PressMint: Towards Interoperable Corpora of Historical Newspapers

Tomaž Erjavec[1], **Matyáš Kopp**[2], **Maciej Ogrodniczuk**[3], **Petya Osenova**[4], **German Rigau**[5]

[1] Department of Knowledge Technologies, Jožef Stefan Institute
[2] Charles University [3] Institute of Computer Science, Polish Academy of Sciences
[4] Sofia University "St. Kl. Ohridski" and IICT-BAS
[5] UPV/EHU

tomaz.erjavec@ijs.si, kopp@ufal.mff.cuni.cz, maciej.ogrodniczuk@ipipan.waw.pl, petya@bultreebank.org, german.rigau@ehu.eus

Abstract

This paper presents PressMint, an ongoing initiative to compile a multilingual, comparable, annotated, translated, and interoperable collection of European historical newspaper corpora. Spanning 17 countries and covering 15 languages, the project addresses a key shortcoming of existing newspaper resources: their lack of interoperability, which limits cross-lingual and transnational research. Building on the infrastructure and experience of the ParlaMint projects, the project adapts established encoding guidelines, validation workflows, and open-source tools to historical newspaper data. We outline the overall project architecture, the corpus encoding scheme, and the GitHub-based framework supporting collaborative development and quality control. The paper further describes the sample linguistic annotation pipeline, including OCR correction, text normalisation, and annotation within the Universal Dependencies framework, with attention to challenges posed by historical language varieties. The resulting FAIR, openly available corpora are intended to support comparative, diachronic research across the humanities and social sciences.

1. Introduction

Historical newspapers are of interest to historians and historical linguists, as well as social scientists, ethnologists, anthropologists, media and communication scholars, and cultural studies scholars. All of these are fields where contemporary digital resources, tools and methods (e.g. ”distant reading”) are still underutilised. On the other hand, corpora of historical newspapers already exist for a number of languages and countries (Fišer et al., 2018; Fišer and Lenardič, 2018; Walcher et al., 2023) to a large extent, as they are out of copyright, and the images, and often OCR, are available via national libraries. However, these corpora are not interoperable, which precludes methods for their comparison, as well as any translingual and transnational research, an especially important consideration, as statehood and nationhood are highly dynamic in Europe in the period to be covered by the project corpora.

PressMint aims to improve this situation by compiling a multilingual, comparable, annotated, translated and interoperable set of corpora of European historical newspapers, centered around the start of the 20th century. The corpora will be openly available, both for download in a variety of instances and formats, as well as via several on-line corpus analysis tools. The project will proactively disseminate and foster the use of the corpus collection.

The project heavily relies on leveraging the infrastructure and experiences reported by the two ParlaMint projects (Erjavec et al., 2023, 2025) which compiled interoperable, multinational and multilingual corpora of parliamentary debates. While the two text types are not identical, they are similar enough for ParlaMint to be adapted to PressMint.

2. The Project Infrastructure

2.1. GitHub-based Framework

PressMint adapts the building blocks of the ParlaMint infrastructure. The encoding guidelines and schema, the validation, conversion, extraction and enrichment scripts, and the functionality of GitHub, in particular for version management of all the documentation and scripts, with corpus samples, use of issues for reporting problems and requests, GitHub pages for displaying the documentation (in particular encoding guidelines), and GitHub actions for automatic validation of samples on commit. This will significantly lower the cost of the project setup and technical coordination and reuse the ParlaMint framework, still being developed in the ParlaCAP project[1] (Ljubešić et al., 2025), which may also lead to infrastructural synergies between ParlaCAP and PressMint.

[1] https://clarinsi.github.io/parlacap/

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 1–5

2.2. Encoding Guidelines and Schema

Like ParlaMint, PressMint also uses the Text Encoding Initiative (TEI) Guidelines (TEI Consortium, eds, 2024) for encoding, but parameterised for newspaper data rather than parliamentary debates. Here, we built on previous work connected with historical corpora, in particular the IMP language resources of historical Slovene (Erjavec, 2015) which, inter alia, contain a corpus with hand-corrected transcriptions, which is linguistically analysed, page-aligned with the facsimile and TEI encoded.

To date, the initial TEI ODD (One Document Does it all) has been written; the ODD customises TEI for a particular project or purpose and should also contain the prose annotation guidelines, while the element and attribute specifications are accompanied by explanatory prose and examples. While the schema customisation is fairly simple for newspaper data, the main effort was invested into the prose part, i.e. the annotation guidelines, which give detailed explanations and examples of the overall corpus structure and metadata, the metadata annotation of corpus components (i.e. individual texts), and the annotation of the texts themselves. As in ParlaMint, we distinguish two variants of the corpora, the so-called ”plain-text” version, with all the metadata and structural annotations and running text, and the linguistically annotated version, which adds automatic linguistic annotations to the plain-text version.

While the schema and guidelines are fully functional, we do envision further changes when the partners' source corpora are analysed and their various metadata and encodings preserved in the project's schema.

2.3. Linguistic Annotation

For linguistic annotation, UDPipe (Straková et al., 2019) will be used. It is a maintained, trainable pipeline for tokenisation, tagging, lemmatisation and dependency parsing for which models for all European languages already exist, and which have already been extensively used in ParlaMint. By default, the named entity tagging will be done by NameTag (Straková and Straka, 2025), which covers a number of European languages, although not all of them. For the missing languages, we will train NameTag to generate models from them as well. It should be noted that the tools will most likely have lower performance on historical texts than they do on contemporary ones. To somewhat mitigate this, we also plan to support word modernisation for languages where the language of the older newspapers differs significantly from the contemporary standard. Modernisation allows the use of linguistic annotation tools that have been trained on contemporary texts and facilitates searching the corpora. To this end, we plan to use the open-source cSMTiser (Ljubešić et al., 2016), a trainable tool for word normalisation, which has been successfully applied to a number of different normalisation scenarios. Text classification according to topic, currently developed in ParlaCAP, with existing models for newspaper texts already available (Kuzman and Ljubešić, 2025), will be used as well.

3. The Source Corpora

The size, time-span and type of newspapers covered differs across the languages, although the main intention is to cover newspapers from around the turn of the 20th century. Table 1 provides information about the corpora which are planned to be included in the project resource set. As can be seen, some partners still have to determine which source(s) they will use for preparing their corpora. This is mainly due to complications in determining accessibility of the sources and evaluating their encoding. Note also that the table describes the sizes and time-spans of the sources, which might — and typically will — differ from those of the corpora of the described project, as partners can filter the corpora to exclude documents that are too old or of bad quality.

4. Corpus Encoding Procedure

This section describes the exemplar procedure that will be employed in the corpus development. Currently, one corpus was processed to test the schema, conversion and validation scripts, namely the Slovenian one.

The Slovenian source data is the sPeriodika corpus (Dobranić et al., 2023, 2024) of historical Slovenian periodicals from the period 1771–1914. sPeriodika consists of OCR-processed PDF and TXT files obtained directly from the Digital Library of Slovenia (dLib[7]), a service of the National and University Library. Each document typically corresponds to a single issue of a periodical or, where available, to an individual article. No manual segmentation into articles was performed beyond what was explicitly present in the source metadata. The raw texts were lightly processed to correct some OCR errors and join end-of-line hyphenated words and were then linguistically annotated with the CLASSLA pipeline (Ljubešić et al., 2024).

The corpus comprises approximately 150,000 texts and 910 million tokens. For the current project, the data were filtered to exclude documents of non-newspaper genres such as yearbooks or magazines, discard texts published before 1850, and

[7] https://dlib.si/

Table 1: Data sources by country

Country	Data source	Data size	Dates
AT	The newspaper *Wiener Abendpost*, a supplement of the *Wiener Zeitung*, provided by the Austrian National Library[2]	TBD	1863–1921
BG	A Collection of Bulgarian newspapers from the Digital Library[3] of National Library Ivan Vazov – Plovdiv	> 100 issues	1863–1944
CZ	Several major Czech periodicals will be selected from the Digital Library[4].	1.2G words, 420k pages	1848–1915
ES	Several multilingual regional corpora based on periodicals dating from the beginning of the Restoration to the end of the Second Spanish Republic that will be selected from various local and national repositories.	TBD	1874–1936
FI	The believed-out-of-copyright Swedish and Finnish pages of the Newspaper and Periodical corpora of the National Library of Finland, OCR by the Library processed by the Language Bank of Finland	800M tokens	1771–1874/1879
FR	The daily newspaper *Le Temps* from the national library of France, for a period to be defined, presumably from 1900 to 1942	TBD	1900–1942
GR	Various newspapers and periodicals from the digital collections of the Library of the Greek Parliament (OCR and plain text)	TBD	1880–1920
HU	A set of Hungarian newspapers and periodicals (e.g., *Pesti Hírlap*, *Pesti Napló*, plus local press such as *Buda és vidéke* and *Magyar Székesfőváros*) will be collected from Arcanum and Hungaricana repositories	TBD	1880–1930
IS	*MC-19: A Corpus of 19th Century Icelandic Texts* (Steingrímsson et al., 2025)	270M tokens	1800–1929
IT	*Excerpts from the Zeit.shift data*[5] *(Walcher et al., 2023) (> 80 newspapers and magazines from the historical region of Tyrol with a focus on the late 19th and early 20th centuries).*	max. 200k pages from ∼20 newspapers	1890–1935
LV	Digitized historical newspaper “Jaunākās ziņas”	TBD	1911–1940
NL	*Couranten Corpus* (version 2.0) (van der Sijs, 2025)	18M tokens	1618–1700
PL	*Microcorpus of Nineteenth-Century Polish* (Bilińska et al., 2018)	300k tokens	1830–1918
	The Interwar Polish Press Corpus	8M tokens	1918–1939
PT	*The Reference Corpus of Contemporary Portuguese* – subcorpus of newspapers from the late 19th to early 20th centuries	TBD	1808–1940
SI	Subset of *Corpus of Slovenian periodicals sPeriodika* (Dobranić et al., 2023, 2024)	910M tokens	1771–1914
UA	Western Ukrainian press (Austro-Hungarian Empire, Poland, Czechoslovakia) from GRAC[6]	10M tokens	1888–1939
	Central-Eastern Ukrainian press (Russian Empire, Ukrainian SSR) from GRAC	10M tokens	1905–1939
UK	A representative selection of articles from the British Newspaper Archive	TBD	1900–1910
ZA	TBD	TBD	1835–1960

those with large estimated OCR noise, leaving us with 84,000 texts with 620 million tokens.

The sPeriodika corpus is distributed in JSON format, with each file representing a document. At the document level, the encoding captures bibliographic and provenance metadata, such as the persistent document identifier (URN), the periodical name and publisher, publication date and year, etc. as well as the correction rate produced by the OCR post-correction tool. The documents are internally structured into pages, each corresponding to a physical page in the original publication. Page-level metadata include page indices, alignment ratios between original OCR output and corrected text, and URLs to page images where available.

Within each page, the transcription is stored in multiple parallel representations, reflecting successive stages of text normalisation and correction. These representations allow both reproducibility of the preprocessing pipeline and selective use of text variants for downstream tasks. The per-page OCR correction quality is also quantitatively assessed using KenLM perplexity scores (mean and standard deviation), which are stored as part of the encoding and were used for assigning the OCR quality to pages, enabling quality-based filtering of the corpus.

The filtered corpus texts were converted from the

[2] https://anno.onb.ac.at/
[3] https://digital.libplovdiv.com/
[4] https://www.digitalniknihovna.cz/
[5] https://zeitshift.eu

source JSON to the project's TEI-based schema with a Perl program; this stage will obviously be corpus-dependent, and most likely developed separately for each corpus, as it depends on the source corpus format and annotations.

Once the corpus was encoded according to the project's schema, we modified the ParlaMint scripts to:

- validate the corpus
- add common and redundant metadata to the corpus (note that the encoding guidelines make explicit which parts of the corpus need to be prepared by the partners and which are automatically added)
- convert the corpus to down-stream encodings, in particular:
  - plain-text format
  - CoNLL-U format
  - vertical format (for concordancers)
  - TSV format with full metadata on individual documents

5. Conclusions

The paper has introduced the first and planned steps in producing a FAIR set of comparable historical newspaper corpora, along with the infrastructure to validate and convert them to down-stream encodings. On this basis, we will also make the corpora available on several on-line tools, which will enable their analysis by SSH scholars.

We aim for an inclusive set of corpora. Although it would be preferable to have a common time span for all the corpora, it turns out that this is not possible given the materials available to the partners. Still, subsets of corpora will overlap, meaning that comparative time-dependent analysis will still be possible, just not with all the corpora. We also do not limit either the minimal nor maximal size of each individual corpus, as we do not want to exclude large and hence maximally usable corpora nor exclude the partners that can currently provide only small amounts of data, as they can gain expertise in the project that will enable them to extend the corpora in the future. The same reasoning applies as regards the metadata included in the corpus: we require only basic metadata consisting of the newspaper name, year of publication, language, and the source of the newspaper (e.g. national library), possibly with its PID/URL giving further metadata. But where available, we will also include further metadata, such as day of publication, publisher, scope, print run etc. and other metadata points contained in the sources. Finally, the actual content of the corpora can be simply the automatically OCR-ed text, but we will cater also for inclusion of facsimiles, per-page alignment of facsimiles, improved transcription, and structural encoding, in particular division into individual articles.

A very important result of the project will also be the expertise gained by the community of partners in the project, related to the processing of historical texts in general and historical newspapers in particular.

We believe that the produced corpora and on-line analysis tools can be used for teaching history not only at universities but also at secondary schools. The encoding schema is also appropriate for encoding contemporary newspapers, and could be in the future used for encoding these as well, thus allowing the study of contemporary times.

6. Acknowledgments

The submission was supported by (1) the PressMint CLARIN Flagship Project, (2) part of the investment: CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2024-2026) funded by the Polish Ministry of Science and Higher Education (Programme: "Support for the participation of Polish scientific teams in international research infrastructure projects"), agreement number 2024/WK/01 and (3) CLARIN-PL, the European Regional Development Fund, FENG programme, agreement number FENG.02.04-IP.040004/24.

7. Bibliographical References

Joanna Bilińska, Monika Kwiecień, and Magdalena Derwojedowa. 2018. Microcorpus of nineteenth-century Polish. In Eric Fuß, Marek Konopka, Beata Trawiński, and Ulrich H. Waßner, editors, *Grammar and Corpora 2016*, pages 377–387. Heidelberg University Publishing.

Filip Dobranić, Bojan Evkoski, and Nikola Ljubešić. 2023. *Corpus of Slovenian periodicals (1771-1914) sPeriodika 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1881.

Filip Dobranić, Bojan Evkoski, and Nikola Ljubešić. 2024. A lightweight approach to a giga-corpus of historical periodicals: The story of a Slovenian historical newspaper collection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 695–703, Torino, Italia. ELRA and ICCL.

Tomaž Erjavec. 2015. The IMP historical Slovene language resources. *Language Resources and Evaluation*, 49(3):753–775.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, Jure Skubic, Peter Rupnik, Tommaso Agnoloni, José Aires, Starkaður Barkarson, Roberto Bartolini, Núria Bel, María Calzada Pérez, Roberts Darģis, Sascha Diwersy, Maria Gavriilidou, Ruben van Heusden, Mikel Iruskieta, Neeme Kahusk, Anna Kryvenko, Noémi Ligeti-Nagy, Carmen Magariños, Martin Mölder, Costanza Navarretta, Kiril Simov, Lars Magne Tungland, Jouni Tuominen, John Vidler, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, and Darja Fišer. 2025. ParlaMint II: Advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*, 59:2071–2102.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2023. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 58:415–448.

Darja Fišer, Jakob Lenardič, and Tomaž Erjavec. 2018. CLARIN's key resource families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1320–1325, Miyazaki, Japan. European Language Resources Association (ELRA).

Darja Fišer and Jakob Lenardič. 2018. CLARIN Resources Families / Newspaper Corpora.

Taja Kuzman and Nikola Ljubešić. 2025. LLM teacher-student framework for text classification with no manually annotated data: A case study in IPTC news topic classification. *IEEE Access*, 13:35621–35633.

Nikola Ljubešić, Luka Terčon, and Kaja Dobrovoljc. 2024. CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. In *Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, pages 251–274, Ljubljana, Slovenia. Institute of Contemporary History.

Nikola Ljubešić, Taja Kuzman Pungeršek, and Daniela Širinić. 2025. ParlaCAP: Comparing agenda-setting across parliaments via the ParlaMint dataset. In *Proceedings of the Annual Conference of the Comparative Agendas Project (CAP)*.

Nikola Ljubešić, Katja Zupan, Darja Fišer, and Tomaz Erjavec. 2016. Normalising slovene data: historical texts vs. user-generated content. In *Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016, Bochum, Germany, September 19-21, 2016*, volume 16 of *Bochumer Linguistische Arbeitsberichte*.

Steinþór Steingrímsson, Einar Freyr Sigurðsson, and Atli Jasonarson. 2025. MC-19: a corpus of 19th century Icelandic texts. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 680–687, Tallinn, Estonia. University of Tartu Library.

Jana Straková and Milan Straka. 2025. NameTag 3: A tool and a service for multilingual/multitagset NER. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–39, Vienna, Austria. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajič. 2019. Neural architectures for nested NER through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. Association for Computational Linguistics.

TEI Consortium, eds. 2024. Guidelines for Electronic Text Encoding and Interchange. `http://www.tei-c.org/P5/`.

Nicoline van der Sijs. 2025. Couranten corpus (version 2.0). [Online Service]. Available at the Dutch Language Institute: `https://hdl.handle.net/10032/tm-a3-c2`.

Johanna Walcher, Andrea Abel, Johannes Andresen, Paolo Brasolin, Isabella Dissertori, Eva Eberwein, Greta Franzini, Silvia Gstrein, Horwath Maritta, Christian Kössler, Barbara Laner, Verena Lyding, Karin Pircher, and Egon Stemle. 2023. On a digital journey into yesterday's future: Zeit.shift – preserving Tyrol's cultural text heritage. In *Proceedings of Austrian Citizen Science Conference 2022 (ACSC 2022)*, volume 407. Sissa Medialab.