-quality results of OCR processing that produced garbled or otherwise unintelligible text. Many of such cases were non-running text embedded into various newspaper graphics.

We examined the score distributions and standard deviation (as depicted in Figure 2). Samples with scores between -80 and -130 were presented to historians and corpus linguists to determine the cut-off point based on the quality of the text and its usefulness and reliability for future research. Together we established the minimum text quality threshold at -100. We included all the documents with scores above the threshold into the corpus, which meant excluding under 5% of documents retained until this point.

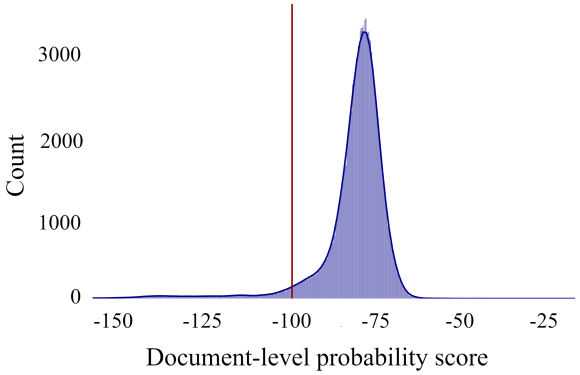

Figure 2: Documents' probability scores as reported by the language model

2.4. Correcting split words

Because of the relatively narrow columns in newsprint, many words are split across lines. A significant amount of OCR-ed files were produced without their merging. We employed a two-step approach to merge the split words. First, we processed all documents with a regular expression joining lines ending in word characters and a word-splitting dash followed by a line beginning with lowercase word characters. [3]

Many split words remained in the corpus after our regular-expression-based cleanup. In order to conservatively fix as many remaining split words as possible, we employed a simple statistical method. Counting the occurrences of all words in our prospective corpus, we checked if there were more occurrences of the "merged" word or each individual word at either end of the dash. If the occurrence of the merged word was higher, we removed the dash and merged the words.

[3]For specifics consult the code in *clean.py* at https://dihur.si/muki/nuknec/.

2.5. Page alignment

As stated in 2.1, we selected the text files without page number metadata data as our text source due to their higher quality compared to their corresponding PDF documents that inherently contain the page number. To preserve page alignment data that allows us to link text to scans of individual pages for inspection and verification, we split the text files according to pages in the PDF.

By iterating through the pages in PDF documents and comparing the text with the corresponding text file, we tried to align the text output with individual pages. Unfortunately, extracting text from PDF files often produced marginally differing outputs compared to our source text files, which prevented us from aligning the documents based on exact text matches.

To successfully align pages, we developed an algorithm based on greedy similarity maximization. The algorithm looks for the minimal character-level difference of line-by-line increments of a text file and its corresponding PDF pages. The incremental search continues as long as the difference between the PDF page text and the collection of lines keeps reducing. You can consult our implementation in the repository.

2.6. Post-OCR Error Correction

Optical Character Recognition (OCR) plays a crucial role in converting scanned documents into machine-readable text. However, OCR systems often introduce errors due to various factors such as degraded source documents, poor image quality, complex fonts, and different techniques and technology for OCR (Traub et al., 2015; Chiron et al., 2017). To address these issues, post-OCR correction techniques have been developed to improve the accuracy of the extracted text (Van Strien et al., 2020; Tong and Evans, 1996). In this section, we present our work on correcting OCR errors using cSMTiser (Ljubešić et al., 2016), a tool for text normalization via character-level machine translation.

cSMTiser works on the character level, and with that offers the capability to correct errors at an arbitrary length of characters, defaulting to word and sentence levels. In our case, we focused on word-level trigrams to strike a balance between capturing sufficient correction context and maintaining computational efficiency.

The pipeline of cSMTiser uses a supervised text normalization model and a statistical language model. For the text normalization model, we selected 300 random paragraphs from our collection with at least 100 characters. These paragraphs were carefully manually corrected by comparing the OCR-ed result to the original scans of the documents. Once annotated, they were fed as trigram

chunks to the normalization model.

For the statistical language model, we used all paragraphs with at least 100 characters from our entire collection. This helped the model to get a good sense of the Slovene language back in the second half of the 18th, 19th, and the beginning of the 20th century. To make sure that this model is not biased toward common OCR errors, we also used the Slovene ParlaMint dataset (Erjavec et al., 2021), which comprises modern parliamentary transcripts without typographical errors. The inclusion of a second dataset with modern language aimed to provide language model stability by counterbalancing the noise introduced by the historical collection.

To assess the effectiveness of our approach, we measured the performance using two widely adopted evaluation metrics: Word Error Rate (WER) and Character Error Rate (CER) (Morris et al., 2004). WER quantifies the percentage of incorrectly recognized words in the OCR-ed texts, while CER measures the percentage of incorrectly recognized characters. The manually labeled subset used for training showed for our data collection to contain a Word Error Rate (WER) of 5.4% and a Character Error Rate (CER) of 1.2% in the OCR-ed texts. These measurements represents the baseline performance that we want to be sure to improve over.

After applying cSMTiser for post-OCR correction, we observed a significant improvement in the quality of the documents. We measured the WER to go down to 4.4%, representing a 19% relative error reduction, and the CER to shrink to 1.0%, which represents a 17% relative error reduction. Furthermore, we evaluated the general quality of the documents by comparing the probability figures of non-normalised and normalised texts given our language model used in Section 2.3. The normalised texts exhibit higher probability values, demonstrating the effectiveness of our approach in reducing the error rates, and indicating a notable enhancement in the accuracy of the text extracted from original scans.

Table 1 presents examples of the post-OCR error corrections. A large majority (roughly 80%) of the corrections are suitable, with the model being able to capture the common OCR errors such as misrecognition of `s` as `a`, `e` as `c` and `l` as `i` or vice versa. Yet there are also other types of corrections that do not really improve the quality of the text, such as ambiguous corrections of punctuation and capitalization, or corrections of heavily damaged text which do not lead to complete recovery. Most importantly, cSMTiser proved to be exceptionally robust by not introducing unsuitable (wrong) corrections in already correct texts.

2.7. Linguistic annotations of the corpora

Our collection is linguistically annotated following the Universal Dependencies formalism, and further enriched with named entity annotations. The annotation was performed automatically by the CLASSLA-Stanza pipeline (Ljubešić and Dobrovoljc, 2019; Terčon and Ljubešić, 2023), a fork of the well-known Stanford Stanza pipeline (Qi et al., 2020). The reason for preferring the CLASSLA pipeline over Stanza is that CLASSLA models are based on a larger training dataset, use large inflectional lexicons, and have support for Named Entity Recognition (NER). Standard Slovenian language models achieve performance scores of 94–97% for morphosyntactic tagging, 98-99% for lemmatization, and 87%-94% for dependency parsing. It is to be expected that the CLASSLA performance on this collection is somewhat lower, due to two reasons: archaic texts and OCR mistakes. Manual inspection of the annotations has shown that the performance of linguistic annotation is satisfactory and that these annotations will serve both improved search, as well as potential corpus linguistic studies on this collection.

3. Corpus distribution

The size of the distributed corpus is presented in Table 2. It presents the corpus size in terms of tokens, sentences, and documents, but also the number of named entities by type.

In order to make our collection useful for researchers from the digital humanities and social sciences, the corpora were converted and mounted on the noSketch Engine concordancer (Kilgarriff et al., 2014), maintained by CLARIN.si, accessible at `https://www.clarin.si/ske/#concordance?corpname=speriodika`. The tool supports metadata-based subcorpus creation, configurable concordances, frequency, keyword and collocation lists, and a RESTful interface and API. It uses the Manatee (Rychlỳ, 2007) back-end, which enables complex queries over large and richly annotated corpora.

The collection can also be downloaded from the CLARIN.si repository from `http://hdl.handle.net/11356/1881`. The corpus is available in the following formats: (1) an all-data-and-metadata JSON file, (2) a collection of per-document all-metadata TSV files, (3) a collection of CoNLL-U files, which, besides morphosyntactic and lemma annotation, also include named entity annotations in IOB2 format, and (4) a vertical file as prepared for the concordancers, including the registry file, so that the corpus can be mounted on any other noSketch Engine installation.

Table 1: Examples of post-OCR error corrections and their types.

OCR-ed trigram	Corrected trigram	Correction	Type of correction
ki ae je	ki se je	a → s	suitable
vcčkrat pa 12	večkrat pa 12	c → e	suitable
bii bi na	bil bi na	i → l	suitable
Izdajatelj In odgovorni	Izdajatelj in odgovorni	I → i	suitable
nočejo oirok pustiti	nočejo otrok pustiti	i → t	suitable
Štev. 8.	štev. 8.	Š → š	ambiguous
Članov in obresti	članov in obresti	Č → č	ambiguous
Številke po 4	številke po 4	Š → š	ambiguous
Aadaiko - \fomkm	Aadaiko - fomkm	\ →	unsuccessful
Ui aa|-dovrženeHa kar	Ui aaj-dovržena kar	| → j; H → l	unsuccessful
Družtvo sv. Jožefa	Društvo sv. Jožefa	ž → š	historical difference

Table 2: Size of the final corpus in terms of number of tokens, sentences, documents and various named entities

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

For every page in the corpus, a JPG image of the scan is published and available for download at `https://nl.ijs.si/inz/speriodika/<URN>-<PAGE_NUMBER>.jpg`. The image URLs are supplied in the metadata accompanying the annotated texts, as well as linked in the concordancer interface.

Additionally, we measured language-model-based probability for each individual page applying the same sampling procedure as we did for full documents (c.f. section 2.3). After additional manual inspection, we decided to label the 25% of pages in the corpus with the lowest probability as "*low quality*" to indicate texts that are not suitable for distant reading tasks. These pages often contain a disproportionately large number of (unique) pictorial material (e.g. advertisements, tables, illustrations), which was partially processed by OCR software into (mostly) meaningless strings of characters.

4. Content Analysis

We utilised topic modeling and named entity analysis to provide an overview of the released collection. This approach allowed us to identify prominent themes and important entities, giving us a bird's-eye view of the dataset. We provide a preliminary insight into the diverse content as we set the stage for a more comprehensive understanding of the collection's composition and facilitate deeper analysis of its contents.

4.1. Topics

We apply an LDA-based topic modeling using the Mallet toolkit (McCallum, 2002). We use default hyper-parameters and detect ten topics on a random sample of 10k documents from our corpus. Table 3 shows a list of the most distinguishing keywords for each topic.

4.2. Named Entity Analysis

While named entity tagging is not perfect, especially when dealing with historical texts containing entities and formulations underrepresented in contemporary training sets it performs well enough to make comparative statements about the corpus. Since we know that the recall of our named entity recognition is not perfect, we can not claim exactly how many times a given entity is mentioned in the corpus, but we can use the results to draw relative inferences and compare entities between themselves. For an initial overview analysis, we present the top locations and people annotated in the corpus.

Figure 4 presents the top most common locations in and around Slovenia. It shows Ljubljana with far the most mentions, which can be attributed to addresses of presses and publishers located there. Vienna comes second, followed by Trieste and then an array of towns and cities across territories culturally relevant to Slovenia's history. While the corpus is biased towards centres of power it mentions a diverse set of locations of administrative, economic, and cultural importance.

Figure 3 lists the most common named entities labeled as persons. The list was obtained by listing the top 50 named person entities and then disambiguating and joining duplicates by hand. We can observe that the most common persons named entities are names for nationalities (Slovenian, Ger-

Table 3: Detected Topics with LDA

Topic ID	Keywords	Description	Documents
1	cerkev, človek, beseda, otrok, gospod, življenje, mesto, ljudstvo, ljubezen, molitev, dežela, papež, resnica, zemlja, veselje, podoba, pismo, postava, pravica	Religious Life	901
2	društvo, cerkev, okraj, jezik, prilika, general, pravica, glava, szvoj, otrok, navada, morje, duhovnik, narod, iorica, barka, nptranslation, i, stera	Social Organizations	242
3	učitelj, knjiga, društvo, narod, jezik, pisatelj, učiteljstvo, beseda, otrok, mesto, učenec, življenje, pesem, vrsta, mladina, slika, zgodovina, profesor, zvezek	Education and Literature	1491
4	vlada, narod, poslanec, gospod, mesto, odbor, društvo, dežela, volitev, jezik, stranka, predlog, minister, pravica, država, zbornica, cesar, beseda, predsednik	Politics and Governance	2357
5	društvo, stranka, odbor, poslanec, gospod, mesto, volitev, vlada, krona, zveza, človek, delavec, nedelja, družba, ljudstvo, narod, cesta, občina, predsednik	Social and Political Organizations	1877
6	človek, gospod, glava, življenje, otrok, mesto, beseda, zemlja, stran, obraz, misel, sreča, vrata, ljubezen, prijatelj, konec, cesta, pesem, morje	Human Experience	1297
7	ulica, zaloga, društvo, blago, vrsta, edinost, občinstvo, obleka, tovarna, lekarna, trgovina, oglas, stvar, krona, narod, mesto, prodaja, jezik, predmet	Commerce and Industry	428
8	sodišče, zakon, pravica, sodnik, pravo, slučaj, tožba, razsodba, razlog, strošek, priča, patron, pogodba, določba, toženec, razprava, postopanje, znesek, sklep	Law and Judiciary	211
9	posestnik, živina, zadruga, zemlja, družba, mleko, vrsta, drevo, krava, bolezen, žival, rastlina, sadje, odbor, posojilnica, drevje, pridelek, kmetovalec, hrana	Agriculture and Farming	861
10	mesto, vlada, vojak, vojna, država, vojska, mesec, armada, cesar, kralj, avgust, stran, poročilo, oktober, ladja, častnik, general, narod, september	Government and Warfare	334

man, Slav, Croatian, Italian...) and religious figures (God, Jesus, Christ, Mary...) followed by political nouns and last names of prominent politicians and common first names (emperor, Hribar, Tavčar...).

Figure 3: Word cloud of the most common personal named entities in the corpus.

5. Conclusion

The lightweight approach to building a giga-token corpus of historical periodicals outlined in this paper provides a cost-effective and efficient means for generating giga-token-sized corpora from historical texts, derived from scans and OCR outputs, within a few months, as opposed to a conventional timeline of a few years. The presented techniques are especially applicable to under-resourced languages and make pragmatic trade-offs that ensure the creation of high-quality corpora from digitised sources of uncertain or variable quality. Through the key stages of alignment, filtering, and post-OCR error correction, we ensure the viability of automatic linguistic annotations in the realm of historical research, even when working with digitised materials way below today's contemporary technological standards.

Using the proposed approach, we prepared and published a linguistically annotated corpus of Slovenian periodicals of almost a billion words and more

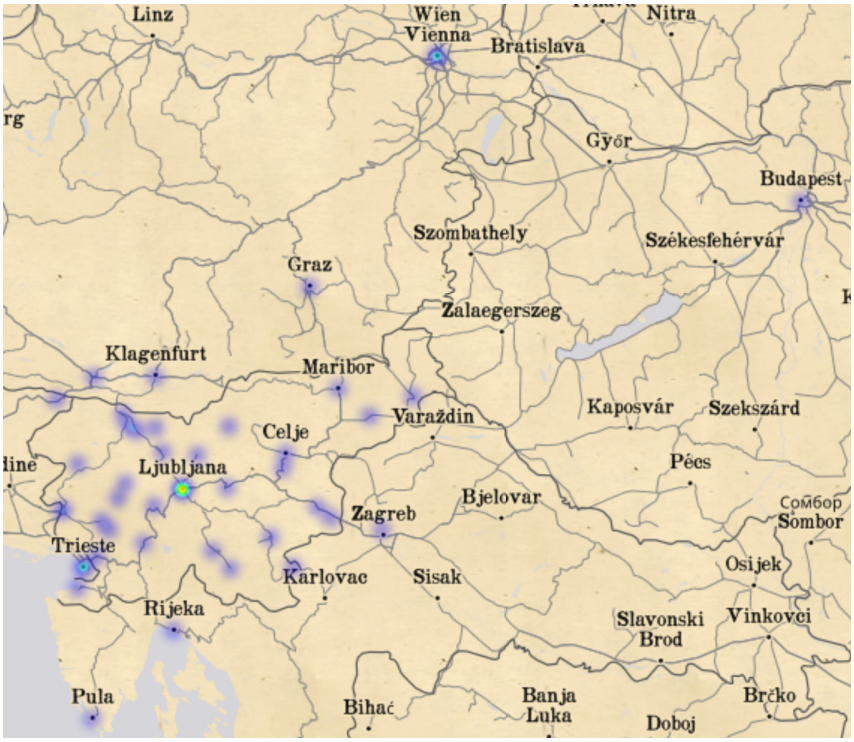

Figure 4: Geographic heat map of Slovenia and neighboring regions. It marks the most common locations detected in the corpus.

than 150k documents. A bird's-eye analysis of the published corpus confirms the robustness of our pipeline and its suitability for preparing corpora for distant reading. We present an overview of the corpus's topic distribution and its contents concerning named entities. As anticipated, both named persons and locations are locally relevant and point to a vibrant and diverse historical life covered in periodical publications of the time.

While the presented corpus construction methodology has many strong points, there are also potential limitations of such an approach. With removing a small portion of the content, we surely add to the overall high bias regarding the spotty availability of data in historical text collections. Errors in the digital version of the text as well as in its enrichment with linguistic and named entity annotation are also not kept at their necessary minimum, which surely includes additional potential caveats in the proper usage of the data collection for answering research questions. Further downstream usage of the corpus will have to investigate potential limitations in the usefulness of the presented resource and construction methodology that go beyond limitations already present in collections that were built in a more traditional way. Our expectation is, however, that there is minimum additional bias present in the data, as already depicted in the successful and useful bird's-eye analyses of the newly constructed dataset.

6. Acknowledgements

The work described in this paper was funded by the Slovenian Research and Innovation Agency (ARIS) research programmes P6-0436: "Digital Humanities: resources, tools, and methods", P6-0411: "Language resources and technologies for Slovene", the research project J7-4642: "Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language", and the DARIAH-SI and CLARIN.SI research infrastructures.

7. Bibliographical References

Ruth Ahnert, Emma Griffin, Mia Ridge, and Giorgia Tolfo. Collaborative historical research in the age of big data: Lessons from an interdisciplinary project. ISBN: 9781009175548 9781009175555 Publisher: Cambridge University Press.

Austrian National Library. retrieved 2023. ANNO - AustriaN newspapers online | ANNO - AustriaN newspapers online. https://anno.onb.ac.at/.

Kaspar Beelen, Jon Lawrence, Daniel C S Wilson, and David Beavan. 2022. Bias and representativeness in digitized newspaper collections: Introducing the environmental scan. *Digital Scholarship in the Humanities*, 38(1):1–22.

BnF. retrieved 2023. Gallica – the BnF digital library. https://www.bnf.fr/en/gallica-bnf-digital-library.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Jože Bučar. 2017. Manually sentiment annotated slovenian news corpus SentiNews 1.0. Slovenian language resource repository CLARIN.SI.

Arindam Chaudhuri, Krupa Mandaviya, Pratixa Badelia, Soumya K Ghosh, Arindam Chaudhuri, Krupa Mandaviya, Pratixa Badelia, and Soumya K Ghosh. 2017. *Optical character recognition systems*. Springer.

Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of OCR errors on the use of digital libraries: towards a better access to information. In *2017 ACM/IEEE joint conference on digital libraries (JCDL)*, pages 1–4. IEEE.

Ryan Cordell, M. H. Beals, Isabel Galina Russell, Julianne Nyhan, Ernesto Priani, Marc Priewe, Hannu Salmi, Jaap Verheul, Raquel Alegre, and Tessa Hauswedell. 2017. Oceanic exchanges.

Ryan Cordell and David Smith. 2022. Viral texts: Mapping networks of reprinting in 19th-century newspapers and magazines.

Deutsche Digitale Bibliothek. retrieved 2023. Deutsche Digitale Bibliothek - Kultur und Wissen online. https://www.deutsche-digitale-bibliothek.de/.

Digitalna knjižnica Slovenije dLib. dLib.si - periodika.

Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Ströbel, and Raphaël Barman. 2020. Language resources for historical newspapers: the impresso collection. In *Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, 2020. European Language Resources Association (ELRA), 958-968.*, pages 958–968. European Language Resources Association (ELRA).

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Vladislava Grigorova, Michał Rudolf, Andrej Pančur, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Henk van der Pol, Griet Depoorter, Jesse de Does, Bart Jongejan, Dorte Haltrup Hansen, Costanza Navarretta, María Calzada Pérez, Luciana D. de Macedo, Ruben van Heusden, Maarten Marx, Çağrı Çöltekin, Matthew Coole, Tommaso Agnoloni, Francesca Frontini, Simonetta Montemagni, Valeria Quochi, Giulia Venturi, Manuela Ruisi, Carlo Marchetti, Roberto Battistoni, Miklós Sebők, Orsolya Ring, Roberts Darģis, Andrius Utka, Mindaugas Petkevičius, Monika Briedienė, Tomas Krilavičius, Vaidas Morkevičius, Sascha Diwersy, Giancarlo Luxardo, and Paul Rayson. 2021. Multilingual comparable corpora of parliamentary debates ParlaMint 2.1. Slovenian language resource repository CLARIN.SI.

Karez Hamad and Kaya Mehmet. 2016. A detailed analysis of optical character recognition technology. *International Journal of Applied Mathematics Electronics and Computers*, (Special Issue-1):244–249.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. 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*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Nikola Ljubešić, Katja Zupan, Darja Fišer, and Tomaž Erjavec. 2016. Normalising Slovene data: historical texts vs. user-generated content. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 146–155.

Jani Marjanen, Elaine Zosa, Simon Hengchen, Lidia Pivovarova, and Mikko Tolonen. 2020. Topic modelling discourse dynamics in historical newspapers. *arXiv preprint arXiv:2011.10428*.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. *http://mallet.cs. umass. edu*.

Franco Moretti. 2013. *Distant Reading*. Verso Books.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER

and WIL: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

Nacionalna i sveučilišna knjižnica u Zagrebu. retrieved 2023. Digitalizirane novine i časopisi. http://dnc.nsk.hr.

Eva Pfanzelter, Sarah Oberbichler, Jani Marjanen, Pierre-Carl Langlais, and Stefan Hechl. 2021. Digital interfaces of historical newspapers: opportunities, restrictions and recommendations. *Journal of Data Mining & Digital Humanities*, (HistoInformatics).

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages.

Pavel Rychlỳ. 2007. Manatee/bonito-a modular corpus manager. In *RASLAN*, pages 65–70.

Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin. 2012. A survey of OCR applications. *International Journal of Machine Learning and Computing*, 2(3):314.

Luka Terčon and Nikola Ljubešić. 2023. CLASSLA-Stanza: The next step for linguistic processing of south slavic languages.

Xiang Tong and David A Evans. 1996. A statistical approach to automatic OCR error correction in context. In *Fourth workshop on very large corpora*.

Myriam C Traub, Jacco Van Ossenbruggen, and Lynda Hardman. 2015. Impact analysis of OCR quality on research tasks in digital archives. In *Research and Advanced Technology for Digital Libraries: 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015, Proceedings 19*, pages 252–263. Springer.

Daniel Van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the impact of ocr quality on downstream nlp tasks.