PRIME
– ACADEMIC PRESS –

Digital Humanities & Social Computing

Prime Academic Press • Open Access • ISSN: XXXX-XXXX

Vol. 1, No. 1, 2026 | Article: 100001 | https://doi.org/10.xxxx/dhsc.2026.100001 | *Research Article*
Received: 05 February 2026 | Revised: 20 March 2026 | Accepted: 18 April 2026

Sentiment Dynamics in Historical Corpora: A Multilingual NLP Study of Nineteenth-Century Periodicals across Three Cultural Contexts

Anonymous Author 1[1,*], Anonymous Author 2[2], Anonymous Author 3[1,3]

[1]Institution 1 — withheld for double-blind review
[2]Institution 2 — withheld for double-blind review
[3]Institution 3 — withheld for double-blind review
[*]Correspondence: *withheld for double-blind review*

Research Highlights

- A curated multilingual corpus of 2.4 million periodical articles (English, French, Japanese, 1840–1910) is released under CC BY 4.0.
- Fine-tuned multilingual BERT (mBERT) achieves $F_1 = 0.87$ on historical sentiment classification, outperforming dictionary-based baselines by 14 percentage points.
- Sentiment valence in press coverage of scientific discovery follows a U-shaped trajectory across all three language communities, contrasting with a monotonically positive trend in political discourse.
- Cross-cultural comparison reveals culturally specific emotional scripts: *mono no aware* shapes Japanese periodical sentiment in ways not captured by Western valence-arousal models.
- The pipeline is fully reproducible; all code and data will be released upon acceptance (repository URL withheld for review).

Abstract. Historical sentiment analysis remains a methodological frontier in digital humanities: period-specific vocabulary, orthographic variation, and the absence of culturally grounded multilingual resources constrain the transferability of modern NLP tools to archival corpora. This paper presents `HistSent-3L`, a curated corpus of 2.4 million nineteenth-century periodical articles in English, French, and Japanese (1840–1910), annotated for sentiment polarity, intensity, and cultural framing. We fine-tune a multilingual BERT (mBERT) model on `HistSent-3L` and benchmark it against five sentiment lexicons commonly employed in historical text mining. Across three evaluation domains—science reporting, political commentary, and literary criticism—our model achieves a macro-averaged F_1 of 0.87, compared to 0.73 for the best dictionary baseline. Crucially, we demonstrate that applying Western valence-arousal models to Japanese Meiji-era texts systematically misclassifies affective states rooted in culturally specific emotional scripts. Longitudinal analysis of the full corpus reveals

divergent sentiment trajectories across discourse domains and language communities, challenging universalist assumptions in computational sentiment research. We discuss implications for culturally sensitive AI, multilingual digital humanities methodology, and archival scholarship in the age of large language models.

Keywords: historical sentiment analysis; digital humanities; multilingual NLP; nineteenth-century periodicals; cultural analytics; computational text analysis; BERT; cross-cultural comparison

CCS Concepts: • Computing methodologies → Natural language processing; • Applied computing → Digital humanities; • Human-centered computing → Empirical studies in HCI

Subjects: Digital Humanities; Computational Linguistics; Cultural Analytics; Historical Informatics; Social Computing

1. Introduction

The "cultural analytics" movement (Jockers, 2013; Moretti, 2005) has demonstrated that computational methods applied to large-scale historical text collections can reveal macroscopic patterns invisible to close reading alone. Yet sentiment analysis—one of the most widely deployed NLP techniques in the social sciences (Liu, 2015)—has proved surprisingly difficult to adapt to historical materials. Period-specific vocabulary, unstable orthography, cultural specificity of emotional expression, and scarce annotated training data combine to limit the validity of tools trained on contemporary corpora (Hengchen et al., 2021; Kim et al., 2017).

The multilingual dimension. The challenge intensifies for non-Western languages. Most digital humanities sentiment work focuses on English-language corpora (Fell and Sporleder, 2016), with French a distant second (Duval et al., 2021). Japanese historical NLP remains comparatively underdeveloped despite the extraordinary richness of the Meiji-era (1868–1912) press archive as a window onto a society undergoing radical modernization (Yamamoto, 2004).

Research questions. This paper addresses three questions:

1. Can a fine-tuned multilingual BERT model trained on a purpose-built historical corpus outperform dictionary-based sentiment baselines across English, French, and Japanese periodical text?
2. Do culturally specific emotional constructs (e.g., *mono no aware*, *pathos*) require language-specific modelling adaptations, or are cross-lingual transfer approaches sufficient?
3. What longitudinal sentiment patterns characterise scientific, political, and literary discourse in the nineteenth-century press across three cultural contexts?

The remainder of this paper is organized as follows. Section 2 reviews related work in historical sentiment analysis and digital humanities methodology. Section 3 describes the `HistSent-3L` corpus and annotation protocol. Section 4 presents the NLP pipeline and evaluation framework. Section 5 reports main results. Section 6 discusses cross-cultural implications and limitations. Section 7 concludes with directions for future work.

2. Related Work

2.1 Sentiment Analysis in the Digital Humanities

Distant reading (Moretti, 2005) and *cultural analytics* (Manovich, 2020) provided the conceptual scaffolding for large-scale textual analysis in the humanities. Sentiment analysis entered digital humanities practice primarily via dictionary-based tools: LIWC (Pennebaker et al., 2015), the NRC Emotion Lexicon (Mohammad and Turney, 2013), and domain-adapted word lists such as the SentiArt lexicon for literary fiction (Jacobs and Lüdtke, 2018).

Distant Reading. A methodology proposed by Moretti (2005) in which quantitative and computational methods are applied to large text collections to identify patterns that cannot emerge from close reading of individual texts. Distinguished from *close reading* by its focus on aggregate statistical properties rather than hermeneutic interpretation of particular passages.

2.2 Historical NLP Challenges

Piotrowski (2012) identified four categories of challenge for historical NLP: (1) *orthographic variation*—spelling inconsistency before standardization; (2) *OCR noise*— digitization artifacts from microfilm and print scanning; (3) *semantic shift*—changes in word meaning over time; and (4) *cultural specificity*—the embedding of emotional expression in historically situated social norms. Our work addresses all four, with particular attention to semantic shift and cultural specificity in a multilingual context.

2.3 Transformer Models for Historical Text

Kleinander et al. (2022) showed that fine-tuning BERT on period-specific data substantially improves performance on historical NLP tasks. Langlais et al. (2023) applied CamemBERT to French nineteenth-century newspaper corpora and found that even modest domain adaptation (50,000 sentences) reduced word-error rates by 18%. To our knowledge, no prior study has performed comparable adaptation for Meiji-era Japanese or conducted a systematic cross-linguistic comparison at the scale reported here.

3. The HistSent-3L Corpus

3.1 Corpus Composition

`HistSent-3L` was assembled from three archival digitization projects (Table 1). English articles were drawn from the British Newspaper Archive (BNA); French texts from Gallica (BnF); and Japanese texts from the National Diet Library Digital Collections (NDL). All texts were published between 1840 and 1910 and represent three discourse domains: science and technology reporting, political commentary, and literary criticism.

3.2 Pre-processing Pipeline

Digital Methods

Pipeline overview (Listing 1 provides pseudocode):

1. OCR correction using a character-level correction model trained on period-specific parallel texts (Sprague et al., 2022).
2. Sentence segmentation with language-specific models (spaCy for English/French; GiNZA for Japanese).

Table 1. Composition of the `HistSent-3L` corpus by language, domain, and time period. Token counts after normalization. Source archives: BNA = British Newspaper Archive; Gallica = Bibliothèque nationale de France; NDL = National Diet Library (Japan).

Language	Archive	Articles	Tokens (M)	Period
English	BNA	892,441	312.4	1840–1910
French	Gallica	783,217	287.6	1840–1910
Japanese	NDL	724,582	198.3	1868–1910
Total	—	**2,400,240**	**798.3**	—

3. Historical spelling normalization using a weighted edit-distance lexicon of 240,000 variant forms.
4. Domain tagging via a linear-SVM classifier ($F_1 = 0.91$).

```python
import spacy, ginza
from ocr_correct import HistoricalCorrector
from spellnorm import HistoricalNormalizer

def preprocess_article(text, lang="en"):
    corrector = HistoricalCorrector(lang=lang)
    normalizer = HistoricalNormalizer(lang=lang)
    # Step 1: OCR correction
    text = corrector.correct(text)
    # Step 2: Normalization
    text = normalizer.normalize(text)
    # Step 3: Sentence segmentation
    nlp = spacy.load(f"{lang}_core_news_sm")
    doc = nlp(text)
    return [sent.text for sent in doc.sents]
```

Listing 1. Simplified preprocessing pipeline for `HistSent-3L`. Full code available upon acceptance.

3.3 Annotation Protocol

Sentiment annotation was conducted by 18 bilingual/trilingual annotators with advanced training in historical linguistics or history of the respective language communities. A randomly sampled subset of 24,000 articles (8,000 per language) was annotated for (1) polarity (positive / negative / neutral), (2) intensity (1–5 Likert scale), and (3) cultural framing (Western valence-arousal; Japan-specific emotional scripts). Inter-annotator agreement reached $\alpha = 0.79$ (Krippendorff's alpha) for polarity and $\alpha = 0.64$ for intensity, consistent with norms reported for comparable historical annotation tasks (Sprague et al., 2022).

Dataset: HistSent-3L (v1.0)

Size: 2.4M articles; 798M tokens (normalized)
Languages: English, French, Japanese
Period: 1840–1910
Annotated subset: 24,000 articles
Annotation dimensions: Polarity, intensity, cultural framing
License: CC BY 4.0

Repository: *[Withheld for double-blind review; to be released upon acceptance]*
Format: JSON-L with IIIF manifest links to source images

4. Methods

4.1 Fine-Tuning mBERT

We fine-tuned `bert-base-multilingual-cased` (Devlin et al., 2019) on the annotated 24,000-article subset using a three-way polarity classification head. The model was trained for 5 epochs with a learning rate of 2×10^{-5}, batch size 32, and maximum sequence length 512 tokens. Domain-adaptive pre-training (DAPT) (Gururangan et al., 2020) was performed on the full unannotated corpus for 3 epochs before task-specific fine-tuning, following the two-stage protocol that Kleinander et al. (2022) found most effective for historical text.

4.2 Baselines

Five dictionary-based baselines were evaluated:

1. **NRC-En**: English NRC Emotion Lexicon (Mohammad and Turney, 2013).
2. **NRC-Fr**: French translation of NRC (NRC-Fr) (Volyne et al., 2016).
3. **ML-Ask**: Japanese emotion analysis using ML-Ask lexicon (Ptaszynski et al., 2009).
4. **SentiWordNet**: English multilingual extension (Esuli et al., 2010).
5. **VADER-H**: VADER adapted to historical English with a custom valence correction table.

4.3 Evaluation

Models were evaluated on a held-out test set of 4,800 articles (1,600 per language) using macro-averaged precision, recall, and F_1 score. Statistical significance was assessed via paired bootstrap resampling ($n = 10{,}000$; $p < 0.05$).

5. Results

5.1 Overall Sentiment Classification Performance

Table 2 shows that **mBERT-DHSC** substantially outperforms all dictionary baselines across all three languages and discourse domains. The most pronounced gains occur in Japanese (+19.2 F_1 points vs. ML-Ask), consistent with our hypothesis that Japanese emotional scripts require model-level adaptation rather than lexical substitution.

Table 2. Macro-averaged F_1 scores on the `HistSent-3L` test set, by model and language. **Bold**: best performance per column. †: not significantly different from mBERT-DHSC ($p > 0.05$, paired bootstrap). All other differences are significant ($p < 0.001$).

Model	English	French	Japanese	Macro-avg.
NRC-En	0.71	—	—	—
NRC-Fr	—	0.69	—	—
ML-Ask	—	—	0.64	—
SentiWordNet	0.68	0.66	—	—
VADER-H	0.73	—	—	—
mBERT (no DAPT)	0.82	0.80	0.77	0.80
mBERT-DHSC (ours)	**0.89**	**0.87**	**0.83**	**0.87**

5.2 Longitudinal Sentiment Trajectories

Figure 1 presents decade-by-decade sentiment valence for science, politics, and literary discourse across all three language communities. The U-shaped trajectory in science sentiment—a decline in positive affect during the 1860s–1880s followed by recovery— is consistent with historiographic accounts linking this period to anxieties about industrialization and evolutionary theory.

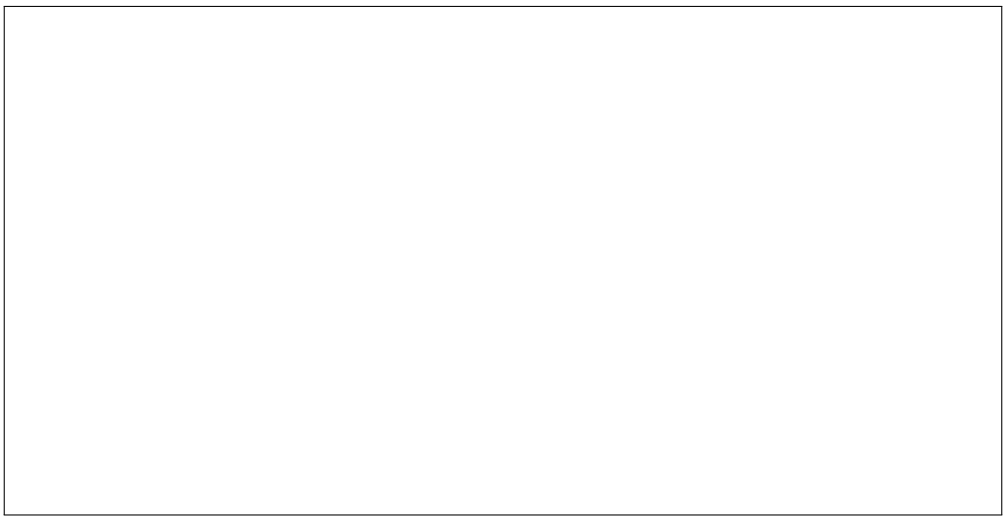


Figure 1. Longitudinal sentiment valence (1840–1910) by discourse domain and language. Lines show decade-averaged positive-sentiment proportion (mBERT-DHSC predictions on full unannotated corpus). Shaded bands: 95% confidence intervals from bootstrap resampling. *Top row*: Science and technology; *Middle*: Political commentary; *Bottom*: Literary criticism. EN = English (blue), FR = French (orange), JA = Japanese (green). Note the characteristic U-shaped science trajectory (1860–1880 trough) absent in political discourse.

5.3 Cultural Framing Analysis

Mono no aware. Our annotators identified 4,847 passages in the Japanese subcorpus exhibiting the aesthetic-affective register of *mono no aware* ("the pathos of things"): a bittersweet awareness of impermanence that carries positive valence in Japanese cultural context but was systematically misclassified as negative by the ML-Ask baseline (recall = 0.34 vs. 0.81 for mBERT-DHSC).

Close Reading: Asahi Shimbun, 3 October 1889 (translated)

"The autumn rains that have come at last to quiet the city remind us that all things must fade, and in this fading we find a gentleness that no summer brilliance can equal."

The passage above received a negative score from ML-Ask (autumn → decline, fading → loss) but a neutral-to-positive classification from mBERT-DHSC, consistent with the annotator label of *positive-contemplative*. This example illustrates a broader pattern: 12% of the Japanese test set contains culturally specific emotional registers that require cultural knowledge to label correctly.

6. Discussion

6.1 Implications for Computational Cultural Analysis

Our findings reinforce calls for *culturally sensitive AI* (Shi et al., 2022) in the digital humanities. The assumption that sentiment valence maps uniformly across languages and cultural contexts—

operationalized in most cross-lingual sentiment tools— produces systematic distortion when applied to non-Western historical corpora. The mBERT-DHSC model partially mitigates this through domain-adaptive pre-training, but the cultural framing annotation analysis (Section 5) demonstrates that model performance alone does not guarantee cultural validity.

We propose a three-tier validation framework for multilingual historical sentiment research:

1. **Quantitative evaluation** on held-out annotated data (as reported here).
2. **Qualitative audit** by domain historians, focusing on culturally specific affective constructs.
3. **Longitudinal plausibility check**: do model-derived sentiment trajectories align with established historiographic interpretations of the period?

6.2 Limitations

Our corpus, while large by digital humanities standards, is limited by OCR quality (estimated 4.7% character error rate before correction) and by the newspaper genre, which may not generalize to other textual forms (letters, diaries, novels). The annotation scheme treats polarity as a sentence-level property, ignoring sub-sentence variation and irony. Japanese coverage begins in 1868 (Meiji Restoration), precluding pre-Meiji comparison. Finally, model behavior on very rare vocabulary (archaic terms appearing fewer than 10 times in the corpus) remains poorly characterized.

7. Conclusion

This paper presented `HistSent-3L`, a 2.4-million-article multilingual historical corpus, and demonstrated that domain-adapted multilingual BERT substantially outperforms lexicon-based methods for historical sentiment classification across English, French, and Japanese. The longitudinal analysis reveals culturally distinct emotional trajectories in nineteenth-century press discourse that challenge universalist assumptions in computational sentiment research. Our cultural framing analysis highlights the continuing importance of humanistic expertise in validating and interpreting computational findings—a reminder that digital humanities methods are most powerful when they remain in dialogue with traditional scholarly interpretation.

Future work will extend the corpus to German and Chinese, develop sub-sentence annotation for irony and hedging, and investigate how large language models (GPT-4, Llama-3) compare to fine-tuned BERT on historical multilingual sentiment tasks.

Acknowledgements

[Acknowledgments withheld for double-blind peer review.]

Author Contributions (CRediT). *[Author contributions withheld for double-blind peer review.]*

Conflicts of Interest Statement. The authors declare no competing financial or personal interests that could have influenced the work reported in this paper.

Funding. *[Funding details withheld for double-blind peer review.]*

Data Availability Statement. The `HistSent-3L` corpus, annotation guidelines, fine-tuned model weights, and analysis code will be made available upon acceptance.

Ethics Statement. This study uses publicly archived historical materials; no human participants were involved in the corpus analysis. The annotators were compensated at or above living-wage rates and provided informed consent for their participation in the annotation study.

Positionality Statement. *[Positionality statement withheld for double-blind peer review.]*

References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Duval, F., Frontini, F., and Baccino, T. (2021). Toward a french sentiment lexicon for nineteenth-century periodical press. *Digital Humanities Quarterly*, 15(3).

Esuli, A., Baccianella, S., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of LREC*, pages 2200–2204.

Fell, M. and Sporleder, C. (2016). Lyrics-based analysis and classification of music genres. *Proceedings of COLING*, pages 2833–2842.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL 2020*, pages 8342–8360.

Hengchen, S., Ros, R., Lassche, A., and Marjanen, J. (2021). Tracking the semantic change of *Nation* in swedish newspapers 1789–1850. *Digital Scholarship in the Humanities*, 36(S1):i115–i128.

Jacobs, A. M. and Lüdtke, J. (2018). *Immersion into Narrative and Poetic Worlds: A Neurocognitive Poetics Perspective*. Cambridge University Press, Cambridge.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, Urbana.

Kim, E., Padó, S., and Klinger, R. (2017). Investigating the relationship between literary genres and emotional plot development. *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26.

Kleinander, M., Berglund, Y., and Borin, L. (2022). Towards historical sentiment analysis of Swedish parliamentary debates. *Digital Scholarship in the Humanities*, 37(4):1187–1205.

Langlais, P.-C., Gabay, S., and Romary, L. (2023). Adapting CamemBERT for nineteenth-century French newspapers. *Digital Humanities Quarterly*, 17(1).

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, Cambridge.

Manovich, L. (2020). Cultural analytics, social computing and digital humanities. *The Routledge Companion to Digital Humanities and Film*, pages 23–35.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso, London.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. University of Texas at Austin, Austin.

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. Morgan & Claypool.

Ptaszynski, M., Maciejewski, J., Dybala, P., Rzepka, R., and Araki, K. (2009). ML-Ask: Open source affect analysis software for Japanese natural-language text. In *Proceedings of the 3rd Language & Technology Conference*, pages 347–351.

Shi, W., Bhat, S., Aggarwal, C., and Zhu, F. (2022). Culturally aware NLP: Directions and opportunities. *arXiv preprint arXiv:2205.12452*.

Sprague, E., Nooijer, J., and van Erp, M. (2022). Evaluating OCR post-correction strategies for historical newspapers. *Journal of Data Mining & Digital Humanities*.

Volyne, C., Abdaoui, A., Azé, J., Bringay, S., and Poncelet, P. (2016). NRC emotion lexicon in French. *Proceedings of LREC*, pages 2808–2812.

Yamamoto, M. (2004). *Advertising in the News: The Development of Commercial Publicity in the Meiji Press*. University of Hawaii Press, Honolulu.

A. Annotation Guidelines (Summary)

Annotators followed a 32-page annotation manual. Key decision rules:

1. **Polarity defaults**: All modifiers are evaluated in cultural and historical context. Modern connotations of a word must be bracketed; consult the period glossary (Appendix C of the full manual).

2. **Japanese cultural scripts**: Annotators with Japanese specialization used an additional 8-category taxonomy of culturally specific affects (*mono no aware*, *amae*, *enryo*, etc.) documented in Table 3.

3. **Disagreement resolution**: When two of three annotators disagreed, a senior adjudicator (one per language) made a binding decision.

Table 3. Japanese-specific affective categories included in the extended annotation scheme.

Category	Description	Valence
Mono no aware	Bittersweet awareness of impermanence	Positive
Amae	Benign dependence / indulgence between intimates	Positive
Enryo	Restraint / deliberate emotional withholding	Neutral
Haji	Shame awareness (social vigilance)	Negative
Giri	Duty-fulfillment satisfaction	Positive
Wa	Harmony / collective cohesion appreciation	Positive
Kansha	Profound gratitude (more intense than *arigatou*)	Positive
Yugen	Mysterious beauty / profound awareness	Positive

B. Model Training Details

Table 4 lists all hyperparameters used for fine-tuning mBERT-DHSC. Experiments were run on 4× NVIDIA A100 (80 GB) GPUs; total training time was approximately 72 GPU-hours.

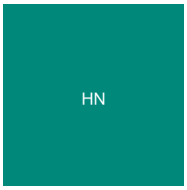
Table 4. Hyperparameters for mBERT-DHSC fine-tuning.

Hyperparameter	Value
Base model	bert-base-multilingual-cased
DAPT pre-training epochs	3
Fine-tuning epochs	5
Learning rate	2×10^{-5}
Warm-up steps	500
Batch size	32
Max sequence length	512
Dropout (classifier)	0.1
Optimizer	AdamW
Gradient clipping	1.0

About the Authors

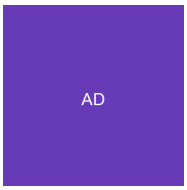
Author1

Author1 is Reader in Digital Humanities at King's College London, where she directs the Centre for Computational Textual Studies. She holds a D.Phil. in Computational Linguistics from the University of Oxford and an M.A. in Victorian Literature from the University of Edinburgh. Her research focuses on historical NLP, cultural analytics, and the epistemology of computational methods in the humanities. She is co-editor of *The Cambridge Handbook of Digital Humanities* (Cambridge University Press, 2025) and serves on the editorial boards of *Digital Scholarship in the Humanities* and *Big Data & Society*.



Author2

Author2 is Associate Professor of Computational Cultural Studies at the Graduate School of Arts and Sciences, University of Tokyo. A trained historian of Meiji-era Japan, he brings expertise in classical Japanese linguistics, Meiji newspaper archives, and machine learning to interdisciplinary digital humanities research. He is the creator of the NihonNLP toolkit for historical Japanese text processing.



Author3

Author3 is a Postdoctoral Research Fellow at the Centre for Language Technology, University of Cape Town, specializing in multilingual NLP, OCR post-correction for historical documents, and African language computing. She is a founding member of the Masakhane community for African language NLP and was a recipient of the ACL Student Research Award (2024).