LREC 2026

First Workshop on Creating Interoperable Corpora of Historical Newspapers

Workshop Proceedings

Editors
Maciej Ogrodniczuk, Petya Osenova and Tanja Wissik

May 16, 2026

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026)

ISBN 978-2-493814-71-5

Preface

Welcome to the 1st Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint), held on May 16, 2026, in Palma de Mallorca, Spain, as part of the 15th biennial Language Resources and Evaluation Conference (LREC 2026).

While historical newspapers are a goldmine for researchers in linguistics, history, and social sciences, they often suffer from a lack of interoperability. This workshop represents a milestone for the PressMint project – a CLARIN flagship initiative – which aims to build a multilingual, TEI-encoded, and standardized set of corpora covering European press from around the start of the 20th century. By bringing together over 20 partners across Europe, the project aims to bridge the gap between isolated national archives and a unified, transnational research infrastructure.

For this inaugural edition, we received 17 submissions. Following a rigorous double-blind peer review process, the program committee decided to accept 14 papers, resulting in an 82% acceptance rate. These contributions reflect a high standard of scholarship, with an average score of approximately 3.81/5 across the accepted submissions. Three papers were presented orally, including the project overview paper, one paper representing data modelling and another one discussing data processing. 11 papers were presented as posters.

The program of the workshop highlights the geographic and thematic breadth of the PressMint community. The authors represent a diverse array of countries, including Austria, Bulgaria, Denmark, Germany, Hungary, Italy, Poland, Portugal, Slovenia, Spain, Switzerland and Ukraine. The presented papers explore critical dimensions of historical data processing, including corpus construction, interoperability and standards, technical challenges and user perspectives.

We are especially honored to welcome our invited speaker, Maud Ehrmann from Ecole Polytechnique Fédérale de Lausanne (EPFL), whose pioneering work in the digital humanities provides a vital context for our efforts in newspaper processing.

The success of this workshop is due to the hard work of many. We are deeply grateful to our Program Committee for their detailed and constructive reviews, which ensured the high quality of these proceedings. We also thank CLARIN ERIC and the PressMint project for their ongoing support. Finally, we would like to thank the workshop participants for joining us.

We hope these proceedings serve as a valuable resource for the community and inspire further collaboration in the preservation and analysis of our shared European history.

We hope you enjoyed the workshop and Mallorca as much as we did!

— Maciej Ogrodniczuk, Petya Osenova and Tanja Wissik

Organizers and Reviewers

Organizing Committee

Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences, PL

Tanja Wissik, Austrian Academy of Sciences, AT

Petya Osenova, Sofia University ”St. Kl. Ohridski” & IICT-BAS, BG

Program Committee

Emanuela Boros, University of La Rochelle, FR

Jesse de Does, Instituut voor de Nederlandse Taal, Leiden, NL

Maud Ehrmann, École Polytechnique Fédérale de Lausanne, CH

Tomaž Erjavec, Jožef Stefan Institute, SI

Aritz Farwell, HiTZ Basque Research Center for Language Technology, EHU, ES

Maria Gavriilidou, Institute for Language and Speech Processing, Athena Research Center, GR

Normunds Grūzītis, University of Latvia, LV

Maarten Janssen, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. Charles University, CZ

Matyáš Kopp, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. Charles University, CZ

Noémi Ligeti-Nagy, Hungarian Research Centre for Linguistics, HU

Nikola Ljubešic, Jožef Stefan Institute, SI

Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences, PL

Petya Osenova, Sofia University ”St. Kl. Ohridski” and IICT-BAS, BG

Adam Pawłowski, University of Wrocław, PL

Stelios Piperidis, Athena Research Centre, GR

Claudia Resch, Austrian Academy of Sciences, AT

German Rigau, HiTZ Basque Research Center for Language Technology, EHU, ES

Maria Shvedova, National Technical University "Kharkiv Polytechnic Institute" UA / University of Jena, DE

Inguna Skadiņa, Institute of Mathematics and Computer Science, University of Latvia, LV

Steinþór Steingrímsson, The Árni Magnússon Institute for Icelandic Studies, IS

Egon W. Stemle, Institute for Applied Linguistics, Eurac Research, IT

Tanja Wissik, Austrian Academy of Sciences, AT

Invited Talk

Beyond Borders: Connecting Historical Media at Scale with Impresso

Maud Ehrmann

Abstract

Mass digitisation has transformed historical research by producing vast machine-readable corpora – notably historical newspapers and broadcasts – making full-text search and text mining-based enrichments standard tools for retrieval and exploration. Yet digitised collections remain fragmented, siloed by media type, language, institution, and national boundary, and the transformative potential of studying them together, at scale and across borders, remains unrealised.

Overcoming this fragmentation requires more than aggregation. Connectivity is the core ambition of *Impresso — Media Monitoring of the Past*: not simply consolidating collections, but semantically enriching and linking them, and ensuring legally grounded access through interfaces designed for historical research.

This talk presents Impresso's strategy for connecting historical media collections from over twenty European cultural heritage institutions along four interdependent dimensions. First, data governance: cross-border access to copyright-protected collections demands frameworks that balance institutional control with researcher needs, amid concerns around AI and data reuse. Second, data interoperability: processing and indexing diverse collections at scale requires reconciling heterogeneity in format, levels of granularity and quality, and archival structures. Third, data processing: connecting information requires semantic enrichment and linking across languages and modalities within a multilingual embedding space. Fourth, interface design: transnational and transmedia discovery requires versatile research affordances – from human-centred exploration to fully programmatic, data-driven inquiry – as provided by the Impresso WebApp and Datalab.

Together, these efforts outline a model for connecting historical media responsibly at scale, ensuring that access extends beyond institutional borders, semantic enrichment spans languages and modalities, and scholarly inquiry is supported by versatile interfaces.

Speaker Bio

Maud Ehrmann is a research scientist and lecturer at the Digital Humanities Laboratory (DHLAB) of the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Her research sits at the intersection of natural language processing, digital humanities, and historical document processing, with a focus on named entity recognition and linking, semantic enrichment, and the large-scale processing of historical media collections. She works on Impresso, a series of two interdisciplinary research projects that uses machine learning to advance the processing, semantic enrichment, exploration, and study of historical media across modalities, time, languages, and national borders. She has co-initiated and co-organises the series of HIPE shared tasks on historical document processing, originally centered on named entity processing, and now extending to relation extraction and OCR post-correction.

Table of Contents

Workshop Program

Saturday, May 16, 2026

Opening Remarks

09:00–09:05 *Opening and Welcome*
Maciej Ogrodniczuk, Petya Osenova and Tanja Wissik

Invited Talk

09:05–10:00 *Beyond Borders: Connecting Historical Media at Scale with Impresso*
Maud Ehrmann

Project Overview

10:00–10:30 *PressMint: Towards Interoperable Corpora of Historical Newspapers*
Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova and German Rigau

10:30–11:00 *Coffee Break*

11:00–12:00 **Poster Session**

The Polish PressMint Corpus
Maciej Ogrodniczuk, Dariusz Czerski and Adam Pawłowski

PressMint QuickCheck: Operationalising Readiness Diagnostics for Interoperable Historical Newspaper Corpora
Elena Battaner Moro, Almudena Caballos Villar, María Cuevas Riaño, Marina Miguez Lamanuzzi and Dolores Romero López

PressMint-PT — Compiling a Portuguese Historical Newspaper Corpus
Jose Aires and Amalia Mendes

Historical Newspapers in the General Regionally Annotated Corpus of Ukrainian (GRAC): Current State and PressMint Integration Prospects
Maria Shvedova and Arsenii Lukashevskyi

CLARIAH-ES PressMint: Building Interoperable Corpora of Historical Press in Spain
Ainara Estarrona, Aritz Farwell, German Rigau and Xabier Goenaga

Towards an Interoperable Corpus of Austrian Historical Newspapers: The case of PressMint-AT
Tanja Wissik, Jona Hassenbach, Hannes Pirker, Claudia Resch and Stefan Resch

11:00–12:00 **Poster Session**

A Growing Literature of the Public Sphere: Fiction in Danish Newspapers (1666–1850)
Pascale Feldkamp, Alie Lassche, Kit Morgenstjerne, Kristoffer Nielbo, Johan Heinsen and Yuri Bizzoni

Towards an Interoperable Hungarian Historical Newspaper Corpus
Noémi Ligeti-Nagy and Henrietta Szabó

Towards a Bulgarian Historical Newspaper Corpus – Construction of Reading Order over the Text in Searchable PDFs
Nikolay Paev, Stefan Marinov, Ivan Kratchanov, Petya Osenova and Kiril Simov

A Survey of the Digitisation of German Newspapers in interwar Lithuania (1918–1940)
Lina Plaušinaitytė and Heike Zinsmeister

Toward Interoperable and Scalable Representations of Complex Heterogeneous Digitized Historical Media
Pauline Conti, Simon Clematide and Maud Ehrmann

12:00–12:30 **Oral Presentation: Data Modelling**

Data Matters: Looking for High-Quality Corpora to Build Robust and Reliable Models for Humanists
Jaione Macicior and Ana García-Serrano

12:30–13:00 **Oral Presentation: Data Processing**

Thematic Landscapes of the Past: Analysing Slovene Historical Periodicals With Topic Modeling
Filip Dobranić, Uroš Šmajdek, Oliver Pejić, Ciril Bohak, Vojko Gorjanc, Tina Munda and Darja Fišer

Closing Remarks

13:00–13:05 *Closing the workshop*
Maciej Ogrodniczuk, Petya Osenova and Tanja Wissik

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.

The Polish PressMint Corpus

Maciej Ogrodniczuk[1], **Dariusz Czerski**[1], **Adam Pawłowski**[2]
[1] Institute of Computer Science, Polish Academy of Sciences
[2] University of Wrocław
dariusz.czerski@ipipan.waw.pl, maciej.ogrodniczuk@ipipan.waw.pl, adam.pawlowski@uwr.edu.pl

Abstract

This article presents the Polish contribution to the PressMint project, a CLARIN initiative aimed at creating a pan-European, multilingual corpus of historical newspapers. The Polish dataset consists of three subcorpora spanning 110 years (1830–1939). The first two components are drawn from the Microcorpus of Nineteenth-Century Polish (its short press texts and journalistic texts subcorpora), each containing 200 samples of brief news items and journalistic articles from diverse periodicals. The third component, the InterWar Corpus, covers the period 1918–1939 and comprises approximately 6.5 million words from complete newspaper issues, representing the territory of the interwar Republic of Poland. The authors argue for the scholarly value of historical press, highlighting its precise chronological dating as a key advantage for diachronic research despite challenges such as heterogeneous content and anonymous authorship. The conversion pipeline maps source metadata to a standardized TEI format and enriches texts with linguistic annotation using the Hydra NLP tool, providing lemmatization, part-of-speech tagging (mapped to Universal Dependencies), dependency parsing, and named entity recognition. The resulting openly accessible dataset enables cross-linguistic comparison and distant reading of historical press materials on a European scale.

Keywords: historical press corpora, Polish language, TEI encoding

1. Introduction

PressMint[1], one of CLARIN flagship projects, intends to address critical gaps in transnational historical research by constructing a pan-European, multilingual, and interoperable corpus collection of 19th- and early 20th-century newspapers. Designed to overcome the fragmentation of existing resources, it will provide a standardized, richly annotated, and translated dataset that is openly accessible for download and via online analysis tools. The PressMint consortium includes seventeen national partners.

The Polish contribution to the project consists of two corpora. The first is the Microcorpus of Nineteenth-Century Polish (Bilińska et al., 2016, 2018), while the second is the InterWar Corpus — a corpus of Polish press texts from the interwar period, when Poland regained independence after 123 years of political dependence on Russia, Prussia, and Austria (1918–1939).

The Microcorpus of Nineteenth-Century Polish (Bilińska et al., 2018) comprises two subcorpora: Short Press Texts and Journalistic Texts, both covering the period from 1830 to 1918. By contrast, the InterWar Corpus contains a representative selection of materials published between 1918 and 1939. Chronologically, these datasets are consistent and together represent nearly 110 years of the development of the Polish press. The criteria for material selection differ slightly between the two periods (pre-1918 and post-1918).

In the following chapters, we first outline the concept of the PressMint project and then provide a detailed description of the structure and content of the corpus, as well as of the conversion process used to encode the data in the TEI format.

2. Is historical press worthy of scholarly attention?

There are several reasons why historical newspapers and periodicals have not been regarded so far as an attractive object of study in natural language processing and the digital humanities to the same extent as books. Newspapers and magazines are heterogeneous in terms of content, encompassing multiple topics and styles, and they are characterized by irregular and discontinuous editorial structures that hinder automatic processing (the merging of articles divided into sections printed on different pages is a challenge for an NLP processing systems). The frequent practice of publishing unsigned texts further complicates the creation of satisfactory metadata, in contrast to the situation in scholarly journals. Additional difficulties arise from issues of quality. The emphasis on topicality results in texts that are written rapidly and schematically and that do not represent enduring values extending beyond the specific context of time and place. Last but not least, the press traditionally occupies a lower cultural status than the book: it becomes

[1] https://www.clarin.eu/pressmint

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 6–10

obsolete very quickly and, shortly after publication, is often reduced to secondary material of little value or simply discarded.

However, our experience with processing the language of the press of the Polish People's Republic, gained through CLARIN-PL consortium while building the ChronoPress press text corpus (Pawłowski, 2021, current coverage: 1945–1972), leads us to a different conclusion: "old newspapers" contain a substantial and largely untapped potential of information and knowledge (Pawłowski, 2023). First, historical newspapers constitute a form of mass registration of individual facts (sometimes accompanied by commentary), many of which have been forgotten but are valuable for research in history and cultural anthropology. Second, when approached from a big data perspective, the press may reveal enduring, timeless content which is unnoticeable when reading individual texts. Viewed within a broad stream of daily information and across long temporal spans, these materials provide an exceptionally rich representation of social and/or political reality. Such content can be explored diachronically because, despite the many shortcomings of the press discussed above, it has one crucial advantage over literary texts: it is always precisely dated. Even if the authors of press texts are often unknown and proper names or even common words may be printed with spelling errors, it is always possible to assign a date to these texts and map them onto a chronological axis.

The press thus offers remarkable and still underappreciated opportunities for the exploration of informational resources along the chronological axis, studied using distant reading methods and presented to users through powerful visualization tools, such as time series, semantic maps of concepts, or the projection of extracted terms (e.g. named entities) onto other databases. An invaluable and unprecedented feature in the history of the humanities offered by the PressMint project is the possibility of automatic text translation, which effectively overcomes the barrier of the "foreign" language. This provides users with virtually unlimited opportunities for conducting comparative analyses on a European scale.

3. The Polish PressMint Corpus

The first Polish component of the PressMint corpus is grounded in the *Microcorpus of Nineteenth-Century Polish (1830–1918)*, abbreviated as FXIX, the first balanced, tagged and verified diachronic corpus developed to support linguistic research on historical Polish, in particular morphological analysis. The corpus comprises one million tokens, organized into 1,000 samples of 1,000 tokens each, and evenly distributed across five stylistic subcorpora: scientific texts for the general public, press news, feuilletons (journalism), fiction, and drama.

Texts included in FXIX originate from first printed editions written originally in Polish and published between 1830 and 1918. The sampling strategy ensures temporal balance, with each year represented by at least five and no more than twenty samples. Source materials were primarily obtained from major Polish digital libraries. Where machine-readable text layers were unavailable, optical character recognition (OCR) was applied, followed by manual verification and correction.

Each corpus sample consists of three elements: a fragment of continuous text, a structured metadata file, and a facsimile of the source document (PDF, DjVu, or image format). The texts preserve original nineteenth-century spelling and orthography. No linguistic annotation is provided in the released version.

3.1. 19th Century Press: Short Press Texts

The main part of FXIX included in the PressMint dataset comes from the Short Press Text subcorpus. This material primarily consists of brief news items and reports published in daily and non-daily newspapers in major Polish urban centers, as well as by smaller local printing houses where daily publication was not available.

This subcorpus includes 200 samples of 1,000 tokens each. The texts are organized into samples corresponding to newspaper issues or sections. Authors are often anonymous, and the texts typically represent concise, informational styles.

3.2. 19th Century Press: Journalistic Texts

Another component of FXIX consists of texts selected from the journalistic subcorpus of the microcorpus. This dataset includes texts published in newspapers, journals, and books, so this material was only partially included in PressMint, with books excluded from the resulting dataset.

A characteristic feature of this material is the frequent anonymity of authorship: almost half of the texts are unsigned or signed only with initials, pseudonyms, or collective author names.

The Journalistic Texts dataset included in PressMint contains 200 samples. The total volume of the dataset is approximately 1,418,000 characters and 204,000 words. The texts originate from 80 distinct periodicals published in 29 different locations, covering the full temporal span of the microcorpus from 1830 to 1918.

	Short Press	Journalistic	20th century
Files	200	200	750
Characters	1,454,941	1,418,060	40,000,000
Words	206,470	203,944	6,500,000
Years covered	1830–1918	1830–1918	1918–1939
Distinct years	89	89	22
Publication places	37	29	5
Periodicals	115	80	6

Table 1: Quantitative summary of the three Polish PressMint subcorpora.

3.3. 20th Century Press: Complete Issues

In the case of the InterWar Corpus no distinction was made between functional styles (short notes, journalistic articles). A representation of complete issues of newspapers and magazines was developed instead. The subcorpus covering the period 1918–1939 consists of a representative sample of the Polish press with a total size of approximately 6.5 million words. The texts were selected so as, first, to represent the entire territory of the Polish Republic (including press titles published in Kraków, Lwów, Poznań, Warszawa, and Wilno, as well as one nationwide newspaper). In addition, the press samples are evenly distributed over time (approximately 300,000 words per year). Owing to the discontinuous and unpredictable structure of printed newspapers, the selection and preparation of texts were carried out manually; consequently, the corpus does not include complete annual volumes. The texts were annotated also manually, as automatic recognition of author and title metadata was not feasible. A maximum of three hierarchical levels of annotation was applied (section title, article title within a section, and the title of a thematically distinct part of an article).

It should be emphasized that the structural differences between the 1918–1939 InterWar Corpus and the earlier subcorpora do not affect the efficiency of data processing and the quality of services for future users of the PressMint resources, as its primary purpose is to provide a workspace for text mining tasks, not linguistic analysis. Therefore, the correctness of the content and syntactic and semantic annotation are important. One problem that has not yet been fully resolved and that affects the effectiveness of data retrieval (and other functionalities) is the historical orthography of older texts. We are currently testing the performance of NLP tools originally trained on contemporary language when applied to texts printed before 1939. In addition, we are evaluating the effectiveness of automatic translation of such orthographic forms, taking into account the fact that large language models are trained primarily on contemporary language data.

3.4. Corpus statistics

Table 1 presents a quantitative summary of the three Polish PressMint subcorpora. The Short Press Texts and Journalistic Texts subcorpora are comparable in size, each containing 200 samples from the nineteenth century. The third component contains more text samples and textual material, resulting in a total word count that exceeds the combined volume of the two FXIX subcorpora.

4. Encoding Samples in the PressMint Format

The source corpora store metadata in two different formats. The FXIX subcorpora use plain text files with a "key: value" structure, where keys are Polish metadata labels. The InterWar subcorpus uses plain files with metadata embedded in the initial paragraphs of each file. In both cases, all available metadata from the source corpus are retained without modification.

The conversion pipeline maps the source metadata fields to the corresponding TEI elements defined in the PressMint schema. Table 2 presents this mapping. Each row shows the original Polish field names as they appear in the two source corpora and the TEI element to which they are mapped in the output.

In the InterWar data, the periodical name and issue number are not stored as separate metadata fields. Instead, they are automatically extracted from the title field during the conversion process. Additional source metadata fields such as editor, section title, style, and notes are preserved internally but are not mapped to TEI elements, as the current PressMint schema does not include corresponding elements for them.

5. Linguistic annotation

The PressMint conversion process enriches the source texts with sentence and token segmentation, lemmatization, part-of-speech tagging, dependency parsing, and named entity annotation, while preserving the original historical spelling and orthographic variation.

Microcorpus field	Chronopress field	TEI element
autor	—	`author`
tytuł	tytuł	`title level="a"`
data wydania	data	`date@when`
miejsce wydania	miejsce wydania / druku	`pubPlace`
tytuł gazety, czasopisma, serii wyd.	(derived from tytuł)	`title level="j"`
nr	(derived from tytuł)	`biblScope unit="issue"`
wydawnictwo	wydawca / drukarz	`publisher`
źródło	lokalizacja oryginału	`idno type="source"`
link	adres www	`idno type="URI"`

Table 2: Mapping of source metadata fields to TEI elements in the PressMint schema.

As noted in Section 3.3, we are currently testing the performance of NLP tools originally trained on contemporary language when applied to historical texts. However, to achieve consistent linguistic annotation across all three subcorpora — spanning from 1830 to 1939 — we selected HYDRA (Krasnowska-Kieraś and Woliński, 2024) as the unified processing back-end for linguistic analysis. HYDRA is a state-of-the-art Polish NLP model that integrates morphological analysis, dependency parsing, and named entity recognition in a single processing pipeline. It is particularly well-suited for historical Polish texts, as it builds upon the National Corpus of Polish (NKJP) (Przepiórkowski et al., 2012) training data, providing robust performance on both contemporary and archival material. Its ability to handle non-standard orthography and historical word forms is essential for processing the nineteenth- and early twentieth-century press texts in our corpora.

The HYDRA model uses the NKJP tagset for morphological analysis, which employs a rich, fine-grained system of part-of-speech classes specific to Polish (e.g. *subst* for nouns, *praet* for past-tense verbs, *ppron12* for first/second-person pronouns). To ensure cross-linguistic interoperability within the PressMint consortium, these NKJP tags are mapped to the Universal Dependencies (UD) UPOS tagset (Nivre et al., 2020), following the conventions established by the Polish PDB treebank. The mapping covers all major NKJP classes: nominal categories (*subst*, *depr*, *ger*) → NOUN, verbal forms (*fin*, *praet*, *inf*, etc.) → VERB, adjectival and participial forms → ADJ, and similarly for adverbs, adpositions, conjunctions, pronouns, particles, and numerals. The original NKJP tags are preserved in the XPOS column of the CoNLL-U output, while the standardized UPOS tags are placed in the UPOS column.

The resulting CoNLL-U annotations include ten-column records for each token: word form, lemma, UPOS tag, XPOS tag (NKJP), morphological features, dependency head, dependency relation, and named entity annotations in the MISC column using the IOB2 scheme (e.g. `NER=B-persName`, `NER=I-placeName`). HYDRA identifies entities of several types, including person names (*persName*), place names (*placeName*), and organization names (*orgName*), which are embedded directly into the CoNLL-U output rather than stored in a separate annotation layer.

The annotations are subsequently embedded in the PressMint TEI output, where each sentence is encoded as a `<s>` element and each token as a `<w>` element carrying the lemma, UPOS, and XPOS attributes. Named entities are wrapped in the corresponding TEI elements (`<persName>`, `<placeName>`, `<orgName>`).

6. Conclusions

The three Polish PressMint subcorpora provide historically grounded press material spanning the period from 1830 to 1939. The Short Press Texts and Journalistic Texts subcorpora from the Microcorpus of Nineteenth-Century Polish contribute diverse nineteenth-century press material from numerous periodicals and locations. The InterWar subcorpus extends the temporal coverage into the interwar period, adding a substantial volume of early twentieth-century press texts. This time frame is consistent with the dynamics of historical changes in Poland, where World War II (and not the Great War 1914–1918) marks the main milestone of modern times.

The conversion pipeline developed for the Polish data handles two distinct input formats — plain text with separate metadata files and multi-article TXT documents with embedded metadata and automatic author detection — and produces standardized PressMint TEI output. The use of HYDRA (Krasnowska-Kieraś and Woliński, 2024) as the linguistic processing back-end ensures high-quality morphological analysis, dependency parsing, and named entity recognition, while the deterministic NKJP-to-UPOS mapping guarantees interoperability with the Universal Dependencies framework used across the PressMint consortium. All metadata from the source corpora are preserved in the process. The resulting datasets are suitable for cross-linguistic comparison within the PressMint

consortium and for research on diachronic press language.

7. 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; (3) CLARIN-PL, the European Regional Development Fund, FENG programme, agreement number FENG.02.04-IP.040004/24, and (4) Digital Research Infrastructure for the Humanities and Arts DARIAH-PL (Programme: A2.4.1 Expanding Research Capacity under the National Recovery and Resilience Plan), agreement KPOD.01.18-IW.03-0013/23.

8. Bibliographical references

Joanna Bilińska, Magdalena Derwojedowa, Monika Kwiecień, and Witold Kieraś. 2016. Mikrokorpus polszczyzny 1830-1918 [EN: Microcorpus of Nineteenth-Century Polish]. *Komunikacja Specjalistyczna/Communication for Special Purposes*, (11):149–161.

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.

Katarzyna Krasnowska-Kieraś and Marcin Woliński. 2024. Parsing Headed Constituencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 12633–12643. ELRA and ICCL.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Adam Pawłowski. 2023. Korpus prasy polskiej ChronoPress jako infrastruktura i narzędzie badań medioznawczych [EN: The ChronoPress Polish press corpus as infrastructure and a tool for media studies]. *Annales Universitatis Paedagogicae Cracoviensis. Studia ad Bibliothecarum Scientiam Pertinentia*, (21):379–393.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

9. Language Resource References

Bilińska, Joanna and Kwiecień, Monika and Derwojedowa, Magdalena. 2018. *Microcorpus of Nineteenth-Century Polish*.

Pawłowski, Adam. 2021. *ChronoPress – Korpus Tekstów Prasowych*. DOI: 10.34616/139101.

PressMint QuickCheck: Operationalising Readiness Diagnostics for Interoperable Historical Newspaper Corpora

Elena Battaner Moro[1], Almudena Caballos Villar[2], María Cuevas Riaño[2], Marina Miguez Lamanuzzi[2], Dolores Romero López[2]

[1]Universidad Rey Juan Carlos (Madrid, Spain), [2]Universidad Complutense de Madrid (Spain)
elena.battaner@urjc.es, a.caballo@ucm.es, mmcuevas@ucm.es, marimigu@ucm.es, dromerol@ucm.es

Abstract

PressMint QuickCheck is a lightweight, reproducible readiness diagnostic for historical newspaper collections. Given a candidate dataset (ZIP export or IIIF manifests), it detects which components are present, identifies interoperability-critical metadata gaps, and applies lightweight OCR sanity checks. It produces three standardised artefacts: a human-readable readiness report, a minimal normalised manifest (CSV), and a tentative v1 scorecard (suitability_score 0-4) for prioritisation across collections. The workflow is delivered as a Colab-first notebook (no installation required). A key design decision treats content_language and metadata_language declarations as first-class interoperability signals, reflecting the multilingual scope of PressMint and ParlaMint corpora projects

Keywords: historical newspapers; interoperability; readiness diagnostics; metadata quality; OCR triage; PressMint

1. Introduction

Historical newspapers support research on language change, political history, cultural circulation and everyday life. However, digitisation programmes have often prioritised access over computational interoperability, resulting in heterogeneous "library exports": PDF issues with embedded OCR, OCR text directories, ALTO/METS-ALTO packages, partial TEI encodings, or IIIF manifests with descriptive metadata and image pointers.

National digitisation programmes illustrate this variety: some portals distribute PDF+OCR exports with embedded full text, others use METS/ALTO packages with Dublin Core metadata, and some university heritage repositories expose IIIF Presentation API 3.0 manifests. Such variability complicates a basic but essential first step: deciding whether a collection is ready for interoperability-oriented processing and what remediation should occur before investing in conversion and annotation.

PressMint seeks to compile interoperable corpora of historical newspapers and enable comparable processing across collections, following the model of projects such as impresso (Ehrmann et al., 2020). Onboarding candidate collections remains resource-intensive, especially for small DH teams and libraries without dedicated engineering capacity. QuickCheck addresses this gap by providing a low-barrier readiness diagnostic that makes early assessment transparent, reproducible, and actionable.

2. Problem Statement and Contribution

We target an onboarding problem prior to model choice or annotation schemes: given a candidate collection, (1) what components are present (OCR, images, structured metadata), (2) which interoperability-critical metadata are missing or inconsistent (e.g., dates, stable identifiers, provenance, rights, language declarations), and (3) whether OCR exhibits obvious risk signals (empty or low-signal items) that would undermine downstream NLP. OCR quality in historical collections is known to vary considerably (Springmann and Lüdeling, 2017). In many projects this information is produced ad hoc and cannot be compared across collections.

Our contribution is threefold:

- A staged readiness checklist that operationalises these concerns into concrete checks, with explicit treatment of content_language and metadata_language declarations as first-class interoperability signals.
- A lightweight reference implementation (Colab-first notebook) that produces three standardised outputs: a readiness report, a minimal normalised manifest (CSV), and a tentative v1 scorecard for prioritisation.
- A governance-aligned schema and scoring design intended as a starting point for community co-design within the PressMint ecosystem, with configurable weights and thresholds subject to community review.

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 11–15

3. PressMint QuickCheck Method

3.1 Workflow overview

The workflow ingests a single ZIP package or a folder of IIIF manifest JSON files, detects which components are present, applies staged checks, and writes three artefacts: a readiness report (HTML), a minimal manifest (CSV), and a scorecard (JSON with optional per-check breakdown CSV). The pipeline is issue-level: each row in the manifest corresponds to one newspaper issue, defined as a single PDF file, a folder of OCR/ALTO files, or a unit identified heuristically from filenames.

3.2 Inputs and graceful degradation

QuickCheck detects which components are present and applies only the corresponding checks. This design supports realistic library deliveries: collections may include PDF+OCR but no structured metadata, or IIIF manifests with rich descriptive fields but no OCR export. When inputs are missing or partial, QuickCheck still produces an inventory and a minimal manifest while explicitly reporting gaps via structured flags.

3.3 Readiness checklist

Checks are grouped into four lightweight families: Inventory and structure: file tree, sizes, detected components, and missing-component flags.

- Metadata completeness and consistency: field coverage, parseable dates (ISO 8601), stable identifiers, content_language and metadata_language declarations (presence check, basic BCP 47-like syntax validation; missing declarations flagged as a high-impact P1 readiness gap), duplicates, and other high-impact missing fields.
- OCR sanity (if present): coverage, empty or low-signal items, abnormal character ratios, and extreme-length outliers (triage indicators only; not a measure of OCR accuracy).
- IIIF checks (if present): basic manifest validity, canvas counts, label presence, and consistency of identifiers and links, following the IIIF Presentation API 3.0 specification (IIIF Consortium, n.d.).

3.4 Prioritisation: needs_priority and needs_actions

Each issue receives a structured flag list (issues_flags) from which two fields are derived. needs_priority classifies urgency at issue level: P0 (blocker — the issue cannot be onboarded; triggered when no usable text or structure is present), P1 (required — onboarding possible after resolving specific gaps), or P2 (desirable — does not block onboarding). The final priority is the most severe flag present: a single P0 flag makes the issue P0 regardless of other signals.

needs_actions translates each flag into a concrete remediation step. Table 1 shows the flag-to-action mapping.

Table 1: Flag-to-action mapping (v2.0).

Flag	Priority	Action
P0:no_text_or_struct	P0	Provide OCR text (TXT/ALTO), IIIF manifests, or structured metadata exports
P1:missing_language_declaration	P1	Add declared content_language (and metadata_language) using BCP 47 tags
P1:unparseable_date	P1	Provide/normalise issue date (ISO 8601) or encode it in filenames/metadata
P1:missing_rights	P1	Add rights/licence statement at collection/issue level
P1:missing_provenance	P1	Add provider/source provenance label for cross-collection comparison
P1:ocr_empty_or_low_signal	P1	Check OCR extraction quality; consider re-OCR or alternative exports
P1:iiif_invalid	P1	Validate IIIF manifests (id/type/items/labels)
P2:generated_id	P2	Stable external ID

Flag	Priority	Action
		not found; generated from filename/date
P2:fields_from_defaults:X,Y	P2	Fields X, Y filled from notebook defaults, not declared by provider. Verify accuracy and request explicit metadata declarations.

The P2:fields_from_defaults flag (introduced in v2.0) addresses a transparency gap identified during external testing (see Section 4): when a user sets notebook-level defaults for content_language, rights, or source_provenance, these values populate the manifest but the provider has not declared them. In v1.9, this produced an empty needs_actions despite fields being synthetic — a misleading signal. In v2.0, the flag is added with the list of affected fields, ensuring needs_actions always reflects the actual state of provider-declared metadata.

4. Minimal manifest (CSV) for harmonisation

The manifest schema is intentionally minimal and designed as a harmonisation starting point for PressMint onboarding workflows. Language-related fields (content_language, metadata_language, language_present) are first-class outputs because PressMint and ParlaMint-style corpora are intrinsically multilingual and require explicit language signalling for interoperability. The suitability_score (0–100) and suitability_level (0–4) are derived from a weighted checklist; weights and thresholds are configurable and subject to community review aligned with PressMint governance decisions. The manifest schema is designed to align with PressMint operational requirements for onboarding: fields such as stable identifiers, rights statements, and language declarations correspond directly to minimum metadata requirements under discussion within the PressMint community.
Table 2 shows the minimal manifest schema. Fields marked “Always” are present in every output row; “If present” fields are populated when the source collection declares them.

Table 2: Minimal manifest schema (v1).

Field	Status	Notes
item_id	Always	Original stable ID if present; generated otherwise (P2:generated_id flag)
date	Always	Issue date parsed to ISO 8601; empty with flag if unparseable
content_language	Always	Declared language of OCR/full-text content; BCP 47-like validation; missing declaration flagged as P1
metadata_language	If present	Declared language of descriptive metadata fields; BCP 47-like validation where present
language_present	Always	Boolean: true if at least one language declaration found; false triggers P1 flag
title_or_masthead	If present	Title or masthead string from metadata
source_provenance	If present	Provider/source label; key for cross-collection comparison
rights	If present	Rights/licence information when declared
files_present	Always	Detected components (PDF, OCR, ALTO, IIIF…)
issues_flags	Always	Structured flags for missing fields, broken

Field	Status	Notes
		references, low-signal OCR, etc.
suitability_score	Always	0–100 integer; derived from weighted checklist. Weights are v1, tentative, configurable
suitability_level	Always	0–4 coarse category: 0=Not usable; 1=Text-only triage; 2=Candidate; 3=Ready for harmonisation; 4=PressMint-ready
needs_priority	Always	P0/P1/P2 derived from issues_flags.
needs_actions	Always	Short remediation steps per flag (see Table 1).

4.1 Limitations

QuickCheck v1 does not attempt full PressMint TEI conversion, article segmentation, heavy NLP enrichment, or automatic language identification. It is explicitly a pre-flight diagnostic for onboarding, not a pipeline component. The suitability scorecard measures onboarding readiness only; it does not measure OCR accuracy, linguistic quality, scholarly value, or corpus relevance. The language_guess field (optional, v1) is a lightweight triage hint based on stopword frequency and is not a substitute for declared BCP 47 tags. Issue boundaries are determined heuristically from filenames when explicit boundaries are absent, which may misgroup files in collections with non-standard naming conventions. Scoring weights and suitability thresholds are v1 and tentative; they are configurable and subject to community governance rather than fixed as final technical decisions.

When collections lack reliable metadata — a question raised during peer review — QuickCheck still produces an inventory and a minimal manifest while flagging all gaps explicitly (P0/P1/P2). Notebook-level defaults can supply provisional values for missing fields, but v2.0 explicitly flags these as P2:fields_from_defaults to avoid masking the underlying metadata gap.

5. Demo and Use Case

The demo illustrates the end-to-end workflow on two sample packages representative of common library delivery patterns: (i) a set of PDF issues with embedded OCR from a national digital newspaper portal (PDF+OCR sample); and (ii) a set of IIIF manifests from a university heritage repository exposing IIIF Presentation API 3.0 (IIIF sample). We show: component detection and inventory across both delivery patterns, metadata coverage and identifier/date/language consistency, OCR sanity indicators on the PDF sample, and export of a unified minimal manifest covering both collections.

The demo then shows how the manifest supports (a) selecting a “first subset” for harmonisation, (b) prioritising remediation tasks such as missing language tags, unstable identifiers, or absent provenance/rights fields, and (c) producing comparable collection-level summaries across the two input types.

6. Discussion: Positioning within PressMint

QuickCheck is intended to complement, not replace, PressMint’s conversion and annotation pipelines. By standardising an early assessment step, it makes onboarding decisions more transparent and comparable across partner collections, supporting the interoperability goals pursued by initiatives such as Europeana Newspapers (Neudecker and Antonacopoulos, 2016) and the CLARIN PressMint flagship project (CLARIN ERIC, n.d.).

The minimal manifest schema is designed to align with PressMint’s emerging operational requirements for onboarding. Fields such as stable identifiers, rights statements, and language declarations correspond to minimum metadata requirements under discussion within the PressMint community, and the suitability_level 4 threshold (“PressMint-ready”) is intended to reflect those requirements as they are formalised through governance. This alignment is intentionally configurable rather than fixed, since PressMint’s operational decisions are still evolving.

Language field coverage is one check among others in the readiness checklist; it is included because missing or inconsistent language tags are a recurring interoperability gap in multilingual corpora. This is particularly valuable for small teams and library collaborations where engineering capacity is limited, and format heterogeneity is the norm. A shared readiness report and minimal manifest also provide a concrete basis for discussion between content providers and technical teams, helping to align

expectations and define realistic remediation plans.

The schema, checklist, and scoring are presented as a pathway for community co-design within the PressMint ecosystem. We do not claim prior endorsement by PressMint or ParlaMint governing bodies.

7. Availability

The notebook (in English and Spanish, other languages and formats are foreseen) and sample outputs are available in the following public repository URL: https://github.com/ebattanermoro/PressMint-Quickcheck.

8. Conclusion

PressMint QuickCheck operationalises a lightweight readiness diagnostic for historical newspaper collections and provides a reproducible, low-barrier reference workflow. By producing three standardised artefacts (a readiness report, a minimal harmonisation manifest, and a tentative v1 scorecard for prioritisation) it supports faster and more transparent onboarding of candidate collections into the PressMint ecosystem. Language declarations are treated as first-class interoperability signals, reflecting the multilingual scope of PressMint and ParlaMint corpora. The v2.0 release addresses a transparency gap identified during external testing, ensuring that notebook-level defaults are always explicitly flagged rather than silently masking missing provider metadata. We would like to thank three anonymous reviewers and the PressMint workshop team at LREC 2026 for their useful and encouraging comments and suggestions.

9. Bibliographical References

CLARIN ERIC. (n.d.). PressMint. https://www.clarin.eu/pressmint (accessed 25 March 2026).

Ehrmann, M., Romanello, M., Clematide, S., Strobel, P. B., and Barman, R. (2020). Language Resources for Historical Newspapers: the Impresso Collection. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 958-968. Marseille, France. European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.121.pdf (accessed 25 March 2026).

IIIF Consortium. (n.d.). IIIF Presentation API 3.0. https://iiif.io/api/presentation/3.0/ (accessed 25 March 2026).

Neudecker, C. and Antonacopoulos, A. (2016). Making Europe's Historical Newspapers Searchable. In Proceedings of the 12th International Conference on Document Analysis Systems (DAS 2016), pp. 405-410. Santorini, Greece. IEEE. https://doi.org/10.1109/DAS.2016.83 (accessed 25 March 2026).

Springmann, U. and Lüdeling, A. (2017). OCR of Historical Printings with an Application to Building Diachronic Corpora: A Case Study Using the RIDGES Herbal Corpus. Digital Humanities Quarterly, 11(2). https://dhq.digitalhumanities.org/vol/11/2/000288/000288.html (accessed 25 March 2026).

Statement on the use of AI: Claude Sonnet 4.6 and ChatGPT 5.2. were used for this work to review texts and code.

PressMint-PT — Compiling a Portuguese Historical Newspaper Corpus

José Aires, Amália Mendes
University of Lisbon, School of Arts and Humanities, Centre of Linguistics
Alameda da Universidade, 1600-214 Lisboa, Portugal
jagc@edu.ulisboa.pt, mendes@edu.ulisboa.pt

Abstract

We present a new European Portuguese corpus of newspapers from the 19th and early 20th centuries, integrated in the recent PressMint project, whose goal is to provide a set of comparable newspaper corpora for European languages in that time frame. We discuss the raw data that was previously available, as well as new data specifically compiled for the project, and the challenges involving OCR, text recognition and different orthographical norms. We describe the pipeline setup for XML encoding and annotation, partially based on work developed for the ParlaMint corpora. The corpus is currently under development and will be made freely available at the end of the project, as part of the PressMint corpora.

Keywords: newspaper corpus, historical corpus, text recognition

1. Introduction

Although corpora composed of European Portuguese are increasingly available for research, finding historical newspaper corpora in Portuguese is still a challenge. The existing initiatives either do not comprise newspaper texts and focus on earlier stages of the Portuguese language, such as the medieval corpus *Corpus Informatizado do Português Medieval*[1], or they only comprise epistolary writing, such as the Post Scriptum corpus[2]. Other corpora (see section 2) frequently do not separate European from Brazilian Portuguese, treat the 20th century as a single period, or follow their own standards.

The situation applies to other European languages, and as a result, the PressMint project aims to compile a multilingual, comparable, annotated, translated, and interoperable set of corpora of European historical newspapers. This initiative follows the work already completed to compile a comparable multilingual corpus of Parliamentary data from several European countries under the ParlaMint project (Erjavec et al., 2025). The PressMint-PT corpus follows the common standards setup in the project for this specific genre and builds upon the expertise gained in Portuguese XML file processing for the ParlaMint-PT project.

The period of the late 19th and first half of the 20th centuries was extremely prolific in terms of newspaper titles. A list, in the form of a dictionary, of the more than three hundred daily newspapers published between 1900 and 2000 in Portugal can be found in Lemos (2020), together with a summary of the history of each newspaper, details of where it can be consulted and a study on the History of the Portuguese Daily Press in the 20th Century.

The evolution of newspapers is directly related to the political and social changes that occur in Portugal (Tengarrinha, 1971). So, a historical newspaper corpus is crucial for historical studies, as it would make available data from a period that covers the Monarchy, then the first Republic, and finally the dictatorship of the Estado Novo. The comparisons that an interoperable set of European corpora offer are valuable for analyzing the political and social events that structured the beginning of the 20th century. It will also be of interest to a historical perspective on Communication Studies, enabling a direct view of the changes that affected the newspaper genre. These are areas of knowledge that are frequently unaware of the available natural language processing methods, or simply underuse them, instead accessing historical newspapers in an unfriendly image format. We are confident that these areas of knowledge that deal with digitized versions of paper-based formats will be greatly enhanced by using the PressMint-PT to access relevant data. Additionally, such a corpus is especially interesting for observing changes that occurred in early contemporary Portuguese from a Historical Linguistics point of view, comparing this period with later stages of the language.

We will present in section 2 other newspaper corpora for Portuguese and in section 3 the data that we have included so far in the PressMint-PT corpus. The actual corpus processing is discussed in section 4, and the XML encoding and annotation pipeline is discussed in section 5, before concluding in section 6.

[1] https://cipm.fcsh.unl.pt
[2] http://teitok.clul.ul.pt/postscriptum/

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 16–20

2. Related Work

There are three main corpora of newspaper texts in European Portuguese. One is the CETEMPublico corpus, which comprises 190 million words from the Portuguese newspaper *Público*[3] (Santos and Rocha, 2001). It covers contemporary Portuguese from the 1990's to the 2000's, and falls outside the goals of the PressMint corpus. The *Corpus do Português* (Corpus of Portuguese) contains 45 million words from European and Brazilian Portuguese taken from the 14th to the 20th century[4] (Mark Davies, 2016) (Mark Davies and Michael Ferreira, 2006). The 19th century section covers several genres of European and Brazilian Portuguese and comprises around 10 million words. The 20th century section includes around 3 million words, but most fall into the 1990's and 2000's (an example would be the newspaper Público, which started in the 1990's). The corpus is available for online queries.

The other corpus containing newspaper texts from the 19th and 20th centuries is the *Corpus de Referência do Português Contemporâneo* (Reference Corpus of Contemporary Portuguese). Since we will use part of this corpus for the constitution of the PressMint corpus, the corpus will be presented in detail in section 3.

3. Raw Corpus

For the compilation of a Portuguese newspaper corpus, we relied on the data included in the Reference Corpus of Contemporary Portuguese (CRPC). The corpus focuses mainly on European Portuguese data from the last quarter of the 20th century, but also includes smaller sections on the other varieties of Portuguese spoken in the world, namely data from Brazil, Angola, Cape Verde, Guinea-Bissau, Mozambique, Sao Tome and Principe, Macau, and Timor. As there is very little newspaper historical data in the CRPC, we explored adding new data available in image format to expand our collection. We describe the already existing data in section 3.1 and the new collected data in section 3.2.

3.1. The CRPC corpus and the VARPORT subsection

The CRPC corpus comprises 311 million words of written and spoken texts of different varieties of Portuguese in the world. The written subcorpus includes texts from different genres: newspapers and magazines, fiction, didactic and scientific texts, parliamentary data, law and court rulings, letters, and brochures (Généreux et al., 2012). The written part of this corpus covers 309,812,943 tokens, compiled from 356,208 documents, mostly from 1970 to 2008, although texts from 1800 forward are also included. The corpus was developed at the Center of Linguistics of the University of Lisbon[5] and is available for online queries on CQPweb[6]. The newspapers section of the CRPC, restricted to Portugal, contains 98,579,946 tokens and 160,289 texts. For the PressMint project, we restricted this newspaper section to the period from 1808 to 1945. We provide the number of tokens and the number of files for 30-year periods in Table 1. The table shows that only a small subset of 63,178 tokens fits the time window that was set for the PressMint collection, and highlights the difficulty of finding and processing historical newspaper data.

A large part of the CRPC data mentioned in Table 1, covering the period of the 19th century and early 20th century, was compiled in the framework of the VARPORT project[7]. We will refer to the data that originates from the CRPC corpus and its VARPORT subset as the CRPC/VARPORT corpus.

Table 1: Number of newspaper files and tokens per time period in the European Portuguese CRPC/VARPORT subcorpus

time period	no. of tokens	no. of files
1800-1829	6,239	28
1830-1859	21,245	170
1860-1889	9,682	69
1890-1919	13,344	93
1920-1949	12,668	76
Total	63,178	436

The time period selected for the PressMint corpus is determined by the available material in the CRPC/VARPORT corpus. Each 30-year phase includes three types of data: newspaper articles, editorials, and advertisements/announcements. We decided to keep the latter type in the PressMint corpus for two reasons: first, some announcements are related to information provided by institutions and companies and are not strictly advertising texts; second, even the advertisements can be relevant for future applications, such as the geolocation of companies. The transcriptions of the newspaper texts were performed through digitization, OCR recognition, and manual revision. It should be noted that we kept the original orthography in the

[3] https://www.linguateca.pt/CETEMPublico/

[4] https://www.corpusdoportugues.org

[5] https://www.clul.ulisboa.pt

[6] gamma.clul.ul.pt/CQPweb

[7] VARPORT is a joint venture between the Center of Linguistics of the University of Lisbon (CLUL) and the Federal University of Rio de Janeiro (UFRJ). Project website: https://varport.letras.ufrj.br; supervision: Sílvia Brandão (UFRJ) and Antónia Mota (CLUL)

texts of this set of files.

3.2. Enlarging the corpus

The data in the CRPC/VARPORT corpus is available in digital format and has been previously revised, making it immediately suitable for the next processing step, but the number of tokens is relatively small. To enlarge the corpus, we explored publications from the late 19th and the beginning of the 20th centuries available on the Hemeroteca website[8], in image format. We initially selected three potential newspapers:

- *A Capital*, a daily newspaper;
- *O Ocidente*, a three-times-monthly publication;
- *Revista Universal Lisbonense*, a weekly magazine.

The time period and the number of images for each publication are shown in Table 2.

Table 2: Details from selected Hemeroteca publications

publication	time period	no. of files
A Capital	1910-1938	19,086
O Ocidente	1878-1915	10,845
Rev. Uni. Lisb.	1841-1853	7,042
Total	1841-1938	36,973

The digitized images of these publications raise several challenges to natural language processing methods: not only are some images damaged with tears, wrinkles, or old-age stains, but also the orthography used at the time is different from the current norms, which affects the quality of the OCR process and the accuracy of the linguistic annotation tools. In addition, both the publication in columns and the presence of embedded images make it harder to extract the text in its correct sequence. Such issues will be addressed in the following section 4.

The example in Figure 1 provides the front page of the Lisbon newspaper *A Capital* on its edition of August 1st of 1910, illustrating its compact design and the different orthographical norms where, for instance, *oficiaes* would now be spelled *oficiais*, *monarchico* would be *monárquico*, and *teem* would be *têm*.

4. Corpus Processing

To simplify the production of the required XML documents, we processed the sources mentioned in section 3 above, since even though such text files

[8] https://hemerotecadigital.cm-lisboa.pt


A CAPITAL

DIARIO REPUBLICANO DA NOITE

LISBOA—Segunda-feira, 1 de Agosto de 1910

Campanha eleitoral

CONTRABANDO

Outra assembleia do Credito Predial

O sr. João Albino "despeja o sacco"

Blasco Ibañez

Figure 1: A front page of the newspaper *A Capital*.

were already compiled, they still required additional processing, which will become clear in the following subsections.

4.1. Processing the CRPC/VARPORT Documents

As mentioned before, the CRPC/VARPORT files were already available in text format, but because the purpose for which they were originally created did not account for a clear separation among some data elements relevant to the PressMint project, the available metadata was not quite normalized and had some minor issues, so additional processing was necessary to deal with situations like the following:

- different terms used to identify the same field;
- dates in different formats;
- more than one element in the same line;
- elements spanning more than one line;
- some orthographical errors;
- missing accents;
- extra spaces;

- different orthographies;
- mismatched letter casing.

As an example, we can see below a situation in which the date (Data) and location (local) could be split into separate lines. Also, the date includes the weekday, which is generally unnecessary.

Data e local: Segunda-feira, 18.12.1837 - Lisboa

Some other examples of date normalization include the following:

- domingo, 4 de janeiro de 2026;
- 4 de janeiro de 2026;
- 4-1-2026;
- 4 jan 26 00:00;
- 4.1.2026.

The examples above are all represented by 2026-01-04 after normalization.

The following example shows several fields concatenated and split into several lines in which some fields are even empty.

Jornal: A Capital Número: 8679 Data: Sábado, 13 de Novembro de 1995 Local/Edição: Lisboa Secção: Página: Coluna: Autor: Título: Ficheiro: acordo.txt Introdução: Suporte magnético Revisão:

Such normalization methods required the use of regular expressions, date parsing, and expression substitutions, which greatly simplified the final metadata encoding, described in section 5 below.

4.2. Processing the Additional Image Documents

These documents are the most difficult to process since they are only available as images, requiring an OCR stage to produce the corresponding text. Such stage might even be preceded by a segmentation stage in which the location of the actual text within the image is determined before being presented to the text recognition process.

Taking into account the challenges mentioned in section 3.2, we have conducted a few initial experiments using ImageMagick (ImageMagick Studio LLC, 2024) to modify the images, and tesseract (Ooms, 2026) with its best (most accurate) model to produce the text contained in the corresponding images, with the purpose of finding the best parameters with which to apply the OCR, such as image enhancement, contrast, anti-alias, gamma, and resizing, confirming those have an impact on quality, but so far we have obtained mixed results, gaining quality in some sample areas but losing it in others, so we have still not found a clear winner.

Additionally, we have also prepared a small set of verified examples, with which we fine-tuned the previous tesseract best model, but the images used must also go through some image processing, as well as the number of examples apparently has to be increased before we are able to achieve any relevant positive impact.

Finally, we fear that we might need to tune the process to the different publications since they have some graphical styling differences between them.

To tackle all these difficulties, we plan to start by using a more simplistic programmatic approach, in which we only consider the number of total words and the number of out-of-vocabulary words, hoping to at least be able to make obvious the parameters that will be less likely to produce good results and therefore discard them. Once we have a smaller set of parameter candidates, it will be easier to check their results individually, ideally using a smaller set of random samples to be verified by a human.

5. XML Encoding

Once we normalized the metadata from the CRPC/VARPORT texts, as described in section 4, it became easier to produce their XML and their annotated XML counterpart, as explained in the following subsections.

5.1. XML Documents Generation

The contents of each document took into account not only the actual text, but also its metadata, which became very simple thanks to the processing stage described above. Additionally, in order to apply the naming convention, the files were organized according to their publication dates, followed by their publication source, and finally followed by a number to distinguish files with the same publication date and publication source.

5.2. Annotated XML Documents Generation

The annotation information for each XML file produced above, consisting of lemma, POS, UDR and NER, was obtained using UDPipe 2 (Straka, 2018) and NameTag 3 (Straková and Straka, 2025). However, given the language's old orthographical norm, the quality of the annotation might be suboptimal, so we are considering the implementation of an additional preceding stage in which the old orthography is modernized before being presented to the annotation tools.

5.3. Main XML Documents Generation

The main XML documents include references to the individual XML documents of each text, as well as the sum of several element amounts of each document, such as the number of paragraphs, words, and punctuation marks.

6. Final Remarks

The compilation of the corpus of the additional data in image format to be included in the PressMint-PT set is still underway, which again is the most challenging part of the project. The next step is to evaluate the results of the OCR and text recognition, as well as checking to what extent the annotation tools are capable of (fully or at least partially) automatically dealing with these historical texts, which might require, for instance, an additional orthography modernization stage.

Also, considering the significant progress achieved in the AI field, we intend to explore other alternatives for image text processing, like segmentation with fine-tuned YOLO or Meta's SAM, and models like unsloth/gemma-3-4bit or dots.ocr, to name a few. The segmentation stage will hopefully provide some ideas as to how the articles could be presented, since those are generally non-linear and their borders are not easily determined.

An evaluation of the difficulty of processing individual historical publications will ultimately affect the selection of newspapers to include in the corpus.

7. Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and suggestions, which have actually provided some ideas we have included and which we intend to explore further. This work was partially supported by CLARIN ERIC PressMint-Interoperable corpora of historical newspapers, and by Fundação para a Ciência e a Tecnologia as part of the project of Centro de Linguística da Universidade de Lisboa (UID/214/2025).

8. Bibliographical References

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 Dargis, 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*, volume 59. Springer Netherlands.

M. Généreux, I. Hendrickx, and A. Mendes. 2012. Introducing the Reference Corpus of Contemporary Portuguese On-Line. In *LREC'2012 – Eighth International Conference on Language Resources and Evaluation*, pages 2237–2244, Istanbul, Turkey. European Language Resources Association (ELRA).

ImageMagick Studio LLC. 2024. *ImageMagick*.

Mário Lemos. 2020. *Jornais Diários Portugueses do Século XX – um dicionário*.

Mark Davies. 2016. Corpus do português: Web/dialects.

Mark Davies and Michael Ferreira. 2006. Corpus do português: Web/dialects.

Jeroen Ooms. 2026. *tesseract: Open Source OCR Engine*. R package version 5.2.5.

Diana Santos and Paulo Rocha. 2001. Evaluating cetempublico, a free resource for portuguese. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 450–457, Toulouse, France. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

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.

José Tengarrinha. 1971. *História da Imprensa Periódica Portuguesa*. Portugália Editora.

Historical Newspapers in the General Regionally Annotated Corpus of Ukrainian (GRAC): Current State and PressMint Integration Prospects

Maria Shvedova[1,2], **Arsenii Lukashevskyi**[1]
[1] National Technical University "Kharkiv Polytechnic Institute"
Kyrpychova str. 2, 61002, Kharkiv, Ukraine
[2] Friedrich Schiller University Jena
Fürstengraben 1, 07743 Jena, Germany
mariia.shvedova@khpi.edu.ua
arsenii.lukashevskyi@sgt.khpi.edu.ua

Abstract

This paper presents the historical newspaper collection of the General Regionally Annotated Corpus of Ukrainian (GRAC) and outlines its prospective integration into the PressMint infrastructure. The collection comprises 117 newspaper titles published before 1950, totaling 23.6 million tokens, and reflects the political fragmentation, regional variation, and orthographic diversity of Ukrainian-language press from the late nineteenth to mid-twentieth century. We describe the corpus composition, temporal and geographic distribution, and metadata architecture. Special attention is given to morphosyntactic annotation challenges arising from the historical Western Ukrainian orthography (Zhelekhivka), as well as issues related to annotating historical texts using the rule-based TagText parser and neural UDPipe2 models. The paper compares GRAC's vertical format and metadata system with the TEI-based PressMint standard, identifying technical and conceptual harmonization challenges. Integrating GRAC newspapers into PressMint will facilitate comparative research on language policy, regional standardization, and media discourse within a broader European context.

Keywords: newspaper corpus, Ukrainian language, historical newspapers, corpus annotation, metadata standards, GRAC, PressMint

1. Introduction

The General Regionally Annotated Corpus of Ukrainian (GRAC) (Maria Shvedova (2017–)) is a large representative corpus of standard Ukrainian covering the 19th-21st centuries and effectively serving as the national corpus of the Ukrainian language. The newspaper collection is an important part of GRAC, enabling researchers to track and accurately date language change and study the influence of language policy and ideology. At the same time, newspaper texts are challenging to process, not only because of access and digitization issues, but also (in the case of Ukrainian) because of different orthographies and language norms at different times and in different territories. Tools designed for the modern Ukrainian language are of limited use for parsing historical texts.

Integration into PressMint: Interoperable Corpora of Historical Newspapers (CLARIN ERIC, 2025) will enable the study of Ukrainian newspapers in comparison with press materials from neighboring linguistic and political contexts, opening up new opportunities for research into language contact and shared European historical discourse.

This paper is structured as follows: Section 2 provides the historical and linguistic context of Ukrainian press from the late nineteenth to the twentieth century; Section 3 describes the GRAC newspaper collection's composition and sources; Section 4 presents the metadata architecture that captures regional and orthographic variation; Section 5 addresses morphosyntactic annotation challenges; Section 6 demonstrates research applications of the collection; Section 7 analyzes the alignment between GRAC and PressMint metadata standards; and Section 8 offers concluding remarks.

2. Historical and Linguistic Context

Before World War I, Ukrainian-speaking territories were politically divided between the Russian and Austro-Hungarian empires, and thereafter between the Ukrainian Soviet Socialist Republic (Ukrainian SSR), Poland, Czechoslovakia, and Romania[1]. By 1945, following the post-war territorial settlements, almost all these territories had been incorporated into the Ukrainian SSR. Policies regarding the Ukrainian language varied from country to country. Ukrainians in Austria had a fairly developed press since the end of the 19th century (in 1900 there were 25 periodicals published in Galicia and six in Bukovina (Shevelov, 2008)), while in the Russian Empire, Ukrainian-language publishing was severely restricted and a Ukrainian-language press did not exist until 1905.

[1] Historical map of Ukraine from (Magocsi, 1987).

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 21–26

Until the end of World War II, the Ukrainian language was shaped by political partition. The language of editions published in the Russian-controlled and later Soviet part of Ukraine (with its main cultural center in Kyiv, and from 1919 to 1934 in Kharkiv) differed substantially from the language of Western Ukraine, which was mostly culturally oriented toward Lviv. Researchers describe distinct regional variants of literary Ukrainian for this period, which developed under the influence of local dialects and different dominant languages (Hrytsenko, 1993; Franko, 1995; Matvijas, 1998), with differences in lexical and grammatical norms and different orthographic standards until the 1920s.

The Ukrainian language coexisted in each territory with other languages that enjoyed greater social prestige: in the large cities of Central and Eastern Ukraine it was Russian, in Galicia it was Polish, in Bukovina it was German and/or Romanian, and in Transcarpathia it was Hungarian (Shevelov, 2008). Therefore, in early Ukrainian newspapers, we observe a significant influence of dominant languages both in the Ukrainian language itself, saturated with borrowings at the level of vocabulary and syntax (Shvedova and von Waldenfels, 2021), and in the form of code-switching (since the newspaper audience was predominantly bilingual). Some early 20th-century Ukrainian newspapers published entire texts or columns in other languages. For example, the Ukrainian newspaper *Rada* (Kyiv, 1906-1919) contained advertisements in Russian. In the Ukrainian SSR before World War II, some newspapers published articles in different languages within a single issue (Ukrainian-Polish *Radianska Volyn* 'Soviet Volyn' (1924, Zhytomyr), Ukrainian-Yiddish-Russian *Chervona Shvachka* 'Red seamstress' (1932, Kyiv)).

This linguistic heterogeneity directly shapes the metadata and annotation challenges addressed in the following sections.

3. The GRAC Newspaper Collection: Composition and Sources

The GRAC newspaper collection contains only texts in Ukrainian (although there may be some cases of code-switching that are not currently specifically tagged). The collection consists of digitized newspaper texts: OCR output subsequently verified by human annotators. This paper focuses on the historical component: 117 newspaper titles published before 1950, totaling 23.6 million tokens.

The distribution of old newspaper texts across macroregions reveals significant temporal and geographic imbalances in corpus coverage (Figure 1). Western Ukrainian newspapers (macroregion W) constitute the earliest materials in the collection, with substantial coverage beginning in the 1880s and continuing through the interwar period. In contrast, newspapers from the Kyiv region (macroregion KYV) enter the corpus only after 1905, due to censorship in Russian Empire. The bulk of KYV materials dates from 1910 onward, when orthographic practices became more standardized and closer to modern conventions, facilitating corpus processing and linguistic annotation. The 1920s exhibit more diverse geographic representation, coinciding with the relatively liberal Ukrainization policy in Soviet Ukraine, which encouraged Ukrainian-language publishing, with materials from Central (C), Eastern (E), Northern (N), and Southern (S) regions appearing alongside continued coverage of Western and Kyiv publications. The World War II period (1941–1945) is particularly well represented, with substantial materials from both German occupation newspapers and underground press, as well as Soviet publications. The immediate post-war years (1945–1946) maintain strong coverage, particularly in Western Ukraine.

Quantitatively, Western Ukraine dominates our pre-1910 materials, with peaks exceeding 700,000 tokens in the early 1890s and consistent coverage through the interwar decades. Our collection for 1924-1925 shows exceptional volume across multiple regions, with Northern Ukraine contributing over 1.3 million tokens and Eastern Ukraine nearly 450,000 tokens. The WWII years represent our best-documented period, with approximately 4.4 million tokens assembled across all regions.

Newspaper texts in GRAC have been collected from multiple sources. The core Western Ukrainian collection derives from the historical press archive curated by Orest Drul on the Zbruch portal (zbruc.eu), featuring Galician newspapers from the late 19th–mid-20th centuries (Drul, 2014–). Most texts from the 1910s–1940s were prepared by university students during research practicum projects, working from scans provided by LIBRARIA, a digital archive of Ukrainian periodicals (Arkhivni Informatsijni Systemy, 2017–), or downloaded from the Archive of Old Newspapers (Old, 2010–2016). The WWII collection was systematically compiled by Anna Bordovska, encompassing newspapers from German occupation authorities, Soviet publications, and underground OUN-UPA press (Bordovska, 2024). The shape of the collection reflects not only historical publishing activity but also, to a large extent, the priorities of the digitisation projects from which these materials were sourced, and the specific research interests of the corpus compilers.

These diverse sources and the resulting uneven temporal-geographic distribution must be considered when designing comparative studies or assessing the representativeness of linguistic patterns across regions and periods.

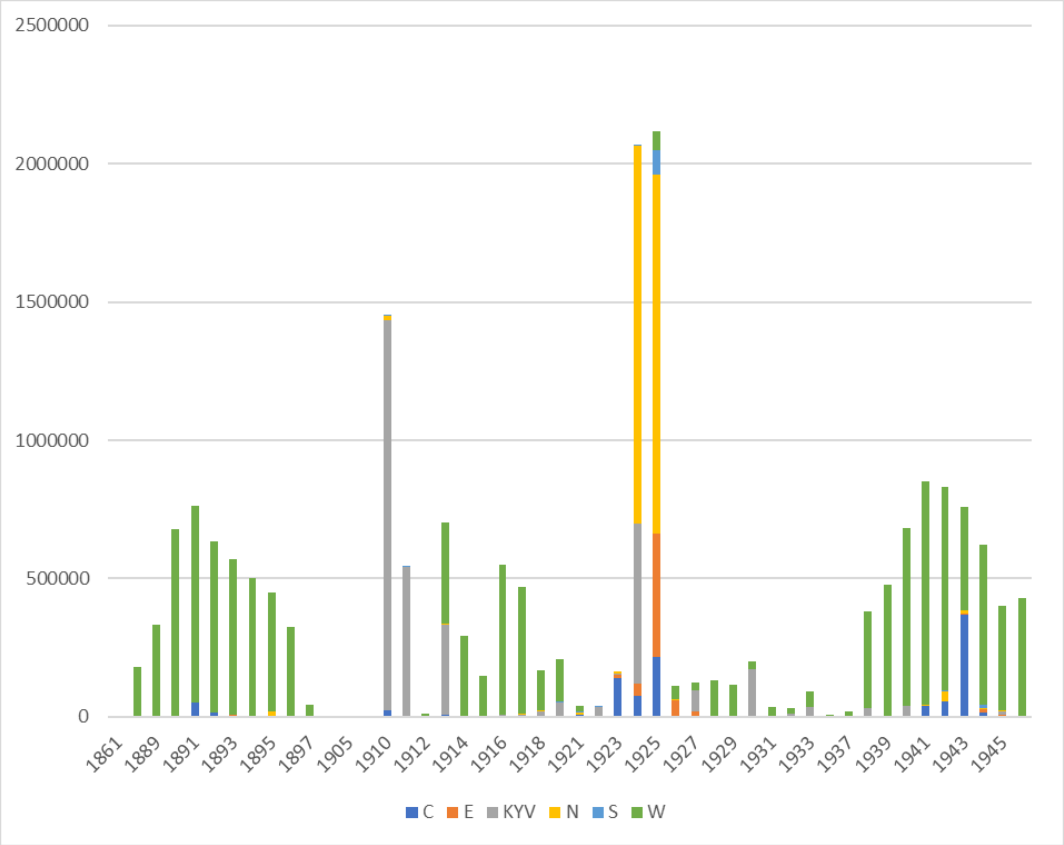


Figure 1: Tokens of old newspaper texts by year and macroregion: West, East, Center, South, North, Kyiv.

4. Metadata Schema in GRAC

The GRAC metadata schema for periodicals, described in detail in (Shvedova, 2020) and on the GRAC site, captures the complex political and administrative landscape of Ukrainian-language press across multiple states and regimes, reflecting the sociolinguistic heterogeneity described in Section 2. Periodicals are classified by type through the `doc.mediaType` attribute (*newspapers* and *magazines* for pre-1950 materials, with additional categories for later periods), with newspapers representing the most diverse category.

All periodicals in GRAC are annotated for the political entity that controlled the publication. This classification is encoded in the `doc.mediaAdmin` attribute and reflects the political fragmentation of Ukrainian territories throughout the corpus timespan, see Table 1.

Regional attribution follows different logic depending on whether individual authorship is available at the article level: when authors are known and annotated, regional tags reflect the author's origin; when author information is unavailable (as is common in newspaper materials) regional attribution is assigned based on the place of publication (Shvedova and von Waldenfels, 2021).

Text segmentation and the associated metadata vary by time period: pre-WWII newspapers were added article-by-article, resulting in text-level metadata (author, title, and regional attribution based on author's origin, where known). Newspapers from the WWII period onward were predominantly digitised as complete issues (with the exception of Western Ukrainian publications from 1945–1946, which retain article-level metadata), and therefore lack text-specific metadata fields.

The resulting metadata schema provides the foundation for regional and diachronic comparative

Code	Administrative Entity
AHI	Austro-Hungarian Empire
RUI	Russian Empire
POL	Interwar Poland
CZE	Interwar Czechoslovakia
ROM	Interwar Romania
ZUNR	West Ukrainian People's Republic
MAX	Makhno Movement
SOV	Ukrainian SSR
RUK	Reichskommissariat Ukraine
OUN	Organization of Ukrainian Nationalists
UKR	Contemporary Ukraine
DIA	Diaspora

Table 1: Administrative and political classification codes for periodicals in GRAC.

analyses, though its internal heterogeneity must be accounted for when formulating research queries or comparing findings across regions and time periods.

5. Morphosyntactic Annotation Challenges

Morphological annotation in GRAC is performed automatically using the TagText program, a dictionary-based tagger for modern Ukrainian based on the VESUM open-access dictionary (Starko and Rysin, 2022). The primary challenge for automatic morphosyntactic analysis arises from orthographic variation in Western Ukrainian materials. Newspapers from Austrian Galicia and interwar Poland (approximately half of the historical newspaper subcorpus) represent the Western Ukrainian regional variant of the literary standard, characterized by distinctive lexicon, some specific grammatical forms, and orthographic system that differ from both the modern standard and Eastern Ukrainian practices, leading to reduced morphosyntactic annotation accuracy when processed with tools developed for contemporary Ukrainian. Western Ukrainian historical newspapers in GRAC are represented in the Western Ukrainian orthographic standard of the late 19th–early 20th century (Zhelekhivka)—partly in its original form and partly in reconstructed form (the oldest Zbruč collection materials (Drul, 2014–), originally published in the etymological orthography, or Maksymovychivka, have been converted to Zhelekhivka). To handle Zhelekhivka texts, GRAC employs an additional rule-based module that adapts the standard morphological analyzer. However, this approach does not yield optimal results, partly due to inherent variability within Zhelekhivka itself (Chemerys et al., 2023).

UDPipe2 (Milan Straka and Hajič (2016–)), a neural network-based multilingual morphosyntactic parser, is planned for use in PressMint corpus annotation. Both Ukrainian models within UDPipe2 were trained on modern Ukrainian orthography data. As a neural network-based system, UDPipe2 handles out-of-vocabulary items and orthographic variation more robustly than dictionary-based approaches, mostly successfully processing the distinctive lexicon and spelling conventions of historical texts. However, this approach differs from TagText in lemmatization strategy: TagText's rule-based normalization module maps historical orthographic variants to standardized lemma forms, while UDPipe2 generates lemmas that preserve the input orthography. This results in distinct lemma representations for orthographic variants of the same lexeme—for instance, modern *svit* 'world' and historical Western Ukrainian *svït* 'world' would be lemmatized as separate entries. While this preserves orthographic information, it fragments lexeme frequencies across spelling variants and complicates cross-period lexical queries without post-processing normalization.

A more significant challenge for UDPipe2 arises from differences in word segmentation conventions between historical and modern orthographies. Zhelekhivka wrote clitics separately from their host words, as in *byty mut' sja* (modern *bytymut'sja* 'will fight'), where the reflexive marker *sja* and future tense marker *mut'* appear as independent tokens. Word boundaries for adverbs and prepositions also differ systematically: Zhelekhivka wrote *do domu* 'homeward' and *v oseny* 'in the fall' as two words (modern *dodomu*, *voseny*), while *vkinci* 'at the end' was written as one word (modern *v kinci*). The preposition *popry* 'despite' exhibits additional variability, appearing in Zhelekhivka as *popry*, *po-pry*, or *po pry*. Particles *že*, *ž*, *by*, *b* were attached directly to the preceding word in Zhelekhivka, preventing accurate recognition of both the host word and the particle.

Empirical evaluation confirms these challenges: manual verification of UDPipe2 part-of-speech tags in a 572-token sample from *Bil'shovyk Poltavshchyny* (Poltava, 1924) versus a 504-token sample from *Dilo* (Lviv, 1889, Zhelekhivka) revealed 96.7% accuracy for Soviet Ukrainian text but only 92.7% for the Lviv text, with errors concentrated in non-standard spelling and segmentation.

These substantial linguistic differences suggest that historical Western Ukrainian newspapers should be treated as a distinct subcorpus within the PressMint framework. For optimal annotation quality, we plan to develop a dedicated UDPipe2 model trained specifically on Zhelekhivka-orthography texts, capable of capturing the linguistic and orthographic features of this historical regional standard. In the longer term, further harmonization in the treatment of orthographic and segmentation variation across varieties, including normalization of lemma forms, would be particularly beneficial for enabling consistent use of both Ukrainian corpora in linguistic research.

6. Use Cases and Research Applications

The GRAC newspaper collection has enabled diverse corpus-based studies investigating linguistic variation in Ukrainian press during critical periods of political transformation.

Shvedova (2021) compared three newspaper subcorpora (Soviet 1919–1933, Western Ukrainian 1937–1943, and Western Ukrainian Soviet 1939–1946) to trace the formation of a new journalistic lexical norm in the 1940s. Analysis of 117 synonymous sets demonstrated that the new norm had

an Eastern Ukrainian basis, with minimal Western Ukrainian influence. The integration into PressMint would allow extending this analysis beyond the Ukrainian component of the multilingual Galician discourse of the interwar period: systematic comparison with contemporary Polish-language press would enable distinguishing lexical items specific to Ukrainian journalistic norm from those shared with and reinforced by Polish.

Bordovska (2024) examined World War II newspapers of different political orientations (German occupation, Soviet, and underground OUN-UPA press), revealing systematic orthographic and lexical differences correlated with political ideology. German and underground newspapers gravitated toward 1928 Orthography norms, while Soviet publications consistently applied 1933 norms. At the lexical level, Soviet periodicals systematically preferred international vocabulary, while underground press favored Ukrainian equivalents.

Hleba (2025) investigated regional variation in spatial prepositions *pry*, *bilja* and *kolo* 'near' across Lviv and Kyiv newspapers and fiction texts, using profile-based and collocational analysis. Fiction texts showed 75% uniformity (indicating distinct regional differences), with Lviv preferring *pry* (Polish influence) and Kyiv preferring *bilja* and *kolo*. Newspapers, however, showed 98% uniformity, suggesting genre-specific standardization effects.

These studies relied on GRAC's detailed metadata (including `doc.mediaAdmin` and regional annotation) and full-text search, with linguistic patterns identified through manual analysis.

However, some documented patterns of regional variation reflect not only internal Ukrainian factors but also substantial influence from neighboring languages. Integration into PressMint would not only enable consistent morphosyntactic analysis across regional varieties through improved annotation tools, but also allow for systematic comparison of GRAC newspapers with comparable corpora from neighboring linguistic and political contexts from the same time period, revealing shared public terminology across multilingual discourses and patterns of cross-linguistic influence at multiple linguistic levels.

7. GRAC Metadata vs. PressMint Standards

Harmonization of GRAC with PressMint involves both technical and sociolinguistic challenges. Structurally, the two formats differ substantially. GRAC uses vertical files optimised for NoSketch Engine, where tokens carry positional attributes (word form, lemma, morphological tag, semantic annotation), and metadata is encoded as attributes of the <doc> element. PressMint, in contrast, follows the Text Encoding Initiative (TEI) XML standard, widely used for encoding historical documents in digital humanities, completely separating metadata from the text and supporting a richer set of structural elements, including article types, column divisions, and links to facsimiles. Most positional attributes and basic structural elements (`<doc>`, `<s>`) can be converted to TEI-XML without conceptual loss.

More significant difficulties arise from the historical and sociolinguistic complexity of Ukrainian press. The GRAC attribute `doc.mediaAdmin`, designed to reflect political affiliations of Ukrainian-language periodicals (Table 1), has no direct equivalent in PressMint and will require the development of a language-specific taxonomy. Regional metadata also requires normalization. All metadata must be translated into English for interoperability. Publication dates, currently often recorded at the year level, should be refined to the day of issue where possible.

Named entity information in GRAC is encoded at the token level as semantic features within morphological tags, with separate labels for personal name components (given names, patronymics, surnames), geographical names, and other proper nouns. PressMint requires multi-token named entities with explicit boundaries (e.g., *Taras Hryhorovych Shevchenko* as a single PERSON entity rather than separate tokens). Converting to PressMint's four-class system (PERSON, LOCATION, ORGANIZATION, MISC) will require post-processing to identify entity boundaries and merge multi-token entities.

At the morphosyntactic annotation level, orthographic heterogeneity must be explicitly addressed. We plan to treat Zhelekhivka materials as a dedicated subcorpus with a specialized UDPipe2 model trained on historical Western Ukrainian texts, which will improve annotation accuracy while preserving orthographic information for diachronic research.

8. Conclusion

The GRAC historical newspaper collection reflects the political fragmentation, regional diversity, and orthographic heterogeneity of Ukrainian press from the late nineteenth to the mid-twentieth century. While structural conversion to PressMint is technically feasible, successful integration depends primarily on careful metadata normalization, taxonomy development, and adaptation of annotation layers to account for historical variation.

Addressing these challenges will enable Ukrainian historical newspapers to function as a fully interoperable component of the PressMint infrastructure, supporting comparative research on media discourse, language policy, and regional standardization processes across Europe.

9. Acknowledgements

The authors are grateful to Orest Drul for his assistance in integrating his large Western Ukrainian newspaper collection into GRAC. The authors also thank the anonymous reviewers for their valuable comments, which helped to significantly improve the paper. The authors benefited from discussions within the Universal Dependencies community and participation in COST Action CA21167 "UniDive". The preparation of the interwar Soviet and WWII newspaper collections for GRAC was partially funded by Friedrich Schiller University Jena (2019–2023), with the support of Prof. Ruprecht von Waldenfels.

10. Bibliographical References

2010–2016. Archive of old newspapers. Digital archive of historical Ukrainian newspapers.

Arkhivni Informatsijni Systemy. 2017–. LIBRARIA: Archive of Ukrainian periodicals.

Anna Bordovska. 2024. Orthographic and lexical variation in the journalism of the Second World War period: Based on GRAC data. *Movni i kontseptual'ni kartyny svitu*, 1(75):36–59. In Ukrainian.

Yurij Chemerys, Olesia Nakhlik, Andriy Rysin, and Maria Shvedova. 2023. Normalization of a historic Western Ukrainian orthographic system Zhelekhivka in the Ukrainian language reference corpus (GRAC). In *Proceedings of the IEEE 18th International Conference on Computer Sciences and Information Technologies (CSIT)*, Lviv, Ukraine.

CLARIN ERIC. 2025. Pressmint: Interoperable corpora of historical newspapers. Accessed: 2026-02-28.

Orest Drul. 2014–. Western Ukrainian historical press collections. Zbruč Digital Archive. Three chronological collections: *125 Years Ago* (1897–1902), *100 Years Ago* (1922–1927), *75 Years Ago* (1947–1952).

Zynovija Franko. 1995. Variation or territorial distinction of the Ukrainian literary language. *Ukrans'ka istorična ta dialektna leksyka*, (2):169–173. In Ukrainian.

Anastasiia Hleba. 2025. Regional variation of the spatial Ukrainian prepositions in the 1920s–1930s: a corpus-based study. In *Synsémantické slovní druhy ve slovanských jazycích*, pages 147–166. Institute of Slavonic Studies of the Czech Academy of Sciences, Prague.

Pavlo Hrytsenko. 1993. Some remarks on the dialectal basis of the Ukrainian literary language. In *Philologia slavica: To the 70th Anniversary of Academician N.I. Tolstoy*, pages 284–294. Moscow. In Russian.

Paul Robert Magocsi. 1987. *Ukraine: A Historical Atlas*. University of Toronto Press, Toronto.

Ivan Matvijas. 1998. *Variants of the Ukrainian Literary Language*. Kyiv. In Ukrainian.

Yurij Shevelov. 2008. The Ukrainian language in the first half of the twentieth century (1900–1941): State and status. In *Selected Works: In 2 volumes. Vol. 1: Linguistics*, pages 26–279. Kyiv-Mohyla Academy, Kyiv. In Ukrainian.

Maria Shvedova. 2020. The General regionally annotated corpus of Ukrainian (GRAC, uacorpus.org): Architecture and functionality. In *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS)*, pages 489–506, Lviv, Ukraine.

Maria Shvedova. 2021. Lexical variation in the language of the Ukrainian press of the 1920s–1940s and the development of a new lexical norm: A corpus-based research. *Movoznavstvo*, (1):16–35. In Ukrainian.

Maria Shvedova and Ruprecht von Waldenfels. 2021. Regional annotation within GRAC, a large reference corpus of Ukrainian: Issues and challenges. In *Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS)*, pages 32–45, Kharkiv, Ukraine.

Vasyl Starko and Andriy Rysin. 2022. VESUM: A large morphological dictionary of Ukrainian as a dynamic tool. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS)*, pages 71–80, Gliwice, Poland.

11. Language Resource References

Maria Shvedova, Ruprecht von Waldenfels, Sergey Yarygin, Andriy Rysin, Vasyl Starko, Tymofij Nikolajenko, Arsenii Lukashevskyi et al. 2017–. *GRAC: General Regionally Annotated Corpus of Ukrainian*.

Milan Straka, Jana Straková and Jan Hajič. 2016–. *UDPipe Web Service (LINDAT/CLARIAH-CZ): Trainable Pipeline for Tokenization, Tagging, Lemmatization and Parsing*. LINDAT/CLARIAH-CZ, Institute of Formal and Applied Linguistics, Charles University.

CLARIAH-ES PressMint: Building Interoperable Corpora of Historical Press in Spain

Ainara Estarrona, Aritz Farwell, Xabier Goenaga, German Rigau
HiTZ Center - University of the Basque Country (EHU)
{ainara.estarrona, aritz.farwell, xabier.goenaga, german.rigau}@ehu.eus

Abstract

This paper describes CLARIAH-ES's contribution to PressMint in Spain as a distributed effort across regional nodes (e.g., Catalonia, Madrid, Basque Country, Galicia, Canary Islands, Alicante), each developing manageable corpora in partnership with key repositories such as ARCA, Patrimonio Digital Complutense, Euskariana, Jable, Galiciana, and the BVMC periodicals portal. A central technical challenge is heterogeneous legacy OCR quality, motivating experiments with AI/LLM-assisted OCR renewal, normalization layers, and linguistic enrichment (e.g., NER and entity linking). This effort is situated alongside ongoing dissemination and the EOSC Mesh "historical newspapers" use-case work aimed at scalable discovery, access, and federated computation over interoperable historical press data.

Keywords: CLARIAH-ES, corpus, historical newspapers, PressMint

1. Introduction

PressMint is a CLARIN Flagship initiative designed to produce a pan-European, interoperable, multilingual corpora of European historical newspapers (mostly from the late nineteenth to early twentieth centuries) that are comparable across countries and languages. A key motivation is that, although many national libraries already provide digitized newspaper collections, these datasets are typically not interoperable, which limits comparative "distant reading" approaches and broader European-scale analyses. PressMint addresses this by delivering a shared, FAIR-aligned resource and actively promoting its uptake among historians, linguists, media scholars, and other social sciences and humanties (SSH) communities. The project runs from June 2025 to May 2027 and is organized as a distributed effort involving multiple national consortia.

PressMint's objective is interoperability by design. Participating corpora are encoded to a common PressMint schema (a customisation of the TEI Guidelines) and supported by shared scripts and workflows—explicitly building on infrastructure and practices developed in the earlier ParlaMint project (Erjavec et al., 2023; Erjavec et al., 2025). In this manner, the same processing pipeline can be applied across corpora even when their source characteristics differ. In addition to TEI XML, PressMint plans to provide downstream formats (e.g., TSV, CoNLL-U, JSON) and to make the corpora openly available for download and through online corpus analysis tools, lowering the barrier for researchers to explore, compare, and reuse historical newspaper data at scale.

Within this European context, CLARIAH-ES[1], Spain's national research infrastructure for CLARIN and DARIAH, is actively contributing to PressMint in Spain by coordinating TEI encoding efforts, partnering with national and regional digital libraries, and experimenting with AI methods (including LLM-based approaches) to improve OCR and downstream processing. The following discussion describes CLARIAH-ES, its approach to building corpora for PressMint, and where the process of corpora creation for PressMint in Spain is at the present time.

2. CLARIAH-ES

CLARIAH-ES is a distributed digital research infrastructure created to coordinate Spain's participation in the two main European social sciences and humanities research infrastructure consortia, CLARIN and DARIAH (Riudavets et al., 2024). CLARIN's mission is to make language data, tools, and expertise accessible for research, while DARIAH focuses on digitally enabled research and teaching in the arts and humanities. Together, the two infrastructures form a complementary ecosystem of standards, services, and research communities.

CLARIAH-ES grew out of the earlier INTELE strategic network (2020–2022) and was a central player in Spain becoming a full member of CLARIN and DARIAH (Iruskieta et al., 2022). In practice, CLARIAH-ES exists to support the digital transformation of SSH research in Spain, enabling computational work with textual, visual, and audio materials by connecting researchers to shared infrastructures, methods, and tools. A core objective is to reduce the "digital divide" by promoting multilingualism, interoperability, resource sustainability and reuse, and open science practices through coordinated services and community building.[2]

[1] https://www.clariah.es/

[2] The network is funded by the Ministry of Science,

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 27–33

Organizationally, CLARIAH-ES brings together a multidisciplinary consortium that includes experts in language technologies, AI, HPC, library science, and SSH scholarship.[3] It comprises twelve nodes located across Spain, with its administrative and coordinating office based at HiTZ,[4] the Basque Center for Language Technology at the University of the Basque Country (EHU).

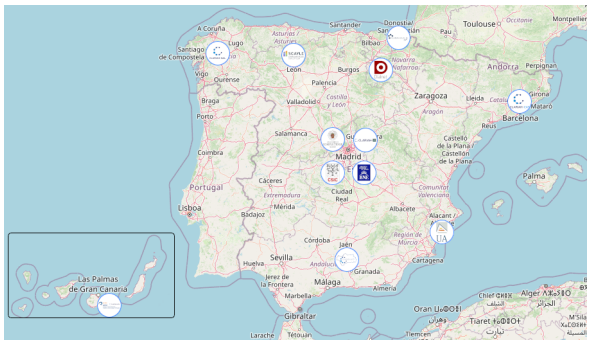


Figure 1: Map of CLARIAH-ES

3. Corpora For PressMint

CLARIAH-ES has taken the strategic decision to produce several separate corpora for PressMint according to regional nodes within the infrastructure, in part due to the presence of regional languages that were used in historical press in addition to Spanish. Currently, these nodes include CLARIAH-CAT (Catalonia), CLARIAH-CM (Madrid), CLARIAH-EUS (Basque Country), CLARIAH-GAL (Galicia), CLARIAH-IATEXT (Canary Islands), and CLARIAH-UA (Alicante). Other nodes are likely to participate in the future as well. Several of these nodes, in addition to CLARIAH-UNED (Madrid) and CLARIAH-SCAYLE (León),[5] will also experiment with AI techniques or provide expertise in language technology.

The distribution among these regional nodes in CLARIAH-ES has enabled several teams across the infrastructure to focus on building smaller, discrete, and more manageable corpora within the guidelines set out by PressMint. At the moment, these corpora are in different phases of development and experimentation, as described in the following subsections: 3.1. Compiled Corpora, 3.2. Corpora Under Construction, and 3.3. Language and AI Technology.

Innovation and Universities (RED2024-154077-E).

[3] https://www.clarin.eu/blog/introduction-clariah-es

[4] https://www.hitz.eus/en

[5] CLARIAH-SCAYLE will offer compute to perform OCR experiments.

3.1. Compiled Corpora

CLARIAH-GAL is currently the only node that already possesses a corpus of historical texts that may be adapted for use in PressMint[6] The collection, titled "Tesouro Informatizado da Lingua Galega" (TILG),[7] brings together over 3,000 documents written in Galician between 1612 and 2013, including historical press (which would be extracted for the PressMint corpus), totaling roughly 30 million word forms that are lemmatized and morphosyntactically tagged.

With respect to PressMint, one option under consideration is to take advantage of the existing annotation in the TILG and map its tags to Universal Dependencies (UD), which would avoid having to apply OCR and normalization from scratch. A drawback to this approach, however, is that TILG applies orthographic regularization, so the resulting corpus would be less faithful to the graphic and lexical variation of the original texts.

A second tentative alternative is to build a separate corpus from the historical press available in Galiciana,[8] the digital repository maintained by the Biblioteca de Galicia. This approach would preserve the original language as found in the historical texts, but also require CLARIAH-GAL to correct the largely unreviewed OCR for most of those texts, build a normalization layer, and perform any added annotation from scratch. The result would be a richer corpus from a historical perspective, but the effort would be considerably greater.

3.2. Corpora Under Construction

The latter option outlined above reflects the current status of the remaining corpora to a greater or lesser degree. As discussed in the following subsections, all other contributing nodes in CLARIAH-ES are at different stages in the process of constructing corpora in collaboration with regional repositories.

3.2.1. CLARIAH-CAT

CLARIAH-CAT,[9] coordinated by the Barcelona Supercomputing Center,[10] integrates research groups from every Catalan public university (UB, UAB, UPF,

[6] CLARIAH-GAL (https://www.clariah.gal/) is the Galician node of the Spanish CLARIAH-ES infrastructure, coordinated by the Instituto da Lingua Galega (https://ilg.usc.gal/en) with the collaboration of CiTIUS at the Universidade de Santiago de Compostela (https://citius.gal/).

[7] https://ilg.usc.es/TILG/

[8] https://biblioteca.galiciana.gal/gl/inicio/inicio.do

[9] Website currently under construction (https://clariah.cat/).

[10] https://www.bsc.es/

UdG, UdL, URV, UPC, UOC), as well as research centers such as the Computer Vision Center (CVC), the Artificial Intelligence Research Institute (IIIA-CSIC), and the Catalan Institute of Classical Archaeology (ICAC). In addition, the network includes the Biblioteca de Catalunya, the civil society organization SoftCatalà, and the companies LaTempesta and Avenir-Cultura.

CLARIAH-CAT has contacted the Arxiu de Revistes Catalanes Antigues (ARCA)[11] about the opportunity to build a corpus utilizing its collection of historical press. ARCA is a collaborative portal promoted by the Biblioteca de Catalunya[12] that provides centralized and open access to digitized Catalan historical newspapers and magazines.

3.2.2. CLARIAH-CM

CLARIAH-CM,[13] led by the Universidad Complutense de Madrid (UCM), acts as a regional coordination hub that brings together researchers and research groups from the six public universities in the Madrid area (the Universidad Politécnica, the Universidad de Alcalá), the Universidad Autónoma de Madrid), the Universidad Rey Juan Carlos), the Universidad Carlos III), and the Complutense). CLARIAH-CM expects to build its corpus with material provided by the Patrimonio Digital Complutense (PDC) and Biblioteca Digital de la Comunidad Madrid.

The PDC[14] is the UCM Library's portal for digitized cultural heritage. Its wide-ranging holdings include historical periodicals from the nineteenth and twentieth centuries. The Biblioteca Digital de la Comunidad Madrid[15] aggregates holdings from the Biblioteca Regional de Madrid, the Real Academia Española, the Real Academia de la Historia, and the Fundación Lázaro Galdiano. It provides access to mostly public-domain works dating roughly from the fifteenth to the twentieth century, with particular emphasis on materials related to Madrid and its culture.

3.2.3. CLARIAH-EUS

CLARIAH-EUS,[16] led by HiTZ, seeks to strengthen Basque language- and Basque culture–related digital humanities research (Alkorta et al., 2025). In addition to HiTZ, the node is made up of various research groups from the University of the Basque Country and from other institutions and organizations, such as the Soziolinguistika Klusterra[17], UEU-GOI[18], Badalab[19], and the Research Centre for Basque Language and Texts (Iker–administered by the CNRS, the University Bordeaux Montaigne, and the University of Pau and Pays de l'Adour).[20] CLARIAH-EUS is collaborating with Euskariana to construct its PressMint corpus and hopes to expand this collaboration to other regional libraries and repositories in the near future. Euskariana, the Basque Government's collaborative digital portal for Basque culture and cultural heritage, is managed by the Biblioteca Digital de Euskadi and gathers together contributions from various partners, including public administrations, municipalities, universities, cultural institutions, archives, libraries, and museums.

This material encompasses a large collection of historical press and Euskariana has agreed to provide CLARIAH-EUS and HiTZ with close to 180 periodicals –images, pdf, txt, and metadata– that cover the period 1874-1939. Published mostly in the Basque Country, these periodicals run the gamut of political ideologies and include newspapers printed in Basque, English, French, and Spanish. All the objects are OCRed, but it is likely that many will benefit from a renewed OCR effort given that most were processed over fifteen years ago. HiTZ is currently considering how best to create datasets for evaluation (a gold standard) and exploring methods to apply OCR using LLMs. The node also expects to enrich the corpus with normalization, NERC, and entity linking.

3.2.4. CLARIAH-IATEXT

CLARIAH-IATEXT[21] is the Canary Islands node of CLARIAH-ES, led by the Instituto Universitario de Análisis y Aplicaciones Textuales (IATEXT) at the Universidad de Las Palmas de Gran Canaria (ULPGC). In similar fashion to its counterparts, CLARIAH-IATEXT serves as a regional infrastructure for Canarian research communities working in the fields of digital humanities and social sciences. CLARIAH-IATEXT expects to build its PressMint corpus in collaboration with the Museo Canario and the Biblioteca de la ULPGC.

The Museo Canario's repository of historical press[22] contains a continuously growing archive

[11] https://arca.bnc.cat/arcabib_pro/en/inicio/inicio.do

[12] The Biblioteca de Catalunya is Catalonia's national library (https://www.bnc.cat/).

[13] https://www.ucm.es/clariah-cm-en

[14] https://patrimoniodigital.ucm.es/s/patrimonio/page/inicio

[15] https://bibliotecavirtualmadrid.comunidad.madrid/bvmadrid_publicacion/es/inicio/inicio.do

[16] https://www.clariah.eus/en

[17] https://soziolinguistika.eus/en/

[18] https://www.goi-institutua.eus/

[19] https://badalab.eus/

[20] https://iker.cnrs.fr/?lang=en

[21] https://iatext.ulpgc.es/es/clariah_iatext

[22] https://www.elmuseocanario.com/en/

that aims to gather all periodicals published in the Canary Islands, complemented by titles produced by Canarian communities abroad. It is regarded as the most complete repository in the archipelago, holding periodicals dating from the eighteenth century to the present day, including fin-de-siècle newspapers, when the Canary Islands became one of the five Spanish provinces with the highest newspaper production. The periodical repository at the Biblioteca de la ULPGC is largely delivered through Jable,[23] the Canary Islands Digital Press Archive, a long-running initiative that provides access to historical and modern periodicals published in the Canary Islands.

3.2.5. CLARIAH-UA

CLARIAH-UA[24] is the University of Alicante's research infrastructure node for digital humanities within the CLARIAH-ES consortium. The node is closely associated with the Biblioteca Virtual Miguel de Cervantes (BVMC), whose mission includes advancing DH research, designing technologies for the humanities, and providing access to Hispanic cultural material. The BVMC curates a dedicated periodicals portal[25] that gives researchers and the general public structured access to pre-1930 historical press, comprising about 275 titles. CLARIAH-UA will draw on this collection to construct its corpus for PressMint.

3.3. Language and AI Technology

Much of the material that will be utilized to build corpora for PressMint across CLARIAH-ES's nodes possess OCR that was applied several years before. Recent advances in OCR technology that leverages LLMs have significantly improved and facilitated the application of OCR to historical texts, including periodicals. In addition to other AI-related tools and approaches, CLARIAH-ES is experimenting with how these OCR techniques may be applied to the respective corpora it is producing for PressMint.

Although several nodes within CLARIAH-ES will participate in this initiative, CLARIAH-UNED and CLARIAH-EUS have already begun to explore how best to approach the OCR problem. CLARIAH-UNED,[26] is doing so as part of the GRESEL-UNED project,[27] which investigates Spanish-language historical press between 1850 and 1945 by using LLMs to analyze discourses about nation, identity, feminism, literature, and international relations. The project's core goal is to build a RAG research assistant that is reliable for scholarly inquiry, but to make the assistant effective, the project is developing and adapting linguistic resources and NLP/IR workflows, including training or tailoring models, extracting and classifying historically meaningful entities, and improving factual grounding through retrieval so researchers may run better question-answering and exploratory analyses over the corpus.

CLARIAH-EUS and HiTZ are investigating how Latxa,[28] the largest and best-performing LLM available for Basque, may be harnessed to help build its corpus for PressMint. Early experiments with OCR involve tests utilizing PERO OCR (Kodym and Hradiš, 2021) and ScribbleSense.[29] For normalization, both statistical (Phonetisaurus,[30] cSMITiser) and AI techniques are under evaluation (Ljubešić et al., 2016; Scherrer and Ljubešić, 2016).

4. PressMint-Related CLARIAH-ES Events and Dissemination

Members of CLARIAH-ES have organized or participated in several events that have included a focus on historical press or PressMint. Below is a list of some of these as well as several more that will take place in the coming months.

4.1. PastReader

PastReader[31] was an IberLEF 2025 shared task designed to enable automatic transcription of digitized Spanish historical press using newspapers housed at the digital repository of Spain's National Library. The PDFs that where utilized often include OCR but the extracted text can be unreliable because of degraded scans, complex layouts, and historical typography. To address these challenges, PastReader organized evaluation around two core problems: (1) OCR error correction and (2) end-to-end "curated" text extraction directly from scanned images, encouraging the use of multimodal approaches as well as robust post-processing. Overall, the task's goal was to reduce human effort in mass digitization workflows and improve the accessibility, retrieval, and

newspaper/

[23] https://jable.ulpgc.es/

[24] https://clariah-ua.cervantesvirtual.com/

[25] https://www.cervantesvirtual.com/portales/hemeroteca/r

[26] https://clariah.uned.es/

[27] The GRESEL initiative (PID2023-151280OB-C22) is funded by Spain's Ministry of Science, Innovation and Universities (https://gresel-uned.hypotheses.org/).

[28] https://www.hitz.eus/en/node/340

[29] https://scribblesense.cz

[30] https://github.com/AdolfVonKleist/Phonetisaurus

[31] https://sites.google.com/view/pastreader2025

long-term preservation of Spanish newspaper heritage by benchmarking and promoting more accurate, efficient transcription systems.

4.2. FDS Seminar “Humanidades Digitales en Acción"

The Fundación Duques de Soria seminar “Humanidades digitales en acción: herramientas para el análisis de prensa histórica en español,”[32] held in Soria on July 2-4, 2025, was an award-winning international initiative that gathered together a multidisciplinary team to discuss how to make Spanish historical newspapers easier to access, digitize, and analyze through digital methods, Conceived as a bridge between three professional communities —library science specialists, humanists, and computer scientists— the seminar was structured around core thematic axes that combined historical and literary approaches to newspapers with hands-on digital workflows for digitization/OCR, discoverability, and computational analysis.

4.3. Fourth CLARIAH-EUS Workshop

The Fourth CLARIAH-EUS Workshop (“Humanitate Digitalak eta Gizarte Zientziak gaur egungo Hizkuntza Teknologiak aplikatuta”)[33] took place on November 28, 2025 in Vitoria-Gasteiz. A CLARIAH-EUS community event focused on applying current language technologies to digital humanities and the social sciences, it included a poster session that showcased tools, corpora, and RAG-oriented work undertaken by members of the infrastructure. A poster dedicated to PressMint, “PRESSMINT: Egunkari Historikoen Corpus Elkarreragingarriak,” introduced the project to the Basque audience.

4.4. II Ciclo CLARIAH-CM

The “II Ciclo CLARIAH-CM: formación en Prensa Histórica (Herramientas y metodologías digitales),”[34] an in-person training series launched by the CLARIAH-CM node, provides regular, hands-on workshops plus short theoretical introductions on digital methods for humanities research. The current cycle focuses on building, processing, and analyzing Spanish historical press corpora. The aim is to help researchers streamline workflows and deepen methodological expertise in this field and consists of three sessions that combine guided practice with the option to work on participants’ own materials. The sessions include Label Studio for OCR dataset creation (February 27, 2026), historical press corpus processing with Sketch Engine (March 23, 2026), and AI-based exploration of historical press (April 27 2026).

4.5. I Jornada CLARIAH-ES en Bibliotecas

The "I Jornada CLARIAH-ES en Bibliotecas: Infraestructuras Digitales y Ciencia Abierta"[35] (Madrid, March 3, 2026) is an outreach-oriented event hosted at the Instituto Cervantes (Madrid) that brings together librarians, archivists, and researchers to discuss how digital research infrastructures can help transform traditional collections into open, reusable, and more visible resources within the broader ecosystem of Open Science. The program is organized into three thematic blocks: an introduction to infrastructures (CLARIN, DARIAH, CLARIAH-ES, SSH Open Marketplace, and EOSC), a section devoted to workflows and projects, and a closing block focused on data-driven library services and community building. A key highlight is the PressMint session, which introduces PressMint as an example of how libraries and infrastructures can collaborate to improve large-scale access, processing, and reuse of historical newspaper collections.

4.6. IA y Humanidades Digitales para la Prensa Histórica

The CLARIAH-UNED and CLARIAH-EUS organized event, “IA y humanidades digitales para la prensa histórica”[36] (Madrid, April 20, 2026) is a forum devoted to how AI and digital humanities can improve the digitization, transcription, and analysis of historical newspapers. The morning sessions are dedicated to the AI-driven exploration of historical press (particularly methodological challenges such as verification, prompting strategies, and AI analysis of multilingual press) and a library roundtable that brings together perspectives from major cultural heritage institutions, including the "Hemeroteca Digital" at Spain’s National Library and the "Biblioteca Virtual de Prensa Histórica," maintained by Spain’s Ministry of Culture. The afternoon features a discussion on research design for multimodal OCR inference, domain-specific entity recognition, and knowledge-graph/semantic exploration approaches, followed a session dedicated to PressMint, connecting the event to broader

[32] https://fds.es/seminario-humanidades-digitales-en-accion-herramientas-para-el-analisis-de-prensa-historica-en-espanol

[33] https://www.clariah.eus/eu/4.workshopa

[34] https://www.ucm.es/clariah-cm/ii-ciclo-clariah-cm-formacion-en-prensa

[35] https://www.clariah.es/en/node/50

[36] https://gresel-uned.hypotheses.org/jornada-ia-y-humanidades-digitales-para-la-prensa-historica

European efforts to build interoperable historical newspaper corpora.

4.7. III Xeira CLARIAH-GAL

The III Xeira CLARIAH-GAL (May 31, 20026) is an annual event organized by the Instituto da Lingua Galega and supported by CiTIUS. It is designed as a networking and dissemination space to connect Galician projects and research groups working in digital humanities, arts, social sciences, and language-technology supported research. This year, one of the talks will be devoted to PressMint.

4.8. First CLARIAH-ES Summer School

The first CLARIAH-ES Summer School, "Impulsando las Humanidades Digitales en la era de la IA generativa," to take place in June in Donostia-San Sebastian, will also devote space to historical press. The two-day course will feature a keynote talk on new LLM-based approaches for digital humanities using historical newspapers.

5. EOSC Mesh Use Case

CLARIN, DARIAH, HiTZ (CLARIAH-EUS), and the Universidad de Jaén (CLARIAH-AND) are all involved in the EOSC Mesh project. EOSC Mesh is a Horizon Europe project designed to reduce fragmentation in Europe's research data and service landscape by expanding the EOSC Federation with seven interoperable nodes working together as an EOSC Mesh Hub. Overall, its objective is to make EOSC more resilient and scalable, accelerate node onboarding via a Node Operator Framework, and enable large-scale discovery and reuse of research objects across domains.

Several use cases have been designed within the project. One, "Discovery, access and integration of data from historical newspapers," focuses on enabling datafication and computational analysis workflows for textual cultural heritage, specifically digitized historical newspapers. Part of the objective is to overcome several of the main obstacles with respect to interoperability caused by the disparate nature of the collections, such as discovering relevant content across multiple trusted repositories, gaining access, integrating heterogeneous sources, and selecting suitable tools/execution environments. This use case targets cultural heritage institutions (libraries, archives, museums) and humanities/DH researchers (estimated at 500,000 in Europe) and anticipates infrastructure needs of around 150 TB to store and process corpora alongside key knowledge bases (e.g., Wikidata and GeoNames). It also leverages EOSC Mesh core capabilities, generic capabilities (cloud compute, notebooks, online storage), and the SSHOC Marketplace as the thematic entry point, with development steps that include deploying workflows for new service creation and building federated Retrieval-Augmented Generation (RAG) systems over indexed datasets produced by the datafication pipelines.

6. Conclusion

PressMint provides a timely and structured response to the rapid growth of digitized historical newspapers in Europe. It seeks to make these historical newspaper collections comparable across countries and languages by pursuing interoperability by design. PressMint does so through a shared TEI-based schema, reusable scripts/workflows, and dissemination in multiple downstream formats to lower barriers for "distant reading" and large-scale reuse. Overall, PressMint is not merely as a corpus-building project, but also a shared European methodology for transforming historical newspapers into an interoperable network, unlocking the potential for new comparative research questions while encouraging collaboration among technologists, humanists, and cultural heritage institutions.

In this context, CLARIAH-ES contributes to Spain's involvement in PressMint by aligning partners, repositories, and technical practices across its distributed national infrastructure. The decision to develop several regional corpora has enabled teams to work with manageable collections while still converging on common standards: assembling corpora through partnerships with major digital libraries and experimenting with AI-based improvements to OCR and normalization. Across these efforts, the main technical challenge remains the heterogeneity and quality of legacy OCR, often produced many years ago, making robust correction, normalization layers, and consistent linguistic enrichment central to producing reliable, comparable corpora.

Ongoing experimentation with AI tools and techniques will ideally help improve transcription quality and scholarly utility, while dissemination activities will aid in increasing PressMint's uptake across library, humanities, and language technology communities. Finally, alignment with broader initiatives such as the EOSC Mesh use case on historical newspapers points toward scalable discovery, access, and federated computation over interoperable historical press data, supporting reproducible, cross-regional and cross-lingual research on Europe's cultural and political history at an unprecedented scale.

7. Acknowledgements

PressMint is funded by CLARIN ERIC. The CLARIAH-ES infrastructure is funded by the Ministry of Science, Innovation and Universities (RED2024-154077-E). EOSC Mesh is funded by the European Commission.

8. Bibliographical References

Jon Alkorta, Aritz Farwell, Joseba Fernandez de Landa, Begoña Altuna, Ainara Estarrona, Mikel Iruskieta, Xabier Arregi, Xabier Goenaga, Jose Mari Arriola, Inma Hernáez, and David Lindemann. 2025. CLARIAH-EUS: A strategic network helping basque country researchers to participate in european research infrastructures. In *Selected papers from the CLARIN Annual Conference 2024. Linköping Electronic Conference Proceedings 216 (eds. Vincent Vandeghinste and Thalassia Kontino)*.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Matyáš Kopp, Steinþór Steingrímsson, Sigrún Helgadóttir, Črtomir Grobol, et al. 2023. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57(1):415–448.

Tomaž Erjavec, Maciej Ogrodniczuk, Agnes Pisanski Peterlin, Simon Krek, and Andrej Pančur. 2025. Parlamint ii: advancing comparable parliamentary corpora across europe. *Language Resources and Evaluation*, 59(3):1–25.

Mikel Iruskieta, Ainara Estarrona, Aritz Stephen Farwell, and Germán Rigau. 2022. INTELE: promoviendo la participación en las infraestructuras eric clarin y dariah. *Boletín de la ANABAD*, 72(2):63–91.

Oldřich Kodym and Michal Hradiš. 2021. Page layout analysis system fornbsp;unconstrained historic documents. In *Document Analysis and Recognition – ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II*, page 492–506, Berlin, Heidelberg. Springer-Verlag.

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.

Francisco J. Carreras Riudavets, Ainara Estarrona, Aritz Farwell, Mikel Iruskieta, Manuel Marco Such, Maite Melero, Arturo Montejo-Ráez, Daniel Riaño, German Rigau, Dolores Romero, Salvador Ros, Elena Sánchez, and Xulio Sousa. 2024. CLARIAH-ES: Strategic network for the integration in the european research infrastructures in social sciences and humanities. In *SEPLN (Projects and Demonstrations)*, volume 3729 of *CEUR Workshop Proceedings*, pages 30–35. CEUR-WS.org.

Yves Scherrer and Nikola Ljubešić. 2016. Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 248–255.

Towards an Interoperable Corpus of Austrian Historical Newspapers: The case of PressMint-AT

Tanja Wissik, Jona Hassenbach, Hannes Pirker, Claudia Resch, Stefan Resch

Austrian Centre for Digital Humanities
Austrian Academy of Sciences, Vienna, Austria
{Tanja.Wissik, JonaMarie.Hassenbach, Hannes.Pirker, Claudia.Resch, Stefan.Resch}@oeaw.ac.at

Abstract

In this paper the PressMint-AT project is presented, which aims to create a historical newspaper corpus based on the *Wiener Abendpost*. The quality of automatic text recognition (ATR) is a key factor in creating historical newspaper corpora. Therefore, the performance of established ATR tools, multimodal large language models (LLMS), and existing full-text transcriptions provided by the Austrian National Library via ANNO is evaluated in order identify the most suitable approach for the *PressMint-AT* project. Even though recent research has demonstrated promising results for OCR tasks using multimodal LLMs, the experiments presented in this paper show, that PERO OCR achieves the best performance for the *PressMint-AT* dataset.

Keywords: historical newspaper corpora, OCR, multimodal LLMs

1. Introduction

Since the early 2000s, regional and national libraries, alongside transnational organisations and commercial providers, have invested substantially in the digitisation of historical newspapers converting newspaper content into machine-readable text by means of optical character recognition (OCR). As a result, millions of newspaper facsimiles, together with their corresponding textual transcriptions, have been produced at regional, national, and international scales (Ehrmann et al., 2023). These developments have led to numerous historical newspaper corpora (Fišer and Lenardič, 2018). Examples include corpora containing several newspapers such as *sPeriodika 1.0* (Dobranić et al., 2024) for Slovenia or the *Couranten Corpus 2.0* (2025) for Dutch, as well as corpora containing a specific newspaper such as the Corpus *Wienerisches Digitarium (Resch and Kampkaspar 2020*). However, several challenges remain regarding historical newspaper corpora: first, the varying quality of OCR-generated full texts provided by libraries, and second, the lack of interoperability between corpora resulting from differing encoding standards.

The CLARIN funded project *PressMint*[1], aims to address these issues by compiling a multilingual, comparable, annotated, translated and interoperable set of corpora of European historical newspapers from around the start of the 20th century. Within the Austrian sub-project, *PressMint-AT*, a corpus of the *Wiener Abendpost,* a supplement of the *Wiener Zeitung,* published from 1863 to 1921, will be created.

This paper examines and discusses different approaches (including the use of multimodal LLMs) that can contribute to producing automatic high-quality transcripts of Fraktur typeface. These transcriptions form the basis for subsequent processing steps within the *PressMint* project, such as part-of-speech tagging, named entity recognition (NER), entity linking, and TEI encoding.

2. Related Work on Austrian Historical Newspapers

In recent years, archives and libraries in Austria have made substantial progress in the digitisation of historical newspapers, providing full-text search functionality based on textual transcriptions generated through OCR. For example, the Austrian National Library is providing access to the full texts of 27 million pages from historical newspapers and magazines.[2] In practice, however, the quality of these full texts varies considerably, ranging from low-quality or “noisy” OCR output to manually corrected OCR results. Therefore, several research projects, working with historical newspapers, have created new full texts by applying different OCR approaches: The full text digitization project of the *Wiener Zeitung* between 1703 and 1798 (Resch, 2023), which trained a dedicated Transkribus model for Fraktur typeface (Resch and Kampkaspar, 2020). The CLARIAH-AT funded Esperanto Newspaper Excerpt project used the open-source software Tesseract OCR in order to create full texts of digitized newspaper excerpts about Esperanto, which are preserved in the Department of Planned Languages and Esperanto Museum of the Austrian National Library (Mayer, 2024). Also, the JobAds Project produced OCR and corrected several thousand job advertisements from 14 different newspapers between 1850 and1950 provided by ANNO (Venglarova et al., 2024). For the OCR task Tesseract with the frak2021 model (M. U. Library, 2021) was used and then the

[1] https://www.clarin.eu/pressmint

[2] https://www.onb.ac.at/en/

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 34–39

automatic OCR output was manually corrected within Transkribus.

As these examples demonstrate, most earlier projects used well-established tools or platforms for the transcription and processing of historical documents such as Transkribus or Tesseract. But recently, with the rise of multimodal LLMs, LLMs are also used for OCR tasks with promising results (Greif et al., 2025). Early research found that LLM-based text recognition often outperforms state-of-the-art pipelines. Multimodal LLMs often transcribe unseen manuscripts zero-shot for printed and even handwritten documents with better results than well-established tools like Transkribus (Humphries et al., 2024) or Tesseract (Kim et al., 2025).

However, using multimodal LLMs for automatic text recognition also has limitations, and poses challenges such as hallucinations (Li, 2024; Boros et al., 2024) and that their performance for English texts is better than for other languages (Corsillia et al., 2025).

Accordingly, this paper evaluates both established automatic text recognition tools and multimodal LLMs for OCR tasks using *PressMint-AT* data.

3. Source Material

As source material for the *PressMint-AT* corpus the *Wiener Abendpost* was selected. It was a newspaper supplement of the *Wiener Zeitung*, published and printed by the *Wiener Zeitung* between 1st July 1863 and 31st Dezember 1921. The *Wiener Abendpost* was published 6 days a week (except Sundays) and had on average between 4 and 8 pages. It could be subscribed to separately or together with the Wiener Zeitung. In terms of content, the *Wiener Abendpost* contained a daily news section, a feuilleton section and advertisements. The newspaper had a three-column layout and was written in Fraktur typeface (see Figure 1). The *Wiener Abendpost* is accessible via ANNO (AustriaN Newspaper Online)[3], a virtual reading room for digitized newspapers maintained by the Austrian National Library. There, the scanned images, as well as full-text transcriptions are available, with all the limitations of automatic OCR, created some years ago without manual corrections (Resch & Kampkaspar, 2019). Since the *Wiener Abendpost* was digitized as a supplement of the *Wiener Zeitung*, there are no separate entries in ANNO, the *Wiener Abendpost* pages are just part of the *Wiener Zeitung*. Consequently, no dedicated metadata for the *Wiener Abendpost* is available. Therefore, a separate workflow to extract the *Wiener Abendpost* pages, needed to be set up (see section 4.2). For the *PressMint-AT* Corpus more than 17,450 newspaper issues will be processed containing more than 90,300 pages.


Wiener Abendpost.

1917. — Nr. 296. Beilage zur Wiener Zeitung. Freitag, 28. Dezember.

Figure 1: First page of the *Wiener Abendpost* issue 296 from 28. December 1917 (ANNO/Österreichische Nationalbibliothek) made available by the Austrian National Library via ANNO[4].

4. Data Preparation

4.1 Ground Truth

The ground truth data consist of four issues of the *Wiener Abendpost*, published between 1914 and 1918, totalling 20 pages. Given the limited variation in print quality and newspaper layout across the source period, issues were selected largely based on their historical significance, while allowing for the inclusion of some lower-quality pages (e.g. a line through the text as shown in Figure 1).

The ground truth was created using Transkribus (READ-COOP SCE, n.d.). Layout and text regions were manually annotated, followed by automatic transcription using Transkribus' *Text Titan I* text recognition model. There was no additional fine-tuning of models, neither for layout nor for text recognition. Lastly, extensive manual correction was applied, in accordance with the following guidelines: The transcription aims to reproduce the original print as faithfully as possible in terms of pages, lines, and characters. The texts have undergone multiple rounds of collation, have been carefully checked for quality, and reflect the historical state of the language without alteration. The original typography has largely been preserved; that is, *u* and *v* remain unchanged, as does the alternation between Fraktur and Antiqua typefaces, as well as the use of bold and italic print - except for the distinction between round s and long s, and for ligatures for which no Unicode representation exists. Abbreviations in the text have not been expanded. Original line-end hyphenation was preserved. The ground truth is thus aligned at the line level, with each text line transcribed individually within

[3] https://anno.onb.ac.at/node/15

[4] https://anno.onb.ac.at/cgi-content/annoshow?call=wrz|19171228|17|100.0|0

annotated text regions. While higher-level layout elements (e.g., titles, paragraphs, headers) were annotated, they were not considered in the computational evaluation. The ground truth data have been made available in the project's GitHub repository[5].

4.2 Identification of relevant pages

The *Wiener Abendpost* was published as a supplement to the *Wiener Zeitung,* which is accessible via ANNO, but the metadata on ANNO lacks information on the exact location of the *Wiener Abendpost* within each issue of the *Wiener Zeitung* (see Section 3). Therefore, it is necessary to provide a method for reliably spotting the *Wiener Abendpos*t by identifying the pages, which mark the starts and end of the supplement.

A first attempt consisted of identifying the *Wiener Abendpost* on a textual basis, i.e. by searching the transcriptions of the *Wiener Zeitung* provided by ANNO for stable indicators. This approach was abandoned because the quality of the existing OCR was deemed too low.

Instead, a visual approach using image recognition and classification was taken. For a human observer it is simple to "spot" the relevant supplement by viewing the thumbnail representation of a complete *Wiener Zeitung* issue. The first page of the *Wiener Abendpost* is indicated by its prominent title area (see Figure 1). The last page of the *Wiener Abendpost* is either indicated by the title page of another supplement called *Amtsblatt* (official gazette), or just by the end of the whole issue. This capability was emulated by training two classifier models: one for spotting the title page for the *Wiener Abendpost* and one for the *Amtsblatt*. Transfer learning was applied using a pretrained ResNet-18 model from PyTorch's torchvision library, fine-tuned for binary page classification by replacing the final layer with a single sigmoid output. Training used stratified splits, BCEWithLogitsLoss with optional class-weight balancing, and early stopping via an Adam optimizer with adaptive learning rate scheduling. To provide the necessary training samples for this supervised method, a simple browser plugin was created which allows to manually point out the title pages of the relevant supplements directly on ANNO's webpage (see Figure 2). The same GUI is used for visualising classifier predictions and enable immediate correction of faulty results. The overall process performed very satisfactorily. With an initial training set of 50 positive examples, the error rate was already below 10%. With the visualisation and correction GUI it took minimal effort to increase the amount of training data. The classifiers reached 0% error rate with only 500 positive samples.

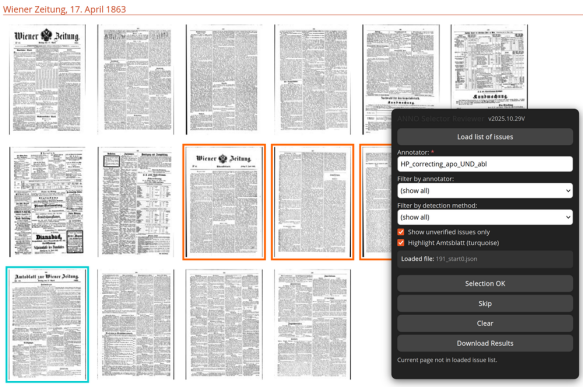


Figure 2: Screenshot of ANNO with a complete issue of the *Wiener Zeitung* and the *Wiener Abendpos*t Annotation plugin in place. Orange frames indicate the extent of the *Wiener Abendpos*t supplement, the turquoise frame depicts the title page of the *Amtsblatt* supplement.

5. Experiments

Different automatic text recognition tools, as well as several large language models (LLMs), were compared to evaluate their suitability for historical newspaper transcription. The systems evaluated include the OCR tools Google Cloud Vision OCR, Churro OCR, (Semnani et al., 2025), dots.ocr (Li et al, 2025), and PERO OCR (Kodym & Hradiš, 2021) using the German Fraktur modell as well as LLMs such as Anthropic's Claude Sonnet 4, DeepSeek, OpenAI's GPT-4o, and Google Gemini. A zero-shot procedure was employed, where the entire page image was provided as a single input. This approach maximises the context information but risks hallucination (Levchenko, 2025). The original, uncorrected Transkribus output obtained during ground truth creation (see Section 4.1), and the transcriptions provided by ANNO, the digital newspaper database of the Austrian National Library, were likewise included in the comparison.

Multiple prompts were tested, including system prompts, if available, (e.g. for dots.ocr or Churro OCR), and longer, hand-crafted prompts in both German (e.g. "Das ist ein Scan einer deutschen historischen Zeitung aus dem frühen 20. Jahrhundert. Bitte führe OCR darauf aus, also extrahiere den Text und behalte dabei die Leserichtung bei. Beachte auch, dass die Schrift in Fraktur gehalten ist. Der Output soll nur der Text alleine sein"), and in English ("This is a scan of a historic german newspaper from the early 20th century. Please do OCR on it, extract all the text and keep the reading order. Also keep in mind that the writing is in german 'Fraktur'. The output should only be text."), as the language of the prompt might have an influence on model performance (Kmainasi et al. 2025). All different

[5] https://github.com/acdh-oeaw/pressmint-OCR-AI-evaluation/tree/main/data/texts/transkribus_corrected

prompts used are documented in the open GitHub repository[6].

For evaluation, standard OCR metrics were employed, namely Character Error Rate (CER) and Word Error Rate (WER), both of which are based on the Levenshtein distance. To complement these metrics, the similarity between continuous text sequences was assessed using the Ratcliff/Obershelp pattern recognition algorithm (Ratcliff and Metzener, 1988), here referred to as DIFFLIB. Quantitative evaluation was further supplemented by qualitative error analysis focusing on recurring challenges in Fraktur script.

SYSTEM	WER	CER	DIFFLIB
PERO	0.13	0.09	0.94
DOTS.OCR_1	0.38	0.11	0.91
DOTS.OCR_4	0.4	0.14	0.89
ANNO	0.35	0.17	0.88
TRANSKRIBUS	0.37	0.25	0.85

Table 1: Scores per workflow, ranked by mean performance across all three evaluation metrics. Highest score per metric is displayed in bold.

Results for the five highest-performing systems are presented in Table 1. All system–prompt combinations with a CER above 0.25 were excluded from further investigation. However, the complete set of results is documented in the GitHub repository[7]. The remaining scores include PERO OCR, dots.ocr with different system prompts (system prompt 1 and system prompt 4), the existing ANNO transcriptions, and the automatically generated Transkribus output. Across all three metrics, PERO OCR achieved the highest performance, followed closely by both dots.ocr systems, the existing transcriptions published in ANNO, and the automatically generated Transkribus output from the automatic text recognition feature (without manual corrections). All remaining systems underperformed relative to these results, exhibiting widely varying levels of error. Consistent with these findings, the four top-performing outputs showed low variance across all metrics, indicating stable and reliable performance at their respective levels.

Additional qualitative analysis of the data obtained through PERO OCR, dots.ocr, ANNO, and Transkribus further corroborated the observed pattern. In the case of Transkribus, the primary source of error consists of column crossings, i.e., instances in which the system fails to follow the correct reading order and instead merges lines from different columns into a single line. This issue was not observed in the other three highest-performing systems. In these cases, differences in performance are almost entirely attributable to word- and character-level OCR quality: PERO OCR produced the cleanest transcription; dots.ocr contained some misspellings typical of OCR for Fraktur texts, such as confusion between *f* and the Fraktur long *s*, or the faulty transcription of *tz* as *ß*, *b*, *z*, *ft*, *szt*, or *gt*; the transcriptions provided by ANNO contained numerous non-systematic misspellings, often involving non-alphabetic characters (e.g., #, », «, ‘), rendering some words unrecognizable.

In conclusion, its superior performance, together with its computational efficiency, renders PERO OCR in combination with the German Fraktur model the most suitable system for the task at hand; it was therefore selected for further experimentation.

6. Future Work

The next step will involve generating automatic full-text transcriptions for all *Wiener Abendpost* issues using PERO OCR, followed by TEI encoding in accordance with the *PressMint* schema and subsequent data processing (e.g., POS tagging and semantic enrichment) as outlined in the *PressMint* project work plan.

7. Conclusion

In this paper we described the *PressMint-AT* project, the Austrian subproject within *PressMint,* that aims at creating an Austrian historial newaspaper corpus for the *Wiener Abendpost*. The submission discusses and evaluates different approaches for creating full-text transcription via ORC, including methods based on LLMs. We compared several OCR tools such as Google Cloud Vision OCR, Churro OCR, dots.ocr, PERO OCR, Transkribus (without manual corrections) as well as generic LLMs such as Anthropic’s Claude Sonnet 4, DeepSeek, OpenAI’s GPT-4o, and Google Gemini alongside the already existing transcriptions provided by ANNO. Among these approaches, PERO OCR achieved the best results for our dataset. Considering both transcription quality and computational efficiency, PERO OCR appears to be the most suitable system for the OCR task within the *PressMint-AT* project.

8. Acknowledgments

The submission was supported by the PressMint CLARIN Flagship Project and CLARIAH-AT.

[6] https://github.com/acdh-oeaw/pressmint-OCR-AI-evaluation

[7] https://github.com/acdh-oeaw/pressmint-OCR-AI-evaluation/tree/main?tab=readme-ov-file#comparison-results-plot

9. Bibliographical References

Boros, E., Ehrmann, M., Romanello, M., Najem-Meyer, S. and Kaplan, F. (2024). Postcorrection of historical text transcripts with large language models: An exploratory study. In Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), pages 133–159, St. Julians, Malta. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.latechclfl-1.14.

Crosilla, G., Klic, L. and Colavizza, G. (2025). Benchmarking large language models for handwritten text recognition. https://arxiv.org/pdf/2503.15195.

Ehrmann, M, Bunout, E. and Clavert, F. (2023). "Digitised Historical Newspapers: A Changing Research Landscape". Digitised Newspapers – A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology, edited by Estelle Bunout, Maud Ehrmann and Frédéric Clavert, Berlin, Boston: De Gruyter Oldenbourg, 1-22. https://doi.org/10.1515/9783110729214-001.

Fišer, D. and Lenardič, J. (2018). CLARIN Resources Families / Newspaper Corpora. https://www.clarin.eu/resource-families/newspaper-corpora.

Humphries, M., Leddy, L. C., Downton, Q., Legace, M., McConnell, J., Murray, I. and Spence, E. (2024). Unlocking the archives: Using large language models to transcribe handwritten historical documents.

Kmainasi, M.B., Khan, R., Shahroor, A.E., Bendou, B., Hasanain, M. and Alam, F. (2025). Native vs Non-native Language Prompting: A Comparative Analysis. In: Barhamgi, M., Wang, H., Wang, X. (eds) Web Information Systems Engineering – WISE 2024. WISE 2024. Lecture Notes in Computer Science, vol 15440. Singapore: Springer. https://doi.org/10.1007/978-981-96-0576-7_30.

Kim, S., Baudru, J., Ryckbosch, W., Bersini, H. and Ginis, V. (2025). Early evidence of how llms outperform traditional systems on ocr/htr tasks for historical records. https://arxiv.org/abs/2501.11623.

Li, L. (2024). Handwriting Recognition in Historical Documents with Multimodal LLM. In CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark https://arxiv.org/html/2410.24034v1.

Li, Y., Yang, G., Liu, H., Wang, B. and Zhang, C. (2025). *dots.ocr: Multilingual document layout parsing in a single vision-language model*. arXiv. https://arxiv.org/abs/2512.02498.

Mayer, S. (2024). Reviving History: Reconstructing Esperanto Newspaper Excerpts from the Hachette Collection (1898-1915). ESF Connected https://esfconnected.org/2024/10/14/reviving-history-reconstructing-esperanto-newspaper-excerpts-from-the-hachette-collection-1898-1915/.

Ratcliff, John W. and Metzener, D. (1988). "Pattern Matching: The Gestalt Approach". Dr. Dobb's Journal (46).

READ-COOP SCE. (n.d.) Transkribus. Innsbruck: READ-COOP SCE, [30.07.2025]. https://transkribus.eu/.

Resch, C. (2023). Volltextoptimierung für die historische Wiener Zeitung Mit einem Anwendungsszenario aus der germanistischen Sprachgeschichte. In Bunout, Ehrmann, M., Clavert, F. (Eds.), Digitised Newspapers – A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology. Berlin, Boston: De Gruyter, 89-111. https://doi.org/10.1515/9783110729214.

Resch, C. and Kampkaspar D. (2019). DIGITARIUM – Unlocking the Treasure Trove of 18th-Century Newspapers for Digital Times. In: Wallnig, T, Romberg M. and Weis J. (Eds.): Digital Eighteenth Century: Central European Perspectives. Wien / Köln / Weimar: Böhlau Verlag, 49-64.

Semnani, S., Han Zhang, H., Xinyan He, X., Tekgurler, M. and Lam, M. (2025). CHURRO: Making History Readable with an Open-Weight Large Vision-Language Model for High-Accuracy, Low-Cost Historical Text Recognition. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34777–34824, Suzhou, China. Association for Computational Linguistics.

Kodym, O. and Hradiš, M. (2021). Page Layout Analysis System for Unconstrained Historic Documents. ICDAR, https://arxiv.org/abs/2102.11838.

Greif, G., Griesshaber, N. and Greif, R. (2025). Multimodal LLMs for OCR, OCR Post-Correction, and Named Entity Recognition in Historical Documents. https://arxiv.org/abs/2504.00414

Venglarova, K., Adam, R., Mölzer, W., Balasubramanian, S., Kleinert, J., Füllsack, M. and Vogeler, G. (2024). Who Advertises in Newspapers? Data Criticism in Mining Historical Job Ads. In: Proceedings of the Computational Humanities Research Conference 2024. Aarhus, Denmark. CEUR. 2024. 788-801.

10. Language Resource References

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

Dobranić, F., Evkovski, B. and Ljubešić, N. (2024). A Lightweight Approach to a Giga-Corpus of Historical Periodicals: The Story of a Slovenian Historical Newspaper Collection. LREC-COLING 2024.

Resch, C. and Kampkaspar, D. (Eds) (2020). Wienerisches DIGITARIUM - Digitale Ausgabe des "Wien[n]erischen Diarium" (322 Ausgaben im Volltext).
Resch, C. and Kampkaspar, D. (Eds) (2022). German Fraktur 18th Century – WrDiarium_M9. https://www.transkribus.org/models/german-fraktur-18th-century.
Weil, S. (2021). Tesseract OCR models for historic prints based on Latin script. Zenodo. https://doi.org/10.5281/zenodo.10125246.

A Growing Literature of the Public Sphere: Fiction in Danish Newspapers (1666-1850)

Pascale Feldkamp*, Alie Lassche*, Rie Eriksen*, Kit Morgenstjerne*, Johan Heinsen‡, Kristoffer Nielbo*, Yuri Bizzoni*

*Center for Humanities Computing, Aarhus University, ‡MASSHINE, Aalborg University
{pascale.feldkamp, a.w.lassche, yuri.bizzoni}@cas.au.dk, heinsen@dps.aau.dk

Abstract

Digitized literary corpora of the 19th century largely focus on standalone volumes, sidelining the broader and more diverse literary production of the period. Fiction published in less enduring formats – such as novellas and serialized pieces in newspapers – remains underexplored, particularly for low-resource languages like Danish, despite the growing availability of digitized newspaper archives. This paper addresses that gap by identifying and tagging fiction in Danish newspapers (1666–1850). We (1) present a manually annotated dataset of 1,831 articles with both binary (fiction/nonfiction) and fine-grained subcategories (travelogue, biography, essay), and (2) evaluate a document-embedding classifier that achieves an F1-score of up to 0.89 for the fiction/nonfiction distinction. Building on this pipeline, we further provide two resources for future research: (a) fiction probability scores for nearly five million newspaper articles ($n = 4,898,084$), and (b) a small, cleaned, and curated subset of newspaper fiction ($n = 139$), intended as a growing resource.

Keywords: literature, historical newspapers, embeddings, serialized fiction

1. Introduction and Related Works

Only a small fraction of novels are widely studied, while most of literary production remains what Franco Moretti (2000) famously called "the great unread". Recent computational and statistical approaches have begun to expand the literary horizons (Underwood, 2019; Algee-Hewitt et al., 2016), examining lesser-known texts and opening the way for a kind of literary sociology that traces the circulation and diversity of literary production.

However, this ambition falls short in practice. In most digitized corpora – especially for under-resourced languages like Danish – the standalone novel remains the dominant form, overshadowing the diverse genres and formats that animated the 19th-century print market (Stangerup, 1936).[1] As the literary market consolidated (Bourdieu, 1996), much literary culture circulated through ephemeral media: newspapers, periodicals, and serials. The 19th century marked a turning point: literature became a mass-cultural product in the everyday media landscape (Easley, 2024; Horstbøll, 1999). While newspapers were central to the formation of the public sphere and the emerging nation-state (Habermas, 1989; Anderson, 2006), they also provided a venue for literary experimentation. Serialized fiction circulated across and beyond national borders, connecting readers, creating new genres, and attracting subscribers (Lehrmann, 2018). But despite some recent digital initiatives[2], there are few available datasets of these more impermanent forms of 19th-century literature. This is a critical gap. When writing literary history, we lose a great deal of information about the real scope of literary production, but we also lose an opportunity to understand literature's role in public life during this formative period of the emerging public sphere and nation-state. If most corpora overlook the transient life of literature, newspapers offer a way to restore it. Luckily, recent years have seen a significant digitization effort for historical newspapers. Detecting fiction directly within them allows us to glimpse the everyday circulation of stories that once animated public life and to recover lost dimensions of 19th-century literary culture.

Recent efforts have effectively distinguished fiction and nonfiction (Qureshi et al., 2019), also in the noisy environments of historical newspapers (Feldkamp et al., 2025; Repo, 2024). That is despite the task being theoretically complex: the distinction between fiction and nonfiction is notoriously difficult to pin down, with some arguing it depends more on reader framing than textual profile (Culler, 2002; Fish, 2003; Stockwell, 2002). This ambiguity is especially pronounced in historical sources: 19th-century literature and journalism competed to depict social reality and assert social truths (Lepenies

[1] Many corpora index novels published as standalone volumes, such as the Chicago Corpus, the ELTEC corpora, or Common Library 1.0; and, for Danish, the MeMo corpus (Bjerring-Hansen et al., 2022). In contrast, computational studies of *modern* (and English) literature focus on exploring alternative, impermanent forms such as Wattpad and fanfiction (Pianzola et al., 2020; Jacobsen and Kristensen-McLachlan, 2025).

[2] For instance, the Ciphers project: https://libraryponders.github.io/index.html

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 40–49

and Plard, 1995), while the modern ideal of journalistic objectivity emerged gradually (Schudson, 2001). Writers like Zola moved between literary and journalistic modes, and narrative techniques circulated across registers, with 'realism' shaping early novelistic forms (Watt, 2001). Furthermore, authorization strategies – claims of eyewitness accounts, discoveries of hidden documents, and use of documentary conventions – have long been used to enhance verisimilitude in fiction (Panayotakis et al., 2010). Nevertheless, linguistic research identifies features distinguishing fiction from nonfiction, including adjective/adverb ratios, pronouns, type-token ratios, nominalizations, and syntactic complexity, with nonfiction generally exhibiting higher information density (Qureshi et al., 2019; Kazmi et al., 2022; Kubát and Milička, 2013; Sadeghi and Dilmaghani, 2013; Vicente et al., 2021; Dijk, 2009). Recent studies show that semantic document embeddings outperform surface and sentiment features in detecting narrative segments (Repo, 2024; Laippala et al., 2019; Feldkamp et al., 2025), suggesting that fiction's profile is not only stylistic or sentiment-based but also semantic.

Here, we present a pipeline for extracting fiction from noisy historical Danish newspapers using text embeddings, along with a small, curated set of serialized fiction as a proof of concept.[3]

2. Methods

2.1. Data

2.1.1. Newspapers

The corpus used in this study consists of Danish newspapers published in the conglomerate state of Denmark-Norway between 1666 and 1850.

Due to OCR difficulties with *fraktur*, thin paper, and complex layouts, the corpus was re-digitized by the ENO group at Aalborg University[4] using custom transcription models in Transkribus (Kahle et al., 2017). Articles were segmented via line-level classification using `SetFit` and `RandomForest` models. The resulting corpus spans 28 Danish newspapers, printed in both Danish and Norwegian towns, covering both center and periphery, with a total of almost five million individual articles.

Notably, the corpus includes early examples of children's literature in the *Adresseavis for Børn* (1779-1782).[5] This was the first Danish newspaper explicitly aimed at a young readership, featuring a variety of genres including moral tales, travelogues, didactic essays, letters, and brief news items. While most of the newspaper's content consists of short pieces, a smaller part follows a serial model, with stories spanning several editions. We hypothesize that the inclusion of *Adresseavis for Børn* broadens the stylistic and social range of the dataset, as it may exhibit a distinct topical and linguistic profile compared to the rest of the corpus.

2.1.2. Annotated set

The articles for annotation were partly randomly selected from the entire period, and partly gathered to capture serialized novels, with batches of fiction and nonfiction identified using search terms such as "to be continued".[6] Each article was tagged by at least two expert annotators.[7] In the annotated set, we tagged both the coarse fiction/nonfiction distinction and fine-grained categories ('biography', 'travelogue', 'essay', 'poem', 'anecdote', 'narrative nonfiction', 'play'). To focus on the main distinction, we excluded hybrid or distinct forms such as 'narrative nonfiction', 'anecdote', 'poetry', and 'play'. Still, to maintain task complexity, we retained subcategories 'biography' and 'travelogue' that occur under both fiction and nonfiction, as annotators could reliably distinguish the higher-level register.[8] Only articles with agreement from at least two annotators were kept, resulting in 1,962 tagged articles, of which 1,831 exceed 20 words. They span 1759–1874.[9]

When clear story signals – titles, tags, prologues, formatting – were present in the inspected newspaper scans, texts were annotated as fiction; otherwise, annotation focused on distinguishing registers. Given the fluid boundary between early news and fiction, fiction is arguably better treated as a register rather than a fixed category (Repo, 2024), consistent with literary theory that frames literariness as a mode of communication rather than a bounded product (Jakobson, 1981; Bachtin et al., 2014; Jauss and Benzinger, 1970). So in cases where there was no clear signal but where, for ex-

[3] The anonymized repository for the code underlying this paper is available here: `https://github.com/centre-for-humanities-computing/feuilleton_dataset`, including links to all resources presented in this paper.

[4] `https://hislab.quarto.pub/eno/`.

[5] Later called *Avis for Børn* (Newspaper for Children).

[6] Since the pipeline aims to identify literary segments as they appear in practice, we retained editorial markers like "to be continued", as they likely reflect conditions of downstream application.

[7] We had three annotators, two with a university background in Literary Studies and one in the Study of Religion. The annotators were trained in close reading and had comprehensive knowledge of textual culture in 18th and 19th-century Denmark.

[8] The annotated set is available here: `https://huggingface.co/datasets/chcaa/fiction-nonfiction-testset`

[9] This period reflects peak newspaper output, not annotation choices.

Features	Class	Precision	Recall	F1-score
TF-IDF (max. 5,000)	*Fiction*	0.89 ± 0.03	0.85 ± 0.06	0.87 ± 0.02
	Nonfiction	0.87 ± 0.04	0.91 ± 0.04	0.88 ± 0.01
Document embeddings (768 dimensions)	*Fiction*	0.88 ± 0.02	0.88 ± 0.05	0.88 ± 0.03
	Nonfiction	0.89 ± 0.04	0.89 ± 0.02	0.89 ± 0.02

Table 1: Average classification performance and SD across 5 folds. Precision, recall, and F1-score are reported per class.

Category	N articles	N words
All	1,831	569,543
Nonfiction	951	225,146
Fiction	880	344,397
Biography	170	66,017
Travelogue	80	28,786
Essay	78	32,285

Table 2: Description of the annotated set (1759-1874). Note that these are the numbers after we removed articles with less than 20 words. A lot of fiction articles have the general fiction tag.

ample, a biographical text leaned heavily on literary devices (such as first-person or internal focalization, storylines, tropes, etc.) that were characteristic for fiction in the period, we annotated it as fiction. This is complicated by article fragmentation, which was overcome by filtering out short texts and, finally, on annotator agreement.

2.2. Classification

Embeddings: We tested six embedding models for the task of differentiating fiction and nonfiction against a baseline of TF-IDF representations (max. 5,000 features), summarized in Appendix subsection 3.5. The best performing model was `Old_News_Segmentation_SBERT_V0.1`,[10] which is why we used it for making document embeddings.[11]

Model: Some fiction texts (longer running pieces) appeared as multiple article fragments or feuilletons serialized across editions. We assigned a unique ID to each serialized piece or fragment and used dummy values for missing or incomplete IDs to maintain consistent grouping. For evaluation, we applied a `StratifiedGroupKFold` cross-validation scheme that preserved both class balance and ID integrity – ensuring that fragments or installments of the same piece were never split between train and test sets. Using 5-fold cross-validation, we trained a `LogisticRegression` classifier with balanced class-weight to account for label imbalance.[12] The classifier was trained on each fold and evaluated on the held-out set, computing precision, recall, and F1-scores per class and averaging across folds. This procedure was applied to both TF-IDF features and embeddings. Results are shown in Table 1. Note how closely embeddings and TF-IDF perform, suggesting that both semantic and lexical patterns are informative for the fiction/nonfiction distinction.[13]

3. Results

3.1. Tagged corpus

We applied the best-performing (embedding-based) model to all articles in the newspaper corpus (1677-1849) and assigned fiction probability scores. To get a sense of the distribution of fiction tags across the corpus and how the makeup of the newspaper landscape (high/low heterogeneity) changes it, we plotted percentages of fiction tags across the period (Figure 1).

Importantly, among articles tagged as fiction ($n = 77,513$; probability > 0.5), predicted probabilities show no correlation with publication date or text length.[14] While fiction tags appear more common in earlier newspapers (see Figure 1), this is unlikely to reflect model bias, as the training set included no examples before 1750. The spike in the 1680s is better explained by the newspaper *Danske Mercurius* (the first red spike in Figure 1), which presented news in flowing alexandrines accompanied by short poetic reflections. Upon manual inspection, it seems our classifier assigns high fiction probability to poetry, but also tended to misclassify obituaries as fiction, perhaps for their distinct poetic and emotional tone. The misclassifications

[10] https://huggingface.co/JohanHeinsen/Old_News_Segmentation_SBERT_V0. It was developed on the same Danish historical newspaper corpus for article segmentation.

[11] Procedure as outlined for the model comparison in Appendix subsection 3.5: texts were split into max input chunks, and a mean embedding was computed by averaging chunk embeddings.

[12] Stratification and classifier from Scikit-learn.

[13] Downscaling TF-IDF to 768 features or selecting the top 768 using `SelectKBest` with a chi-squared (χ^2) test yields similar results, indicating that TF-IDF performance is not simply due to its larger feature space.

[14] Spearman's $\rho < 0.07$, $p > 0.5$.

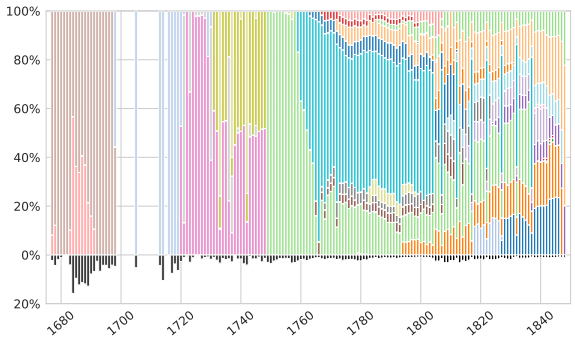


Figure 1: Distribution of newspapers as a percentage of total articles per year (top, in color) and histogram of the share of articles labeled fiction per year (bottom). Fiction is slightly more prominent in earlier newspapers, but overall remains low and stable (average 1.47%), despite a more heterogeneous publishing landscape.

highlight both a limitation and a strength of our approach to annotating fiction as a matter of literary register. While the model foregoes publishing categories, it picks up on linguistic mode – stylistic and tonal cues that cross formal category boundaries. Remember that we purposely excluded poetry and hybrid categories like anecdotes and narrative nonfiction. As such, rather than pointing to the model, these results seem to tell us something about formal categories, e.g., that poetry or obituaries appear in a fiction register.

3.2. Curated dataset

To better understand serialized fiction and ensure the reliability of story-level analyses, we created a curated dataset by manually inspecting a subset of high-probability predictions (probability > 0.99). The threshold was chosen to maximize confidence in the classifier's assignments, since fully automated grouping across installments and article fragments remains challenging. From this subset, we collected all fragments of the same story across the corpus and grouped serialized fiction into installments (a, b, c, etc.), assigning each story a unique ID. The resulting dataset contains 139 unique IDs, of which 69 consist of multiple installments (see Table 3). Additional cleaning and tagging ensured stories were spell-checked and assigned to general or more specific categories.[15] Some categories, like installments from the same series or travelogue and biography, show clustering (see Appendix Figure 3). Children's literature, by contrast, is widely dispersed, indicating it does not form a semantically consistent category in our data. While relatively small, this curated set serves as a proof-of-concept resource, allowing us to explore serialized fiction at the story level and providing a foundation for future, larger-scale expansions.

General	
N words	361,344
N articles	266
N stories	139
N stories $>$ 1 part	69
Stories per category	
Biography	35
Travelogue	18
Lovestory	12
Children's literature	17
Dialogue	2
Satire	2

Table 3: Description of the curated set (1763-1841).

3.3. Conclusion and Discussion

Our classification results suggest that fiction and nonfiction in newspapers differ in both word-usage patterns (TF–IDF) and semantic representations. In fact, it is certainly possible to distinguish between the two, even in short texts: probabilities show no correlation with length, suggesting that short literary fragments are just as recognizably "fictional" as longer ones. This makes the approach promising for other newspaper datasets. Although the share of fiction in any given year is modest, it's share is stable across time and still represents a substantial body of material ($\sim$77,500 articles).

The curated dataset illustrates both the feasibility of extracting literary corpora from historical newspapers and the character of the literary culture they contain, shaped by translation, transmission, and republication. Many stories are translations – from German, English, and French – and are frequently republished across newspapers, pointing to active (transnational) exchange. While Danish novelistic literature may have remained largely insular, newspaper fiction participated in wider literary circuits (Lehrmann, 2018). Authorship and transmission were often fluid: names are frequently pseudonymized, abbreviated, or absent, and the provenance of many texts is uncertain. For example, *The Maidens War* derives from a German source claiming Latin origins, while *The Labyrinth* allegedly comes from Zend, though the originality of such texts was questioned even at the time (Rask, 1826). This anonymity and cross-border mobility reflect the ephemeral, fast-moving nature of the medium. Where authorship can be identified, ca. 20% of names are female, matching proportions in contemporaneous novel publication (Degn et al., 2025), though translators – often women – were frequently uncredited (Nøding, 2017).

[15] This dataset is available here: https://huggingface.co/datasets/chcaa/Press-and-Plot

Unlike long, standalone volumes, these stories seem designed to move across authors, languages, and borders as smoothly as possible. Transmission and republication strategies reveal a highly dynamic, transnational literary space that both connected and shaped the emerging public sphere. Future research could trace these trajectories to better understand how fiction contributed to cultural exchange, reading practices, and the formation of modern national literary life.

Ethical considerations & limitations

Several limitations should be noted. First, annotator subjectivity remains a factor: even when depending on agreement and textual signals in the original scans, the fiction/nonfiction distinction is partly interpretive, reflecting the fluidity of literary registers in historical newspapers. Second, temporal coverage is restricted to 1666–1850; textual profiles may differ before and after this period. Notably, fiction reportedly increased toward the late 19th century, coinciding with the expansion of printed mass communication and the literary market (Feldkamp et al., 2024; Horstbøll, 1999). Future extraction of fiction in late-19th-century newspapers could therefore yield particularly rich literary corpora.

Historical newspapers reflect the socio-political biases of their time, including gender, class, and colonial perspectives. At the textual level, language referring to marginalized groups does not meet modern standards and should be understood in its historical context. At the production level, many female translators remain uncredited (Nøding, 2017), reflecting persistent inequities in attribution. While the dataset is historical and publicly available – minimizing privacy/copyright concerns – researchers should engage with it carefully, remaining attentive to both historical context and representational biases.

Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser. 2016. *Canon/Archive. Large-scale Dynamics in the Literary Field*. Stanford Literary Lab.

Benedict Anderson. 2006. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso.

Michail Michajlovič Bachtin, Michael Holquist, and Vern McGee. 2014. *Speech Genres and Other Late Essays*. University of Texas Press, Austin.

Jens Bjerring-Hansen, Ross Deans Kristensen-McLachlan, Philip Diderichsen, and Dorte Haltrup Hansen. 2022. Mending Fractured Texts. A heuristic procedure for correcting OCR data: 6th Digital Humanities in the Nordic and Baltic Countries Conference, DHNB 2022. In *CEUR Workshop Proceedings*, volume 3232, pages 177–186, Uppsala, Sweden.

Pierre Bourdieu. 1996. *The Rules of Art: Genesis and Structure of the Literary Field*. Stanford University Press.

Jonathan D. Culler. 2002. Literary competence. In *Structuralist poetics: structuralism, linguistics and the study of literature*, pages 131–152. Routledge, London. OCLC: 56560333.

Kirstine Nielsen Degn, Jens Bjerring-Hansen, Ali Al-Laith, and Daniel Hershcovich. 2025. Unhappy Texts?: A Gendered and Computational Rereading of the Modern Breakthrough. *Scandinavian Studies*, 97(1):1–24.

Teun A. van Dijk. 2009. *News as discourse*. Routledge, New York. OCLC: 868975895.

Alexis Easley, editor. 2024. *British writers, popular literature and new media innovation, 1820-45*. Nineteenth-century and neo-Victorian cultures. Edinburgh University Press, Edinburgh.

Pascale Feldkamp, Alie Lassche, Katrine Frøkjær Baunvig, Kristoffer Nielbo, and Yuri Bizzoni. 2025. Fact from fiction: Finding serialized novels in newspapers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 695–707, Vienna, Austria. Association for Computational Linguistics.

Pascale Feldkamp, Alie Lassche, Jan Kostkan, Márton Kardos, Kenneth Enevoldsen, Katrine Baunvig, and Kristoffer Nielbo. 2024. Canonical status and literary influence: A comparative study of Danish novels from the modern breakthrough (1870–1900). In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 140–155, Miami, USA. Association for Computational Linguistics.

Stanley Eugene Fish. 2003. *Is there a text in this class? the authority of interpretive communities*, 12. print edition. Harvard Univ. Press, Cambridge, Mass.

Jürgen Habermas. 1989. *The structural transformation of the public sphere : an inquiry into a category of bourgeois society*. MIT Press.

Henrik Horstbøll. 1999. *Menigmands medie: det folkelige bogtryk i Danmark 1500 - 1840: en kulturhistorisk undersøgelse*. Number 19 in Danish Humanist Texts and studies. Det Kongelige

Bibliotek, Museum Tusculanums Forlag, København.

Mia Jacobsen and Ross Deans Kristensen-McLachlan. 2025. Beyond Style: Rethinking Computational Fanfiction Research. *Journal of Data Mining & Digital Humanities*, NLP4DH:16414.

Roman Jakobson. 1981. Linguistics and poetics. In *Linguistics and Poetics*, pages 18–51. De Gruyter Mouton.

Hans Robert Jauss and Elizabeth Benzinger. 1970. Literary History as a Challenge to Literary Theory. *New Literary History*, 2(1):7.

P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger. 2017. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pages 19–24.

Arman Kazmi, Sidharth Ranjan, Arpit Sharma, and Rajakrishnan Rajkumar. 2022. Linguistically Motivated Features for Classifying Shorter Text into Fiction and Non-Fiction Genre. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 922–937, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Miroslav Kubát and Jiří Milička. 2013. Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics*, 20(4):339–349.

Veronika Laippala, Roosa Kyllönen, Jesse Egbert, Douglas Biber, and Sampo Pyysalo. 2019. Toward multilingual identification of online registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 292–297, Turku, Finland. Linköping University Electronic Press.

Ulrik Lehrmann. 2018. Føljetonromanen og dansk mysterie-litteratur i 1800-tallet. *Passage - Tidsskrift for litteratur og kritik*, 33(79):31–46. Number: 79.

Wolf Lepenies and Henri Plard. 1995. *Les trois cultures - entre science et littérature, l'avènement de la sociologie*, 0 edition edition. MSH PARIS, Paris.

Franco Moretti. 2000. The Slaughterhouse of Literature. *Modern Language Quarterly*, 61(1):207–228.

Aina Nøding. 2017. *Periodical Fiction in Denmark and Norway before 1900*. Oxford University Press.

Stelios Panayotakis, Maaike Zimmerman, and Wytse Hette Keulen, editors. 2010. *The ancient novel and beyond*. Number 241 in Mnemosyne, bibliotheca classica Batava 0169-8958. Supplementum. Brill, Leiden, Netherlands Boston.

Federico Pianzola, Simone Rebora, and Gerhard Lauer. 2020. Wattpad as a resource for literary studies. Quantitative and qualitative examples of the importance of digital social reading and readers' comments in the margins. *PLOS ONE*, 15(1):e0226708. Publisher: Public Library of Science.

Mohammed Rameez Qureshi, Sidharth Ranjan, Rajakrishnan Rajkumar, and Kushal Shah. 2019. A simple approach to classify fictional and non-fictional genres. In *Proceedings of the Second Workshop on Storytelling*, pages 81–89, Florence, Italy. Association for Computational Linguistics.

Rasmus Rask. 1826. *Om Zendsprogets og Zendavestas Ælde og Ægthed*. Andreas Seidelin.

Liina Repo. 2024. Towards automatic register classification in unrestricted databases of historical English. In *Linguistics across Disciplinary Borders: The March of Data*, 1 edition, pages 97–126. Bloomsbury Publishing Plc.

Karim Sadeghi and Sholeh Karvani Dilmaghani. 2013. The Relationship between Lexical Diversity and Genre in Iranian EFL Learners' Writings. *Journal of Language Teaching and Research*, 4(2):328–334.

Michael Schudson. 2001. The objectivity norm in American journalism. *Journalism*, 2(2):149–170. Publisher: SAGE Publications.

Hakon Stangerup. 1936. *Romanen i Danmark: Romanen i det Attende Århundrede*. Levin & Munksgaards Forlag.

Peter Stockwell. 2002. *Cognitive poetics: an introduction*. Routledge, London.

Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, Chicago, IL.

Marta Vicente, María Miró Maestre, Elena Lloret, and Armando Suárez Cueto. 2021. Leveraging Machine Learning to Explain the Nature of Written Genres. *IEEE Access*, 9:24705–24726.

Ian Watt. 2001. *Rise of the Novel, Updated Edition*. University of California Press, Berkeley, CA.

Appendix A

3.4. Distribution of articles over time

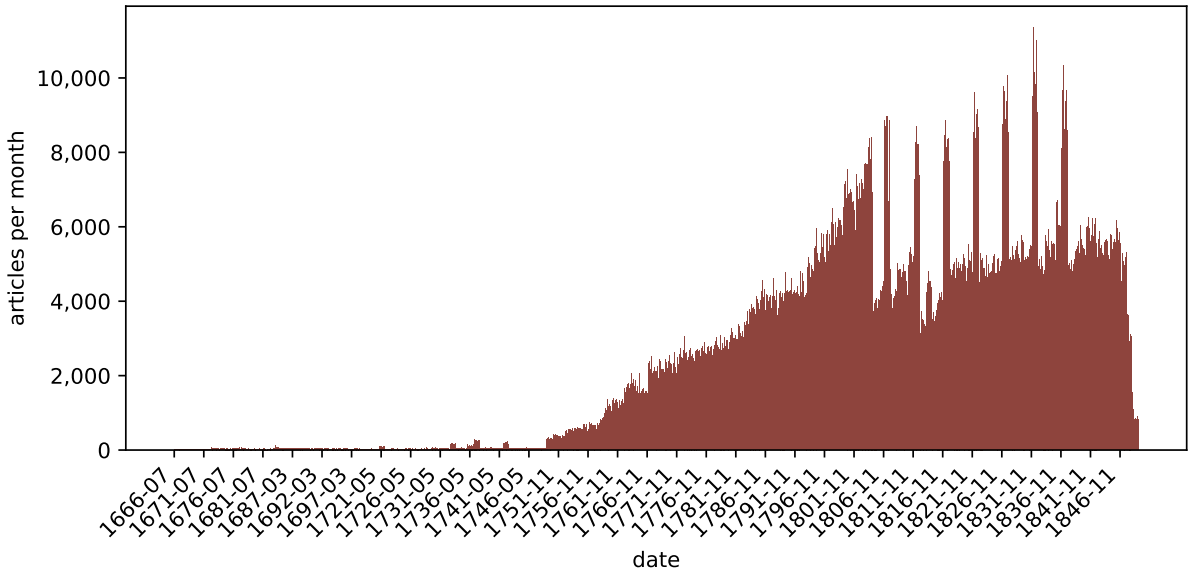


Figure 2: Distribution of the full corpus, number of articles per month.

3.5. Model comparison

Model	Precision		Recall		F1-score		Accuracy
	fiction	non-fiction	fiction	non-fiction	fiction	non-fiction	
TF/IDF	0.887	0.850	0.840	0.905	0.860	0.874	0.869
`MeMo-BERT-03`	0.873	0.878	0.878	0.875	0.875	0.876	0.876
`Old_News`	0.884	0.886	0.886	0.887	0.885	0.886	0.885
`e5`	0.888	0.876	0.872	0.892	0.879	0.884	0.882
`jina`	0.863	0.868	0.868	0.865	0.865	0.867	0.866
`bge-m3`	0.861	0.871	0.869	0.864	0.864	0.867	0.866
`gemma`	0.816	0.813	0.808	0.821	0.812	0.817	0.816

Table 4: Performance (averaged across 5 folds) of TF/IDF and six embedding models on the fiction/non-fiction classification task, evaluated using 5-fold `StratifiedGroupKFold` cross-validation with group-preserving splits. Metrics reported are precision, recall, F1-score (per class), and overall accuracy. Note that the second-best model (`e5`) sometimes has a higher precision or recall for one of the classes, while `Old_News` performs more consistently (i.e., has slightly higher F1 for fiction).

Of the models tested, three had shown potential in earlier studies with Danish historical corpora: among these, two were fine-tuned on nineteenth-century Danish texts, and one was multilingual. Full model names and URLs are shown in Table 5. We selected three other state-of-the-art multilingual models for their strong performance in the Multilingual Text Embedding Benchmark (MTEB), manageable size ($< 1B$ parameters), non-instruction-tuned nature, and high maximum input length (8,194 tokens).

Pooling: For all models except `jina-embeddings-v3`, `bge-m3` and `embeddinggemma-300m`, the maximum input length was limited to 512-514 tokens. In all cases, each feuilleton text was split into chunks of up to the maximum number of tokens, and a mean embedding was computed by averaging the resulting chunk embeddings.

Model	Max tokens	Dimensions	Layers	Source
MeMo-BERT-03	514	768	12	https://huggingface.co/MiMe-MeMo/MeMo-BERT-03
Old_News_Segmentation_SBERT_V0.1	512	768	12	https://huggingface.co/JohanHeinsen/Old_News_Segmentation_SBERT_V0.1
bge-m3	8,192	1,024	24	https://huggingface.co/BAAI/bge-m3
embeddinggemma-300m	2,048	768*	24	https://huggingface.co/google/embeddinggemma-300m
jina-embeddings-v3	8,192	1,024*	24	https://huggingface.co/jinaai/jina-embeddings-v3
multilingual-e5-large	514	1,024	24	https://huggingface.co/intfloat/multilingual-e5-large

Table 5: Overview of full model names. We also show the maximum input context length, final embedding dimension size, number of hidden layers, and HuggingFace urls. The order of models is by language (Danish models on top) and alphabetical.

3.6. Classifier comparison

For our final classification with the Old_News embeddings, we tested various classification models. We find that Logistic regression performs consistently well (as does Logistic Elastic Net), which was why we used it for the final tagging of articles.

Classifier	Class	Precision	Recall	F1-score
LogisticRegression	*Fiction*	0.88 ± 0.03	0.88 ± 0.06	0.88 ± 0.02
	Non-fiction	0.89 ± 0.05	0.89 ± 0.04	0.89 ± 0.01
LogisticElasticNet	*Fiction*	0.88 ± 0.03	0.88 ± 0.05	0.88 ± 0.02
	Non-fiction	0.89 ± 0.04	0.89 ± 0.03	0.89 ± 0.01
LinearSVC	*Fiction*	0.87 ± 0.03	0.85 ± 0.05	0.86 ± 0.02
	Non-fiction	0.87 ± 0.04	0.88 ± 0.03	0.87 ± 0.02
RandomForest	*Fiction*	0.90 ± 0.02	0.85 ± 0.06	0.87 ± 0.02
	Non-fiction	0.87 ± 0.04	0.91 ± 0.03	0.89 ± 0.01

Table 6: Average classification performance and standard deviation across 5 folds. Precision, recall, and F1-score are reported per class for each classifier. Logistic Regression and ElasticNet just slightly outperform the others based on F1-score.

3.7. Misclassifications

We include an analysis of the false positives and negatives in our fiction/non-fiction classification. Note that all subcategories can be labeled both fiction or non-fiction in the gold standard. Analysis of misclassifications (Table 7) shows that certain subcategories consistently sit near the boundary of the fiction register. Fictional biographies and travelogues are often misclassified as nonfiction, suggesting that these genres share stylistic features with nonfiction, whereas essays and general nonfiction are occasionally misclassified as fiction, likely because they adopt literary or narrative elements. Overall, the patterns reveal which subcategories most closely overlap with the linguistic and stylistic cues the model associates with fiction, that is, biographies and travelogues make up the largest share of false negatives, while essays account for the largest share of false positives, indicating that the model tends to read fictionalized life-writing as nonfiction and stylistically marked essays as fiction.

Subcategory	False Negatives			False Positives		
	Count	% of subcategory	% of total FNs	Count	% of subcategory	% of total FPs
Biography	37	21.76	34.91	3	1.76	2.91
Travelogue	21	23.86	19.81	12	13.64	11.65
Essay	1	1.28	0.94	24	30.77	23.30

Table 7: False negatives (fiction predicted as non-fiction) and false positives (non-fiction predicted as fiction) by subcategory. Percentages show both the proportion relative to the subcategory and relative to the total number of false negatives/positives.

3.8. Semantic space of the curated set

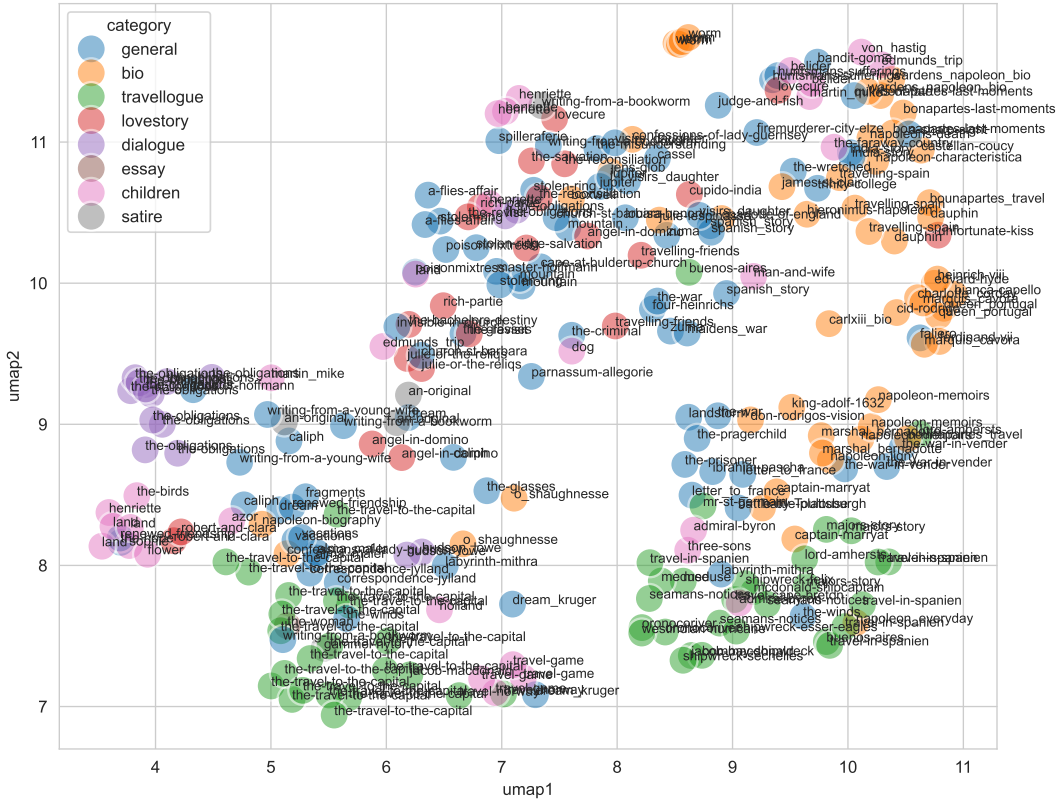


Figure 3: UMAP of the curated set of fiction all individual pieces/installments. Here, we see some tendential clusters. (a) of installments from the same series, and (b) of travelogue and biography tags. For example we see stories involving Napoleon clustering at the upper right corner. However, this is not the rule: consider the relative spread of, e.g., childrens literature across the whole semantic space, indicating that children's literature is not a semantically consistent category (in our data).

Towards an interoperable Hungarian historical newspaper corpus

Noémi Ligeti-Nagy, Henrietta Szabó
ELTE Research Centre for Linguistics
Benczúr u 33. Budapest Hungary
{surname.firstname}@nytud.elte.hu

Abstract

PressMint is a CLARIN initiative that aims to build multilingual, comparable and interoperable corpora of historical newspapers. For Hungarian, the main challenge is not a lack of material but fragmentation: newspapers are distributed across several portals, with heterogeneous metadata, access paths and OCR quality. This extended abstract reports the current status of the Hungarian PressMint subcorpus, focusing on the 19th century and the early 20th century (roughly 1800–1920). We describe two project artefacts already used in practice: a structured source inventory and a validation-driven repository. We summarise source scouting across Europeana, Hungaricana, OSZK–EPA, DiFMOE and related portals, including a curated 12-title Hungaricana manual-download pilot list with explicit target coverage periods. We then outline a reproducible pipeline for acquisition, OCR, layout analysis and conversion to PressMint-compatible TEI with facsimile linkage. Finally, we specify near-term deliverables for a first Hungarian release candidate and the evaluation steps planned for OCR and layout processing.

Keywords: historical newspapers, interoperability, TEI, OCR, CLARIN, PressMint, Hungarian

1. Workshop context and scope

PressMint aims to provide a common format and comparable linguistic annotation for multilingual corpora of historical newspapers that overlap, at least partly, in time. For the Hungarian contribution, we adopt a strict time window: the 19th century and the early 20th century (roughly 1800–1920). This period is highly valuable for historical research and is usually less constrained by copyright and reuse conditions than later newspaper material. It also aligns well with the turn-of-the-century perspective that often motivates cross-national press comparison.

Hungarian brings two PressMint-relevant characteristics. First, its diacritics-rich orthography and the frequent hyphenation found in narrow newspaper columns mean that OCR noise can directly affect tokenisation, sentence splitting and later annotation. Second, Hungarian-language newspapers were historically published both within and beyond the borders of present-day Hungary, so provenance and cross-border coverage are essential selection criteria.

Our main design principle is to keep *selection* separate from *processing*. Selection must remain transparent and revisable as new sources are identified and coverage gaps become visible. Processing must remain reproducible and validation-driven. This is why the project is organised around two central artefacts already in daily use: a structured inventory and a repository that turns inventory records into buildable TEI releases.

2. Related work, existing resources and standards

PressMint follows a well-established CLARIN pattern: multilingual partners converge on a shared, validation-driven interchange format while keeping national acquisition and preprocessing pipelines partly heterogeneous. ParlaMint is a strong point of reference here, demonstrating that shared TEI modelling, stable identifiers and coordinated release governance can work at scale (Erjavec et al., 2023).

In the historical newspaper domain, several projects have shown that interoperability depends on profiling and conversion rather than on a single upstream format. Europeana Newspapers promoted METS/ALTO profiling for exchange and highlighted the practical cost of harmonising provider metadata (Mühlberger, 2014; Freire et al., 2019). Neudecker's overview of searchable historical newspapers documents the diversity of OCR outputs and the need for conversion layers (Neudecker and Antonacopoulos, 2016). Impresso is a prominent example of combining facsimile-linked OCR text with curated metadata and stand-off enrichments (Ehrmann et al., 2020). OCR quality is also not a secondary technical issue: user-oriented studies show measurable effects of OCR quality on the perceived usefulness of historical newspaper collections (Kettunen et al., 2022).

For Hungarian, important digitised historical newspaper holdings already exist in Hungaricana, OSZK–EPA, Arcanum, ANNO and smaller institutional collections. These resources are highly valuable, but they are usually delivery platforms or

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 50–55

digitised holdings rather than interoperable corpora distributed in a shared TEI representation with stable identifiers and reproducible conversion steps. The Hungarian PressMint work therefore builds on existing digitisation efforts rather than duplicating them: the goal is to turn a selected subset of available material into a corpus that is consistent with PressMint-wide requirements.

3. Source scouting: what is available for Hungarian (and how)

The Hungarian subcorpus is *not* currently driven by a single ready-to-ingest provider. Instead, our work begins with explicit source scouting and a living inventory, and we only commit to titles once (i) the acquisition path is stable and reproducible and (ii) the time window is satisfied.

3.1. Scouted portals and current extraction status

Our scouting currently covers Europeana (Willems and Atanassova, 2015), Hungaricana (Hungaricana, n.d.), OSZK–EPA (Országos Széchényi Könyvtár (OSZK), 2026), DiFMOE (Digital Forum Central and Eastern Europe, 2023), ANNO (Müller, 2004; Austrian National Library, n.d.), Arcanum (Arcanum, n.d.), the National Archive press interface and related portals. Because these resources are not equally well suited for reproducible ingestion, we record both *availability* and *acquisition feasibility*.

Four scouting results are already directly relevant for PressMint planning. First, Europeana, a pan-European aggregation platform, yielded 69,290 Hungarian-language *text-type* items through API harvesting. This set is useful for discovery, but it still requires substantial filtering because the newspaper filter was not reliable enough for Hungarian items. Second, OSZK–EPA, the National Széchényi Library's electronic periodicals archive, contains many relevant collections; our current manual survey lists at least 142 regional, 54 cross-border and 50 newspaper collections, with overlaps still to be resolved. Third, DiFMOE, a portal that includes digitised periodicals from Central and Eastern Europe, yielded eight Hungarian-language periodicals already entered into the inventory with URLs; seven of them fall within the PressMint time window. Fourth, Hungaricana, the main Hungarian digital heritage portal, is now represented by a curated manual-download list of 12 titles with title-level collection URLs, holding institutions, full available coverage periods and explicitly selected in-window target periods.

The Hungaricana list is important because it moves that source from a vague manual option to a concrete acquisition pilot. The working sheet is already detailed enough to support title-level planning: for each candidate it records a stable internal source ID, the holding institution, the Hungaricana collection title, the full available coverage range, the selected PressMint target period and the title-level URL. At the same time, it is still an acquisition manifest rather than a full issue-level inventory: issue counts, page counts, download progress and notes on source-specific constraints are not yet populated.

This title-level detail matters for planning because the selected target periods are not uniform. Most titles are currently capped at 1915 to stay safely within the PressMint window, but some have narrower selections: *Eger – hetilap* ends in 1914, *Szatmári Értesítő* is represented by a single year (1862), *Váczi Közlöny* ends in 1895, and *Felsőmagyarországi Hírlap* is currently restricted to 1903–1907. The Hungaricana pilot list therefore provides not only a count of candidate titles, but an explicit acquisition plan for what should actually be downloaded first.

An important methodological consequence is that the inventory – not the processing scripts – is the main coordination artefact. It records what is known, what is reproducible and what still remains uncertain.

3.2. Selection criteria for the PressMint tranche

Within the time window, our planned inclusion criteria are: (i) stable and citable identifiers for title and issues/pages, (ii) at least page images (PDFs or images) that allow facsimile linkage, (iii) sufficient metadata to support cross-corpus comparison (publication place, years, language), and (iv) a realistic acquisition path at scale. We also prioritise cross-border Hungarian press and regionally diverse publication places, because these are especially relevant for comparative work on public discourse and regional vocabulary.

3.3. Inventory as coordination layer

The Hungarian inventory currently contains more than 150 candidate records with a fixed column schema covering provider, title, place of publication, available years, selected target coverage, acquisition method, persistent identifiers, scan/OCR fields and ingestion bookkeeping. This inventory is validated and aligned with controlled vocabularies in the repository (Section 4). As a result, selection decisions are transparent and reversible, and downstream processing can be driven by structured manifests rather than ad-hoc manual tracking.

Source	Current evidence base	In-window titles	Main open issue
Europeana	69,290 Hungarian-language text items harvested via API	discovery set only	newspaper-specific filtering and metadata cleanup
OSZK–EPA	manual survey of regional, cross-border and newspaper collections	to be cleaned	overlap resolution and automated harvesting
DiFMOE	eight Hungarian-language periodicals entered in inventory with URLs	7	mostly manual acquisition, then OCR/TEI conversion
Hungaricana	curated manual-download list with title-level URLs, institutions and selected coverage periods	12	issue counting and download workflow still manual

Table 1: Current status of the main scouting tracks for Hungarian historical newspapers.

Title	Holding institution	Available years	Selected target coverage
Békésmegyei Közlöny	Békés Megyei Könyvtár	1877–1938	1877–1913
Dunántúli Protestáns Lap	Jókai Mór Városi Könyvtár (Pápa)	1890–1945	1890–1915
Eger – hetilap	Bródy Sándor Megyei és Városi Könyvtár (Heves megye)	1863–1914	1863–1914
Esztergom	Helischer József Városi Könyvtár (Esztergom)	1895–1932	1895–1915
Esztergom és Vidéke	Helischer József Városi Könyvtár (Esztergom)	1879–1944	1879–1915
Felsőmagyarországi Hírlap	MNL BAZ Megyei Levéltárának Sátoraljaújhelyi Fióklevéltára	1903–1917	1903–1907
Független Budapest (Az Erzsébetváros)	Erzsébetváros Önkormányzata	1906–1938	1906–1915
Nyírvidék	Móricz Zsigmond Megyei és Városi Könyvtár (Szabolcs-Szatmár-Bereg)	1867–1942	1867–1915
Szatmári Értesítő	Móricz Zsigmond Megyei és Városi Könyvtár (Szabolcs-Szatmár-Bereg)	1862	1862
Váci Hírlap	Katona Lajos Városi Könyvtár (Vác)	1887–1942	1887–1915
Váczi Közlöny	Katona Lajos Városi Könyvtár (Vác)	1881–1895	1881–1895
Zemplén	MNL BAZ Megyei Levéltárának Sátoraljaújhelyi Fióklevéltára	1886–1937	1886–1915

Table 2: Curated Hungaricana manual-download pilot list

4. Repository, conventions and validation

The project already has an enforceable engineering substrate. The `pressmint-hu` repository contains: (i) CI validation (`.github/validate-inventory.yml`), (ii) documentation (`docs/hu_source_inventory.ods`, `docs/conventions.md`), (iii) an inventory schema and controlled vocabularies (`inventory/schema.json`, `inventory/sources.yaml`, `inventory/vocabularies.yaml`), (iv) a TEI header template (`inventory/tei_header_template.xml`), and (v) scripts for crawling (Europeana, Hungaricana collection-name extraction) and utilities (IDs, TXT→ODS import, YAML→TEI export. The repository also stores intermediate scouting artefacts (e.g. `hu_europena_corpus.json`, `hungaricana_collections.txt`).

This structure operationalises interoperability in three ways:

- **Deterministic IDs:** title, issue and page identifiers are generated from inventory fields, which keeps them stable across rebuilds.
- **Constrained metadata:** controlled vocabularies reduce drift in source descriptions, access methods, scan formats and OCR flags.
- **TEI by construction:** TEI headers are generated from structured YAML, which reduces

manual editing and supports validation-driven release builds.

The Hungarian workflow also aligns with the broader PressMint logic. Upstream steps such as source acquisition, OCR and some normalisation details may differ by national partner, but the target representation, mandatory metadata, identifier logic and facsimile linkage are shared. This allows local heterogeneity in acquisition while keeping downstream comparison possible.

The repository structure is designed so that schema files, vocabularies and conversion code can be released publicly with minimal changes. Public release of every project artefact, however, will depend on cleaning project-internal notes and checking source-specific redistribution constraints.

5. Processing pipeline: ingestion, OCR and TEI conversion

5.1. Ingestion and normalisation

Ingestion starts with a provider-specific acquisition step that yields a local artefact (typically PDFs or page images) plus a manifest entry in the inventory. The local storage layout is organised by Source/Title/Year, mirroring the inventory keys and making partial re-ingestion possible when target coverage changes.

For near-term work, manually tractable sources are the most realistic starting point. DiFMOE provides a small set of periodicals with item-level URLs already recorded in the inventory. Hungaricana now provides a second concrete pilot track through the curated 12-title manual-download list. In parallel, Europeana remains a metadata-first discovery track: it helps us locate candidate titles and estimate coverage, but acquisition feasibility still has to be checked provider by provider.

5.2. OCR and layout analysis: from prototype to evaluation

Our OCR strategy is driven by the fact that newspaper layout is not a marginal issue but a central source of downstream error. An initial prototype log, based on a single test page image, showed that faint column separators can break automatic segmentation and thereby damage reading order. Although this first test page lies outside the final PressMint time window, the lesson is still relevant: for historical newspapers, OCR must be evaluated as a *layout-aware pipeline*, not only as plain text recognition.

At the current stage we do not commit to a single OCR engine. The material is too heterogeneous for premature tool lock-in, and the key question is not only character recognition accuracy but also preservation of columns, reading order and article boundaries. We therefore plan a small but systematic evaluation on manually corrected pilot pages. The evaluation will compare at least a page-level OCR baseline with a layout-aware pipeline and will measure: (i) character and word error rates on sampled pages, (ii) preservation of column structure and reading order, (iii) systematic error classes particularly relevant for Hungarian newspaper print, such as diacritics and line-break hyphenation, and (iv) robustness across different scan qualities and page layouts.

The pipeline itself is organised in four steps: (i) page image extraction and normalisation, (ii) layout analysis to detect columns and reading order, (iii) OCR on layout-aware regions, and (iv) conversion into PressMint-compatible TEI with explicit facsimile linkage. This architecture keeps open the possibility of article- or section-level segmentation where the layout evidence is strong enough, without making it a hard requirement for the first release.

5.3. Quality assurance and comparability

To ensure that the Hungarian tranche remains comparable across PressMint partners, we plan quality monitoring at three levels:

- **Metadata QA:** validation against controlled vocabularies and mandatory fields such as title, years, place and identifiers.
- **OCR QA:** periodic sampling with character error rate estimation and tracking of systematic error classes such as hyphenation, diacritics and ligatures.
- **Structural QA:** validation that each TEI document preserves facsimile linkage and that identifiers remain stable across rebuilds.

Where feasible, a small manually corrected gold set will be prepared to calibrate OCR and layout components, following broader recommendations for multilingual and historical OCR research agendas (Smith and Cordell, 2018).

5.4. TEI conversion and optional stand-off layers

Following PressMint interoperability goals, TEI is the hub representation for text, metadata and provenance. We implement TEI generation in two layers:

1. **Guaranteed page-level TEI** (issue $\rightarrow$ pages) with stable identifiers and facsimile pointers.
2. **Optional article or section segmentation** when layout cues allow reliable inference.

To support comparative NLP while preserving reprocessability, linguistic annotations (tokenisation, sentence boundaries and, optionally, POS or lemma information) will be kept as stand-off layers whenever feasible, so that OCR and normalisation updates do not invalidate previously released base TEI.

6. DH use cases enabled by the Hungarian tranche

The Hungarian PressMint tranche is meant to support both close and distant reading. Concrete use cases include: (i) tracing the vocabulary of industrialisation, public health or administration across the long 19th century; (ii) comparing how the same events were reported in Budapest and in cross-border or regional newspapers; (iii) studying lexical and stylistic differences across publication places such as Esztergom, Kassa, Nyíregyháza, Sátoraljaújhely or Vác; and (iv) tracking named entities, institutions and quoted actors in a way that remains anchored in facsimile-linked evidence.

These use cases motivate our emphasis on provenance-rich TEI for citation and on an OCR/layout pipeline that preserves reading order and, where possible, section boundaries.

7. Deliverables and roadmap

The immediate goal is a Hungarian release candidate that (i) fits the 19th/early-20th century time window and (ii) can be rebuilt reproducibly from inventory manifests. Near-term deliverables for 2026 are:

- a cleaned, deduplicated and newspaper-focused Europeana-derived candidate list starting from the harvested 69,290 Hungarian-language text items;
- a first TEI pilot release for the seven in-window DiFMOE titles and a pilot subset of the 12 curated Hungaricana titles, all with facsimile-linked pages;
- a documented OCR and layout benchmark on manually corrected sample pages, including a sampling protocol for error analysis and ground-truth creation;
- a harmonised ingestion and export toolchain (YAML→TEI) integrated into repository validation.

8. Conclusion

The Hungarian PressMint subcorpus is being built around explicit engineering artefacts – a structured inventory and a validation-driven repository – and around the assumption that OCR and layout are central interoperability constraints, not secondary implementation details. Current scouting shows a fragmented but usable landscape. Europeana is effective for large-scale discovery, while DiFMOE and the curated Hungaricana pilot list provide concrete, manually tractable starting points for acquisition. By treating selection, ingestion, OCR and TEI conversion as reproducible and testable processes, we aim to deliver a Hungarian tranche that is aligned with PressMint-wide requirements and useful for both close and distant reading research.

9. Bibliographical References

Arcanum. n.d. Arcanum newspapers. https://adt.arcanum.com/. Accessed: 2026-03-30.

Austrian National Library. n.d. Historical newspapers and periodicals of the austrian national library. https://www.onb.ac.at/en/departments/department-of-manuscripts-and-rare-books/holdings/rare-books/historic-newspapers-and-magazines. Accessed: 2026-03-30.

Digital Forum Central and Eastern Europe. 2023. Digital library of the digital forum central and eastern europe. https://www.copernico.eu/en/online-resources/digital-library-digital-forum-central-and-eastern-europe. Accessed: 2026-03-30.

Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. 2020. Language Resources for Historical Newspapers: the Impresso Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 958–968, Marseille, France. European Language Resources Association.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Çağrı Çöltekin, Tommaso Agnoloni, Orsolya Ring, et al. 2023. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57:415–448. Published online 2022.

Nuno Freire, Antoine Isaac, Twan Goosen, Daan Broeder, Hugo Manguinhas, and Valentine Charles. 2019. Opening Digitized Newspapers Corpora: Europeana's Full-Text Data Interoperability Case. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *Open Access Series in Informatics (OASIcs)*,

pages 22:1–22:14, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Hungaricana. n.d. Library | hungaricana. https://library.hungaricana.hu/en/. Accessed: 2026-03-30.

Kimmo Kettunen, Heikki Keskustalo, Sanna Kumpulainen, Tuula Pääkkönen, and Juha Rautiainen. 2022. OCR quality affects perceived usefulness of historical newspaper clippings – a user study. arXiv:2203.03557.

Günter Mühlberger. 2014. METS/ALTO Profile (ENMAP) – Deliverable D5.2 (Draft). Technical report, Europeana Newspapers Project. Accessed 2026-03-03.

Christa Müller. 2004. A N N O – Austrian Newspapers Online: Historische österreichische Zeitungen und Zeitschriften online. Eine Digitalisierungsinitiative der Österreichischen Nationalbibliothek. In Hartmut Walravens, editor, *Newspapers in Central and Eastern Europe / Zeitungen in Mittel- und Osteuropa: Papers presented at an IFLA conference held in Berlin, August 2003*, pages 141–148. K. G. Saur, Berlin and Boston.

Clemens Neudecker and Apostolos Antonacopoulos. 2016. Making Europe's Historical Newspapers Searchable. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 405–410, Santorini, Greece. IEEE.

Országos Széchényi Könyvtár (OSZK). 2026. EPA – Elektronikus Periodika Archívum. https://epa.oszk.hu/. Accessed 2026-03-03.

David A. Smith and Ryan Cordell. 2018. A Research Agenda for Historical and Multilingual Optical Character Recognition. Technical report, Northeastern University, NULab for Texts, Maps, and Networks. Accessed 2026-03-03.

Marieke Willems and Rossitza Atanassova. 2015. Europeana Newspapers: searching digitized historical newspapers from 23 European countries. *Insights*, 28(1):34–39.

Towards a Bulgarian Historical Newspaper Corpus — Construction of a Reading Order over the Text in Searchable PDFs

Nikolay Paev[1], **Stefan Marinov**[1], **Ivan Kratchanov**[2], **Petya Osenova**[1], **Kiril Simov**[1]
[1] Artificial Intelligence and Language Technology
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
Bulgaria,
{nikolay.paev, stefan.marinov}@iict.bas.bg, {petya, kivs}@bultreebank.org

[2] National Library "Ivan Vazov" - Plovdiv,
Bulgaria,
ivankra@gmail.com

Abstract

The first task in the process of preparing historical newspaper corpora is to determine the reading order of texts. These texts are often extracted from searchable PDFs produced by some OCR software. In this paper we present an algorithm for generating the original reading order of the text blocks selected from the corresponding PDF. We also performed a tuning of the algorithm parameters. The optimization provides an improvement of 10 %.

Keywords: Bulgarian Historical Periodicals, OCR Models for Bulgarian, Large Language Models for Combining Segments of Bulgarian Texts

1. Introduction

Old periodicals are a source of information to support research in many areas of social sciences and humanities (SS&H) —- ranging from linguistics and literature to history, ethnography, journalistic studies, etc. The research in SS&H is in the stage of its active digitization period, in which physical artifacts in archaeology, libraries, archives, museums, art galleries, and others are converted into digital objects via audio and video recording, scanning and OCR processing, 3D scanning and modeling. In parallel to digitization of the artifacts, methods for linguistic processing have also been developed. These methods provide new perspectives on the existing research in SS&H, such as natural language processing services, AI technology, etc. In this paper, we present the first steps in the digitization of old Bulgarian newspapers in the period from the reinstatement of Bulgarian state (1878) to the middle of twentieth century.

The first step refers to detecting the text in the document with an OCR system. During the OCR-ing the important task is to ensure a very good quality with respect to the period of publishing, and to the typography of the newspaper edition — the fonts, selected font sizes, line lengths, line spacing, letter spacing, and formatting of the page. All these factors appear to be very important since the included period is one of great dynamics in the area of publishing, technology, and typographic style in Bulgaria. Also, the standards of spelling have not yet been well-established. From 1870 to 1945, eight spelling forms were proposed and applied in various ways – either officially, or non-officially. These are:

- the Marin Drinov's spelling
- the spelling project by the philological committee from 1893
- the spelling project from 1895
- the Ivan Vazov's spelling system from 1899
- the Drinov-Ivanchevski spelling system from 1899
- the Omarchevski spelling system (1921 - 1923)
- the spelling project of the Historic-philological branch of the Bulgarian Academy of Sciences (1899)
- the Tsankov's spelling system (1923)
- the contemporary system (1945)

The most prevalent spelling system until 1945 (with small interruptions) remains the Drinov-Ivanchevski one from 1899, which to great extent adheres to the traditional/etymological spelling. The linguistic phenomena that usually show heterogeneous features across various spelling systems are as follows:

- *phonetic*: changing of *ya* to *e*; order of the assembly of a *shwa* and *r* or *l*; reduction of the open vowels to non-open ones; the usage of the small *er* at the end of the nouns.

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 56–64

- *grammatical*: long and short forms of the definite article in masculine singular nominals; deverbal nouns ending in *-ne* and *-nie*; vocative; pronouns.
- *other*: small and capital letters; spelling of foreign words; spellings with a dash; punctuation.

After 1945 Bulgarian lost some letters, among which the yat, the ers in the end of the words, the nasal 'big nosovka'. Also, full and short articles started to be used in masculine to mark the subject and oblique argument positions, respectively.

To handle these problems, we retrained the OCR model for the peculiarities of each newspaper. The OCR provides the recognized text in a formatting that resembles the formatting of the physical appearance of the scanned documents. In the case of newspapers, the formatting includes various types of information —- titles, date, publisher information, columns of text (containing articles, advertisements, notes), images, page numbers, etc. The OCR-ed result represents this information in the form of separate text blocks with coordinates related to the respective sheet of paper.

After the formal division, our next task is to interpret the content structurally - detecting the page layout, the titles, the paragraphs, the articles, the advertisements, and others. This appeared to be a very non-trivial task. For example, different editions of periodicals have almost arbitrary page layouts, which change on each page. The publishers tried to accommodate as much text from different articles as they could on the first page and also to fill all page spaces with some information. They also employed various layouts which changed the reading order of articles in many ways. All these issues led us to create complex algorithms to extract and then order the text in the complete articles.

The structure of the paper is as follows: the next section focuses on the related work on newspaper article extraction. Section 3 describes the preparation of the dataset with searchable PDFs. Section 4 introduces the processing of the dataset in order to define a reading order over blocks of text. Section 5 presents the evaluation of how well the reading orders are generated. Section 6 outlines the manual post-editing framework. Section 7 concludes the paper.

2. Related Work

Here we present a non-exhaustive list of related works that show good practices for other languages.

In Bourne (2025) Pixtral 12B (OCR using an image-to-text model) is applied to the Nineteenth Century Serials Edition (NCSE) archive, containing about 84,000 pages from six British periodicals. The process includes: automatic detection and post-processing of the layout (bounding boxes) and subsequent text processing. We deal with the same process, using OCR trained/prepared for Bulgarian and for the specific fonts of the printed publications. However, in this paper, we pay more attention to the post-processing of the text, as it is inevitable, but still directly dependent on the quality of the OCR.

Ding and Huang (2014) examine the digitization of historical newspapers in China, focusing on the production of PDF files and the practices used in projects such as the DaChengLaoJiu database, the digitization of Dazhong Daily, and solutions by Beijing companies. They extract text from PDFs and then try different OCR approaches. Not surprisingly, the same difficulties with the extracted text have been encountered as ours, due to the complexity of the OCR processing. They recognize the need for human verification and editing. We also became aware of the necessity for human intervention. Thus, we propose an approach to reduce the amount of human labor required and try to optimize the editing process.

Schultze et al. (2025) presents the Chronicling Germany Dataset – the largest currently annotated corpus of historical German newspapers (801 pages), mainly from the period of the Austro-Prussian War (1866). The data contains detailed annotations of page layouts (paragraphs, titles, tables, images, separators) and more than 371,000 text lines with manually corrected transcriptions according to the OCR-D standard. A complete automated pipeline has been developed that contains the following steps:

- Layout segmentation (U-Net) – recognizes types of regions;
- Baseline detection (U-Net) – detects text lines;
- OCR (LSTM and Transformer models) – recognizes text.

In our paper also the need for a subsequent manual processing of the text is emphasized, together with the optimization of this process being our main focus.

Isaacs et al. (2024) presents the creation of three large humanitarian corpora in English, French, and Spanish, compiled from the ReliefWeb platform through its public API. The authors describe a methodology for language identification (since they use multilingual data), noise reduction, and automatic language processing (tokenization, lemmatization, and morphosyntactic annotation), as well as enrichment of the corpora with metadata. In our work, additional linguistic annotation, as well as metadata, will also be necessary. At this stage, however, we focus on the extraction of as accurate text as possible. We aim at ensuring the correct

order of the texts and preventing this order from blended newspaper articles. In this way, the results of any subsequent automatic language processing would improve significantly. Currently, we additionally annotate the metadata embedded in certain places in the print (mainly at the beginning and end) to facilitate their subsequent extraction.

Kjosbakken (2025) describes a practical approach to building a high-quality OCR pipeline, which is a key component in many machine-learning systems (e.g., document classification, RAG systems, receipt processing). The authors recommend testing different OCR systems and their subsequent evaluation. We use a similar approach to the definition of the algorithm settings, experimenting with different parameters and evaluating with a Levenshtein-type algorithm, which is one of the recommended ones by them.

In June (2026), the authors perform OCR using small language models and additional processing in the pipeline, which is also our goal. However, they are processing copy-heavy documents (forms, insurance, government documents, financial statements), which are highly structured and massively repetitive. This is the opposite of our type of documents, which have a very diverse structure and relatively limited repetition. This creates difficulty in implementing “rules” to ensure the quality of the text extracted from the model; therefore, more attention should be paid to post-editing and verification of the extracted text.

Gutehrlé and Atanassova (2021) presents a rule-based method for Logical Layout Analysis (LLA) of historical newspapers presented in an XML ALTO format. This paper also addresses the problem of text extraction from printed publications. The main focus is on the text type classification – especially the headline detection. The approach they used is to extract text formatting information (size, italic, bold) into a special xml format designed to encode such information – XML ALTO. At this stage, we have not determined the type of text pieces, but this is one of the necessary next steps. For us, the important lesson is that text formatting information, according to this paper, can improve the results.

Lee et al. (2020) presents the Newspaper Navigator Dataset, a collection of automatically extracted visual content from historical newspapers. The project uses more than 16.3 million digitized pages from the Chronicling America initiative (Library of Congress), covering the period 1789–1963. The authors develop a pipeline based on deep learning that automatically detects and extracts 7 types of visual content: headlines, photographs, illustrations, maps, comics, editorials, cartoons, advertisements. These findings do not fall directly into the line of our current work, since our focus is exclusively on text extraction. However, in our future work we plan to incorporate the available visual content.

3. Preparation of a Searchable PDF Dataset of Bulgarian Historical Newspapers

In this section, we present our approach towards providing a very high quality of the OCR result.

The newspapers chosen to participate in the dataset, in their original paper form, are stored in the "Ivan Vazov" National Library - Plovdiv. Considering the digitization process, the library fulfills two main tasks:

1. Preparation of high-quality "master" files for long-term archiving, which are typically 24-bit color scans at 300 ppi made directly from the original paper items. These files (usually in .tiff format) are not accessible to the general public.

2. Preparation of web-suitable PDF files of lower quality and size, acquired by running OCR with ABBYY FineReader software directly onto the master files. Then they are uploaded to the Digital Library to provide public access. The masters are compiled into a single PDF per cultural heritage unit, e.g. one PDF for one newspaper issue.

The most valuable public-domain material predates the 1945 orthographic reform. Therefore, the OCR quality is strongly affected by the historical alphabets and spelling conventions, the archaic vocabulary and mixed-language content, as well as the limited dictionary coverage. The age-related document issues (paper darkening, faded ink) and the uses of non-standard fonts inevitably reduce the recognition quality. A method for improving the accuracy of the OCR, provided by ABBYY FineReader, allows a targeted “training” for hard-to-recognize characters. It is particularly valuable for non-standard fonts and for historical Cyrillic characters as well as others that disappeared in the contemporary Bulgarian alphabet. For this task, we prepared a set of training snippet samples from different parts of the pages of the different newspaper issues. We also provided the corresponding machine-readable text within the clippings. Here are some examples in Fig.1:

Ideally, the training snippets must contain all the upper-case and lower-case letters of the alphabet, and all fonts to be found in a typographically stable period of a series of newspaper issues. The number of snippets used depends on the variety of fonts within the respective period. After the training based on the snippets is completed, all the trained symbols can be exported into the form of a *.fbtx*

Figure 1: Training snippets (on the left) and the recognized text (on the right).

file. This file can then be applied to all of the preselected newspaper issues. This is performed by loading the training file into the HotFolder software (part of the FineReader suite) for batch OCR processing. It should be noted that this procedure significantly improves the recognition quality, without resorting to post-OCR manual corrections.

Our plan is to process 309 periodical editions and about 150 000 pages. Each scanned issue is equipped with metadata consisting of: the newspaper title, the publisher, the date, and some other information. During the creation of the corpus, we have been trying to extend the coverage of the metadata given that we find additional information.

4. Extraction of Reading Orders over the Searchable PDF Dataset

In this section, we present our approaches towards the text extraction from Searchable PDF produced by the OCR software. The output of ABBYY FineReader can be stored in different formats. For our tasks, we consider two formats appropriate: *Searchable PDF* format and *MS Word docs* format. Both formats provide segmentation of texts into blocks that resemble the visual formatting of newspaper pages. In the work presented here, we are working with searchable PDF, because, as described in the previous section, there already exist 134 916 pages of scanned newspapers. It is not effective at all to perform the OCR of these newspapers again. At the same time, for the newspapers that have not been OCR-ed yet, we plan to store both formats.

Fig 2 shows a page from the “Belomorec” newspaper, in which the software segments, implemented by us, divide the text into many (relatively) small blocks. They need to be ordered in an appropriate way form a consistent and complete article. As mentioned above, in our current processing, we do not extract the images from the newspapers. This is left for future work.

Our task is after the initial segmentation of the content of the page in blocks to apply some heuristic algorithms to resize some of them and produce the correct blocks. In addition, it is necessary to determine the order of the blocks in such a way that the concatenation of their content forms the complete articles in the newspaper issue.

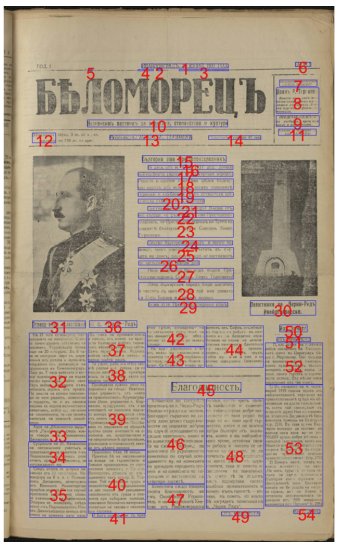


Figure 2: An example of a page where the blocks are ordered with the help of the reading order algorithm. The resulting order is consistent with both — the five-column layout and the partial horizontal separator.

The PDF documents for each newspaper issue in the dataset are presented in a searchable PDF. This means that within the PDF there is a layer of text aligned to the visual format of the paper. Each recognized character is connected to the coordinates that determine the place of the character occurrence. We are using the `pdfminer`[1] library to extract the text from PDFs. [2]

Such libraries use heuristic methods on the coordinates to compile the characters into lines and the lines into blocks. They also try to define a reading order of the blocks based solely on the coordinates. The order of the extracted blocks is often wrong - it fails to capture the layout of the newspapers and jumps between columns, making the text output

[1] https://github.com/euske/pdfminer

[2] We have performed experiments with several libraries for text extraction from searchable PDFs and after manual evaluation we selected `pdfminer`. After creating a gold corpus of 45 corrected and ordered newspapers, we compared the extracted texts from `pdfminer`, `pdfplumber`, `pymupdf`, and `pypdf` to the gold corpus using both normalized insertion and deletion ratio and Levenshtein distance. The scores are as follows: 0.8399 (486 714 edits), 0.4492 (1 520 982 edits), 0.8588 (408 803 edits) and 0.8504 (436 287). The scores are similar (with the exception of `pdfplumber`) and suggest that `pymupdf` is the best option according to these metrics.

incomprehensible. In this paper, we propose an algorithm that more successfully assigns the reading order of the blocks after extraction. We used the off-the-shelf order of the `pdfminer` blocks as a baseline for evaluation.

Our reading order algorithm works with blocks of text. Blocks are defined as a sequence of lines that form a coherent text. They are usually longer than one word, but shorter than an article, and have a rectangular shape in the document. We extracted the initial blocks with the `pdfminer` library. We view them determined by 4 numbers: (x_1, x_2, y_1, y_2) — the 2D coordinates of their bottom left corner — (x_1, y_1) and top right corner — (x_2, y_2). The `pdfminer` library uses coordinates, where (0, 0) is the bottom left corner of the page. Thus, in the representation of a block in this way always $x_1 \leq x_2$ and $y_1 \leq y_2$.

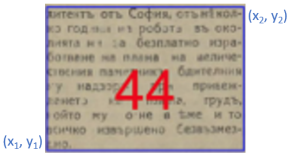


Figure 3: Block 44 from Fig. 2 represented by bottom left corner — (x_1, y_1) and top right corner — (x_2, y_2)

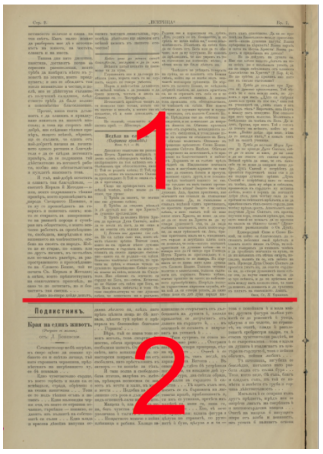


Figure 4: An example of a page segmented into two main areas, containing text from two articles. The areas are segmented horizontally. The red line is the gap between the two subpages, marked here with numbers 1 and 2, respectively.

As a first step, we find the *subpages* of the page. For this purpose, we find the horizontal gaps that expand from the left side to the right side of the page. They should not cross any blocks and should have at least some minimal thickness parameterized as *subpage_gap_threshold*. These gaps between blocks or subpages in the newspapers are marked with lines. However, the OCR software does not encode these lines in the text layer because they do not play any role in the search within the searchable PDF. Thus, we cannot use them for tracing the boundaries of the subpages or blocks. For that reason, we could only use the block coordinates to find such boundaries. The process of finding these gaps is as follows:

1. We initialize a list of candidate gaps from the bottom y coordinate of all blocks — y corresponds to y_1 for each block.
2. For each candidate, we test whether there is another block with a bottom y coordinate lower than the candidate gap and a top y coordinate higher than the candidate line plus a threshold of *subpage_gap_threshold* y coordinates. There is a block which is crossed by the candidate line and the candidate gap line cannot be a subpage boundary.
3. If the candidate gap does not cross a block, it becomes a gap that separates two subpages.

The gaps found in such a way separate the page in subpages and we consider them as boundaries of the subpapes — an illustration of such subpages is depicted in Fig. 4. The subpages are the first criterion for defining the reading order.

After identifying the subpages, our next goal is to find the subpage columns. The boundaries of the columns are recognized as vertical gaps:

1. With a step of *x_step* points we move horizontally from left to right. Each vertical line that spans at least *min_column_page_ratio* points of the page and does not cross a block (with a tolerance of *x_tolerance*), we mark as a candidate column separator.
2. If the column gap is larger, it is possible to get many duplicated separators close to one another. We remove each consecutive separator that is closer than 1.5x*x_step* points to the previous.
3. We also remove any separator that is closer than *min_column_width* points to the previous, thus defining the minimum column width.
4. Each separator defines a column to its right (the beginning of the page is also taken as a separator), and every block is assigned to the column of its closest left separator.

The columns are the second criterion for defining the reading order. Fig. 5 shows all correct separators for the columns.

Analogously to the vertical separators that define columns, there are also horizontal separators that change the reading order between the columns. Such segmentation of a column is depicted in Fig. 6.

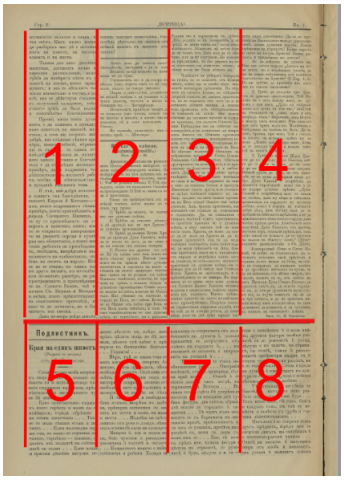


Figure 5: The same page from the Fig 4 additionally segmented vertically in columns which represent the order of the text in each horizontal area.

1. We find them in the same way as the full page horizontal separators by detecting the bottom y coordinates of blocks, which do not cross other blocks for every sublist of columns. For example, if we have 4 columns in the page, we try to find separators in the blocks of the sets of (1, 2, 3, 4), (1, 2, 3), (2, 3, 4), (1, 2), (2, 3) and (3, 4) columns. We define a minimal height of the gap as *partial_gap_threshold* points to account for the accidental occurrence of two adjacent blocks that end in adjacent points.

2. After finding these lines, we unify the adjacent lines with a tolerance of *y_tolerance* points.

3. We define the lines with a y coordinate and the minimum and maximum x coordinates of the blocks in their set of corresponding columns.

4. We also remove lines that are nested in longer lines.

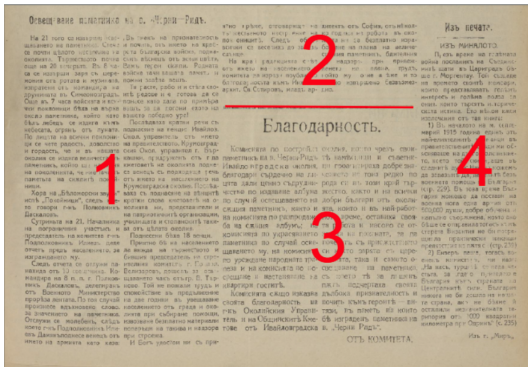


Figure 6: An example of a page with a partial horizontal separator that assigns a reading order to the blocks to the left, to the top, to the bottom, and to the right.

We sort the separators according to their coordinates from the beginning of the page — the bottom left corner of the page. Then, in the sorted order of the separators, each separator adds a new criterion to the final sorting of the blocks:

1. We assign 1 to all blocks left of the separator
2. 2 to all blocks higher than the separator
3. 3 to all blocks lower than the separator
4. 4 to all blocks right of the separator

After these steps, we employ a stable lexicographical sort on all blocks based on the aforementioned criteria. In other words, we sort by:

1. page number
2. subpage number
3. the numbers assigned by each of the partial horizontal separators
4. column number
5. top y coordinate
6. left x coordinate

These provide the reading order of the blocks on the page.

Problems can also arise. For example:

- Bad OCR or small font can make the algorithm skip a separator;
- A larger font can induce a false separator.

As mentioned above, the other output of the OCR is the *MS Word Docs* format. In this case, the ABBYY FineReader performs its own segmentation in blocks and groups them into larger blocks. These blocks are ordered according to the heuristics built in the software. Although this segmentation looks much better than the segmentation from PDFs, it suffers from the same problems: an incorrect union of smaller blocks which need to be reformatted, parasitic blocks within other blocks which need to rearrange; noise blocks. Fig. 7 depicts the two segmentations. We envisage in future to apply our algorithms to the *MS Word Docs* formats as well as to combine the two segmentations in order to exploit their mapping.

5. Evaluation

In this section, we evaluate the quality of the algorithm for determining the reading order of texts within the different blocks on the pages, as described in the previous section. The performance of the algorithm depends

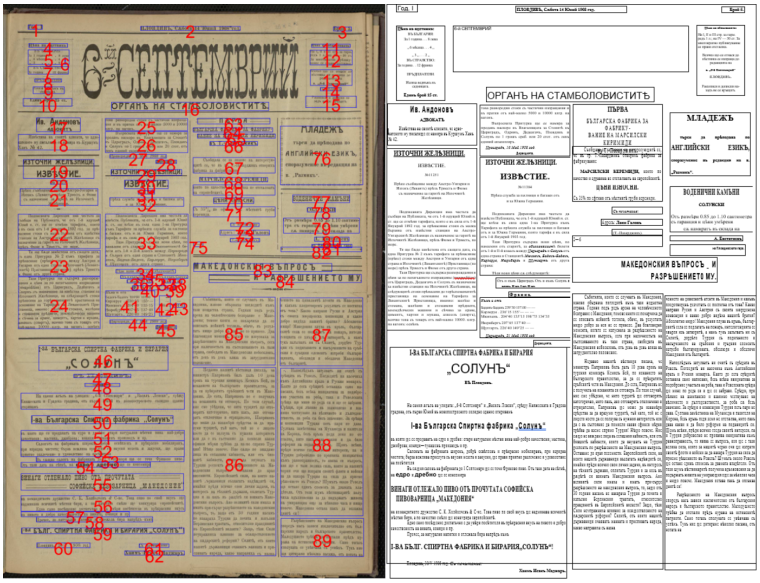


Figure 7: Here we present the segmentation of the PDF format and of the MS Word Docs format. As we could see the segmentation provided by the OCR software and stored in the MS Word Docs format is much more precise. Unfortunately it is not available to us for the already processed newspapers. We use them for the new scanned ones.

on its parameters — *subpage_gap_threshold*, *x_step*, *min_column_page_ratio*, *x_tolerance*, *min_column_width*, *partial_gap_threshold*, *y_tolerance*. The borderline cases of too large or too small values suggest automatic fine-tuning over the parameter values of the algorithm. In order to find the best parameters, we need a gold set of examples and a method of evaluation over the constructed reading order.

5.1. Our Setup

Some initial parameter settings were selected. Then we first processed, then manually checked and corrected the reading orders produced by the algorithm for 259 pages in 45 newspaper issues from different publishers with different typography. The corpus contains 1 878 295 characters, 379 395 tokens, and 10 251 blocks. In this way, we constructed a gold dataset for comparison among the reading orders produced with different parameter settings. As a baseline, we consider the reading order produced by the `pdfminer` library.

Once we have an available gold corpus of newspaper pages with correct reading order for each automatically generated reading order, we need to find a method to compare the two reading orders. For that purpose, we defined different operations over the reading orders such as *change scope of the block — dividing a block into two or more blocks; combining two or more blocks into one block*; change the order of the blocks. We formalized these operations as insert, delete, and replace of blocks in a reading order. When we had to compare two automatically generated reading orders, we decided to calculate the changing steps for each of these reading orders towards the gold standard reading order. Thus, we consider the reading order with the smallest number of steps to be better than the automatically generated reading order.

As a metric for the evaluation of the reading order, we defined a generalized version of the Levenshtein Edit Distance over blocks with respect to the changing operations mentioned above. The edit operations are the insertion and deletion of blocks, as well as the modification of block coordinates. Therefore, we consider two blocks as matching (or identical) if all of their coordinates are up to 5 points different. Using this metric, we can optimize the threshold parameters for all documents in the corpus. A possible approach is to optimize the threshold parameters only for a specific periodical edition, but this direction would require manual work to annotate at least one example.

5.2. Experimental results

We started the experiments for tuning the parameters by comparing both — the edit distance of the baseline ordering, produced by the `pdfminer`, and the ordering, produced by our algorithm with the manual set of parameters, to the ordering in the gold corpus. With initially manually set values of the parameters, we achieved a 33% reduction in the number of edits — from 2874 edits for the baseline, to 1925 edits for the initial version of the algorithm.

For fine-tuning the parameters, we performed

Baseline	Initial	Optimized
2874	1925	1702

Table 1: Edit distance between the manually annotated reading order and: the off-the-shelf library extraction (baseline) (**Baseline**), our page processing algorithm with initial intuitive set of parameters (**Initial**) and our page processing algorithm with optimized parameters over the gold dataset (**Optimized**).

Parameters	Initial	Optimized
x_step	5	5
x_tolerance	10	12
y_tolerance	20	20
subpage_gap_threshold	10	9
partial_gap_threshold	20	26
min_column_page_ratio	0.60	0.55
min_column_width	100	20
Edit distance	1925	1702

Table 2: Optimal values of the reading order parameters of the algorithm according to an extensive grid search over the gold dataset.

a grid search testing over 20,000 combinations of parameters of the algorithm, and compared the minimum edit distance between the manual annotated reading order of the gold set and the automatic reading order of the algorithm. This further reduced the edits to 1702. The test results are presented in Table 1. The optimal values for the parameters are given in Table 2.

Several experiments were also conducted with a subset of specific newspapers, where using certain individual parameter values, the results improved by up to 60% compared to the optimal values of the whole set. These results suggest that optimizing the parameter values to a specific subset of newspapers could be beneficial but requires some manual work.

The final result of the algorithmic extraction is an XML file with the following structure: pages → subpages → columns → blocks → text.

6. Manual Correction of the Extracted Reading Order

The algorithm for assigning the order of the blocks reduces manual work by a large margin. Still, there may be problems (some of them inherited from the block extraction of the `pdfminer` library):

- Noise blocks or whole columns;
- Blocks covering more than one column;
- Wrong block order.

Thus, if we want to produce an extraction of high quality, the manual correction is inevitable. At this stage of processing, our goal is the manual correction to be performed on the level of blocks. This would minimize the need to read all the internal text. Thus, we designed a software application in which users can edit the blocks themselves. The users can delete blocks and change their coordinates to have a better coverage over the text. The users can also change the reading order of the blocks. Our software visualizes changes in geometry and order of blocks immediately on the page. Furthermore, if the text is incorrectly divided into two blocks, it is easier to delete the second block and expand the first rather than merge the text.

Correcting the blocks has one more advantage, which can be beneficial in the long run. When we edit the block coordinates and do not move the text, we still have the mapping between the blocks and the characters inside them. In this way, we can use the correct block order to get the correct word reading order. In the future, one can train a model to produce a correct reading order of the words based on the 2D coordinates of the characters on the page.

For visualization and correction, we created an application that loads the XML file representing the blocks and the reading order as well as the corresponding PDF for presentation of the page with visualized blocks and their numbers — similar to the figures like Fig. 2. The application implements three commands for editing the blocks:

- Classification of blocks as normal, meta, and noise (classifying a block as noise is equivalent to deleting it). The meta and noise blocks will be skipped when extracting articles.
- Correction of the block coordinates — either a specific coordinate or all of them;
- Editing the block order by selecting the blocks to swap.

The application visualizes each change immediately. This is the approach used to create the gold dataset that was used to compare the algorithm performance. The practice shows that these three commands are sufficient to perform all the necessary edits.

7. Conclusion and Future Work

In this paper, we present an approach for defining a reading order over the blocks extracted from searchable PDFs of a Bulgarian Historical Newspaper Dataset. For this task we constructed a specialized gold dataset. We also propose a method that builds on a Levenshtein edit distance to measure the quality of the generated reading order and to

tune the parameters of the algorithm. As it can be seen, the approach is language independent.

Although the approach we presented here reduces the number of errors in the reading order, the amount of newspaper pages to be processed — over 150 000 pages in more than 300 titles, makes the manual inspection a tremendous task. Thus, we will proceed with the development of an automatic method to improve the processing of the data.

On the basis of the gold dataset, we will train several transformer-based language models. We plan to modify and fine-tune a pre-trained encoder transformer model to take as input subwords together with their 2D positions, and to output a text in the correct order.

Another future task is to add annotation of articles and metadata to the gold dataset in order to fine-tune the encoder article segmentation and classification models. In addition, spelling translation and correction will also be also performed with language models.

8. Acknowledgments

The reported work has been supported by CLaDA-BG, *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH*.

9. Bibliographical References

Jonathan Bourne. 2025. Reading the unreadable: creating a dataset of 19th century english newspapers using image-to-text language models. *Digital Scholarship in the Humanities*, page fqaf151.

Xiaowen Ding and Weiqun Huang. 2014. Pdf converter production of historical newspaper digitization: The picture experience of china's dachenglaojiu database. In *Proceedings of the International Newspaper Conference, IFLA*. International Federation of Library Associations and Institutions (IFLA).

Nicolas Gutehrlé and Iana Atanassova. 2021. Logical layout analysis applied to historical newspapers. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 85–94, NIT Silchar, India. NLP Association of India (NLPAI).

Loryn Isaacs, Santiago Chambó, and Pilar León-Araúz. 2024. Humanitarian corpora for English, French and Spanish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8418–8426, Torino, Italia. ELRA and ICCL.

Florian June. 2026. Hybrid OCR-LLM: Not a bigger model, but a smarter pipeline. Online article. AI Exploration Journey (Substack / Medium).

Eivind Kjosbakken. 2025. How to develop a powerful ocr pipeline for machine-learning systems. Towards AI article.

Benjamin Charles Germain Lee, Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, and Daniel S. Weld. 2020. The newspaper navigator dataset: Extracting headlines and visual content from 16 million historic newspaper pages in chronicling america. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3055–3062, New York, NY, USA. Association for Computing Machinery.

Christian Schultze, Niklas Kerkfeld, Kara Kuebart, Princilia Weber, Moritz Wolter, and Felix Selgert. 2025. Chronicling germany: An annotated historical newspaper dataset.

A Survey of the Digitisation of German Newspapers in interwar Lithuania (1918–1940)

Lina Plaušinaitytė, Heike Zinsmeister
Department of German Philology, Institute for German Studies
Vilnius University, University of Hamburg
lina.plausinaityte@flf.vu.lt, heike.zinsmeister@uni-hamburg.de

Abstract

This paper presents a survey of the preservation and digitisation status of the German-language press published in interwar Lithuania, which existed between 1918 and 1940. In the newly established and ethnically diverse Lithuanian Republic, which was operating in accordance with the European Minority Protection Regime, German-language newspapers and other periodicals formed a relevant part of the country's multilingual press. They represent an interesting yet underexplored resource for historical and linguistic research. The survey summarises bibliographic information and the results of earlier digitisation projects. Although systematic digitisation remains future work, this paper notes some challenges for optical character recognition (OCR) within this collection, particularly in relation to typographic variation, and outlines the envisaged format of the digital resource

Keywords: historical German newspapers, interwar Lithuania, digitisation, OCR

1. Introduction

Historical newspapers and other periodicals, such as calendars and almanacs, constitute valuable sources for the study of social, cultural, religious, economic and political life in earlier periods, as well as for tracing various development processes over time (Mills, 1981). They also represent an important resource for linguistic research. As publications addressed to a broad public readership, they generally adhere to conventional linguistic standards; however, subtle geographic and temporal influences can be detected comparing significant amounts of digitally accessible data stratified by place and time of publication. Newspaper-based corpora provide evidence for linguistic structures and their diachronic development. At the same time, newspapers' coverage of a wide range of topics provides a rich basis for analysing vocabulary use, including collocations and lexical change (e.g., Pedrazzini and McGillivray, 2022).

Digitisation of historical newspapers often results in facsimiles and low-quality automatic optical character recognition (OCR) that identifies (sub)strings in the image without actually capturing the semantic zones of the page. Further processing of structure and content of newspapers remains largely a matter of debate. OCR of historical papers in general, and newspapers in particular, presents a distinct set of challenges in comparison with modern papers (see, e.g., Springmann and Lüdeling, 2017 for OCR of historical German books published between 1487 and 1914). These problems are due to the material condition of the paper or ink, as well as to typography and layout complexities. The latter are particularly prominent in newspaper settings, where pages are partitioned into different articles with varying numbers of columns, and interspersed images, figures, etc. (e.g. Barman et al., 2021). Newspaper texts also include different font types and script variations. The historical blackletter Fraktur font (also referred to as 'Gothic' script) is particularly challenging for OCR (e.g. Généreux et al., 2014; Bjerring-Hansen et al., 2022). Advertisements in newspapers distinguish themselves from ordinary articles and often employ creative typographic layouts, including non-horizontal lines. Last but not least, historical newspapers of the diaspora exhibit not only linguistic variation due to diachronic change, but also code-switching and interference from local contact languages. These characteristics pose considerable challenges for OCR and further processing.

Despite these challenges, a number of projects provide search interfaces and download options for historical OCRed German-language newspapers, in particular the *German Newspaper Portal* (DZP), *Austrian Newspaper Online* (ANNO), and *Deutsches Textarchiv* (DTA). The *Digitales Forum Mittel- und Osteuropa* (DiFMOE) specializes in periodicals of Central and Eastern Europe. Recently, several projects have advanced research on the digitisation of historical newspapers and the application of natural language processing (NLP) tools to them; see the compilation in Ridge et al. (2019). The Swiss-based Impresso project (Ehrmann et al., 2020), for example, supports community building. In addition to a corpus collection, the project also hosts a DataLab platform with NLP resources accessible via Jupyter notebooks that can be applied to the user's own research data. Another more recent large international effort is the European project PressMint[1], which aims at compiling com-

[1] https://www.clarin.eu/pressmint

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 65–71

parable, interoperable, and annotated corpora of European historical newspapers for about the last 125 years.

None of the aforementioned resources include historical German-language newspapers from interwar Lithuania, which are the focus of the present study.

2. Historical contextualisation

To contextualize the German-language press in interwar Lithuania between 1918 and 1940, it is necessary to give a brief historical outline of the political development of the region that had a strong influence on languages and publication efforts (see, e.g., Plaušinaitytė (2021) and the literature cited therein). While playing a prominent political role in Eastern Europe in the medieval and early modern periods, Lithuania lost its independence as part of the Russian empire at the end of the 18th century. Only after the end of World War I was its political independence restored, and the formerly suppressed Lithuanian language became the national language again. In accordance with the European Minority Protection Regime, minorities were allowed to use their own languages, such as German, Yiddish, or Polish in public life, churches, schools, and even state administration. This was revolutionary for a region in which schools were only taught in Russian for over a century.[2] Figure 1 shows the political setting of Lithuania in 1939-40 with Germany (former Prussia and German Reich) in the South-West and Polish territory (modern Belarus) in the South-East. The bright yellow area is core Lithuania with Kaunas as its interim capital. The black outline marks the modern state of Lithuania since 1990.

Two of the orange-coloured areas require additional explanation. The Klaipėda region in the West (former German *Memelland*) became part of Lithuania only in 1923. It was a partly German-speaking area, because it had belonged to Prussia (or, later, the German Reich) for many centuries and had never been part of the Russian Empire. The Vilnius region in the South-East including the historical (and modern) capital of Lithuania was disputed after 1918 and came under Polish rule in 1920 until 1939.

The democratic nation of Lithuania turned into a right-wing dictatorship in 1926 and ended in 1940.[3] This is also the end of the German-language publications in the area.

In this survey, we are mainly interested in the German-language press that was published by the German minority in core Lithuania. In contrast to the Klaipėda region, their ancestors were mostly artisans and merchants invited to migrate to Lithuania in the late medieval ages, or Lutheran Christians expelled from the Austrian Salzburg area and invited to stay in neighbouring East Prussia in the 18th century, settling on both sides of the border. In the 19th century, a number of German workers also arrived in Lithuania to work on railway construction and in metal processing factories.

Figure 1: Map of territorial disputes and claims regarding Lithuania in 1939-1940. For the current survey the (yellow) core region is most relevant (image by Renata3, CC BY-SA 4.0 via Wikimedia Commons)

3. Bibliographical preservation

To explore the preservation of relevant newspapers, we consulted bibliographies and libraries' collections in Lithuania and Germany, as well as those of the Library of Congress in Washington, D.C. Table 1 summarizes our findings for German-language newspapers distributed in Kaunas in 1918–1940. Table 2 in the Appendix provides a more detailed report for one of the newspapers. The individual sources are briefly introduced in the rest of this section.

A specialised four-part bibliography of German-language periodicals from Eastern Europe, including Lithuania, has been compiled at the Regensburg *Leibniz Institute for East and Southeast European Studies*. The volumes contain bibliographic information on newspapers and journals (Weber, 2013a), popular calendars, almanacs, and yearbooks (Weber, 2013b), and also research publications on the German press in Eastern Europe (Weber, 2013c). Weber (2013a) mentions 55 German-language newspapers and journals published in Lithuania during the relevant years from 1918 to

[2] Even the printing of Lithuanian books in Latin script had been criminalized in the Russian empire.

[3] It was first under brief Soviet occupation, then occupied by Nazi Germany. From 1944 to 1990, it was part of the Soviet Union.

Newspaper	Years	Source	Comment
Litauische Rundschau / Lietuvos apžvalga	1920–1921, 1924–1929	LNB Vilnius, VL Vilnius, ZDB (and Carlton 1965)	for details on preservation and digitisation, see Table 2
Deutsche Nachrichten für Litauen / Vokieciu Zinios Lietuvoje	1931—1940	LNB, ZDB, Weber 2013a, Carlton 1965 (starting 1930[!])	published by the German-Lithuanian Cultural Association; 151 issues digitized by VU Vilnius
Kownoer Zeitung / Soldatenrat Kowno	1918	LNB Vilnius, ZDB, Carlton 1965	Kowno = Kaunas; initiated in 1916å
Die neue Zeit: Organ des Soldatenrates Kowno	1918—1919	LNB Vilnius, ZDB	
Korrespondenz B	1918	ZDB, Weber 2013a	'Reports [...] from the administrative territory of the Oberbefehlshaber Ost', since 1916
Baltisch-Litauische Mitteilungen	1918	ZDB, Weber 2013a	successor of *Korrespondenz B*

Table 1: German-language newspapers published in Kaunas, the interim capital of Lithuania, between 2018 and 1940 (abbreviations are explained in the main text)

1940, only three of which were published in the interwar capital Kaunas (see Table 1), nine in Vilnius, and the remaining 41 in the Klaipėda region.

Copies of German newspapers have been systematically collected at least since 1912 when the *Deutsche Bücherei* was founded by the German Booksellers Society among others, which developed into the modern *Deutsche Nationalbibliothek* that hosts the German Union Catalogue of Serials (*Zeitschriftendatenbank* ZDB) which "is the largest dedicated database for serials of all kinds, particularly journals and newspapers." Its interface allows filtering according to dates, language, location of distribution, among others. The ZDB points to six relevant newspapers published in Kaunas as shown in Table 1. In addition, it mentions 14 newspapers published in Vilnius and 29 in the Klaipėda region. The ZDB aggregates information about the availability of paper copies and microfilm holdings of newspaper issues in German and Austrian libraries. However, it does not take into account resources in other countries, such as the extensive collection of German-language newspapers held by the *Martynas Mažvydas National Library of Lithuania* (LNB)[4] in Vilnius, which, for example, includes 1,167 issues of the newspaper *Litauische Rundschau* in print (as well as approximately 440 duplicate issues). A smaller collection of this newspaper is held by the Vrublevski Library (VL) of the Lithuanian Academy of Sciences in Vilnius, which also provides facsimiles for eight issues, see Table 2. We also consulted the bibliographic resources of the Library of Congress in Washington, D.C., the world's largest library, in particular Carlton et al. (1965).

While the German newspaper portal (DZP) does not host any of the newspapers published in interwar Lithuania, it offers interesting inter-textual insights into their reception in the German Reich, exemplified in Figure 2.

Figure 2: Mention of *Litauische Rundschau* in Badischer Presse in 1927 (Source: DZP, Visualisation: DFG Viewer)

4. Digitisation

The intended target representation of the digitised material comprises facsimile and OCRed text with positional coordinates, as well as article and zone segmentation, enabling users to switch between

[4] LNB: https://www.lnb.lt/

the text and its position in the facsimile. We intend to use the DFG Viewer tool[5], which requires METS/MODS[6] and structured METS/TEI[7] representations. In addition to plain text access via the DFG Viewer, we aim to compile the newspaper articles into a linguistically annotated corpus, at least enriched with lemmas and parts of speech, and make it searchable for linguists, for example using SpaCy[8] for annotation and ANNIS (Krause and Zeldes, 2016) for search, or the more recent Discourse Analysis Tool Suite (DATS) (Fischer and Biemann, 2025).

4.1. What has been digitised of interwar Lithuanian newspapers?

As mentioned in section 3 and shown in Table 2, the *Vrublevki Library* of the Lithuanian Academy of Sciences offers facsimiles of eight issues of *Litauische Rundschau*, each issue comprising about two to eight pages. A much larger collection of higher quality images is available at *Die Presse der deutschen Minderheit in Litauen 1918-1940* (Nareckaitė and Plaušinaitytė, 2020) that provides facsimiles for two newspapers and two further periodical resources, accompanied by neat introductions that contextualize the respective resources. In particular, these are: *Litauische Rundschau* with 24 issues from 1920 and 127 issues from 1912,[9] and *Deutsche Nachrichten für Litauen* with 151 issues unevenly distributed over six years. The other two resources are *Deutsche Genossenschaftnachrichten* (eight issues) and *Deutscher Kalender für Litauen* (ten annual editions with more than 100 pages on average). Several facsimiles of interwar German press publications have also been published on the Lithuanian digital heritage portal *epaveldas.lt*. Here you can find some issues of *Deutscher Kalender für Litauen* (1922, 1924, 1925, 1932, 1933).

4.2. OCR support

In the process of preparing our own digitisation project, we could rely on an active community concerned with questions of OCR, including information by the special interest group for newspapers and journals of the Association of Digital Humanities in the German-speaking area (DHd)[10], a monthly online OCR consulting hour of experts

Figure 3: Facsimiles of *Litauische Rundschau* from year 1920 exemplifying some of the challenges for OCR: different font sizes, complex page layout, advertisements, bad paper quality with shining-through letters.

from German National Data Infrastructure (NDFI) consortia, and an online OCR recommender[11] that provides detailed recommendations for different OCR software and transcription tools. One of the reviewers also pointed out the extensive resources and community work of the project OCR-D.[12]

Even with professional support, the task of digitising historical newspaper is substantial. Figure 3

[5] https://dfg-viewer.de/en/the-project

[6] https://www.loc.gov/standards/mods/

[7] https://www.loc.gov/standards/mets/METSOverview.v2.html

[8] SpaCy:https://spacy.io/

[9] Litauische Rundschau: https://www.dpl.flf.vu.lt/litauische-rundschau/

[10] https://dhd-ag-zz.github.io/index.html.

[11] OCR recommender by BERD@NFDI: https://wiki.bib.uni-mannheim.de/limesurvey/index.php/996387

[12] https://ocr-d.de/

shows three example page sections of *Litauische Rundschau* from 1920 that exhibit challenging characteristics for OCR. Whiles these pages are printed in Antiqua script, issues from this newspaper published between 1924 and 1929 present the addition challenge of being printed in Fraktur script (not shown here).

5. Conclusion and future work

The paper provides a survey of historical German-language newspapers published in interwar Lithuania. By detailing the political history of the area and distinguishing three different historical regions in the area of the modern Lithuanian state, we try to motivate the selection of newspapers with which we want to work. Our bibliographic search identified relevant newspaper collections in Lithuania and Germany and also collections of facsimiles.

As far as we are aware, there is no machine-readable corpus of interwar Lithuanian German-language newspapers yet. The value of physically inspecting the available paper copies should not be underestimated either. When browsing through the collections, one comes across unexpected discoveries. For instance, whilst examining the collection of paper copies of the *Litauische Rundschau* from 1926 at the National Library of Lithuania (LNB), several issues of another German newspaper from Kaunas, Der *Wächter/Sargas*, were discovered. The title is neither listed in the library catalogue, nor mentioned in Weber's bibliography.

We performed pilot studies with different OCR tools in the context of a Digital Humanities and Linguistics seminar at the University of Hamburg, which resulted in very low-quality OCR texts, mainly because the semantic structures of the pages were not correctly identified. We are aware that this reflects not only the challenges of applying OCR to this type of image, but also our own limitations as novices in the field. In the next step, we will address the challenging task of applying OCR to the facsimiles in a more systematic manner.

Acknowledgements

We would like to say thanks to three anonymous reviewers for their helpful comments and suggestions. We would also like to thank the participants of the workshop "Korpora und Editionen. Ansätze der Digital Humanities und didaktische Perspektiven" at Vilnius University for interesting discussions. Part of this work was funded by the German Academic Exchange Service (DAAD) with funds from the German Federal Foreign Office, project ID 57759025.

Bibliographical References

Raphaël Barman, Maud Ehrmann, Simon Clematide, Sofia Ares Oliveira, and Frédéric Kaplan. 2021. Combining visual and textual features for semantic segmentation of historical newspapers. *Journal of Data Mining & Digital Humanities*, (HistoInformatics).

Jens Bjerring-Hansen, Ross Deans Kristensen-McLachlan, Philip Diderichsen, and Dorte Haltrup Hansen. 2022. Mending fractured texts. A heuristic procedure for correcting OCR data. In *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference, DHNB 2022*, volume 3232 of *CEURS Workshop proceedings*, pages 177–186.

Robert G. Carlton et al. 1965. Newspapers of East Central and Southeastern Europe in the Library of Congress. Slavic and Central European Division, Reference Department, Library of Congress, Washington.

Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. 2020. Language Resources for Historical newspapers: the Impresso Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 958–968, Marseille, France. European Language Resources Association.

Tim Fischer and Chris Biemann. 2025. Semi-automatic Sequential Sentence Classification in the Discourse Analysis Tool Suite. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 151–162, Albuquerque, New Mexico. Association for Computational Linguistics.

Michel Généreux, Egon W Stemle, Verena Lyding, and Lionel Nicolas. 2014. Correcting OCR errors for German in Fraktur font. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014: 9-11 December 2014, Pisa*, pages 186–190. Pisa University Press.

Thomas Krause and Amir Zeldes. 2016. ANNIS3: A New Architecture for Generic Corpus Query and Visualization. *Literary and Linguistic Computing*, 31(1):118–139.

T. F. Mills. 1981. Preserving yesterday's news for today's historian: A brief history of news-

paper preservation, bibliography, and indexing. *The Journal of Library History (1974-1987)*, 16(3):463–487.

Nilo Pedrazzini and Barbara McGillivray. 2022. Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 85–95, Taipei, Taiwan. Association for Computational Linguistics.

Lina Plaušinaitytė. 2021. Der Gebrauch der litauischen Sprache in der Presse der deutschen Minderheit im Litauen der Zwischenkriegszeit. In *Schnittstelle Germanistik: Forum für Deutsche Sprache, Literatur und Kultur des mittleren und östlichen Europas.*, volume 1, pages 31–55. Universitätsverlag WINTER GmbH Heidelberg.

Mia Ridge, Giovanni Colavizza, Laurel Brake, Maud Ehrmann, Jean-Phillipe Moreux, and Andrew Prescott. 2019. The past, present and future of digital scholarship with newspaper collections. *Multi-paper panel in DH 2019 book of abstracts*.

Uwe Springmann and Anke Lüdeling. 2017. OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. *Digital Humanities Quarterly*, 11(2).

Albert Weber. 2013a. Teil 1: Zeitungen und Zeitschriften. In *Bibliographie deutschsprachiger Periodika aus dem östlichen Europa*. Institut für Ost- und Südosteuropaforschung.

Albert Weber. 2013b. Teil 2: Volkskalender, Almanache und Jahrbücher. In *Bibliographie deutschsprachiger Periodika aus dem östlichen Europa*. Institut für Ost- und Südosteuropaforschung.

Albert Weber. 2013c. Teil 3: Fachbibliographie deutschsprachiger Periodika. In *Bibliographie deutschsprachiger Periodika aus dem östlichen Europa*. Institut für Ost- und Südosteuropaforschung.

Language Resource References

ANNO. *AustriaN Newspaper Online*. Austrian National Library (ÖNB).

DiFMOE. *Periodika der Digitalen Bibliothek*. Digitales Forum Mittel- und Osteuropa.

DTA. *Erweiterungskorpus des Deutschen Textarchivs, Genre Zeitung*. Digitalen Wörterbuchs der deutschen Sprache.

DZP. *Deutsches Zeitungsportal, German Digital Newspaper Portal*. German Digital Bibliothek (DDB).

Nareckaitė, Vidmantė and Plaušinaitytė, Lina. 2020. *Die Presse der deutschen Minderheit in Litauen 1918-1940*. Vilnius University.

VL. *Vrublevski Library of the Lithuanian Academy of Sciences*.

Appendix

Year	ID	Format	Issues	Location	Comment
1920	[1]	image (pdf)	1–24	VU Vilnius	based on [3]
	[2]		15, 18, 20–23	VL Vilnius	based on [4]
	[3]	paper	1–24	LNB Vilnius	07/16-10/08
	[4]		*	VL Vilnius	
	[5]		2–84	DNB Leipzig	07/20–12/30
	[6]	microfilm	2–84	NOB Lüneburg, IFA Stuttgart	based on [5]
1921	[7]	image (pdf)	1–127	VU Vilnius	based on [9]
	[8]		80, 87	VL Vilnius	based on [10]
	[9]	paper	1–127	LNB Vilnius	01/01–06/29
	[10]		*	VL Vilnius	01/01–07/17
	[11]		2–146	DNB Leipzig	01/04–07/22
	[12]	microfilm	1–146	NOB Lüneburg, IFA Stuttgart	based on [11]
1922–1923			not published		
1924	[13]	paper	1–157	LNB Vilnius	06/08–12/31
	[14]		1–156	LNB Vilnius	
	[15]		*	VL Vilnius	
	[16]		2–157	DNB Leipzig	06/11–12/31
	[17]	microfilm	2–157	NOB Lüneburg, IFA Stuttgart	based on [16]
1925	[18]	paper	1–293	LNB Vilnius	01/01–12/31
	[19]		*	VL Vilnius	
	[20]		1–293	DNB Leipzig	
	[21]	microfilm	1–293	NOB Lüneburg, IFA Stuttgart	based on [20]
1926	[22]	paper	1–132,134-.279	LNB Vilnius	
	[23]	paper	27–28	LNB Vilnius	
	[24]		1–286	DNB Leipzig	01/01–12/19
	[25]	microfilm	1–286	NOB Lüneburg, IFA Stuttgart	based on [24]
1927	[26]	paper	1–295	DNB Leipzig	01/01–12/31
	[27]	microfilm	1–295	NOB Lüneburg, IFA Stuttgart	based on [26]
1928	[28]	paper	1–215,217–289	LNB Vilnius	
	[29]	paper	1–222,224–249, 251–256,258–289	LNB Vilnius	
	[30]		1–289	DNB Leipzig	01/01–12/30
	[31]	microfilm	1–289	NOB Lüneburg, IFA Stuttgart	
1929	[32]	paper	1–48		01/01–02/27
	[33]		1–145	DNB Leipzig	01/01–06/29
	[34]	microfilm	1–145	NOB Lüneburg, IFA Stuttgart	based on [33]

Table 2: Status of **Litauische Rundschau** = Lietuvos apžvalga ('Lithuanian Review'), published from 1920–1921 and 1924–1929; periodicity: initially two or three times weekly, later daily, irregularly; '*': availability of paper copies not verified; **VU Vilnius**: Nareckaitė and Plaušinaitytė (2020); **VL Vilnius**: Vrublevski Library of the Lithuanian Academy of Sciences; **LNB Vilnius**: Martynas Mažvydas National Library of Lithuania; **DNB Leipzig**: Deutsche Nationalbibliothek; **NOB Lüneburg**: Nordost Institut; **IFA Stuttgart**: Institut für Auslandsbeziehungen; other libraries in Berlin and Marburg also hold some issues of 1924 and 1925.

Toward Interoperable and Scalable Representations of Complex Heterogeneous Digitized Historical Media

Pauline Conti[1], **Simon Clematide**[2], **Maud Ehrmann**[1]
[1]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[2]University of Zurich (UZH), Switzerland
<firstname>.<lastname>@epfl.ch, <firstname>.<lastname>@uzh.ch

Abstract

The value of digitized historical media archives for computational historical research is now well established, yet an underexplored challenge concerns data management itself: how to represent and process, at scale, complex primary sources that vary widely in digitization granularity, refinement quality, and archival organization and curation practices. This paper presents the data representation framework designed for large-scale processing and indexing of historical newspapers and radio broadcasts developed within the Impresso project. Grounded in a structured characterization of the heterogeneity found in digitized historical media collections, it identifies the distinct dimensions along which collections diverge and the challenges they pose for a unified representation and processing framework. The framework navigates the competing demands of machine learning pipelines requiring uniform and lightweight document representations, information retrieval systems requiring well-defined indexable content units, user-facing interfaces requiring fidelity to original sources, and the need to return semantically enriched data to archival holders in interoperable formats. We describe the design principles guiding the framework and discuss how it reconciles these constraints across highly heterogeneous collections into a unified and research-ready corpus.

Keywords: digitized newspapers, digitized broadcasts, data representation and processing, heterogeneity and interoperability, machine learning, information retrieval

1. Introduction

The digitization of historical media collections has accelerated considerably over the past two decades, producing vast repositories of machine-readable content from newspapers and, increasingly, radio broadcasts (Balk and Conteh, 2011; Neudecker and Antonacopoulos, 2016). Held by libraries and cultural heritage institutions across Europe and beyond, these collections have opened new avenues for historical research: from full-text search and browsing to semantic enrichment and, more recently, to semantic indexing enabling similarity-based exploration across large corpora (Neudecker, 2022; Düring et al., 2023). The historical and scholarly value of such resources is now widely recognized (Bunout et al., 2023).

Yet this value is greatly amplified when research can be conducted at scale, across institutional, linguistic, and national boundaries. Historical newspapers and broadcasts are, in practice, fragmented across institutional silos: digitized collections are typically bound to a single institution, language, or country, maintained in distinct formats, and designed primarily for preservation and consultation rather than programmatic processing and analysis. This fragmentation makes large-scale processing difficult, hinders the construction of longitudinal research datasets, and severely limits interoperability across collections (Padilla, 2019; Ehrmann et al., 2023a).

Bridging these silos is one of the main objectives of the 'Impresso - Media Monitoring of the Past' project, which focuses on processing, enriching, and enabling the exploration of large-scale historical media sources to increase their accessibility and usability for digital historical research[1]. In its second phase, the project pioneers the joint exploration of Western European newspapers and radio content across temporal, linguistic, and national boundaries, drawing on collections from more than 20 partner institutions. It develops and applies machine learning-based text and image mining approaches – including named entity recognition, topic modeling, text reuse detection, and embedding-based similarity search – to enrich and index a transnational corpus of facsimiles, OCR and ASR transcripts, and associated metadata. The resulting enriched corpus is made accessible through a graphical web application, the Impresso WebApp, and a programmatic access ecosystem, the Impresso Datalab, to support both human-centred and data-driven historical inquiry with transnational and transmedia perspectives.

Beyond well-documented challenges such as OCR noise, the difficulty of applying NLP to historical text, and digitization bias in collections (van Strien et al., 2020; Ehrmann et al., 2023b; Beelen et al., 2025; Opitz et al., 2026), a less frequently examined challenge concerns the technical dimension of data preparation and collection management. Difficulties arise at three levels: at the level of the historical source, newspapers are complex

[1] https://impresso-project.ch

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 72–81
May 16, 2026.

SOURCES	*Media*	NEWSPAPERS (INCLUDING RADIO MAGAZINES)		RADIO BROADCASTS	
	Medium	Print		Typescripts	Radio records
INPUTS	*Modality*	Image	Text (OCR)	Text (ASR)	Audio
	Metadata	Publication year, place, publisher, author, size or length, copyright, ...			
	Language	—	Dutch, English, German, French, Luxembourgish		
PROCESSING	*Semantic enrichment*	Article and images alignment, image classification	Semantic segmentation, NERC, EL, Keyphrases, Topics, Classes, Opinion, Content reuse		—
	Semantic Indexing	Dense vector representation	(Clustered) multilingual dense vector representation of text and enrichments		—
ACCESS	*Content Retrieval*	Visual search and text search (captions, related articles)	Cross-lingual faceted text search and exploration of enrichments		—
	Displayed Objects	Images	Images & Text (OCR)	Text (ASR)	Audio streams

Table 1: Alignment of sources with the types of input they correspond to, the processing they undergo, and the search and rendering modes supported by the interface.

objects whose value lies precisely in their material and editorial structure: a heterogeneous mix of text, images, tables, and graphical elements, organized across issues, pages, and articles in ways that are historically meaningful. Radio broadcasts share a similar complexity, albeit in less documented and more irregular ways, raising non-trivial questions about how the two media can be aligned and compared. At the level of the digitized record, this complexity is inherited and extended: a digitized newspaper content item is not simply raw text, but a layered object comprising facsimiles, OCR transcripts, layout segmentation, content organization, and bibliographic metadata. At the level of the collection, finally, decades of digitization campaigns across institutions compound the difficulty further: collections differ in file formats, processing granularity and quality, and archival organization. All of this is further heightened by the volume of data involved, the requirements of large-scale machine learning and information retrieval, and the need to remain responsive to how historians work with primary sources.

This raises the question: How can complex historical image-text objects, heterogeneous in origin and digitization practices, be represented, processed and indexed at scale without losing what makes them meaningful as historical sources? This paper presents a data representation model and conversion architecture developed within Impresso, grounded in explicit design principles and a structured characterization of collection heterogeneity. The framework reconciles documentary fidelity with machine learning and information retrieval requirements across heterogeneous historical newspapers and broadcast sources, and is validated on a large transnational multilingual corpus.

The rest of this paper is structured as follows. Section 2 outlines the design principles that shaped the framework, Section 3 details the types of heterogeneity we face with such collections, Section 4 reviews existing representation formats, Section 5 describes the framework, and Section 6 discusses and concludes.

2. Design Principles and Requirements

Impresso collects 500+ digitized newspapers and radio sources from libraries and archives, processes them through a semantic enrichment pipeline, and makes the resulting data accessible via two interfaces. Table 1 gives an overview of the media sources, their input types, and the processing and access scenarios they undergo, from visual search over newspaper images to cross-lingual faceted search over enriched transcripts.

The raw input consists, for newspapers, of page facsimiles, OCR transcripts organized as word- or region-level bounding boxes, and layout segmentation when available; for audio sources, of recordings and ASR transcripts. Both are accompanied by bibliographic metadata, which is not considered further in this paper.

The stewardship of such a corpus — that is, its representation and manipulation at scale — is a complex undertaking, shaped by application-specific requirements from the Impresso context and guided by the FAIR principles (Findable, Accessible, Interoperable and Reusable), with which our framework strives to comply. We describe these requirements in turn below.

Breaking Silos: Heterogeneous Collections at Scale The first guiding principle is to break collection silos, of which we identify three kinds. *Media*

silos: historical newspapers and radio broadcasts are inherently preserved in separate collections, reflect different production and consumption practices, and follow different archival logics — yet both are historical media sources worth studying in conjunction, and our framework must accommodate the representation of both. *Digitization silos*: each institution has conducted its own digitization campaigns over the past thirty years, resulting in collections that differ in file formats, processing granularity — from raw OCR output to fine-grained article-level segmentation — quality, and archival organization. *Language silos*: the multilingual nature of the corpus introduces tokenization differences and variable accuracy of automatic language identification, both of which affect downstream text processing. Integrating these heterogeneous collections into a unified corpus is one of the primary drivers of our framework's design, detailed in Section 3.

Fidelity to the Source A second requirement concerns fidelity to the source. Historical newspapers are not merely containers of text but also a structured arrangement of articles, advertisements, images, tables, and other content types across pages and issues, reflecting editorial choices and historical contexts. Radio broadcasts similarly carry meaning through their program structure, sequencing, and temporal organization. A user navigating the Impresso interface should be able to read an article in the context of its page or a broadcast segment within its program, access the original facsimile or audio recording, and situate it within the broader collection. This requires that our data representation maintain the logical structure of each source along with the coordinates – spatial for newspapers, temporal for radio – linking transcripts back to the original medium.

Amenability to Machine Learning and Information Retrieval Enabling large-scale semantic enrichment and indexing requires a data representation that is uniform and lightweight, properties that are in tension with the source fidelity requirement. For machine learning (ML), data must be prepared in a format that is stripped of archival and layout overhead, and in which text content is reconstructed as running text rather than the word-by-word output of raw OCR. This reconstructed text is what the enrichment pipeline operates on. A complementary requirement concerns the identification of the canonical unit of indexing: what constitutes a document for the purposes of the search engine. Ideally, this unit should follow the natural structure of the source media (an article for newspapers, a broadcast episode for radio) but such document-level segmentation is not always available and must sometimes be approximated from coarser representations.

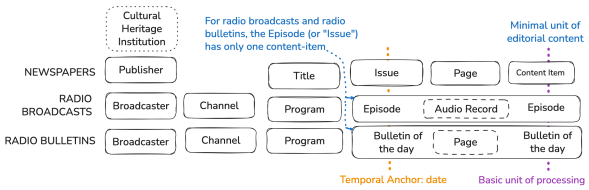


Figure 1: Newspaper and radio source alignment.

Corpus Growth and Framework Modularity Next, the Impresso corpus is not static: new collections are added regularly, and new source types may be integrated as the project evolves. This expectation of growth requires a data representation that is unified yet flexible enough to accommodate new collection variants. First, it must provide a clear and stable scheme for assigning identifiers to each element across the corpus, so that partial updates, such as the re-OCRisation of an existing collection, do not destabilize the broader corpus. Second, it must support a consistent collection organization – a unified file structure, a rigorous protocol for integrating new collections, and a principled versioning scheme covering both staging and releases – allowing new sources to be onboarded without disrupting the existing structure.

Returning Enriched Data to Archival Holders A final principle concerns the return of enriched data to partner institutions. Having provided their collections for processing, institutions expect to receive back the semantic enrichments — named entities, topics, embeddings, and other annotations — in formats compatible with their own systems. This requires that our representations maintain clear links to each institution's original identifiers, directly reinforcing the interoperability and reusability dimensions of FAIR.

Collectively, these principles define a design space shaped by three partly competing demands: fidelity to the source, interoperability, and ML and IR efficiency — respectively requiring preservation of source structure, stable and standardized representations, and lightweight uniform text optimized for large-scale processing. Yet these are requirements set against a complex material reality, which in practice must be understood before it can be managed.

3. A Typology of Heterogeneity in Digitized Historical Media

At scale, heterogeneity is the most pervasive characteristic of digitized historical media collections.

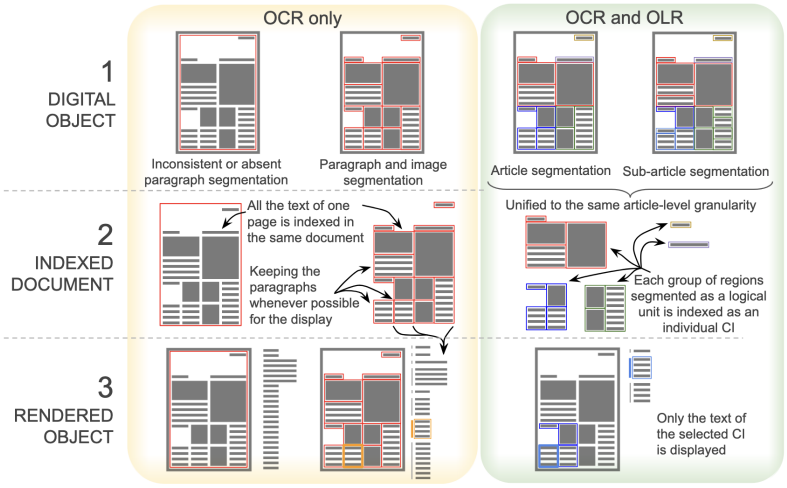


Figure 2: Illustration of the various granularity levels of the OCR and OLR structure present, and how they are indexed and rendered in the Impresso App.

Mapping these dimensions is a prerequisite for organizing the data, as it reveals what must be accounted for. We identify four such dimensions: media sources, digitized records, collection organization and identification, and language and tokenization.

3.1. Heterogeneity of Media Sources

The Impresso corpus spans two broad categories of media sources differing in medium, modality, and archival structure, requiring shared abstractions.

Sources differ in their medium and modality. Newspapers and radio bulletins[2] are print or typescript-based, and their text content is extracted via OCR, sometimes complemented by OLR for structure. Radio broadcasts, by contrast, are audio-based, and their text is produced via automatic speech recognition (ASR). These two extraction processes produce outputs in different formats, each requiring its own processing pipeline, and with characteristic error profiles that affect downstream NLP differently.

Structurally, newspapers follow a simple hierarchy: a title publishes issues at regular intervals, each composed of pages and articles. Radio sources are more irregular: broadcast episodes and bulletins group into recurring programs, with topical diversity introduced at the channel rather than the episode level. Figure 1 illustrates the resulting alignment challenges across source types.

Across this diversity, the framework strives for a common minimal unit of editorial content suitable for indexing, and a shared temporal anchor — the day — across all source types.

[2]Radio bulletins are the typescript scripts read on air by radio presenters.

3.2. Heterogeneity of Digitized Records

Beyond media sources, digitized records are layered objects whose structure and quality vary considerably across collections, along three dimensions: format and schema variants, structural refinement and segmentation, and refinement quality.

3.2.1. Format and Schema Variants

The dominant formats for encoding digitized newspaper content are METS and ALTO[3] — METS for the logical structure of an issue, ALTO for the OCR transcription of individual pages. Both are widely adopted standards yet, in practice, each digitization campaign produces its own flavor: slightly different element names, attribute conventions, nesting structures, and levels of detail. In some collections, only ALTO files are present; page-level OCR is available but not logical grouping of content into articles. Other collections provide both, but with varying degrees of completeness. Additional formats encountered include hOCR and PDF-derived text, each with their own structural conventions. Page images, which serve as the visual anchor for all text coordinates, are similarly provided in varying formats and resolutions. This diversity means that each collection — sometimes each sub-collection within a single institution — requires its own preprocessing and ingestion pipeline, and that any unified representation must abstract away from these format variations.

3.2.2. Structural Refinement and Segmentation

Digitization campaigns differ in how much source structure they recover, distinguishing physical lay-

[3]loc.gov/standards/mets/, loc.gov/standards/alto/

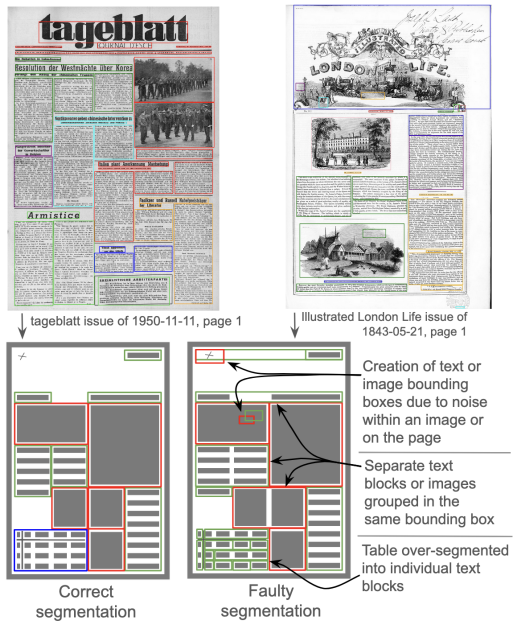


Figure 3: Real and schematic examples of typical OCR/OLR segmentation errors. In the facsimiles, the boxes of one article share their color. In the schemas, colored boxes indicate content type labels: images (red), text (green), tables (blue).

out recognition — identifying page regions such as text blocks, images, and tables — from logical layout recognition, which groups these into meaningful units such as articles or advertisements. We distinguish four levels of increasing structural recovery, illustrated in Figure 2:

No physical layout: OCR embedded in PDFs or older collections where paragraph segmentation is absent or inconsistent, including radio bulletins. *Physical layout only*: text blocks and region boundaries are identified, but not grouped into semantic units; images and tables may be detected, though their relation to surrounding text is not captured. This is the most common level in older campaigns. *Physical and logical layout*: OLR is applied alongside OCR, grouping text blocks into content items — articles, advertisements, obituaries — linking image regions to text, and enabling reconstruction of articles spanning multiple pages. This is the target granularity our framework aims to match. *Extended logical layout*: At the most refined level, content items are further decomposed into labeled sub-elements at paragraph level. Present in only a small subset of collections, the resulting nested structures can be difficult to parse reliably.

Radio sources are generally sparser: logical layout does not apply to audio recordings, and segmentation, when present, reflects program or segment boundaries rather than article-level structure.

The semantic labeling of identified units also varies across collections. Most OLR-processed data features a set of content type labels — article, advertisement, image, table, obituary — but the exact vocabulary, granularity of classes, and consistency of application differ from one campaign to the next. Label recall is often low, particularly when coupled with segmentation errors, and image-caption linking is frequently incomplete.

3.2.3. Refinement Quality

Independently of segmentation level, the quality of digitization outputs varies considerably across and within collections. OCR and ASR quality — affected by image resolution, font style, conservation state, and the age of the material — constitutes a persistent challenge for downstream NLP, though it falls outside the scope of the representation framework itself. OLR quality, by contrast, directly shapes what structural information is available to represent.

OLR errors manifest at the level of region segmentation and classification: columns merged into single blocks, tables hyper-segmented into individual cells, advertisements grouped together, or bounding boxes hallucinated over images or handwritten annotations. Misclassification of content types further reduces the reliability of article-level groupings. Figure 3 illustrates typical examples: the two paragraphs with purple boxes should be green, and the linking between the green-box article and its images and captions was partially lost.

Taken together, these variations mean that a robust framework must provide a unified abstraction over these document element types and granularities, while retaining enough flexibility to account for each collection's specific segmentation characteristics.

3.3. Heterogeneity of Collection Organization and Identification

Each institution organizes its collections according to its own archival logic, resulting in heterogeneous identifier schemes across the corpus. Identifiers exist at each level of the hierarchy — title, issue, page, content item — but vary in scope, permanence, and semantics. Some are persistent and institution-wide, such as ARK IDs resolving to stable URLs[4]; others are internal to a digitization campaign and may be reassigned upon reprocessing. Some encode explicit information — title, date, page number — while others are opaque UUID-like strings requiring standoff metadata.

A more structural dimension concerns the presence of identifiers at the content item level, which is directly tied to segmentation: in the absence of OLR, articles carry no identifiers, creating compatibility and sustainability challenges when segmenta-

[4] arks.org/about/ark-overview/

tion is added later. Such challenges are not merely theoretical: we encountered cases where collections were re-OCRed or re-segmented by partner institutions, causing identifiers to be reassigned and breaking the correspondence with our enrichments.

These variations required careful documentation of each collection's identifier logic before any processing could take place, and informed our own identifier scheme's design, described in Section 5.

3.4. Heterogeneity of Language and Tokenization

A final dimension concerns the textual units produced by OCR and ASR, and the way text is tokenized within them. Each digitization campaign has used different OCR engines and post-processing pipelines, resulting in inconsistent tokenization across collections: word boundaries, hyphenation handling, and the treatment of punctuation — whether encoded as a separate token or not — vary from one collection to the next, and sometimes within the same collection across digitization campaigns.

This variability is further compounded by the multilinguality of the corpus. Each language carries its own orthographic conventions, and the accuracy of language identification — a prerequisite for applying language-specific processing — is itself variable across collections and historical periods. Older sources in particular may use archaic spelling, ligatures, or non-standard character encodings that challenge both language identification and tokenization.

Reconstructing running text from raw OCR output — a requirement for downstream NLP, as discussed in Section 2 — therefore requires language-dependent normalization rules that account for these variations: resolving hyphenation across line breaks, reattaching punctuation, standardizing encoding, and handling script-specific conventions. These rules must be defined and maintained per collection and per language.

Before introducing our data representation model, we briefly survey existing frameworks.

4. Related Work

The *Text Encoding Initiative* (TEI) P5 is a widely adopted XML standard for richly structured and semantically expressive text encoding in the humanities. TEI schemas define an extensive set of elements for capturing detailed structural, editorial, and semantic markup, and projects typically customize the base TEI schema via ODD (One Document Does It All) to match corpus-specific needs (Consortium, 2025; Cummings, 2019). The PressMint initiative, a flagship effort of the CLARIN infrastructure, applies a customized TEI P5 schema to compile interoperable corpora of newspapers, from which multiple downstream formats (e.g., CoNLL-U, JSON) are derived for analysis and tooling (PressMint, 2026).

Models based on TEI P5, such as PressMint, prioritize semantic richness and hierarchical text representation that supports detailed editorial tasks and scholarly interoperability across languages and national contexts. However, because TEI encapsulates a large, optional element space, fully utilizing TEI in computational pipelines often requires extensive schema profiles and transformations to extract *fixed-granularity*, machine-ready units suitable for indexing or machine learning tasks. Customization of TEI vocabularies further implies that distinct corpora may diverge in practice unless governed by shared profiles prior to conversion.

In contrast, our JSONL-based representation defines a flat, uniform, processing-ready document abstraction in which heterogeneous canonical metadata and text segments are systematically consolidated into consistent records optimized for large-scale indexing and downstream machine learning workflows. This design aligns with engineering requirements for scalable processing, trading off some of the expressivity and embedded semantic nuance of TEI's richly structured models in favor of simplicity, consistency, and integration with modern data science tooling.

5. Proposed Framework

The design principles and heterogeneity typology outlined above guided the development of our data representation framework. While it does not resolve every challenge, it provides a robust foundation that accommodates evolving data and user needs.

5.1. Conceptual Model: Two Complementary Representations

Impresso's data representation is based on two complementary data structures which, together, hold the physical and logical content of the source material, each addressing a specific design constraint mentioned in Section 2, namely format uniformity and fidelity to the source.

Canonical Format The canonical format is the first and most source-faithful of the two representations. Its primary objective is to bring all incoming data — regardless of origin, format, or source type — into a single unified representation, retaining only information relevant to the source's logical structure and layout. It is composed of two conceptual data

objects: *Issues* and *Physical Supports*, the latter instantiated as pages or audio records.

Issue objects aggregate all relevant information about their structure: the physical supports they rely on, the list of content items they comprise — editorial units below the page level such as articles, advertisements, images, tables, obituaries, and other content types — and technical metadata. They are uniquely identified by their media title, publication or airing date, and edition, the latter distinguishing multiple editions published on the same day.

Initially based on the newspaper archival object, the issue representation now accommodates other source types — radio broadcasts and bulletins — as illustrated in Figure 1. Mapping radio content to this structure raised the question of what constitutes a radio equivalent to a newspaper issue. Two options were considered: grouping all broadcasts of a given day and channel into a single issue (akin to a daily newspaper edition), or treating each broadcast episode as its own issue. The first option echoes the topical variety and publication regularity of a newspaper issue, while the second reflects the grouping of broadcasts into thematic programs, which become equivalent to newspaper titles. The latter was favored as it better reflects radio's archival and thematic organization.

Physical support objects — pages or audio records — establish the link between the abstract representation and the digitized source (page image or audio record file). Page objects contain basic metadata along with bounding-box coordinates, textual content, and content-item affiliation at region, paragraph, line, and token level. Audio record objects follow the same logic, with timestamps for each speech segment and, where available, speaker turns.

Together, these two object types reduce the storage overhead of layout information while maintaining a close connection to the source and its archival structure, providing a consistent representation across source types for subsequent pipeline steps and interface display.

Rebuilt Format The rebuilt format assembles, for each content item, a self-contained and ML-ready document representation (a content item). Metadata is drawn from the issue object, while full-text content is reconstructed as running text from the token-level bounding boxes or timestamps present in the physical support objects, along with text offsets marking line, paragraph, and region boundaries. The result is a uniform, lightweight media document that abstracts away from source-specific formatting and is ready for downstream processing.

Like the canonical format, the rebuilt format accommodates both page-based and audio-based content items. It constitutes the basic unit of processing for all pipeline steps — semantic enrichment, embedding, and indexing — and provides the fixed granularity level at which content is indexed for information retrieval.

5.2. Design Choices: Addressing the Requirements

Systematically Qualifying the Media Sources To characterize the sources beyond the simple newspaper/radio binary distinction – which does not fully capture their diversity – we define two properties for all data objects: source type and source medium.

Source *type* refers to the specific media of the source (labels on the left of Figure 1), and allows to distinguish between the different types of radio sources. Source *medium* refers to the format in which the source was originally produced — print, typescript, audio recording — as listed in Table 1.

Together, these two properties classify all sources into a fixed set of scenarios that inform how our processing and interface display should adapt to each source. Based on source type and medium values, the set of required and expected attributes in each data object shifts slightly, allowing the processing pipeline to adapt dynamically to each format's specific requirement.

Overall, the attributes of each data object fall into two categories: those required to uniquely identify a document — such as media title and date — and categorical attributes that drive dynamic pipeline adaptation, including source type, source medium, and whether OLR was performed. This defines a shared information structure across all sources, making the framework robust to data heterogeneity.

OCR-Only Data: Pages as Content Items The left part of Figure 2 illustrates the case where OLR was not applied to a collection. As discussed in Section 3, no systematic means exists to determine which paragraphs and images belong together in such cases. The entire text of a given page is therefore treated as a single content item, assigned the type "page" — as opposed to "article" or other semantic types — to indicate that it does not represent a semantically coherent unit of content. These content items are processed in the same way as individually segmented items, and any available paragraphs are rendered in the interface accordingly.

An analogous situation arises for radio sources. ASR transcription does not contain audio chapters or segment boundaries equivalent to newspaper article segmentation: the full transcript of a given audio recording constitutes a single content item for its issue. Similarly, a radio bulletin issue contains

Object	Impresso ID
Issues	`[alias]-[YYYY]-[MM]-[DD]-[ed]`
Pages	`[alias]-[YYYY]-[MM]-[DD]-[ed]-p[page #]`
Audio Records	`[alias]-[YYYY]-[MM]-[DD]-[ed]-r[record #]`
Content Items	`[alias]-[YYYY]-[MM]-[DD]-[ed]-i[ci #]`

Table 2: Impresso identifiers for each format.

a single content item, though it may span multiple page supports.

These cases demonstrate the flexibility of the framework in accommodating sources where segmentation is absent or incomplete. Instead of imposing a single model, it treats segmentation as an issue-level variable while preserving a consistent representation.

Constructing Parsable and Deductible Identifiers Another design choice concerns the identifier scheme assigned to each object in the framework. Identifiers must be unique, parsable, and deductible from basic metadata, establishing a stable mapping between the main representation backbone (issues, physical supports, content items) and other data management components, such as bibliographic metadata.

Each media title is assigned a unique alias, sometimes inherited from the original collection. As shown in Table 2, issue identifiers are composed of the media alias, publication or airing date in year-month-day format, and edition (ed) number. Page, audio record, and content item identifiers are derived by appending a type-specific suffix to the issue identifier, followed by a zero-padded four-digit index indicating the object's position within the issue.

5.3. Implementation

Data Represented Through JSON Schemas All object representations are defined and validated through JSON schemas[5], ensuring that generated data complies with the framework's requirements and specifications. Individual documents are aggregated into JSON-line file archives by title and year. The schemas are publicly available in a GitHub repository[6] and help ensure that the framework remains stable throughout each step of the pipeline.

Combined with the identifier scheme, this supports flexible updates and partial re-runs, fine-grained monitoring of collection statistics, as well as the identification of data inconsistencies or leaks within the pipeline.

Format Converters and Module Architecture In terms of implementation, the publicly available code is separated into two Python submodules[7] : one producing the canonical data from each input format – the *importers*, and one extracting the content items from the canonical format to create our document representation – the *rebuilder*.

The main complexity of the importer module lies in adapting to each input format. It is built around abstract classes for issue and physical support objects, which are then extended as format-specific implementations (classes). The abstract classes ensure that all functions required by the main conversion script are consistently defined, while format-specific classes adapt to each collection format's particularities – for instance, handling the METS/ALTO files of the Berlin State Library differently from the PDF-embedded OCR of the SwissInfo collection. All generated issue and page objects are validated against the corresponding JSON schema.

The rebuilder module operates on the unified canonical representation and focuses on the reconstruction of content items, in a similar way across all collections. Constructing content items in these two steps enables to apply the same processing logic to all data, since the input to the second step is already a unified representation produced by our own processing. This promotes uniformity among the content items – the core unit of information retrieval – while keeping the representation lightweight and independent of each collection's specificities.

Object	2025	2026 (in progress)
Media titles	134	567
Issues	780 186	1 021 869
Pages	7 483 588	9 094 381
Content Items	52 358 158	73 475 049
Images	4 002 089	5 589 521
Tokens	15 652 402 700	$>24 \times 10^9$

Table 3: Impresso Corpus statistics: available in the Interfaces (2025) and in preparation (2026).

Large-scale Processing and Downstream Use The large-scale processing and downstream use of the Impresso corpus provides a concrete illustration of the framework in practice.

Through the canonical format, heterogeneous collections from multiple institutions have been converted into a unified representation, enabling the tenfold growth and diversification of the corpus since the framework's deployment in the first iteration of the project. As shown in Table 3, the 2025 release comprised 134 newspaper titles from 9 institutions, amounting to 52 million content items[8]. Several large collections are currently being prepared for iterative release throughout 2026 and the next release will span collections from 15 institu-

[5] json-schema.org

[6] github.com/impresso/impresso-schemas

[7] github.com/impresso/impresso-text-acquisition

[8] github.com/impresso/impresso-data-release

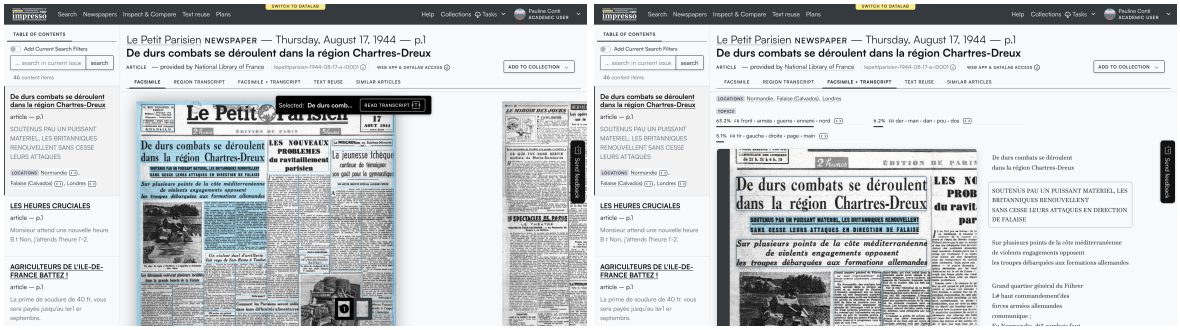


Figure 4: The Facsimile and Transcript views for the same article (`lepetitparisien-1944-08-17-a-i0001`) in the Impresso App, in the case of OCR and OLR data as described in Figure 2.

tions.

The rebuilt format serves as the basic unit of processing and as input to all downstream enrichment steps. This lightweight and uniform representation of content items enables parallelized computation and the application of ML models to an ever-growing collection. Enrichments currently produced include language identification, OCR quality assessment, key phrase extraction, named entity recognition and linking, news agency recognition, topic modeling, text-reuse detection, word and text embeddings, multimodal image embeddings, and image classification.

The resulting enriched transnational, transmedia and multilingual corpus is indexed and made accessible through two interfaces, whose full description lies beyond the scope of this paper. The WebApp offers keyword search, semantic faceted filtering, embedding-based retrieval, comparative and corpus views, and much more. Figure 4 illustrates two source views: the Facsimile view allows navigation between the pages of an issue and selection of an article of interest, while the Transcript view displays the corresponding text alongside bounding boxes on the facsimile. The Datalab offers programmatic access to the corpus, semantic enrichments and models via the Impresso Public API, and hosts notebooks guiding users in the use of these resources. Together, these interfaces represent the foremost output of the project, through which the public and scholars can access data across media, language, and institutional borders.

Crucially, these developments would not have been possible without the foundational work of the data representation framework described in this paper. It is what makes the corpus unified, indexable, and processable at scale: its modular design supports the progressive integration of new collections and source types into a single coherent representation, its lightweight and uniform content items enable large-scale processing and indexing, and its stable identifier scheme accommodates partial updates and re-processing.

6. Discussion and Conclusion

Working with large-scale, heterogeneous historical media collections raises challenges beyond standard data engineering: sources differ in medium, structure, digitization quality, and archival organization, yet must be brought together into a unified, research-ready corpus. The framework presented here navigates these challenges, reconciling documentary fidelity with machine learning usability and archival interoperability — not as a perfect solution, but as a principled and practical one.

The framework presented in this paper is the outcome of an iterative and multidisciplinary process, carried out at the intersection of digital humanities, natural language processing, information retrieval, and archival science. Its design evolved through continuous dialogue with partner institutions, historians, and engineers, reflecting the complexity that such collaboration entails: priorities shift, new source types emerge, and representational choices made early must be revisited as the corpus grows.

Several limitations remain. OCR and ASR quality set a ceiling on downstream ML performance that no framework can fully overcome — the most it can do is accommodate improved transcripts as better models become available. Document size variation, a consequence of heterogeneous segmentation granularity, affects enrichment consistency. Uneven semantic labeling of segmented content further complicates cross-collection comparison.

This work does not propose a standard, but aims to advance the practical possibility of aggregating and processing historical media data at scale, making explicit the design choices that such an endeavor entails. While archival standards are well established, representation frameworks oriented toward large-scale indexing and ML remain comparatively nascent. As collections grow and computational humanities research increasingly demands large-scale, cross-collection data, principled data representation will remain a prerequisite for historical research at scale.

Acknowledgments

The authors warmly thank Matteo Romanello, who greatly contributed to laying the foundations of this data representation framework during the first edition of the Impresso project. This work has been supported by the Swiss National Science Foundation (grant No. CRSII5_213585) and by the Luxembourg National Research Fund (No. 17498891).

7. Bibliographical References

Hildelies Balk and Aly Conteh. 2011. IMPACT: Centre of Competence in Text Digitisation. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, HIP '11, pages 155–160, Beijing, China, USA. ACM.

Kaspar Beelen, Jon Lawrence, Katherine McDonough, and Daniel C. S. Wilson. 2025. Whose news? Critical methods for assessing bias in large historical datasets. *Computational Humanities Research*, 1:e8.

Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, editors. 2023. *Digitized Newspapers - A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology*. Studies in Digital History and Hermeneutics. De Gruyter Oldenbourg, Berlin, Germany.

T. E. I. Consortium. 2025. TEI P5: Guidelines for Electronic Text Encoding and Interchange.

James Cummings. 2019. A world of difference: Myths and misconceptions about the tei. *Digital Scholarship in the Humanities*, 34(Supplement_1):i58–i79.

Marten Düring, Matteo Romanello, Maud Ehrmann, Kaspar Beelen, Daniele Guido, Brecht Deseure, Estelle Bunout, Jana Keck, and Petros Apostolopoulos. 2023. Impresso Text Reuse at Scale. An interface for the exploration of text reuse data in semantically enriched historical newspapers. *Frontiers in Big Data*, 6.

Maud Ehrmann, Marten Düring, Clemens Neudecker, and Antoine Doucet. 2023a. Computational Approaches to Digitised Historical Newspapers (Dagstuhl Seminar 22292). *Dagstuhl Reports*, 12(7):112–179.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023b. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):27:1–27:47.

Clemens Neudecker. 2022. Cultural Heritage as Data: Digital Curation and Artificial Intelligence in Libraries. In *Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022)*, volume 3234 of *CEUR Workshop Proceedings*, Berlin, Germany. CEUR.

Clemens Neudecker and Apostolos Antonacopoulos. 2016. Making Europe's Historical Newspapers Searchable. In *Proceedings of the 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 405–410, Santorini, Greece. IEEE.

Juri Opitz, Corina Raclé, Emanuela Boros, Andrianos Michail, Matteo Romanello, Maud Ehrmann, and Simon Clematide. 2026. CLEF HIPE-2026: Evaluating Accurate and Efficient Person–Place Relation Extraction from Multilingual Historical Texts. In *Advances in Information Retrieval*, pages 354–363, Cham. Springer Nature Switzerland.

Thomas Padilla. 2019. *Responsible Operations: Data Science, Machine Learning, and AI in Libraries. OCLC Research Position Paper.* ERIC.

PressMint. 2026. PressMint Project — CLARIN Flagship for Multilingual Newspaper Corpora. https://www.clarin.eu/pressmint. Accessed 2026-03.

Daniel van Strien, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pages 484–496, Valletta, Malta. SCITEPRESS - Science and Technology Publications.

Data Matters: Looking for High-Quality Corpora to Build Robust and Reliable Models for Humanists

Jaione Macicior-Mitxelena, Ana Garcia-Serrano
Universidad Pública de Navarra (UPNA), ETSI Informática (UNED)
jaione.macicior@unavarra.es, agarcia@lsi.uned.es

Abstract

The digitization of Spanish historical newspapers poses significant challenges due to low scan quality, typographical diversity, complex layouts and linguistic variation from contemporary Spanish. While advances in Optical Character Recognition (OCR) and layout-aware models offer promising results, their effectiveness strongly depends on the quality and consistency of the underlying training corpora. This work focuses on corpus construction and evaluation for historical document processing. Two experiments were conducted. In the first corpus *los101* was used, a manually curated and structurally annotated subcorpus derived from historical Spanish newspapers, designed to ensure coherent ground truth under heterogeneous real-world conditions. This corpus enables systematic experimentation across OCR and document layout analysis tasks. In a second experimental phase, we apply an additional layout-focused corpus characterized by structural regularity and consistent page organization, allowing us to isolate the impact of layout homogeneity on segmentation performance. State-of-the-art OCR models and a layout detection model are evaluated as validation instruments to assess corpus adequacy rather than as primary contributions. Quantitative and qualitative analyses based on (1) relationship between annotation quality, (2) structural variability, and (3) model behavior, show that heterogeneous corpora challenge both transcription and segmentation stability, while layout-consistent data significantly improves structural detection reliability.

Keywords: Digital Humanities, Corpus Construction, OCR, Layout Analysis, Historical Newspapers

1. Introduction

Before the digital era, research on historical newspapers required physical access to archives and printed collections, resulting in labour-intensive and geographically constrained workflows (Tworek, 2024). Since the early 2000s, systematic digitization initiatives, such as those led by institutions like the National Library of Spain (BNE), have enabled remote access to large newspaper repositories and facilitated computational analysis. In parallel, advances in Artificial Intelligence and Natural Language Processing have expanded research possibilities in Digital Humanities, with Large Language Models supporting large-scale processing of unstructured historical texts and enabling applications such as conversational heritage interfaces (Sergeev et al., 2025), sentiment analysis (Jaber et al., 2025), and automated image description (Garcia-Arias and Garcia-Serrano, 2025), as discussed in recent work (Simons et al., 2025; Lastra-Díaz et al., 2021).

However, the effectiveness of these technologies depends fundamentally on the quality and consistency of the underlying corpora. Historical Spanish newspapers present specific challenges: degraded scans, typographical variation, non-standard orthography, multi-column layouts, advertisements, marginal notes, and irregular page structures. In such contexts, model performance is often constrained not only by architectural design but by the reliability and internal coherence of the annotated data used for training and evaluation.

The main objective of this work is to examine how corpus design, particularly layout variability, affects the performance of OCR and layout analysis systems in historical newspapers. To this end, we conduct two experimental phases: first, using a heterogeneous manually curated corpus (*los101*) for joint OCR and layout experimentation; and second, introducing a structurally homogeneous corpus dedicated to layout analysis in order to isolate the impact of layout regularity on segmentation performance. *los101* (Miguez Lamanuzzi et al. (2025)) is a manually curated corpus guided by philological and structural annotation criteria to ensure ground-truth consistency and reproducibility (Torterolo-Orta et al., 2025). Rather than prioritizing model comparison, OCR and layout architectures are employed as diagnostic instruments to examine how corpus design, annotation coherence, and layout variability affect computational performance. Key factors that influence system behaviour are systematically analyzed. Moreover, domain adaptation is limited by the modest size of the annotated dataset, typographical heterogeneity and variable scan quality. Together, both corpora provide complementary experimental conditions that allow us to analyze the relationship between annotation quality, structural variability, and model behavior.

The remainder is organized as follows. Section 2 reviews previous work on OCR and layout analysis for historical document processing. Section 3 details the motivation and construction of the corpora. Section 4 describes the experimental setup and models used and subsequent subsections present

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 82–91

preprocessing strategies and evaluation metrics while section 5 discusses the experimental results. Section 6 concludes with implications for corpus design and future research directions.

2. Related Work

Optical Character Recognition (OCR) converts scanned images into machine-readable text and enables access to digital archives for research and retrieval (Benavent et al., 2010). The accuracy of OCR impacts research outcomes like authorship attribution (Hill and Hengchen, 2019; Garcia Serrano and Menta Garuz, 2022) and user satisfaction in historical documents search (Kettunen et al., 2022). Recent neural approaches have improved transcription quality through automatic correction and normalization (Fleischhacker et al., 2025). Collaborative tools like OCR4all and Impresso further aid this by allowing human-machine interaction for refinement (Düring et al., 2024). Transformer architectures, such as TrOCR, have revolutionized OCR by offering end-to-end transcription with a visual encoder and textual decoder, achieving state-of-the-art results on various datasets (Li et al., 2023; Moreno-Sandoval et al., 2024; Bengio et al., 2013).

Digitizing historical Spanish newspapers poses unique challenges due to varying scan quality, typographical diversity, and complex layouts (Sánchez-Salido et al., 2023; Liebl and Burghardt, 2021). While traditional OCR struggles with these issues, deep learning (CNNs, RNNs) has enhanced robustness by learning visual patterns, though they have limitations with long sequences and parallel computing (Vaswani et al., 2017; Cho, 2014).

Alongside OCR, layout analysis ensures correct reading order and semantic coherence by segmenting elements like columns and titles (Rezanezhad et al., 2023). Early methods improved accuracy by preserving page structure (Chen et al., 2017; Alberti et al., 2017; Zhu et al., 2022). More recently, object-detection frameworks like YOLO, specifically DocLayoutYOLO models, treat structural regions as visual objects, particularly for complex newspaper layouts (Zhao et al., 2024; Santos Júnior et al., 2025; Shen et al., 2021).

Several institutional digitization pipelines have adopted a segmentation-first strategy, where layout boxes are detected and normalized prior to OCR transcription. For example, the Austrian National Library Labs project Esperanto Newspaper Excerpts implements a workflow in which document regions are first identified using object-detection models before text recognition is applied, highlighting the importance of structural preprocessing for historical newspapers (Austrian National Library Labs, 2024).

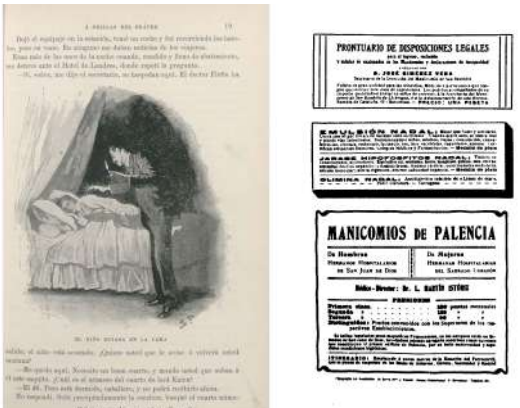
These advancements show a convergence of visual and linguistic modeling in document digitization. However, their effectiveness hinges on high-quality annotated corpora, which are lacking for Spanish historical materials, motivating the creation of the *los101* corpus for this study's experiments.

Nevertheless, while numerous studies evaluate OCR and layout models, fewer works explicitly analyze how corpus structural variability conditions model behavior across tasks. This gap motivates the dual-phase experimental design used in this study.

3. The Need for a Quality Corpus

The experimentation is organized into two phases as introduced in the following paragraphs.

Phase 1: OCR and Layout Experiments with a Heterogeneous Corpus The first phase of this study focused on evaluating OCR systems on historical Spanish newspapers, using a classical approach with Tesseract, an open-source OCR engine maintained by Google. Experiments were conducted within the framework of the PastReader 2025 shared task (IberLEF)[1], which addressed automatic transcription of historical newspapers. The PastReader corpus Montejo-Ráez et al. (2025) consists of over 12,000 scanned pages from eight heterogeneous publications encompassing cultural, scientific, satirical, and literary genres. These materials exhibit diverse typographies, layouts, and states of preservation, presenting a demanding benchmark for OCR systems (view Figure 1).



(a) (b)

Figure 1: Sample pages from the PastReader corpus.

Two configurations of Tesseract were evaluated:

1. **Baseline model (spa):** the default Spanish model (*spa* v5.3), used as an out-of-the-box

[1] https://sites.google.com/view/pastreader2025

reference.

2. **Fine-tuned model (los101):** adapted to the PastReader training set using the `tesstrain` utility. Fine-tuning follows Tesseract's LSTM-based pipeline, requiring aligned pairs of text line images and corresponding UTF-8 transcriptions. Each training sample was segmented line by line, that means that every line needs to be defined by its line coordinates.

Evaluation used standard OCR metrics, including *Character Error Rate (CER)*, *Word Error Rate (WER)*, and semantic measures (*BLEU* and *ROUGE-L*) (some details at subsection 4.1). Results showed that fine-tuning did not improve performance: CER slightly increased from 0.36 to 0.38, and BLEU decreased marginally. Error analysis revealed that the heterogeneous annotations and complex multi-column layouts hindered convergence, particularly for pages with irregular typography or line segmentation inconsistencies.

To overcome these limitations, a dedicated subcorpus with coherent annotations and controlled ground truth were created, ensuring high-quality data for subsequent experiments. From the PastReader dataset, 101 documents were carefully selected based on qualitative analyses and manually transcribed using the Transkribus platform, which supports both textual transcription and structural markup through regions, labels, and inter-region relationships. The corpus (*los101*) was split into 80% training, 10% validation, and 10% test sets Miguez Lamanuzzi et al. (2025).

Annotations followed a unified transcription guide with a *literal modernized* approach, preserving orthography while omitting purely paleographic features (Miguez Lamanuzzi and García Serrano, 2026). In other words, normalization targets graphical variability without altering the underlying linguistic content. Guidelines covered illegible text, marginalia, footnotes, and structural segmentation into paragraphs, headings, and other boxes. Inter-region relationships, such as linking titles to text bodies or images to captions, were included to capture reading order and document layout, supporting multimodal OCR models (Miguez Lamanuzzi et al., 2026).

Phase 2: Layout Analysis with a Homogeneous Corpus Despite its high-quality annotations, *los101* exhibits considerable layout heterogeneity, including variations in column structure, typography, spacing, and overall visual organization.

To investigate the impact of layout regularity, a second corpus (*layout-homogeneous*) was constructed specifically for layout analysis. This corpus consists of digitized pages from nine different newspapers, spanning diverse editorial styles (see Figure 2). It was designed to exhibit more homogeneous page structures, with more consistent column configurations and layout patterns across samples Obispo et al. (2026). It contains layout annotations only, as its purpose is to isolate the effect of structural consistency on segmentation performance rather than to evaluate OCR quality.

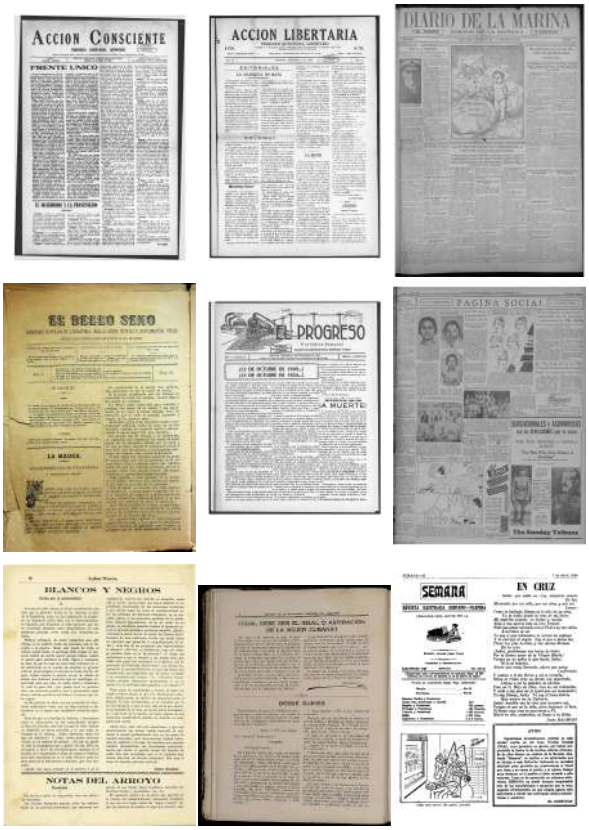


Figure 2: Sample pages from the corpus. Each image is from a different newspaper included in the dataset.

4. The Benchmark for OCR and Layout Tasks

To evaluate the adequacy of the constructed corpora, representative OCR and layout models were employed as diagnostic tools rather than as primary research contributions. The selected systems, Tesseract, TrOCR, Granite, and DocLayoutYOLO, cover classical OCR, transformer-based recognition, multimodal transcription, and object-detection-based layout segmentation. This diversity allows us to observe how corpus characteristics influence performance across different architectural paradigms.

Experiments were conducted in two phases reflecting the corpus design:

- **Phase 1:** OCR and layout experiments on the

heterogeneous *los101* corpus, containing both text and structural annotations.

- **Phase 2:** Layout-segmentation experiments conducted independently on both corpus, using a specialized training configuration focused solely on layout detection to isolate the impact of structural regularity on segmentation performance.

Each model requires a preprocessing pipeline adapted to its architectural assumptions and input constraints. Tesseract, as a classical OCR engine, operates on TIFF images paired with character-level annotations and generates intermediate training representations internally. TrOCR, a transformer-based OCR model, processes JPEG images aligned with their corresponding textual transcriptions, following a sequence-to-sequence learning paradigm. Granite, designed as a multimodal OCR system, requires images to be resized proportionally and embedded within a fixed-size canvas to preserve the aspect ratio and ensure uniform input dimensions. In contrast, DocLayoutYOLO addresses layout segmentation and therefore processes document images together with YOLO-format structural annotations that encode bounding box coordinates and region classes.

Although these preprocessing workflows differ according to architectural design, all models are integrated into a unified experimental framework to guarantee methodological consistency and fair comparison.

4.1. Evaluation Metrics

To assess model behavior, we considered several task-appropriate evaluation metrics. A broader set of measures was explored in (Macicior Mitxelena, 2025) and the code was published on GitHub (Jaione Macicior Mitxelena and Ana Garcia-Serrano., 2026)); however, for the sake of clarity and conciseness, this paper reports and discusses only the most representative ones.

Regarding **OCR subtask** metrics, Character Error Rate (CER) (lower is better), complemented by semantic metrics (BLEU and ROUGE-L) are used. The metrics used for **Layout Analysis subtask** has to take into account segmentation and classification accuracies. Segmentation accuracy is measured using Intersection over Union (IoU), which quantifies the overlap between predicted and ground-truth bounding boxes:

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}}.$$

Classification performance is evaluated using Precision and Recall, in order to get their harmonic mean (F1-score) computed per region type.

In addition, Mean Average Precision at IoU threshold 0.5 (mAP@0.5) and mAP@0.5:0.95 which averages the performance across multiple IoU thresholds (0.5–0.95) are considered. These metrics provide a standard region-detection evaluation framework that jointly captures localization and classification quality under varying levels.

Finally, to explicitly balance spatial and categorical accuracy, we define a new metric: the **composite layout score**:

$$\text{Score}_{\text{layout}} = \alpha \cdot \text{IoU} + (1 - \alpha) \cdot \text{F1-score},$$

where $\alpha \in [0, 1]$ controls the relative weight of localization versus classification. In this study, $\alpha = 0.75$, prioritizing accurate spatial placement while still accounting for correct region labeling.

4.2. Phase 1: Assesment of corpus adaptation

To ensure methodological consistency, the same experimental protocol was applied to the three OCR models (Tesseract, TrOCR, and Granite). Following the preprocessing procedures described, each baseline model was fine-tuned using the training split of the *los101* corpus. Performance was then evaluated on the corresponding train, val and test sets, and results from the fine-tuned models were systematically compared against their respective baselines.

This controlled setup allows us to isolate and measure the specific impact of corpus characteristics on transcription performance. By maintaining identical training and evaluation splits across models, any observed variation between baseline and fine-tuned configurations can be attributed to the influence of the constructed corpus rather than to differences in experimental design. This comparison serves as a diagnostic mechanism to assess how well the corpus supports adaptation across distinct OCR architectures, including classical OCR (Tesseract), transformer-based sequence-to-sequence recognition (TrOCR), and multimodal transcription (Granite).

The initial layout detection and box classification experiments were conducted on the heterogeneous *los101* corpus used a joint configuration that combined layout detection and box classification. (view Table 2).

Fine-tuning on the model DocLayoutYOLO model was performed using a minimal configuration: 30 epochs, batch size of 2, no explicit data augmentation, no layer freezing, and no regularization beyond default optimization parameters. Mixed precision was disabled and validation was enabled, but the training regime did not include any kind of data augmentation or controlled convergence strategies.

Under this setup, the model was optimized jointly for layout detection and box classification, thereby increasing task complexity while operating on a structurally heterogeneous corpus. The limited number of epochs and absence of regularization mechanisms likely exacerbated instability, particularly in the presence of inconsistent page structures.

4.3. Phase 2: Assesment of layout segmentation

As commented previously, in the second phase, a controlled fine-tuning strategy on the model DocLayoutYOLO was applied separately to both corpora, *los101* and *layout-homogeneous*, obtaining the models *los101_augment* and *per_augment* respectively.

To enhance the model's robustness against digitization artifacts, a **geometric data augmentation strategy** was applied, simulating real-world document variability. This included small rotations ($\pm 2.0°$) and lateral shearing ($2.0°$) to replicate page skew and binder-induced curvature, alongside translation (0.1) and scaling (0.3) to account for inconsistent margins. While mosaic augmentation was utilized to improve image feature extraction across multiple scales, it was strategically disabled during the final 15 epochs. This de-augmentation phase allowed the model to refine bounding box precision on the original document distribution, ensuring the final weights were optimized for the undistorted layouts of the target corpora.

The configuration was explicitly adapted to small-data regimes and to isolate the effect of corpus structure: training was restricted to layout detection only (box classification removed), the first 10 backbone layers were frozen to preserve low-level visual priors, perspective distortion and flips were disabled, and Mosaic Augmentation was enabled for most epochs. Dropout (0.15), a learning rate of 0.001 with a 3-epoch warmup, and early stopping (patience = 20) were also employed.

Early stopping was determined based on the detection metrics mAP@0.5 (mean Average Precision at IoU threshold 0.5) and mAP@0.5:0.95 (average mAP across multiple IoU thresholds from 0.5 to 0.95 in steps of 0.05). These metrics reflect the model's detection precision and localization accuracy: mAP@0.5 emphasizes correct object detection with moderate overlap, while mAP@0.5:0.95 provides a stricter evaluation by averaging performance over increasingly demanding IoU thresholds. Incorporating both metrics ensures the model balances accurate box placement with reliable detection confidence.

Despite the initially configured maximum of 300 training epochs, early stopping (patience = 20) was triggered independently for each corpus, indicating that no further performance improvements were observed beyond a certain number of epochs. The stopping behavior suggests convergence of the optimization process and stabilization of detection performance. For *los101*, training stopped at epoch 138, with best performance at epoch 66: $mAP@0.5 = 0.832$, $mAP@0.5 : 0.95 = 0.424$, while for the second homogeneous corpus, training stopped at epoch 138, with the best performance at epoch 118: $mAP@0.5 = 0.768$, $mAP@0.5 : 0.95 = 0.469$.

5. Results

Analysis results are described below.

5.1. Phase 1 results

Table 1 summarizes the results of the OCR experiments on the heterogeneous *los101* corpus. The outcomes highlight the interplay between corpus properties, model architecture, and fine-tuning strategies. Tesseract achieves the lowest baseline error rates (CER = 0.108), confirming its robustness in out-of-the-box scenarios. Interestingly, fine-tuning Tesseract on *los101* leads to a substantial deterioration in performance, with CER more than doubling. This negative impact reflects the influence of heterogeneity in the training corpus: inconsistent line segmentation, variable fonts, and multiple column layouts introduce noise that disrupts the LSTM training process. These results emphasize that classical OCR engines rely heavily on consistent annotations, and their performance can degrade when exposed to heterogeneous historical data, even if the corpus is relatively large.

Transformer-based TrOCR behaves differently. Its baseline performance is comparatively poor given its high error rate (CER = 0.996), largely due to the small size of the training corpus relative to the model's capacity. Fine-tuning on *los101* reduces CER to 0.906, demonstrating that domain adaptation can partially offset the initial lack of specialization. Although the absolute error rates remain high, the positive delta shows that transformer-based models benefit from exposure to domain-specific samples, provided they are cleanly annotated. This indicates that architectural flexibility allows adaptation to historical document variability, albeit limited by the dataset size.

Granite, the multimodal OCR model, maintains intermediate error rates (CER $\approx$ 0.29–0.30) across baseline and fine-tuned versions. Fine-tuning slightly increases CER, suggesting mild overfitting to the small training set. Nevertheless, Granite demonstrates relative stability compared with Tesseract and TrOCR, likely due to its ability to combine visual and textual cues, which makes it more

Model	Approach	CER	△CER	BLEU	ROUGE-L
Tesseract	spa	**0.1080**	–	**0.5998**	**0.8696**
	los101	0.2454	-127%	0.0505	0.4364
TrOCR	baseline	0.9961	–	0.0000	0.0017
	los101	0.9060	**+9.0%**	0.0000	0.0437
Granite	baseline	0.2897	–	0.2577	0.6557
	los101	0.3033	-4.7%	0.2495	0.6390

Table 1: OCR performance on the first phase with corpus *los101* (test set). Relative changes (△) indicate performance variation after fine-tuning.

Configuration	IoU	F1	Composite
Baseline	**0.3099**	0.0000	**0.2324**
Fine-tuned	0.0000	0.0000	0.0000

Table 2: Layout segmentation and box classification results of in Phase 1 on *los101* corpus on test set using DocLayoutYOLO model.

resilient to typographic and layout variation.

Overall, the results in Table 1 highlight that corpus heterogeneity strongly affects OCR performance. Classical engines are highly sensitive to inconsistent annotations, while transformer-based models benefit from fine-tuning only when sufficient clean data is available. Semantic metrics (BLEU and ROUGE-L) reflect the same trends: Tesseract declines sharply after fine-tuning, TrOCR shows modest gains, and Granite remains relatively stable, demonstrating its resilience to layout and typographic variability. In summary, the *los101* corpus itself limits OCR quality, underscoring the need for consistent annotations and controlled layouts. This motivates the second experiment using a new corpora of pages with more homogeneous features.

As far as the layout segmentation subtask is concerned, the model failed to converge to stable spatial representations (view Table 2). Intersection over Union (IoU) and F1 scores collapsed during validation and testing, in some cases approaching zero. Rather than improving segmentation quality, fine-tuning amplified instability, suggesting that structural variability, multiple editorial formats, inconsistent column structures and heterogeneous box organizations introduced excessive noise into the optimization process. These results indicate that the interaction between corpus heterogeneity and multi-task training limits reliable layout learning.

To complement the quantitative evaluation, a qualitative analysis was conducted on the validation set, focusing on files that consistently ranked among the lowest-performing across multiple metrics. The results show that errors are not driven by a single dominant factor, but correlate with several recurring document characteristics. In particular, performance degrades on low-quality scans (e.g., low contrast), non-standard layouts such as diagrams, and pages containing decorative or complex typographic elements, all of which interfere with text segmentation and recognition. Additionally, pages with little or no textual content tend to produce unstable outputs. These factors often co-occur, making error attribution non-trivial. While both baseline and fine-tuned models are affected, the fine-tuned model shows increased sensitivity to visually complex inputs. Overall, these findings indicate that OCR performance is strongly conditioned by visual and structural variability, which is not captured by aggregate evaluation metrics.

5.2. Phase 2 results

The Phase 2 evaluation highlights the impact of augmentation strategies (related to the image-preprocessing) on layout detection performance across the two corpora. The `los101_augment` and `per_augment` consist of models obtained by fine-tuning the baseline model with the *los101* and the *layout-homogeneus* corpus respectively.

For the *los101* corpus (Table 3), `los101_augment` consistently achieves the highest mean IoU across all splits, indicating improved spatial alignment of predicted boxes with the ground truth. It also attains the best mAP50 and mAP50:95 in most splits, suggesting that corpus-specific augmentations help the model better detect and localize layout elements with higher precision. The `per_augment` model shows moderate improvement over the baseline in detection metrics, but its mean IoU remains slightly lower, implying that augmentation strategies designed for general layouts may not fully capture corpus-specific structural patterns. The baseline model maintains reasonable spatial overlap but suffers from lower detection precision and recall, as reflected by its lower mAP50 and mAP50:95.

For the *layout-homogeneous* corpus (Table 4), `per_augment` clearly outperforms other models

	Train			Val			Test		
Model	mean IoU	mAP50	mAP50:95	mean IoU	mAP50	mAP50:95	mean IoU	mAP50	mAP50:95
baseline	0.773	0.231	0.134	**0.788**	0.162	0.105	**0.773**	0.270	0.171
per_augment	0.720	0.293	0.136	0.722	0.283	0.130	0.726	0.362	0.172
los101_augment	**0.799**	**0.696**	**0.464**	0.745	**0.596**	**0.361**	0.747	**0.545**	**0.305**

Table 3: Phase 2 evaluation metrics for the *los101* corpus.

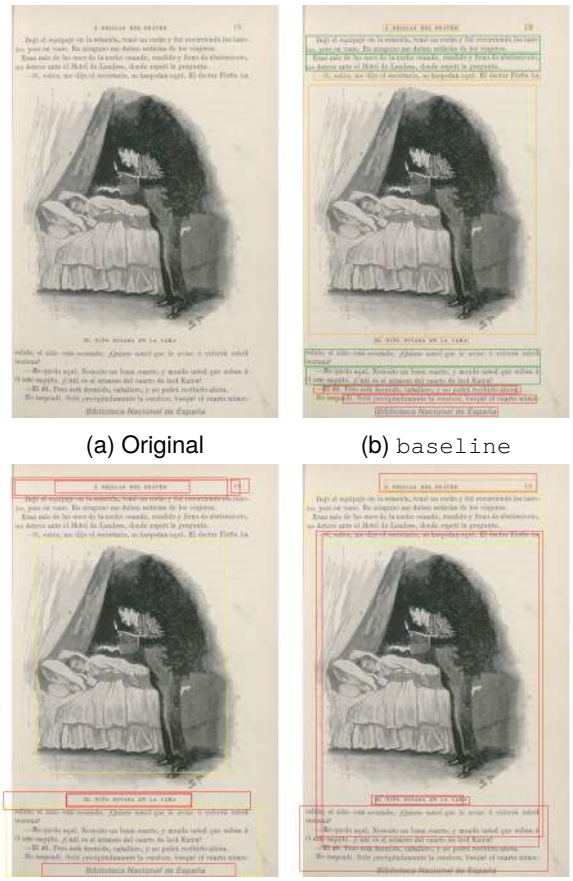
	Train			Val			Test		
Model	mean IoU	mAP50	mAP50:95	mean IoU	mAP50	mAP50:95	mean IoU	mAP50	mAP50:95
baseline	0.730	0.182	0.098	0.740	0.170	0.097	0.734	0.196	0.106
per_augment	**0.841**	**0.755**	**0.560**	**0.810**	**0.652**	**0.436**	**0.804**	0.554	**0.370**
los101_augment	0.749	0.589	0.331	0.769	0.447	0.261	0.741	**0.592**	0.324

Table 4: Phase 2 evaluation metrics for the *layout-homogeneous* corpus.

in all splits and metrics, achieving the highest mean IoU, mAP50, and mAP50:95. This result demonstrates that layout-aware augmentations effectively handle heterogeneous and complex newspaper page structures, improving both box localization and detection confidence. The `los101_augment` model performs well but slightly lags behind `per_augment`, particularly in mAP50:95, suggesting that augmentations optimized for *los101* do not generalize perfectly to different editorial formats. The baseline model remains the weakest across metrics, highlighting the necessity of data augmentation for robust layout learning in highly variable corpora.

To complement the quantitative evaluation, we visualize model predictions in Figures 3 and 4. Bounding boxes are colored according to detection confidence, allowing a quick assessment of model certainty: green indicates high-confidence predictions ($conf > 0.9$), orange represents moderately high confidence ($0.75 < conf \leq 0.9$), yellow corresponds to medium confidence ($0.5 < conf \leq 0.75$), and red highlights low-confidence predictions ($0.2 < conf \leq 0.5$). This visual coding facilitates the inspection of both spatial alignment and model certainty across different corpora and augmentation strategies.

It can be observed that the baseline model generally produces fewer boxes with lower confidence variation, often appearing predominantly green, reflecting a more conservative but consistent detection behavior. In contrast, the fine-tuned models (`per_augment` and `los101_augment` show improved detection coverage and spatial alignment, but their predictions include a higher proportion of orange, yellow, and red boxes. This indicates that while the fine-tuned models are not inherently worse, their confidence is more distributed, reflecting increased sensitivity to diverse page structures and the presence of more challenging layout elements.



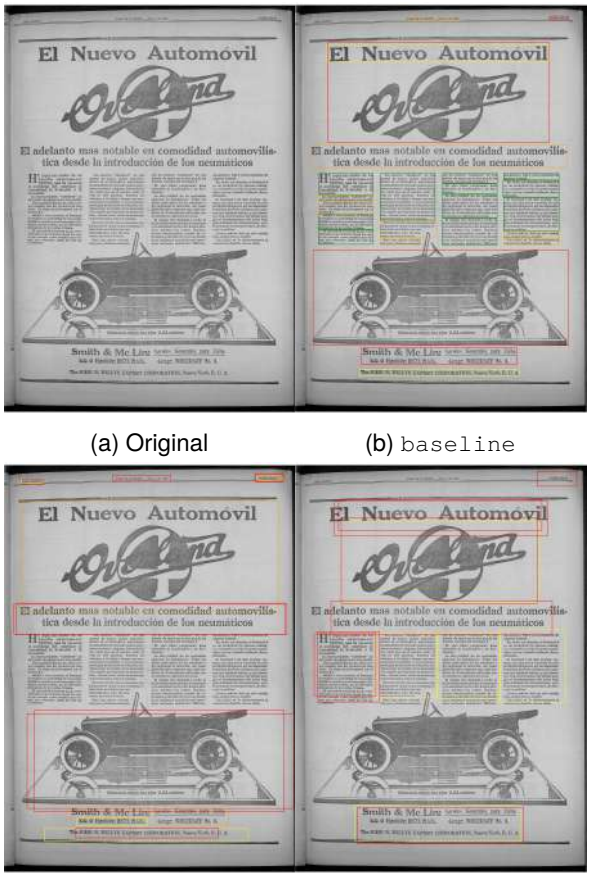
(a) Original (b) `baseline`

(c) `per_augment` (d) `los101_augment`

Figure 3: Visualization of layout detection results in an example of the corpus *los101* corpus.

5.3. Corpus adaptation discussion

Several key trends emerge from the evaluation across both corpora, highlighting the impact of augmentation and the nuances of detection performance. Primarily, the use of augmentation strategies leads to significant improvements in both mean IoU and mAP metrics. This demonstrates that such techniques are highly effective in stabilizing layout learning, likely by providing the model with a more



(a) Original (b) `baseline`

(c) `per_augment` (d) `los101_augment`

Figure 4: Visualization of layout detection results in an example of the new corpus.

diverse and robust set of spatial variations during training.

The results also reveal a trade-off between specialization and generalization. Models optimized for a specific corpus, such as `los101_augment` for the *los101* dataset, consistently achieve the highest performance on their native data. However, these gains do not always translate to other corpora, suggesting that while corpus-specific optimization maximizes local accuracy, it may limit the model's ability to generalize across different layout styles.

Furthermore, the performance discrepancy between mAP50 and mAP50:95 illustrates the complexity of precise localization. The mAP50:95 values are consistently lower across all tests, reflecting the increased difficulty of maintaining high precision across stricter overlap thresholds. By evaluating both spatial alignment through mean IoU and detection precision through mAP metrics, a more comprehensive view of model performance is achieved. This combined approach successfully reveals the inherent trade-offs between achieving accurate box placement and maintaining high overall detection precision.

6. Conclusions and Future Work

In this study of historical document processing, the quality and structural regularity of the underlying corpus are as critical as the choice of model architecture. Our experiments with *los101* reveal that high-quality, manually curated annotations alone are insufficient to overcome extreme layout heterogeneity in classical OCR engines like Tesseract, which suffered a performance drop of over 120% in CER after fine-tuning.

Conversely, the second phase results indicate that structural homogeneity significantly stabilizes layout analysis. By using a corpus with consistent column patterns and refined augmentation strategies, we achieved a marked improvement in spatial localization (mean IoU) and detection precision (mAP50). For the humanist researcher, this implies that a smaller, structurally consistent dataset may be more valuable for model training than a larger, noisier, and highly variable collection. Finally, our results suggest that multimodal models like Granite offer a more resilient middle ground, balancing visual and textual cues to navigate the "noise" of historical digitizations.

Future research will follow three primary directions. First, we intend to explore hybrid training regimes that combine the structural regularity of our second corpus with the linguistic richness of *los101*, starting with simple layouts and gradually introducing complexity. Second, we aim to implement automated layout normalization techniques as a preprocessing step to reduce the "visual noise" before it reaches the OCR engine.

Finally, we plan to expand the *los101* corpus to include a broader range of 19th-century scientific journals, testing the generalizability of our 'literal modernized" transcription approach. This will also involve evaluating the impact of OCR errors on downstream NLP tasks, such as Named Entity Recognition (NER) and Topic Modeling, to quantify exactly how much "annotation matters" for the final historical analysis.

7. Acknowledgements

This work is partially supported by the Ministerio de Ciencia e Innovación/AEI within the framework of the coordinated Spanish National project GRESEL UNED (PID2023-151280OB-C22).

8. Bibliographical References

Michele Alberti, Mathias Seuret, Vinaychandran Pondenkandath, Rolf Ingold, and Marcus Liwicki. 2017. Historical document image segmentation with lda-initialized deep neural networks. In *Proceedings of the 4th international workshop on historical document imaging and processing*, pages 95–100.

Austrian National Library Labs. 2024. Esperanto newspaper excerpts. GitLab repository. Accessed 2026-02-25.

Joan Benavent, Xaro Benavent, Esther de Ves, Ruben Granados, and Ana García-Serrano. 2010. Experiences at imageclef 2010 using CBIR and TBIR mixing information approaches. In *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*, volume 1176 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Kai Chen, Mathias Seuret, Jean Hennebert, and Rolf Ingold. 2017. Convolutional neural networks for page segmentation of historical document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 965–970. IEEE.

Kyunghyun Cho. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Marten Düring, Estelle Bunout, and Daniele Guido. 2024. Transparent generosity. introducing the impresso interface for the exploration of semantically enriched historical newspapers. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 57(1):20–40.

David Fleischhacker, Roman Kern, and Wolfgang Göderle. 2025. Enhancing ocr in historical documents with complex layouts through machine learning. *International Journal on Digital Libraries*, 26(1):3.

Enrique Garcia-Arias and Ana Garcia-Serrano. 2025. Creación de un modelo de descripciones de imágenes especializado en arqueología griega (pending edit). *Procesamiento del Lenguaje Natural*, 75(0).

Ana Garcia Serrano and Antonio Menta Garuz. 2022. La inteligencia artificial en las humanidades digitales: dos experiencias con corpus digitales. *Revista de Humanidades Digitales*, 7:19–39.

Mark J Hill and Simon Hengchen. 2019. Quantifying the impact of dirty ocr on historical text analysis: Eighteenth century collections online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843.

Areej Jaber, Israa Bahati, and Paloma Martínez. 2025. Leveraging pre-trained embeddings in an ensemble machine learning approach for arabic sentiment analysis. *Frontiers in Artificial Intelligence*, Volume 8 - 2025.

Jaione Macicior Mitxelena and Ana Garcia-Serrano. 2026. From Paper To Pixel: Experimental Framework for Access to Historical Spanish Documents. `https://github.com/jaionemacicior/from-paper-to-pixel`. Software/Code.

Kimmo Kettunen, Heikki Keskustalo, Sanna Kumpulainen, Tuula Pääkkönen, and Juha Rautiainen. 2022. Ocr quality affects perceived usefulness of historical newspaper clippings–a user study. *arXiv preprint arXiv:2203.03557*.

Juan J. Lastra-Díaz, Josu Goikoetxea, Mohamed Ali Hadj Taieb, Ana Garcia-Serrano, Mohamed Ben Aouicha, Eneko Agirre, and David Sánchez. 2021. A large reproducible benchmark of ontology-based methods and word embeddings for word similarity. *Information Systems*, 96:101636.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13094–13102.

Bernhard Liebl and Manuel Burghardt. 2021. An evaluation of dnn architectures for page segmentation of historical newspapers. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5153–5160. IEEE.

Jaione Macicior Mitxelena. 2025. Del papel al pixel: Experimentos para la digitalización de documentos históricos españoles. Master's thesis, UNED: Universidad Nacional de Educación a Distancia, Madrid, Spain, September. Directed by Ana Garcia-Serrano. Máster en Tecnologías del Lenguaje. Grade: Sobresaliente (9).

M. Miguez Lamanuzzi and A. García Serrano. 2026. Annotation of historical texts for automatic processing in the gresel-uned project. In *Congreso Internacional de Lingüística de Corpus (CILC 2026)*, Madrid. UAM. Aceptada.

M. Miguez Lamanuzzi, J. Macicior Mitxelena, Y. Torterolo, R. Ortuño Casanova, and A. García Serrano. 2026. Guía de transcripción y anotación para prensa histórica. https://doi.org/10.5281/zenodo.19187624.

Antonio Moreno-Sandoval, Leonardo Campillos-Llanos, and Ana García-Serrano. 2024. Language resources in spanish for automatic text simplification across domains.

Vahid Rezanezhad, Konstantin Baierer, Mike Gerber, Kai Labusch, and Clemens Neudecker. 2023. Document layout analysis with deep learning and heuristics. In *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, pages 73–78.

Eva Sánchez-Salido, Antonio Menta, and Ana García-Serrano. 2023. Seeking information in spanish historical newspapers: The case of diario de madrid (18th and 19th centuries). *DHQ: Digital Humanities Quarterly*, (4).

Eder Silva dos Santos Júnior, Thuanne Paixão, and Ana Beatriz Alvarez. 2025. Comparative performance of yolov8, yolov9, yolov10, and yolov11 for layout analysis of historical documents images. *Applied Sciences*, 15(6):3164.

Alexander Sergeev, Valeriya Goloviznina, Mikhail Melnichenko, and Evgeny Kotelnikov. 2025. Talking to data: Designing smart assistants for humanities databases.

Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. 2021. Layoutparser: A unified toolkit for deep learning based document image analysis. *arXiv preprint arXiv:2103.15348*.

Arno Simons, Michael Zichert, and Adrian Wüthrich. 2025. Large language models for history, philosophy, and sociology of science: Interpretive uses, methodological challenges, and critical perspectives.

Yanco Amor Torterolo-Orta, Jaione Macicior-Mitxelena, Marina Miguez-Lamanuzzi, and Ana García-Serrano. 2025. Transcribing spanish texts from the past: Experiments with transkribus, tesseract and granite.

Heidi JS Tworek. 2024. Digitized newspapers and the hidden transformation of history. *The American Historical Review*, 129(1):143–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. 2024. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *arXiv preprint arXiv:2410.12628*.

Wenzhen Zhu, Negin Sokhandan, Guang Yang, Sujitha Martin, and Suchitra Sathyanarayana. 2022. Docbed: A multi-stage ocr solution for documents with complex layouts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12643–12649.

9. Language Resource References

M. Miguez Lamanuzzi, A. García Serrano, Jaione Macicior Mitxelena, and Yanco Torterolo. 2025. Los101. [Data set].

A. Montejo-Ráez, E. Sánchez Nogales, G. Expósito Álvarez, A. Ureña López, M. T. Martín-Valdivia, J. Collado-Montañez, I. Cabrera de Castro, M. V. Cantero Romero, A. García Serrano, R. Ortuño Casanova, and Y. A. Torterolo Orta. 2025. Pastreader 2025. https://doi.org/10.5281/zenodo.15084265. [Data set].

F. Obispo, R. Ortuño Casanova, L. Garçon, Y. Seyeux, E. Sinardet, D. Villanueva Romero, and E. Vivó Capdevila. 2026. Corpus de prensa histórica con el layout marcado en pagexml. https://doi.org/10.5281/zenodo.18774961. [Data set].

Thematic Landscapes of the Past: Analysing Slovene Historical Periodicals With Topic Modelling

Filip Dobranić[1], **Uroš Šmajdek**[3], **Oliver Pejić**[1], **Ciril Bohak**[3], **Vojko Gorjanc**[1,2], **Tina Munda**[2], **Darja Fišer**[1]
[1] Institute of Contemporary History
Privoz 11, SI-1000 Ljubljana
{filip.dobranic, oliver.pejic, darja.fiser}@inz.si
[2] Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana
{tina.munda, vojko.gorjanc}@ff-uni-lj.si
[3] Faculty of Computer Science, University of Ljubljana Večna pot 113, SI-1000 Ljubljana
{uros.smajdek, ciril.bohak}@fri.uni-lj.si

Abstract

This paper explores the thematic landscapes of three Slovene historical periodicals—*Slovenka*, *Slovenec*, and *Slovenski narod*—from the *sPeriodika* corpus, a comprehensive collection of Slovene press published between 1771 and 1914 (Dobranić et al., 2024). Using BERTopic, we analyse the thematic profiles of these periodicals, enriched with diachronic perspectives. Our study examines the thematic commonalities and specificities of the selected periodicals, highlighting their distinct political orientations, target audiences, and the increasing nationalist polarisation in public discourse. This work contributes to digital humanities by demonstrating the potential of modern topic modelling techniques, such as BERTopic, to advance historical and cultural research.

Keywords: historical newspapers, digital humanities, diachronic analysis, topic modelling, BERTopic, collective identities, sPeriodika

1. Introduction

We present a thematic comparative analysis of three periodicals (Slovenka, Slovenec, and Slovenski narod) in the public domain from the *sPeriodika* corpus, a comprehensive collection of Slovene historical press published between 1771 and 1914 (Dobranić et al., 2023).

Using the popular transformer-based topic modelling framework BERTopic (Grootendorst, 2022), we examine these periodicals to answer the following questions:

1. What are the thematic profiles, commonalities and specificities of the selected periodicals?

2. What is the thematic landscape of the selected newspapers over time, from their inception until 1914 (copyright expiration cutoff)?

By addressing these questions, we explore how these periodicals represented and influenced the cultural and political dynamics of their time.

2. Related Work

Research in historical newspapers has increasingly integrated topic modelling into collaborative digital humanities workflows, bringing together humanities scholars, computer scientists, and information specialists (Oberbichler et al., 2022; Villamor Martin et al., 2023). These interdisciplinary infrastructures are based on processes such as digitization, metadata enrichment, and post-OCR correction, which provide the foundational methods for large-scale computational analyses including topic modelling (Lombardi and Marinai, 2020).

Topic modelling has emerged as one of the key methods for uncovering long-term thematic patterns in such studies, with evaluations of its performance in the context of historical (newspaper) data. For example, Murugaraj et al. (2025b) evaluated topic modelling approaches for newspaper archives, comparing traditional probabilistic Latent Dirichlet Allocation (LDA), matrix factorization-based Non-Negative Matrix Factorization (NMF), and neural-based models such as BERTopic (Grootendorst, 2022). Their findings demonstrate that BERTopic outperforms classical models in all tested aspects, particularly in contextual sensitivity and thematic coherence. In subsequent studies, Murugaraj et al. (2025a,c) highlighted how traditional topic modelling methods often fail to fully capture the dynamic and complex nature of discourse in historical texts.

By contrast, BERTopic proved to be effective in identifying the most relevant topics for specific queries and in restricting retrieval to documents or segments within those topics. Ginn and Hulden (2024) applied both traditional statistical models

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 92–99
May 16, 2026.

(LDA and NMF) and BERT-based models to historical literary texts. Although quantitative metrics tended to favour statistical models, their qualitative evaluation revealed that neural models provided deeper insights into the data, highlighting the potential of modern transformer-based approaches for historical text analysis.

3. Data and Methodology

3.1. Dataset

Our analysis is based on a 1 bn word corpus *sPeriodika* of Slovene periodicals published between 1771 and 1914 (Dobranić et al., 2023), from which we selected three periodicals based on their historical significance:

- *Slovenka* [The Slovene Woman] (1897–1902)
- *Slovenec* [Slovene] (1873–1945)
- *Slovenski narod* [The Slovene Nation] (1868–1943)

Slovenec and *Slovenski narod* were the two most prominent and widely read Slovene-language political dailies during the turn of the century, catering to readerships with opposing political views. While *Slovenec* served as the leading voice of Slovene political Catholicism, *Slovenski narod* was closely aligned with liberal-progressive politics. Both newspapers eventually became the official organs of the Slovene Catholic and liberal parties respectively (Amon and Erjavec, 2011).

The principal difference between the two newspapers laid in their stance towards secularism and the Church's role in society. *Slovenec* campaigned for preserving the Church's independence against state meddling and its supremacy in education and civic life. While it also supported Slovene nationalist demands, its primary discursive enemy was liberalism, which it often equated with German politics (Amon and Erjavec, 2011, 144-147). Conversely, *Slovenski narod*'s rhetoric placed greater emphasis on Slovene nationalism and heavily criticized the unequal status of Austria's non-dominant nationalities. Its main discursive enemies were German nationalism and Slovene political Catholicism, and its core readership consisted of educated professionals as well as wealthier peasants (Amon and Erjavec, 2011, 120-130).

Slovenka began as a supplement to the liberal Slovene newspaper *Edinost* [Unity] from Trieste and later became an independent monthly publication. While its content initially mostly focused on literature, nationalist politics and domestic life, it also published more nuanced commentary on women's and social issues during its last two years. It was published for a much shorter period compared to the other two newspapers but deserves special attention due to its specificity as the first female-oriented and female-edited journal in the history of Slovene journalism (Amon and Erjavec, 2011, 136-138). The size of each dataset in the number of tokens, paragraphs and issues is presented in Table 1.

The corpus itself as well as individual newspaper subcorpora are not evenly distributed through time. *Slovenka* shows a slight decline from its initial to final year, consistent with its start as a biweekly supplement and transformation to a monthly publication in 1900. Opposite to that both *Slovenec* and *Slovenski narod* show a gradual increase of the number of paragraphs through the years, consistent with their development and growth in the second half of the 19th century. In order to control for the uneven distribution of paragraphs in our subcorpora, the analysis presented discusses relative paragraph frequencies when engaging in diachronic thematic analysis in section 5

While all paragraphs were considered in the analysis of *Slovenka*, the analysis of *Slovenec* and *Slovenski narod* excluded a third of the paragraphs from our analysis (see section 3.2). The shape of yearly distribution of excluded paragraphs matches the distribution of all paragraphs in the newspaper which indicates that the paragraphs from merged topics are diachronically representative of the corpus.

	Tokens	Paragraphs	Issues
S. narod	183,294,799	4,404,531	14,039
Slovenec	137,506,802	3,158,842	10,897
Slovenka	1,633,570	56,330	113

Table 1: Corpus size.

3.2. Topic Modelling

We use BERTopic (Grootendorst, 2022) to model the topics in each of the periodicals on individual paragraphs. We use the linguistically annotated texts from sPeriodika in the CONLL-u format and use the paragraph annotations to extract individual paragraphs. These are then assigned metadata from the periodical (issue date, periodical name, text, lemmatised text, paragraph annotations etc.) which is then used for our analysis.

Each of the periodicals is modeled individually using the same set of parameters and random seeds for the UMAP (McInnes et al., 2018) and the topic model. We used the `paraphrase-multilingual-MiniLM-L12-v2` model (Reimers and Gurevych, 2019) for our embeddings and the parameters presented in Table 2.

The model produced 554 topics on *Slovenka*. Due to their larger size, modelling for *Slovenec* and

Parameter name	Parameter value
UMAP `n_neighbors`	15
UMAP `n_components`	5
UMAP `min_dist`	0.1
UMAP `metric`	cosine
Topic model `top_n_words`	100

Table 2: Topic modelling parameters.

Slovenski narod returned two orders of magnitude more topics.

Given these differences, we needed a more manageable number of less fine-grained thematic clusters that would be easier to compare across the periodicals and across time. Furthermore, after manually inspecting the topics we observed the model splitting otherwise thematically-coherent topics based on proper nouns (for example country names). This was expected due to BERTopic's use of cTF-IDF, but ultimately, we were interested in thematic areas of reporting like State administration regardless of the country the paragraph is referring to. In order to merge these topics we tested automated hierarchical clustering but it produced unsatisfactory results. Instead, we decided to group the individual topics into manually curated "themes".

In order to determine the viability of a manual approach we started by manually grouping all 554 topics from *Slovenka* and merged them into the resulting themes. These were created through the manual grouping process by viewing representative words and close reading representative paragraphs of BERTopic-suggested topics. We observed that when the number of paragraphs in the topic starts approaching the minimum, the content and the topics themselves become more likely to contain text fragments and increadingly hard to thematically categorise. The latter, along with our estimates of the labor required to merge all the topics produced for *Slovenec* and *Slovenski narod* led us to consider the 500 most-represented topics in these newspapers, comprising roughly two thirds of all paragraphs present. The rest of the paragraphs were excluded from our analysis. The coarse-grained themes are presented in Table 3 and typeset in Smallcaps. In this paper we use the coarse-grained themes for analysis. However, the original individual topics can still be zoomed in for more in-depth analysis, which is planned as future work.

After grouping the topics, as a filtering criterion, we first excluded out of scope paragraphs (uncategorised topics from *Slovenec* and *Slovenski narod*), followed by textual fragments, and garbled text due to OCR errors, which comprise the theme Outliers and noise and represent 58.3% of all analysed paragraphs in *Slovenka*, 85.1% of *Slovenec*, and 88.4% of paragraphs in *Slovenski narod*. The final set of themes used in our analysis is presented in Figure 1.

	Themes
Slovenski narod	Advertisements and announcements, Art and culture, Countries and nationalities, Criminality and natural disasters, Education, Family, Finance, Food production, Health and mortality, Infrastructure, Narrative, Nature and weather, Newspaper publishing, Outliers and noise, Occupations, Paratext, Political life, Religious practice, Social life, State administration, Travel and communications
Slovenec	Advertisements and announcements, Art and culture, Countries and nationalities, Criminality and natural disasters, Education, Family, Finance, Food production, Health and mortality, Narrative, Nature and weather, Newspaper publishing, Outliers and noise, Non-Slovene text, Occupations, Paratext, Political life, Religious practice, Slovene identity, Social life, State administration, Travel and communication
Slovenka	Abroad, Art, Body and emotion, Culinary arts, Family, Female identities, Group identities, Material culture and objects, Metatext, Morals, Narrative, Nature and environment, Newspaper publishing, Outliers and noise, Non-Slovene text, Occupations, Paratext, Railway, Religious practice, Russian culture, Slovene language, State institutions, Time

Table 3: Topic groups per periodical in alphabetical order.

4. General Thematic Analysis

The themes identified in each of the three periodicals are visualised in Figure 1. Due to *Slovenka* being two orders of magnitude smaller than the other two newspapers, theme size and theme rankings are not directly comparable across periodicals. The population of themes *Slovenka* can be as low as 28 paragraphs for themes such as Slovenian language. For cross-periodical comparisons, we use relative frequencies and shares.

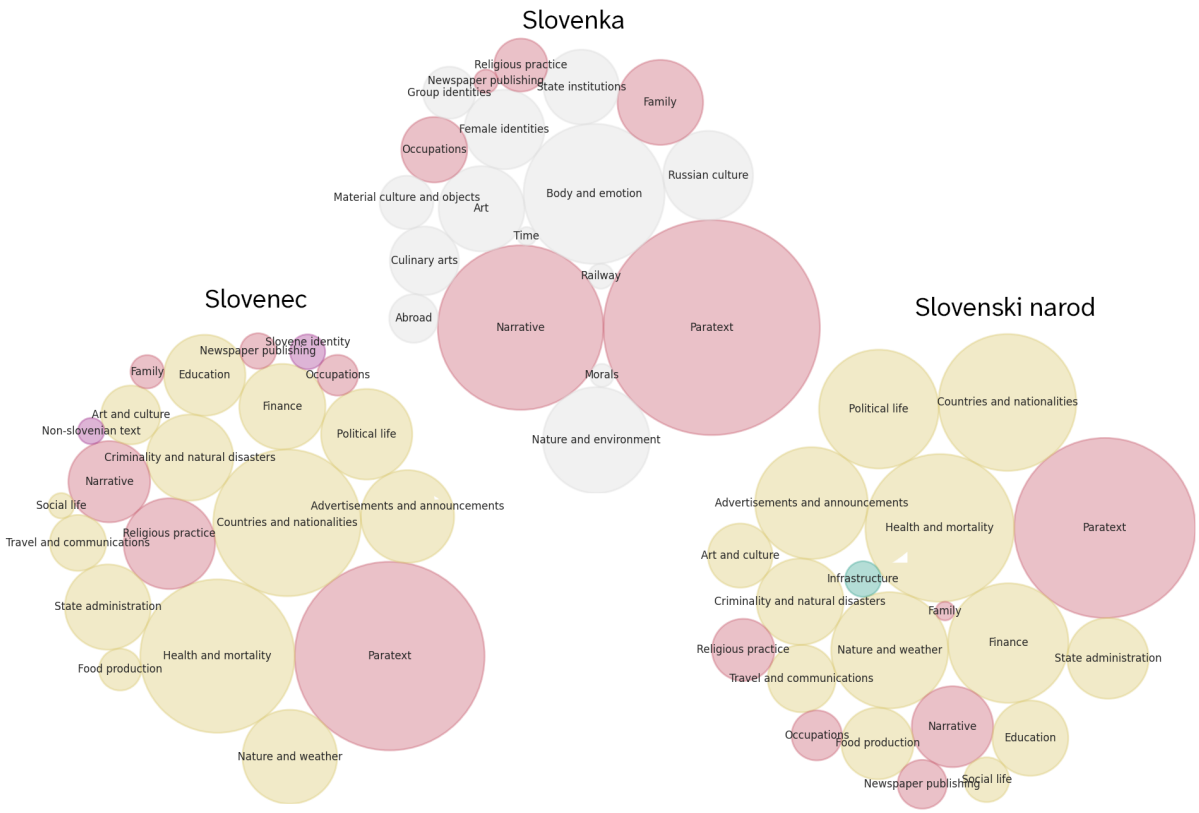


Figure 1: Circle-packed themes (excluding OUTLIERS AND NOISE) per periodical. Circle size indicates share of paragraphs. Themes in red are shared across all 3 periodicals. Themes in yellow are shared between *Slovenec* and *Slovenski narod*. Other colours signify themes unique to that periodical.

4.1. Shared themes

The list of all six themes with absolute and relative paragraph counts that are present in all three periodicals is presented in Table 4.

PARATEXT: Metatextual content, such as titles, attributions, places and dates from correspondents' reports dominate in all three periodicals, which is not surprising, given the journalistic genre of our dataset.

NARRATIVE: Textual fragments describing situations with no specific contextual cues are more frequent in *Slovenka* than in *Slovenec* and *Slovenski narod*, which reflects differences in reporting style, as well as the fact that *Slovenka* contained a much larger share of literary text.

RELIGIOUS PRACTICE: It is three times as likely for a paragraph to be religiously themed in *Slovenec* compared to *Slovenski narod*, which corresponds with the political profiling of the opposing newspapers. While the theme was identified in *Slovenka* as well, references to religion in this newspaper differs almost completely from the other two as it contains almost exclusively listings of Slavic baptismal names in the Catholic calendar, as well as the names of priests, details of their appointments and liturgical duties.

FAMILY: This theme contains references to family relationships, but also home and homemaking. It represents a relatively small portion of *Slovenec* and *Slovenski narod* and even when it does, it is often in texts offering lodging than discussions of familial life. In *Slovenka*, however, we see both a much larger relative presence and a wider array of relationship expressions as well as references to homemaking, motherhood, and care.

OCCUPATIONS: Expectedly, this theme in *Slovenka* is characterised by predominantly feminine occupation variants, contrary to *Slovenec* and *Slovenski narod* where masculine variants are the norm. Another key difference when comparing *Slovenka* with the other two newspapers is that paragraphs about EDUCATION are grouped in this category whereas *Slovenec* and *Slovenski narod* have a standalone theme for it. Not only were discussions on the politics and practice of education so much more prominent in these two newspapers, education-related paragraphs in *Slovenka* are more focused on the day to day profession of teaching as opposed to the politics of it.

4.2. Thematic profiles of *Slovenec* and *Slovenski narod*

While disctinct from *Slovenka*, the thematic profiles of *Slovenec* and *Slovenski narod* are very similar as there are only three marginal themes that are unique to just one of them.

The range of detected themes for *Slovenec* and *Slovenski narod* is much more diverse compared to *Slovenka* and reflects their function as political dailies covering vast aspects of everyday life.

Health and mortality: This is the most represented theme in both newspapers (2.1% vs. 1.4%) and it contains obituaries as well as advertisements for cures and tonics. This is consistent with the structure of newspapers at the time.

Countries and nationalities: Nearly as frequent as the previous one (2% vs. 1.2%), this theme is composed mostly of the reporting on events in foreign lands. As noted in 5.2, while *Slovenski narod* contains fewer paragraphs in this theme on average, we observe much greater spikes in times of significant foreign events.

While both newspapers feature relatively similar amounts of Advertisements and announcements (0.5% ; 1.2%), the liberal *Slovenski narod*, catering to a relatively wealthier (and ultimately progressive) audience, contains more paragraphs in both the Finance (0.5% ; 1.4%) and Political life (0.5% ; 1.3%) themes than the more rural, conservative *Slovenec*.

The topic Nature and weather (0.5% ; 1.3%) in both newspapers primarily features weather forecasts and reports on weather conditions, as well as accounts of interesting or unusual natural phenomena. It also includes reports on animals, ranging from wild animals to pets, with *Slovenski narod* placing a particular emphasis on animal husbandry, which forms part of broader reporting on resource management, such as forestry.

Theme	Slovenka	Slovenec	Slovenski narod
Paratext	2,732 4.84%	64,207 2.21%	60,794 3.06%
Occupations	619 1.10%	3,002 0.10%	4,702 0.24%
Narrative	3,901 6.93%	11,920 0.41%	12,304 0.62%
Newspaper publishing	79 0.14%	2,293 0.08%	4,482 0.22%
Religious Practice	405 0.72%	14,877 0.51%	7,242 0.36%
Family	1,042 0.185%	2,008 0.07%	661 0.03%
Total (all themes)	56,330	2,904,362	1,987,638

Table 4: Absolute and relative paragraph counts for themes present in all three periodicals.

4.3. Thematic profiles of *Slovenka*

Body an emotion and Nature and Environment: These are *Slovenka*'s most prominent unique theme (5% ; 2.9%). They contain references to body parts, emotions, illnesses and similar, which are common in literary texts. The literary character of the content of *Slovenka* is further reinforced by prominent themes such as Art (1.9%) and Russian culture (2%), since the journal also published many translations and discussions of contemporary literature. While Russian culture might as well be subsumed by the Art theme, it was featured so prominently (due to many publications and discussions of Russian literary works) that we decided to present it separately. While we can find mentions of Russian authors in some of the fine-grained topics in Slovenec and Slovenski narod these are nowhere near as prolific as they are in Slovenka.

The second interesting strand of themes relate to content on women's identity, domestic life as well as gender-defined public engagement, which is illustrated by themes such Female identities (1.6%), Culinary arts (1.2%), and Material culture and objects (0.7%), which, while present in the other two newspapers represent a relatively insignificant portion of their content.

5. Diachronic Thematic Analysis

Due to their positioning, we compare *Slovenec* and *Slovenski narod* but analyse *Slovenka* separately due to its shorter publishing period and specificity as a topical women's newspaper.

5.1. Slovenka

When observed diachronically, the distribution of themes confirms established historiographical knowledge regarding a shift in the journal's editorial policy: moving from literary and female-interest content to theoretical feminist and social issues topics.

This shift in editorial policy also manifests itself in the results of our analysis, see Figure 2. The clear predominance of the themes such as Art (2.3% vs. 1.3%), Body and emotion (5.6% vs. 4.3%), Nature (3.3% vs. 2.4%), and Russian culture during Marica Nadlišek Bartol's editorship (1897-1899) point to the overtly literary character of the journal at the time. The relative dominance of Morals (0.2% vs. 0.1%) likewise points to a predominance of literary or didactic content. The first period was also characterized by an emphasis on domesticity-related themes such as Culinary arts (1.6% vs. 0.7%) and Material culture and objects (1% vs. 0.4%).

Conversely, a rise in more elaborate social commentary during Ivanka Anžič Klemenčič's editor-

ship (1900-1902) is illustrated by an increase of themes such as FEMALE IDENTITIES (0.9% vs. 2.5%); STATE INSTITUTIONS (0.7% vs. 2.3%) and OCCUPATIONS (0.8% vs. 1.4%). Some themes take both angles, domesticity and social commentary. FAMILY (2% vs. 1.6%), for example, was more dominant in the early years of *Slovenka*, but experienced a slight uptick in 1901, possibly due to the presence of theoretical texts dealing with women's roles in family life.

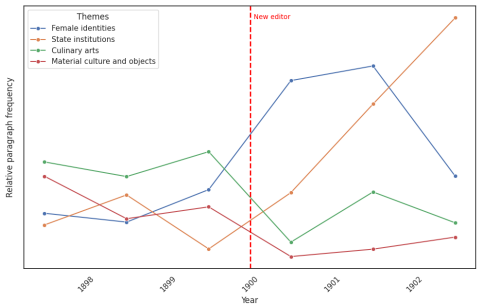


Figure 2: Themes signifying the change of editorial policy of *Slovenka*.

5.2. Slovenec in Slovenski narod

The comparative diachronic thematic analysis of *Slovenec* and *Slovenski narod* is performed by exploring the timelines visualising diachronic behaviour of themes to detect out of the ordinary trends. By zooming in the key terms that characterize the themes, we identify two groups of observations: (1) observations that confirm established historiographical knowledge and align with its expectations, and (2) observations that reveal patterns that could only be discerned by means of distant reading and open potential avenues for deeper historiographic inquiry.

5.2.1. Confirmation of established historiographical knowledge

The diachronic development of the topic COUNTRIES AND NATIONALITIES shows that reporting on developments abroad typically spiked in periods of heightened geopolitical conflict, see Figure 3. The most representative examples include spikes in 1885 (the Mahdist War, more noticeable in *Slovenski narod*), 1900 (the Second Boer War), 1904 (the Russo-Japanese War), 1912-3 (the Balkan Wars), and 1914 (the First World War). As a general tendency, *Slovenski narod* reported with less absolute frequency but with more observable spikes compared to *Slovenec*.

Reporting on the topic RELIGIOUS PRACTICE reveals some insights surrounding the wider ideological framing of the newspapers. As expected, it was more present overall in *Slovenec*, being a Catholic-conservative daily. While the former reported on religious issues with more frequency and discernible annual spikes, the two newspapers nevertheless shared a comparable spike in the years 1903-4, when they both intensely reported on Pope Pius X.

The topic HEALTH AND MORTALITY demonstrates how reporting on epidemic diseases spiked during certain years, e.g. in relation to contemporary cholera and tuberculosis epidemics. While both newspapers reported on this topic with similar frequency, we may observe a disproportionate spike in *Slovenski narod*'s reporting in 1885, which seems to coincide with a cholera epidemic in Trieste. Reporting on CRIMINALITY AND NATURAL DISASTERS likewise reveals spikes that coincide with locally or internationally notorious natural disasters. Some examples of spike years include 1895 (the Ljubljana earthquake); 1906 (likely the San Francisco earthquake, more prominent in *Slovenec*), or 1912 (likely the Mürefte earthquake in Turkey).

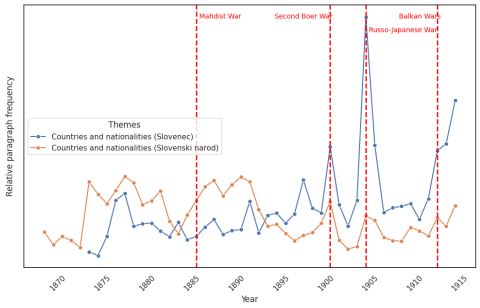


Figure 3: COUNTRIES AND NATIONALITIES theme relative paragraph frequency through time for both *Slovenec* and *Slovenski narod*.

5.2.2. New avenues for research

The topic POLITICAL LIFE reveals some curious differences in reporting among the two newspapers. Throughout the period under investigation and especially before 1907, *Slovenski narod* appears to have reported on political issues more frequently than *Slovenec*. At that point, however, political reporting also increased significantly in *Slovenec* and remained high until the end of our inquiry. One potential explanation for this phenomenon is that the Austrian electoral reform of 1907 marked the beginning of a new period of mass politicization in Austrian society, which could have also manifested itself in heightened political reporting.

Considerable differences are revealed when observing reporting on the topic FINANCE. In *Slovenec*, reporting on the topic began to rise steeply in 1892,

climaxing in 1895, and gradually decreased after stabilization until 1908. Conversely, in *Slovenski narod*, a similar rise began in 1879 and clearly climaxed in 1885, followed by a gradual decline until the end of the graph. The topic ADVERTISEMENTS AND ANNOUNCEMENTS also shows diverging trends in the two newspapers. In *Slovenec*, it rose sharply in 1882, climaxed in 1885 and gradually decreased with localized spikes in the coming decades. In *Slovenski narod*, the presence of this topic gradually increased constantly throughout the entire period, with local spikes in 1874 (perhaps an echo of the Vienna stock exchange crash of 1873?) and 1885.

Finally, reporting on the topic of FOOD PRODUCTION reveals some potentially counter-intuitive insights given the assumed target audiences of the two papers, see Figure 4. While *Slovenski narod* might appear to cater to a more urbanized bourgeois-professional audience, it reported on agricultural products and related topics with double the frequency of *Slovenec*. While *Slovenec* gradually increased its reporting on FOOD PRODUCTION with a climax in 1910, *Slovenski narod*'s reporting on the topic suddenly climaxed in 1885 - perhaps in reaction to the grape phylloxera epidemic - with various localized peaks in the ensuing decades.

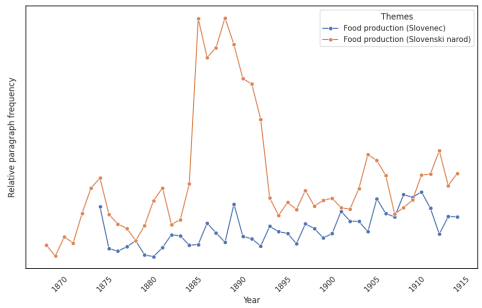


Figure 4: FOOD PRODUCTION theme relative paragraph frequency through time for both *Slovenec* and *Slovenski narod*.

6. Conclusion

This study demonstrates the potential of BERTopic for uncovering thematic and ideological patterns in historical periodicals, offering new insights into the cultural and political landscapes of selected Slovene historical periodicals. By analysing their thematic profiles and diachronic trends, we highlighted the distinct roles these periodicals played in shaping Slovene public discourse.

This research, on the one hand, confirms established historiographical knowledge on topics such as growing nationalist polarisation as well as the impact of key historical events on public discourse. At the same time, it also reveals several novel insights into differences in reporting among the periodicals. These findings underscore the value of computational methods and distant reading in historical and cultural research, particularly in enabling large-scale, data-driven analyses of complex historical texts.

While detailed interpretations of the collected data lie beyond the scope of this study, it provides a preliminary demonstration of how computational distant reading methods can identify phenomena that might otherwise remain elusive through traditional research approaches. Our coarse-grained thematic analysis revealed key reporting patterns and ideological underpinnings; however, a more fine-grained exploration of individual topics is essential to fully understand the nuances of their narrative strategies. Future research could focus on in-depth analyses of individual topics and their linguistic framing, investigating how these periodicals constructed narratives around gender, politics, religion, and other key themes to achieve a deeper understanding of the cultural and political dynamics of the past.

Acknowledgements

This work was supported by the Slovenian Research and Innovation Agency research programme "Digital Humanities: resources, tools and methods" (2022–2027) [grant number P6-0436], the support of the DARIAH-SI research infrastructure, the Slovene Common Language Resources and Technology Infrastructure, CLARIN.SI, and by the project "Large Language Models for Digital Humanities" (2024–2027) [grant number GC-0002].

7. References

Smilja Amon and Karmen Erjavec. 2011. *Slovensko časopisno izročilo: Od začetka do 1918*. Fakulteta za družbene vede, Založba FDV.

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.

Michael Ginn and Mans Hulden. 2024. Historia magistra vitae: dynamic topic modeling of roman literature using neural embeddings. *arXiv preprint arXiv:2406.18907*.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

Francesco Lombardi and Simone Marinai. 2020. Deep learning for historical document analysis and recognition—a survey. *Journal of Imaging*, 6(10):110.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Keerthana Murugaraj, Salima Lamsiyah, Marten During, and Martin Theobald. 2025a. Automating historical insight extraction from large-scale newspaper archives via neural topic modeling. *arXiv preprint arXiv:2512.11635*.

Keerthana Murugaraj, Salima Lamsiyah, Marten Düring, and Martin Theobald. 2025b. Mining the past: a comparative study of classical and neural topic models on historical newspaper archives. In *Proceedings of the 5th international conference on natural language processing for digital humanities*, pages 452–463.

Keerthana Murugaraj, Salima Lamsiyah, Marten During, and Martin Theobald. 2025c. Topic-RAG for historical newspapers: Enhancing information retrieval in humanities research through topic-based retrieval-augmented generation. *Computational Humanities Research*, pages 1–21.

Sarah Oberbichler, Emanuela Boroş, Antoine Doucet, Jani Marjanen, Eva Pfanzelter, Juha Rautiainen, Hannu Toivonen, and Mikko Tolonen. 2022. Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. *Journal of the Association for Information Science and Technology*, 73(2):225–239.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks.

Marta Villamor Martin, David A Kirsch, and Fabian Prieto-Nañez. 2023. The promise of machine-learning-driven text analysis techniques for historical research: topic modeling and word embedding. *Management & Organizational History*, 18(1):81–96.

8. Language Resource References

Filip Dobranić, Bojan Evkoski, and Nikola Ljubešić. 2023. Corpus of slovenian periodicals (1771-1914) sPeriodika 1.0. Slovenian language resource repository CLARIN.SI.

Author Index