

LexiMus-NER: A Named Entity Recognition Dataset for Historical Spanish Musical Press (1918–1935)

María Palacios

Universidad de Salamanca (USAL), Salamanca, Spain, mpalacios@usal.es 

Received 29 April 2026, revised 12 July 2026, accepted 21 September 2026

Abstract

LexiMus-NER is the first Named Entity Recognition (NER) dataset designed specifically for historical Spanish musical press. Drawing on approximately 3,050,000 words from two early twentieth-century publications — *El Sol* (Madrid, 1918–1932) and *Ondas* (Madrid, 1925–1935) — the dataset defines four domain-specific entity types absent from general-purpose Spanish NER resources: COMPOSITOR (composer), INTERPRETE (instrumental performer or conductor), CANTANTE (singer), and AGRUPACION (musical ensemble). It comprises 11,697 annotated entities across 4,243 documents in CoNLL-style TSV/IOB2 format, accompanied by a curated gazetteer of 1,035 unique musical entities. A transformer-based model (BETO fine-tuned via spaCy, v9) achieves 91.9% entity-level detection precision on an independent 73,633-word evaluation corpus and above 90% human-validated classification precision across all four entity classes ($n = 1,183$). All resources are freely available on Zenodo (DOI: 10.5281/zenodo.19429405), Hugging Face, and GitHub under CC BY 4.0, produced within the LexiMus project (‘Léxico y ontología de la música en español’, PID2022-139589NB-C33), funded by the Spanish Ministry of Science and Innovation and the European Union.

Keywords

named entity recognition – digital musicology – historical Spanish press – BETO – spaCy – musical heritage – FAIR data

Related Data Set

“LexiMus-NER” with DOI www.doi.org/10.5281/zenodo.19429405 in repository “Zenodo”

1. Introduction

Historical musical press constitutes a primary source of exceptional density for musicological research. Publications such as *El Sol* (Madrid, 1918–1932) and *Ondas* (Madrid, 1925–1935) document concert programmes, debuts, critical reception, and performer rosters — often the sole surviving evidence when

institutional archives have been lost. At corpus scale, they offer the possibility of tracing how the musical canon was constructed, which figures were systematically excluded, and how critical vocabulary evolved across genres, nationalities, and gender. Answering these questions computationally requires identifying and classifying the musical agents named in the text — a task that standard NLP tools handle poorly.

Generic Spanish NER models (e.g., spaCy ‘es_core_news_lg’; Honnibal & Montani, 2017) detect persons as a flat PER category without distinguishing a composer from a singer, and achieve only approximately 40% precision on historical musical press (roughly 60% of detections being false positives). More fundamentally, historically specific figures of early twentieth-century Spanish musical life — tenors such as Hipólito Lázaro, conductors such as Enrique Fernández Arbós, and ensembles such as the Orquesta Lassalle — are generally absent from the training data of general-purpose models, including large language models (LLMs). When an entity is not represented in the training data, its reliable identification or classification becomes difficult regardless of contextual cues. This limitation creates a systematic blind spot for historical musicology.

LexiMus-NER addresses this gap. It is the first open NER dataset for historical Spanish musical discourse, providing annotated training data, a domain gazetteer, annotation guidelines, and a fine-tuned production model. It was developed within the LexiMus project (*Léxico y ontología de la música en español*, PID2022-139589NB-C33), a coordinated initiative of 47 researchers from 26 institutions across 9 countries, funded by MCIN/AEI and the European Union (NextGenerationEU). This NER also occupies a distinct position within the broader landscape of historical performing-arts data resources. Much recent work in this area — including several data papers published in this journal — has focused on manually curated relational databases of performance records: opera productions, concert programmes, or theatre seasons, reconstructed event by event from archival sources. LexiMus-NER takes a complementary approach: rather than a hand-built database of performance events, it is a natural-language-processing resource — an annotated corpus and a fine-tuned model — that enables the automatic extraction of musical agents directly from unstructured historical press text at scale. Within Spanish- and NER-specific resources specifically, its positioning relative to HIPE 2020 (Ehrmann et al., 2020), CAPITEL, and AnCora is detailed in Table 2.

2. Research Objectives

The working objective of LexiMus-NER is to provide a specialised annotation resource enabling role-level classification of musical agents in historical Spanish press corpora. Where existing tools detect persons as a single

TABLE 1: Entity Types Defined in LexiMus-NER

Label	Definition	Examples
COMPOSITOR	Composer of musical works	Beethoven, Manuel de Falla, Saco del Valle
INTERPRETE	Instrumental performer or conductor	Enrique Fernández Arbós, Ricardo Viñes
CANTANTE	Singer (opera, zarzuela, popular song)	Hipólito Lázaro, Ángeles Ottein, Pepita Embil
AGRUPACION	Orchestra, choir, ensemble, or musical company	Cuarteto Rafael, Orquesta Lassalle

undifferentiated category, the dataset defines four entity types that are both musicologically meaningful and computationally distinguishable from context (see Table 1).

By tagging roles rather than merely detecting persons, the resource allows researchers to pose questions central to contemporary musicology: Which types of musician received the most press coverage and in what contexts? What vocabulary was used to describe composers compared to performers? How does mention of female musicians correlate with specific genres and social spaces? Which artists active in Madrid in the 1920s and 1930s are entirely absent from existing biographical reference works? This last dimension is particularly significant: correct entity classification enables systematic heritage recovery, identifying musicians invisible in modern reference works who appeared frequently in contemporary press.

To position this contribution relative to existing resources, Table 2 compares LexiMus-NER with related historical and Spanish-language NER datasets across five dimensions. To our knowledge, LexiMus-NER is the only resource combining historical Spanish text, domain-specific musical coverage, role-level entity types (rather than a generic PERSON category), and a human-in-the-loop active learning methodology with LLM-assisted pre-review.

3. Methodology

The dataset was constructed through nine iterative annotation and training cycles following a human-in-the-loop active learning approach (Settles, 2012). The main processing phases are illustrated in Figure 1, including the iterative feedback loop across the nine annotation and training cycles.

Corpus compilation. The source corpus comprises approximately 3,050,000 words of musical news from *El Sol* (1,310,000 words, 1918–1932) and *Ondas* (1,740,000 words, 1925–1935), accessed via the Biblioteca Nacional de España

TABLE 2: Comparison of LexiMus-NER with Related Historical and Spanish-language NER Resources

Resource	Language(S)	Temporal Scope	Domain	Entity Granularity/ Annotation Strategy
HIPE 2020 (CLEF-HIPE)	French, German, English	Historical newspapers (19th–20th c.)	General	Fine-grained generic entities (person, location, organisation, product) + entity linking; manual annotation with inter-annotator agreement reported
CAPITEL	Spanish	Contemporary	General (news)	Generic entities (PER, LOC, ORG, OTH); manual, single-pass annotation
AnCorà (Es)	Spanish, Catalan	Contemporary	General (mixed genre)	Generic entities (PER, LOC, ORG, MISC); manual multi-layer annotation (also syntax/ semantics)
LexiMus-NER (this work)	Spanish	Historical (1918–1932/1925–1935)	Domain-specific (musical press)	Role-specific entities (COMPOSITOR, INTERPRETE, CANTANTE, AGRUPACION); human-in-the-loop active learning + LLM-assisted pre-review

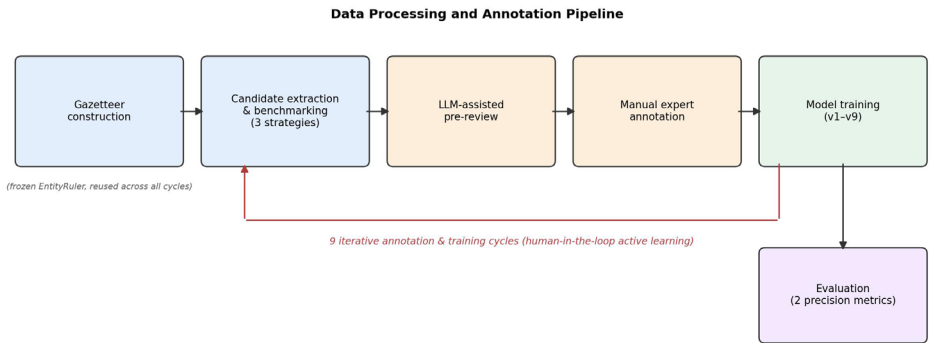


FIGURE 1: Data Processing and Annotation Pipeline (Revised). The workflow proceeds from gazetteer construction and candidate extraction/benchmarking, through LLM-assisted pre-review and manual expert annotation, to iterative model training and a two-metric evaluation. The feedback loop between model training and the earlier annotation stages represents the nine iterative annotation and training cycles described in this section.

open-access portal. A pilot evaluation corpus of 90 music news items from *El Sol* 1926 (73,633 words) was used for benchmarking throughout development. OCR quality was not assessed via a corpus-level error rate; instead,

quality control operated at the entity level, with the gazetteer’s canonical-form field mapping OCR-degraded spelling variants to a single reference entry and the annotation guidelines instructing annotators not to tag visibly broken OCR fragments. This pilot corpus was never used to fine-tune any model version (see Table 3).

Gazetteer construction. A domain-specific CSV of 1,035 unique musical entities was compiled manually from concert programmes and press archives (1918–1935), with normalised and variant forms documented for each entry (317 composers, 305 performers, 290 singers, 123 ensembles). The gazetteer functions as a frozen EntityRuler component in the spaCy pipeline and contributes approximately 20% of all entities detected in corpus-scale inference.

Candidate extraction and benchmarking. Three NER strategies were benchmarked on the pilot corpus before fine-tuning: generic spaCy alone (~40% entity-level precision, ~60% false positives), spaCy combined with the LexiMus EntityRuler (~42%), and EntityRuler-only regex (~94% for known entities, zero coverage of unseen names). This benchmarking confirmed that rule-based

TABLE 3: Benchmark and Evaluation Corpora used During Dataset Construction and Model Evaluation

Evaluation Set	Size	Provenance & Role	Metric Produced
Fixed test set	60 documents	Held constant from v2 onward (test_reannotado.spacy); tracked spaCy training F1 across cycles v1–v7	F1: 0.45 (v1) → 0.87 (v7)
Human validation set	1,200 candidates (300/class); 1,183 retained after exclusions	Extracted from Ondas by v7/v8, re-evaluated with v9; replaced the fixed test set as the primary development benchmark from v8 onward	COMPOSITOR 95.3%, INTERPRETE 92.2%, CANTANTE 91.3%, AGRUPACION 98.0%, global 94.3%
El Sol 1926 pilot corpus	90 documents/ 73,633 words	Fixed subset of El Sol; never used to fine-tune any model version; reused across cycles for strategy benchmarking and the final NER-vs-LLM comparison	Detection precision: 91.9% (v9); Gemini 3 Flash zero-shot: 92.8%
Published test. tsv	1,116 annotation windows (60 base documents + 1,056 windows added from Ondas, v8+)	Union of the original 60-document set and newly annotated Ondas material; deposited on Zenodo	2,158 entities total

approaches alone could not generalise, motivating the development of a fine-tuned transformer model.

LLM-assisted pre-review. Candidate entities were pre-screened using Gemini 3 Flash (Google), which confirmed or rejected each candidate and flagged uncertainty cases. Pre-review outputs were presented to the domain expert in interactive HTML interfaces for full human validation. Gemini was used exclusively as an annotation assistant, not as an inference model: the production pipeline is fully local, deterministic, and open source. On the 1,200-candidate validation set, Gemini achieved 99.5% precision on entities it classified but abstained on 9.8% of candidates, of which 90.2% were valid historical Spanish entities absent from LLM training data. Abstention was highest for AGRUPACION (16.8%) and COMPOSITOR (11.2%) (see Figure 2), confirming that historically specific domain knowledge cannot be recovered from general-purpose language model weights. All final annotations were independently validated by the expert annotator, ensuring that no LLM outputs were accepted without human verification.

Manual expert annotation and annotation guidelines. All final annotation decisions were made by the PI (M.P.N.), a musicologist specialising in Spanish musical press of 1918–1935. Over 2,000 candidate entities were reviewed across iterative cycles using custom HTML review interfaces. Annotation

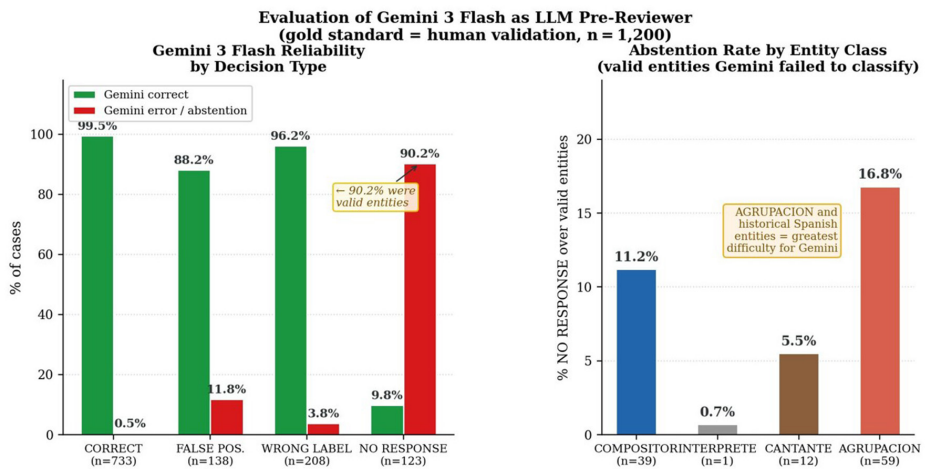


FIGURE 2: Evaluation of Gemini 3 Flash as LLM Pre-reviewer against Human Gold Standard (n = 1,200 Candidate Entities). Left: reliability by decision type — Gemini achieved 99.5% precision on confirmed entities but abstained on 9.8% of candidates, 90.2% of which were valid historical Spanish entities. Right: abstention rate by entity class, highest for AGRUPACION (16.8%) and COMPOSITOR (11.2%), reflecting the historical specificity of Spanish musical institutions and figures absent from LLM training corpora.

criteria are documented in detail in the 46-page [annotation_guidelines.pdf](#) distributed with the dataset, covering: entity type definitions with positive and negative examples; disambiguation rules for persons active in multiple roles; compound entities (e.g., ‘Bach-Busoni’ as arranger–original-composer pair); OCR variant normalisation; and edge cases for ensemble names and historical institutional designations. The guidelines are also available online at <https://leximususal.github.io/leximus-ner-guia-anotacion/>. No formal inter-annotator agreement was computed, since a single domain-expert annotator made all final decisions. This choice reflects the specialised musicological expertise required to identify and disambiguate historical Spanish musical figures largely absent from standard reference works; consistency across cycles was supported by the guidelines above and by cross-checking against LLM pre-review outputs. We identify inter-annotator agreement, computed on a sample by a second domain expert, as a priority for future work.

Model training. Validated annotations were converted to spaCy binary format and used to fine-tune the BETO-based NER component (*es_dep_news_trf*; Cañete et al., 2020; Devlin et al., 2019), with the EntityRuler frozen. Nine model versions were trained across iterative active learning cycles. spaCy F1 on the fixed held-out test set (60 documents, consistent across cycles v2–v7) improved from 0.45 (v2) to 0.56 (v6–v7). From v8 onward, the evaluation set was replaced by a larger human-reviewed corpus (300 documents drawn from the 1,200 manually validated candidates), which is not directly comparable to the prior fixed test set but provided a more demanding and representative benchmark aligned with the evolving annotation scope. Version 9 (v9) is the current production model.

Evaluation methodology. Two complementary precision metrics are reported, and their scope is distinguished explicitly. 1. *Entity-level detection precision on the pilot corpus*: of 707 candidates detected by LexiMus-NER v9 in the *El Sol* 1926 evaluation set (73,633 words), 517 were confirmed correct, 68 required label reclassification (detected as genuine musical entities but initially assigned to a wrong class), and 55 were false positives, yielding a detection precision of 91.9% (false positive rate: 7.8%). 2. *Human-validated classification precision*: across all four entity classes over $n = 1,183$ entities reviewed in the full annotation cycle: COMPOSITOR 95.3%, INTERPRETE 92.2%, CANTANTE 91.3%, AGRUPACION 98.0%, global 94.3% (see Figure 3). Corpus-level recall is not reported: the pipeline-first annotation design — in which human review covers pipeline-detected candidates rather than exhaustive tagging of complete documents — means that true corpus-level recall cannot be computed without independent complete annotation of the full corpus. This approach, standard in active learning for large humanities corpora (Settles, 2012; Ehrmann et al., 2022), trades corpus-wide recall estimability for scalability and expert annotation focus on the most challenging cases. The 73,633-word pilot corpus used

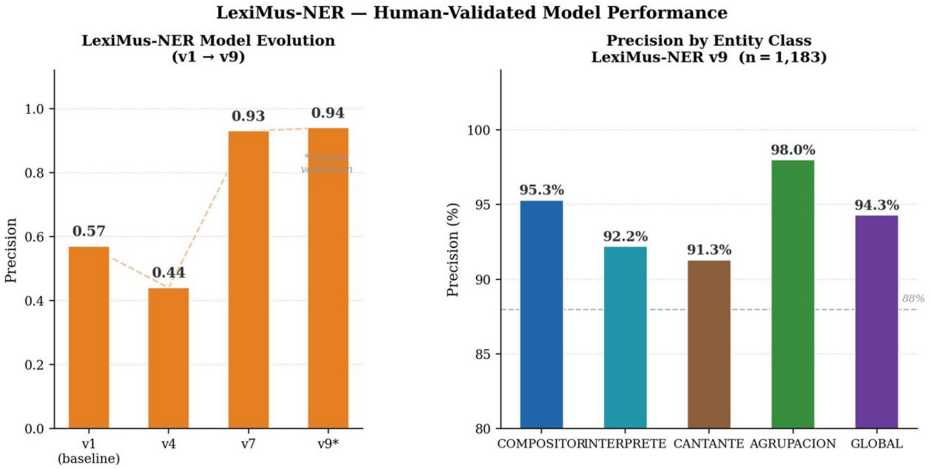


FIGURE 3: LexiMus-NER Model Performance. Left: human-validated precision at key model milestones across nine iterative fine-tuning cycles (v1: 57%; v4: 44%, reflecting scope expansion to INTERPRETE/CANTANTE categories; v7: 93%; v9: 94.3%). These values represent the proportion of pipeline-detected candidates confirmed as correct by the domain expert at each cycle, not spaCy training F1. Right: human-validated classification precision by entity class for LexiMus-NER v9 (n = 1,183 reviewed entities; all classes above the 88% target threshold).

for entity-level detection precision was reused across development cycles to inform methodological decisions rather than used only once; it therefore functions as a recurring development benchmark rather than a single-use, blind held-out test set (see Table 3).

Table 3 summarises every benchmark and evaluation set referenced in this paper, clarifying their relationship, provenance, and non-overlapping roles in the pipeline.

4. Data

The dataset adheres to FAIR data principles. Findability is ensured through DOI registration on Zenodo (Palacios Nieto, 2025); accessibility through open repositories (Zenodo, Hugging Face, GitHub); interoperability through widely used formats (CoNLL TSV, CSV); and reusability through CC BY 4.0 licensing and extensive annotation documentation (see Table 4).

- LexiMus-NER deposited at Zenodo — DOI: www.doi.org/10.5281/zenodo.19429405
- Trained model (LexiMus-NER v9) on Hugging Face: <https://huggingface.co/LeximusUSAL/LexiMus-BETO-per-v1>
- Temporal coverage: 1918–1935
- Licence: CC BY 4.0

TABLE 4: Technical Metadata

Language	Spanish (historical, 1918–1935)
Formats	CoNLL TSV (UTF-8); CSV; PDF
Annotation scheme	IOB2
POS tagset	Universal Dependencies
Licence	CC BY 4.0
DOI	10.5281/zenodo.19429405
Source period	1918–1935
Source corpus size	~3,050,000 words
Related project	LexiMus PID2022-139589NB-C33

The CC BY 4.0 licence applies to the dataset’s original contributions — the entity annotations, the curated gazetteer, the annotation guidelines, the trained model, and the accompanying code — all produced within the LexiMus project. It does not extend to, and cannot grant rights over, the underlying historical newspaper text of *El Sol* and *Ondas* itself. The copyright status of that source text is governed independently by Spanish intellectual property law (Real Decreto Legislativo 1/1996): unsigned items are protected for 70 years from lawful publication (Art. 27.2), a term that has elapsed for material published in 1918–1935, while signed items attributed to a named author remain subject to the general term of 70 years post mortem auctoris and may still be under copyright depending on the author’s date of death. Access to the original digitised text continues to be governed by the terms of use of the Biblioteca Nacional de España’s open-access portal.

File inventory. The dataset deposited on Zenodo includes the files described in Table 5.

Data format. All annotation files use CoNLL-style TSV (Tjong Kim Sang & De Meulder, 2003) with IOB2 encoding (three tab-separated columns: TOKEN, POS, IOB2_LABEL). Part-of-speech tags follow Universal Dependencies conventions produced by spaCy *es_core_news_md*. Documents are separated by blank lines and preceded by a ‘# doc_id =...’ comment. Encoding is UTF-8 throughout. The format is directly compatible with Hugging Face Datasets, spaCy, Flair, and AllenNLP.

Example annotation (El Sol, 14 March 1923):

La	DET	O
Sociedad	NOUN	B-AGRUPACION

Filarmónica	PROPN	I-AGRUPACION
celebró	VERB	O
Hipólito	PROPN	B-CANTANTE
Lázaro	PROPN	I-CANTANTE

Dataset statistics. The dataset is divided into training, development, and test splits (see Table 6). The test split totals 1,116 annotation windows (~150 tokens each): 60 base documents held constant since v2, plus 1,056 additional windows drawn from *Ondas* and annotated from v8 onward (see Table 3 for how this set relates to the other evaluation corpora used during development). Per-source breakdown (*El Sol* vs. *Ondas*) is available in statistics.json.

Gazetteer. The file `entidades_ner_leximus.csv` contains 1,035 unique named entities (317 composers, 305 performers, 290 singers, 123 ensembles) extracted from concert programmes and press coverage of 1918–1935. Each entry records the surface form, entity label, normalised form, instrument or role,

TABLE 5: File Inventory of the LexiMus-NER Dataset

File	Size	Description
train.tsv	2.1 MB	Training split: 2,733 documents, 8,536 entities
dev.tsv	273 KB	Development split: 394 documents, 1,003 entities
test.tsv	755 KB	Test split: 1,116 annotation windows, 2,158 entities
entidades_ner_leximus.csv	—	Gazetteer: 1,035 unique entities with variants and metadata
annotation_guidelines.pdf	—	Full annotation criteria, entity definitions, and edge cases (46 pp.)
statistics.json	—	Entity counts by split, label, and source publication
README.md	—	Dataset documentation
LICENSE	—	CC BY 4.0

TABLE 6: Entity Counts by Split and Entity Class

Split	Documents	Compositor	Interprete	Cantante	Agrupacion	Total
train	2,733	2,343	1,692	1,435	3,066	8,536
dev	394	380	209	210	204	1,003
test	1,116	514	313	449	882	2,158
Total	4,243	3,237	2,214	2,094	4,152	11,697

and canonical form linking spelling variants and OCR-degraded forms to a single reference.

Trained model. LexiMus-NER v9 is a two-layer spaCy pipeline on *es_dep_news_trf* (BETO): a deterministic EntityRuler (gazetteer-based, high precision for known entities) combined with a statistical NER transformer that generalises to unseen entities. Human-validated performance metrics are shown in Figure 3. The model (841 MB) is available on Hugging Face under CC BY 4.0.

5. Concluding Remarks

LexiMus-NER opens new lines of research in digital musicology. The dataset currently focuses on Spanish musical press from Madrid-based publications of markedly different character: *El Sol*, a general daily newspaper founded in 1917 and recognised as one of the most influential Spanish dailies of the period, whose principal music critic was Adolfo Salazar; and *Ondas*, a specialised weekly magazine launched in 1925 and affiliated with Unión Radio, Spain's pioneering private radio broadcaster, combining music journalism (articles, essays, and reviews) with listings of radio programming. Together they offer complementary vantage points on the same musical culture — one from the general press, the other from the emerging medium of radio. Expanding to regional newspapers and earlier periodicals will be a priority in future releases. By enabling automated, role-level classification of musical agents in large digitised Spanish press collections, the dataset supports quantitative analysis of questions that until now required manual reading at impractical scale: canon formation processes, gender distribution by musical role and venue, the vocabulary of critical reception across performer types, and the recovery of musicians absent from modern biographical dictionaries. Correctly classified entities can populate biographical databases of performers effectively invisible in existing reference works.

As a reusable NLP resource, LexiMus-NER is the first domain-specific NER training dataset for historical Spanish musical discourse. It can be used to fine-tune any token classification model supporting CoNLL/IOB2 format. The fixed test split (1,116 windows, 2,158 entities) provides a stable benchmark; the cross-model comparison between LexiMus-NER v9 and Gemini 3 Flash zero-shot on the same evaluation corpus constitutes a published baseline for future systems (see Figure 4). On dimensions most critical to reproducible humanities research — determinism, cost at scale, auditability, and AGRUPACION precision — the fine-tuned local model outperforms zero-shot LLM inference on six of eight evaluated criteria.

The two-layer pipeline architecture (deterministic EntityRuler + statistical NER) is designed for incremental extension: new publications require only

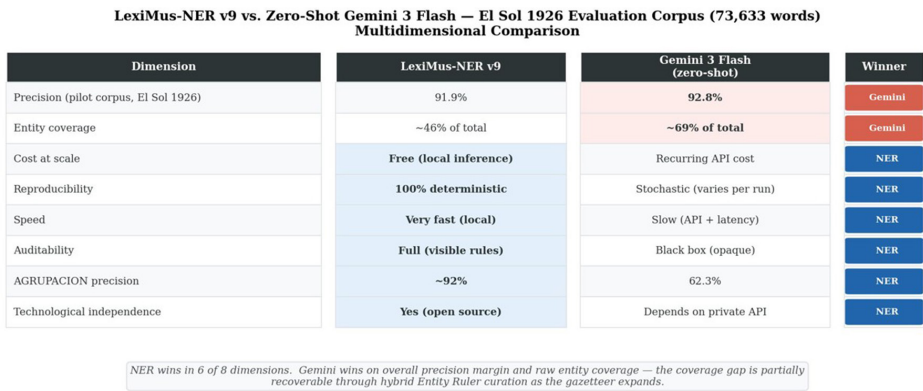


FIGURE 4. LexiMus-NER v9 vs. Zero-shot Gemini 3 Flash on the El Sol 1926 Evaluation Corpus (73,633 words). NER wins on six of eight evaluated dimensions; Gemini wins on overall precision margin (92.8% vs. 91.9%) and raw entity coverage (~69% vs. ~46%). The coverage gap is partially recoverable through hybrid EntityRuler curation.

gazetteer expansion and, optionally, additional annotated documents. The LexiMus project plans to apply the pipeline to its full corpus of 19 Spanish musical magazines (1842–1936, over 11 million words), with the dataset growing incrementally. The documented workflow — LLM-assisted pre-review, expert HTML validation interfaces, iterative active learning, and hybrid local/LLM inference — constitutes a replicable methodology for domain NER annotation in historical humanities corpora, transferable to other languages, periods, and domains.

This NER is, ultimately, an infrastructure for a larger research question: how the historical musical press constructed cultural hierarchies and decided which musicians would form part of the canon and which would remain outside the historical narrative. Without a specialised NER tool capable of correctly classifying the musical agents of the period, a significant portion of this history remains invisible to computational analysis.

Acknowledgements

This work was carried out within the LexiMus project (*Léxico y ontología de la música en español*, PID2022-139589NB-C33), funded by MCIN/AEI/10.13039/501100011033 and the European Union (NextGenerationEU). The project is coordinated by the Instituto Complutense de Ciencias Musicales (IPs: Dr. Álvaro Torrente and Dr. José María Domínguez), with subprojects at the Universidad de La Rioja (IP: Dra. Teresa Cascudo García-Villaraco) and the Universidad de Salamanca. The author thanks the Biblioteca Nacional de España for open-access digitisation of the historical press collections used as source corpus.

Competing Interests

The author declares no competing interests.

References

- Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. *Proceedings of PML4DC at ICLR 2020*. <https://github.com/dccuchile/beto>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). <https://doi.org/10.18653/v1/N19-1423>
- Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., Doucet, A., & Clematide, S. (2022). Named entity recognition and classification on historical documents: A survey. *ACM Computing Surveys*, 56(2), 1–47. <https://doi.org/10.1145/3604931>
- Ehrmann, M., Romanello, M., Flückiger, A., & Clematide, S. (2020). Extended overview of CLEF HIPE 2020: Named entity processing on historical newspapers. In *CLEF 2020 Working Notes*. <https://zenodo.org/records/4117566>
- Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. Explosion AI. <https://spacy.io>
- Palacios Nieto, M. (2025). LexiMus-NER: A Spanish Musical Press NER Dataset (1918–1935). *Zenodo*. <https://doi.org/10.5281/zenodo.19429405>
- Settles, B. (2012). *Active learning*. Morgan & Claypool. <https://doi.org/10.2200/S00429ED1V01Y201207AIM018>
- Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003* (pp. 142–147). <https://doi.org/10.3115/1119176.1119195>