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
May 16, 2026.

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.