

SIGHUM 2026

**10th Joint SIGHUM Workshop on Computational Linguistics
for Cultural Heritage, Social Sciences, Humanities and
Literature**

Proceedings of the Workshop

March 28-29, 2026

The SIGHUM organizers gratefully acknowledge the support from the following sponsors.



©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-373-9

Introduction

Welcome to the 2026 edition of LaTeCH-CLfL!

Whether you are a seasoned participant or attending for the first time, thank you for gathering here. With twenty years of history, our workshop remains a forum for diverse and evolving conversations, and brings together work at the intersection of language technology, computational linguistics, and the Humanities in their widest sense.

This year, we are proud to present a collection of studies that goes beyond the known restrictions of Anglocentric research, and that spans a wide array of different languages. Our collection features several cutting edge works on under-resourced languages: automatic modelling of Armenian vocabulary and Armenian inscriptions, Greek Poetry rhyme generation, a quantitative study of Romanian writings, an analysis of Finnish refugees' interviews, a work on indirect speech in historical Danish and Norwegian corpora, a work on the infamous problem of reported speech in Classical Latin, and even a pipeline for Ottoman-Turkish bibliographies.

High-resource languages that are not English are also present: several works tackle fine-grained linguistic and stylistic problems in both contemporary and historical German, as well as Russian war propaganda and contemporary French literature, and linguistic diversity is central in several brand new resources, such as a cross-lingual corpus of LLM-generated film synopses, a dataset of under-resourced languages for NLP, and a dataset on digital linguistic diversity.

But a rich linguistic diversity is not the only feature of this year's workshop. Our field is still exploring and enriching LLMs' powers of analysis and their relationship with traditional Humanities' problems, applying them to a variety of high complexity tasks, from modelling domain-specific language variation to automatic annotation of Ancient Greek geographical nouns, as well as reflecting on how their development has changed corpus linguistics.

As usual, our workshop also touches several topics relevant for social and historical sciences, such as persuasion simulation, disinformation modeling, the evolution of scientific concepts, entity recognition in historical texts, personality structure retrieval from word embeddings, systems to explore noisy historical data and biographical sources.

Finally, literary concerns remain central, with studies on stylistic transfer, automatic translation of poetry, techniques to quantify translators' interventions, authorship analysis, and of course new advances on LLMs' ability of dealing with full-length novels.

Overall, researchers in our field are trying to expand the reach of computational approaches in the Humanities while keeping their linguistic, cultural, and historical diversity. Whether working with under-resourced traditions, large contemporary corpora, or literary and social-scientific questions, the contributions in this volume sketch a field that is trying to use the neck-breaking evolution of linguistic technologies to illuminate at a new scale human expression in all its complexity.

Any success this workshop enjoys is first and foremost owed to its authors. Thank you for continuing this journey, or for placing your trust in us for the first time! We want to thank our reviewers, whose care and expertise sustain the quality of the discussion. Finally our gratitude goes to the outstanding program committee for their dedication and generosity throughout the process. Enjoy!

Yuri, Stefania, Anna, Janis, Diego, Stan

Program Committee

Chairs

Diego Alves, Saarland University
Yuri Bizzoni, Aarhus University
Stefania Degaetano-Ortlieb, Saarland University
Anna Kazantseva, National Research Council Canada
Janis Pagel, Department of Digital Humanities, University of Cologne
Stan Szpakowicz, EECS, University of Ottawa

Program Committee

Sergei Bagdasarov, Saarland University
Jinyeong Bak, Sungkyunkwan University
Johanna Binnewitt, Federal Institute for Vocational Education and Training
Patrick Brookshire, Academy of Sciences and Literature | Mainz
Paul Buitelaar, University of Galway
Miriam Butt, University of Konstanz
Pierluigi Cassotti, University of Gothenburg
Kent Chang, UC Berkeley
Stefano De Pascale, KU Leuven
Koel Dutta Chowdhury, Saarland Informatics Campus, Saarland University
Mark Finlayson, FIU
Francesca Frontini, Istituto di Linguistica Computazionale A. Zampolli" - ILC Consiglio Nazionale delle Ricerche - CNR
Svenja Guhr, University of California at Berkeley
Matteo Guida, The University of Melbourne
Hans Ole Hatzel, Universität Hamburg
Serge Heiden, ENS de Lyon
Rebecca Hicke, Cornell University
Azwad Anjum Islam, Florida International University
Mia Jacobsen, Aarhus University
Labiba Jahan, Southern Methodist University
Dimitrios Kokkinakis, University of Gothenburg
Stasinou Konstantopoulos, NCSR Demokritos
Maria Kunilovskaya, Saarland University
John Ladd, Washington & Jefferson College
Alie Lassche, Center for Humanities Computing, Aarhus University
Chaya Liebeskind, Jerusalem College of Technology , Lev Academic Center
Thomas Lippincott, Johns Hopkins University
Barbara McGillivray, King's College London
Caio Mello, Luxembourg Centre for Contemporary and Digital History
Craig Messner, Johns Hopkins University
David Mimno, Cornell University
Vivi Nastase, University of Geneva
Pierre Nugues, Lund University
Thijs Ossenkoppele, University of Amsterdam
Andrew Piper, McGill University
Petr Plechac, Institute of Czech Literature CAS

Thierry Poibeau, LATTICE (CNRS & ENS/PSL)
Jelena Prokic, Leiden University
Georg Rehm, DFKI
Nils Reiter, University of Cologne
Pablo Ruiz Fabo, LiLPa, Universit   de Strasbourg
Marijn Schraagen, Utrecht University
Artjoms Sela, Institute of Polish Language (PAN)
Hale Sirin, Johns Hopkins University
Pia Sommerauer, Vrije Universiteit Amsterdam
Julius Steuer, Universit  t des Saarlandes
Elke Teich, Universit  t des Saarlandes
Gaurish Thakkar, University of Zagreb
Ted Underwood, Univ of Illinois
Sowmya Vajjala, National Research Council
Freek Van De Velde, KU Leuven
Marieke Van Erp, KNAW Humanities Cluster
Menno Van Zaanen, South African Centre for Digital Language Resources
Francielle Vargas, University of S  o Paulo
Lorella Viola, Vrije Universiteit Amsterdam
Albin Zehe, University of Wuerzburg
Naitian Zhou, University of Michigan
Heike Zinsmeister, Universitaet Hamburg

Keynote Talk

Between Precision and Scale: Bridging Computational Methods and Humanistic Inquiry in Historical Semantics

Barbara McGillivray
King's College London

Abstract: How do we scale up the study of meaning change across centuries of texts while preserving the nuanced, culturally grounded interpretations that humanistic scholarship demands? This question sits at the heart of my research on Word Sense Disambiguation (WSD) and semantic change in historical languages, where the challenges are both computational and epistemological.

In this talk, I outline several principles that have emerged as central to my approach. First, systematic quantitative analyses of semantic phenomena require disambiguating polysemous words at scale, moving beyond qualitative observations to identify which specific senses are active in each context. Yet historical and low-resource languages resist standard techniques due to sparse data and shifting semantic boundaries. Second, expert annotation is not merely a preliminary step but the essential foundation that enables computational methods to capture culturally and temporally situated meanings. This human-in-the-loop approach ensures that automated WSD algorithms, once trained, can be deployed across large corpora while maintaining interpretive fidelity.

I illustrate these principles through work on nineteenth-century English, where rapid semantic shifts over decades reveal how industrialisation restructured entire semantic fields, and through Computational Corpus Annotation for Quantitative Analysis of Latin Lexical Semantics (COALA), an ERC-selected project that aims to trace gradual evolution across centuries of Latin textual tradition. These contrasting temporal scales demonstrate how integrating computational scalability with philological depth can transform our understanding of the relationship between linguistic and historical change, and offer a model for Digital Humanities research that neither privileges algorithmic automation nor retreats into purely qualitative analysis.

Bio:

Barbara McGillivray is Senior Lecturer in Digital and Computational Humanities in the Department of Digital Humanities of King's College, where she leads the Computational Humanities research group. She is Principal Investigator of the COALA project, successfully evaluated by the ERC. She is Editor-in-Chief of the Journal of Open Humanities Data and convenor of the MA programme in Digital Humanities at King's. Her research focusses on computational methods for the study of language change in both historical languages and contemporary data. As a Turing research fellow at the University of Cambridge and at The Alan Turing Institute she was also co-Investigator of the Living with machine project. Previously she worked as a language technologist in the Dictionaries division of Oxford University Press and as a data scientist in the Open Research Group of Springer Nature. Her most recent book is "Applying Language Technology in Humanities Research. Design, Application, and the Underlying Logic" (Palgrave Macmillan 2020).

Table of Contents

<i>From Corpus to Concept Scheme: Developing a SKOS Vocabulary for Armenian Epigraphic Heritage</i> Hamest Tamrazyan, Kamal Nour and Emanuela Boros	1
<i>Armenian AutoEpiDoc: Automated Extraction and Encoding of Armenian Inscriptions into EpiDoc TEI/XML</i> Hamest Tamrazyan, Emile Cornamusaz and Emanuela Boros	11
<i>Studying Expert-ese: Profiling and Classification of Domain-Specific Language Variation in Architecture with Traditional Machine Learning and LLMs</i> Carmen Schacht and Renate Delucchi Danhier	16
<i>CroCoSyn: A Cross-Lingual and Cross-Model Corpus of LLM-Generated Film Synopses</i> Louis Escouflaire	30
<i>Identity Without Action: Rethinking Collective Action Models in Disinformation Research</i> Lorella Viola	36
<i>Weakly Supervised Named Entity Recognition for Historical Texts</i> Marco Sorbi, Laurent Moccozet and Stephane Marchand-Maillet	48
<i>Invisible Speakers? Gender Disparity in German AI Discourse and Its Reflection in Language Models</i> Milena Belosevic	66
<i>GlobLingDiv: A global dataset linking linguistic diversity and digital support to reveal landscapes with under-resourced languages for NLP</i> Katharina Zeh, Hannes Essfors, Juliane Benson, Lale Tüver, Andreas Baumann and Hannes A. Fellner	80
<i>LLMs Got Rhyme? Hybrid Phonological Filtering for Greek Poetry Rhyme Detection and Generation</i> Sergios Chatzikyriakidis and Anastasia Natsina	87
<i>Style as Signature: Profile-Based Authorship Verification of Mihai Eminescu’s Journalistic Corpus</i> Ioana-Roxana Boriceanu and Liviu Dinu	102
<i>Measuring Social Integration Through Participation: Categorizing Organizations and Leisure Activities in the Displaced Karelians Interview Archive using LLMs</i> Joonatan Laato, Veera Schroderus, Jenna Kanerva, Jenni Kauppi, Virpi Lummaa and Filip Ginter	111
<i>Catalogues as Data: Interpretable NLP Pipelines for Ottoman-Turkish Bibliographies</i> Mark Hill, Ayse Bulus and Paul Spence	128
<i>Stylistic Transfer from Annotator Communities to Large Language Models</i> Jay Chooi	135
<i>Modeling Changing Scientific Concepts with Complex Networks: A Case Study on the Chemical Revolution</i> Sofia Aguilar Valdez and Stefania Degaetano-Ortlieb	146
<i>Speaking on Their Behalf: Detecting Indirect Speech in Historical Danish and Norwegian Texts</i> Ali Al-Laith, Alexander Conroy, Kirstine Degn, Jens Bjerring-Hansen and Daniel Hershovich	157

<i>Harder than Finding the Lost Sheep? Towards Automatically Suggesting Deliberate Metaphor Annotations in German Sermons</i>	
Ronja Laarmann-Quante and Stefanie Dipper	164
<i>Semantic Factor Analysis: Validating Personality Structure Recovery from empirically-mediated Word Embeddings</i>	
Oliver Müller	176
<i>Quantitative Analysis of Rhyme and Metre in LLM-generated Translations of Poetry</i>	
Jan-Felix Klumpp	189
<i>WikiLingDiv: a dataset for quantifying digital linguistic diversity using Wikipedia page views</i>	
Hannes Essfors and Andreas Baumann	202
<i>Modeling Linguistic Imprints of War Propaganda in a Russian Wikipedia Fork: A Comparative Analysis with the Original Wikipedia</i>	
Anastasiia Vestel and Stefania Degaetano-Ortlieb	212
<i>Stylometric Approach to AI-generated Texts. An Analysis of Contemporary French-Language Literature</i>	
Adam Pawłowski and Tomasz Walkowiak	221
<i>Degree Zero of Translation: Using Interlinear Baselines to Quantify Translator Intervention</i>	
Maciej Rapacz and Aleksander Smywiński-Pohl	227
<i>How to Efficiently Explore Noisy Historical Data? Leveraging Corpus Pre-Targeting to Enhance Graph-based RAG</i>	
Donghan Bian, Marie Puren and Florian Cafiero	241
<i>Detecting reported speech as a token classification task: an application to Classical Latin?</i>	
Agustin Dei	251
<i>Narrative in Short German Prose: A Multi-Phenomenon Dataset for Computational Literary Analysis</i>	
Hans Ole Hatzel, Haimo Stiemer, Evelyn Gius and Chris Biemann	257
<i>Sense-Based Annotation of Geographical Nouns in Ancient Greek and Latin: A Diachronic Study with LLMs</i>	
Andrea Farina, Michele Ciletti, Barbara McGillivray and Andrea Ballatore	266
<i>Evaluating Humanities Theory Alignment in Large Language Models: Incremental Prompting and Statistical Assessment</i>	
Axel Pichler and Janis Pagel	280
<i>Too Long, Didn't Model: Decomposing LLM Long Context Understanding With Novels</i>	
Sil Hamilton, Rebecca Hicke, Mia Ferrante, Matthew Wilkens and David Mimno	295
<i>AI Corpus Linguist: More than a Year of Experience</i>	
Jiří Milička and Tomáš Machálek	305
<i>Generative Information Extraction from Biographical Sources</i>	
Robin Winkle, Manfred Stede and Jörn Kreutel	311
<i>WikiFirst: A Genre-Fixed, Content-controlled Corpus for Evaluating Content Effects in Authorship Analysis</i>	
Dung Nguyen, G. Çağatay Sat, Evgeny Pyshkin and John Blake	323
<i>Measuring the Symbolic Power of Languages with LLM-based Multilingual Persuasion Simulation</i>	
Yin Jou Huang and Fei Cheng	328

Program

Saturday, March 28, 2026

14:00 - 14:05 *Welcome*

14:05 - 14:35 *Linguistic Diversity & Resources*

GlobLingDiv: A global dataset linking linguistic diversity and digital support to reveal landscapes with under-resourced languages for NLP

Katharina Zeh, Hannes Essfors, Juliane Benson, Lale Tüver, Andreas Baumann and Hannes A. Fellner

14:35 - 15:30 *Poster Teasers*

WikiLingDiv: a dataset for quantifying digital linguistic diversity using Wikipedia page views

Hannes Essfors and Andreas Baumann

Stylistic Transfer from Annotator Communities to Large Language Models

Jay Chooi

LLMs Got Rhyme? Hybrid Phonological Filtering for Greek Poetry Rhyme Detection and Generation

Stergios Chatzikyriakidis and Anastasia Natsina

Invisible Speakers? Gender Disparity in German AI Discourse and Its Reflection in Language Models

Milena Belosevic

How to Efficiently Explore Noisy Historical Data? Leveraging Corpus Pre-Targeting to Enhance Graph-based RAG

Donghan Bian, Marie Puren and Florian Cafiero

From Corpus to Concept Scheme: Developing a SKOS Vocabulary for Armenian Epigraphic Heritage

Hamest Tamrazyan, Kamal Nour and Emanuela Boros

Armenian AutoEpiDoc: Automated Extraction and Encoding of Armenian Inscriptions into EpiDoc TEI/XML

Hamest Tamrazyan, Emile Cornamusaz and Emanuela Boros

Detecting reported speech as a token classification task: an application to Classical Latin?

Agustin Dei

Saturday, March 28, 2026 (continued)

*CroCoSyn: A Cross-Lingual and Cross-Model Corpus of LLM-Generated Film
Synopsis*

Louis Escouflaire

15:30 - 16:15 *Poster Session (GatherTown)*

16:15 - 16:30 *Closing*

Sunday, March 29, 2026

09:00 - 10:30 *Historical, Diachronic & Cultural Language Data*

Weakly Supervised Named Entity Recognition for Historical Texts

Marco Sorbi, Laurent Moccozet and Stephane Marchand-Maillet

Sense-Based Annotation of Geographical Nouns in Ancient Greek and Latin: A Diachronic Study with LLMs

Andrea Farina, Michele Ciletti, Barbara McGillivray and Andrea Ballatore

Measuring Social Integration Through Participation: Categorizing Organizations and Leisure Activities in the Displaced Karelians Interview Archive using LLMs

Joonatan Laato, Veera Schroderus, Jenna Kanerva, Jenni Kauppi, Virpi Lummaa and Filip Ginter

10:30 - 11:00 *Coffee Break*

11:00 - 12:00 *Invited Talk by Barbara McGillivray: 'LLMs, diachrony, humanities theory, and methodological perspectives'*

12:00 - 12:30 *Identity*

Identity Without Action: Rethinking Collective Action Models in Disinformation Research

Lorella Viola

12:30 - 14:00 *Lunch*

14:00 - 15:00 *Information Extraction & Theory Alignment*

Generative Information Extraction from Biographical Sources

Robin Winkle, Manfred Stede and Jörn Kreutel

Evaluating Humanities Theory Alignment in Large Language Models: Incremental Prompting and Statistical Assessment

Axel Pichler and Janis Pagel

15:00 - 15:30 *Coffee Break*

Sunday, March 29, 2026 (continued)

15:30 - 16:30 *Poster session*

16:30 - 17:30 *Literary Language*

Narrative in Short German Prose: A Multi-Phenomenon Dataset for Computational Literary Analysis

Hans Ole Hatzel, Haimo Stiemer, Evelyn Gius and Chris Biemann

Quantitative Analysis of Rhyme and Metre in LLM-generated Translations of Poetry

Jan-Felix Klumpp

17:30 - 17:45 *Closing Session & SIGHUM Business Meeting*

From Corpus to Concept Scheme: Towards Developing a SKOS Vocabulary for Armenian Epigraphic Heritage

Hamest Tamrazyan

DHLAB, EPFL, Lausanne, Switzerland
hamest.tamrazyan@epfl.ch

Kamal Nour

EPFL, Lausanne, Switzerland
kamal.nour@epfl.ch

Emanuela Boros

DHLAB, EPFL, Lausanne, Switzerland

Abstract

Armenian epigraphy, one of the world's oldest and most diverse inscriptional traditions, remains largely absent from digital research infrastructures due to a lack of basic linguistic and conceptual resources. No machine-readable corpus, standardized terminology, or controlled vocabulary exists for describing Armenian inscription types, preventing indexing and interoperability. This paper addresses this gap by constructing the first dataset of Armenian inscription-type terminology and by developing a computational pipeline for analyzing it at scale. We digitize and preprocess a broad corpus of authoritative printed publications; curate a culturally grounded terminology list; and train transformer-based NER models to identify both attested inscription types and potential terminological variants across unseen texts. The resulting resources form the first empirical foundation for modelling Armenian epigraphic concepts needed for further developing a SKOS vocabulary aligned with, yet culturally distinct from, existing international epigraphic ontologies.

1 Introduction

In the last decade, digital technologies have transformed the way cultural heritage is preserved, accessed, and studied. As inscriptions, manuscripts, monuments, and other heritage artefacts increasingly enter digital ecosystems, the question of how cultural knowledge is organized has become a central concern for the digital humanities (DH) (Yan et al., 2020; Bianchini, 2023). Classification systems such as vocabularies, ontologies, and semantic models, play a decisive role in shaping what becomes visible, searchable, interoperable, or even thinkable within digital infrastructures (Liu et al., 2023). However, the act of classification is far from being neutral and determines *which cultural traditions are accurately represented, which are simplified or misread, and which remain invisible*

altogether. A direct consequence of this shift toward digital heritage is that traditions lacking structured, machine-readable vocabularies cannot be effectively indexed, searched, or linked within modern research infrastructures. This is the case for Armenian epigraphy: *despite the depth and historical significance of its corpus, the field remains only minimally integrated into existing digital standards*.

Thus, the need for a dedicated SKOS vocabulary¹ for Armenian epigraphy arises from the fact that, despite their historical significance (Figure 1), Armenian inscriptions remain largely absent from the semantic infrastructures used in digital humanities (Greenwood, 2004). This gap reflects structural barriers that limit the visibility and preservation of Armenian epigraphic heritage in digital environments (Tamrazyan et al., 2026; Tamrazyan and Hovhannisyan, 2024a,b).

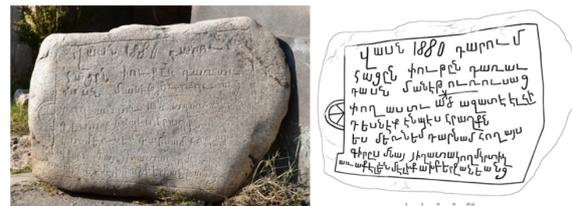


Figure 1: Inscription of Vahravar (Figure 5a, b in (Harutyunyan, 2023a)).

A first challenge is the **lack of standardized, machine-readable descriptive terminology** in Armenian epigraphy, which prevents inscriptions from being systematically indexed, linked, or compared across databases. In the absence of controlled vocabularies, Armenian material cannot be effectively integrated into international corpora, limiting its visibility for researchers working in epigraphy, archaeology, linguistics, and cultural heritage studies.

¹<https://www.w3.org/2004/02/skos/>

Second, the **structural bias of existing epigraphic ontologies toward Greco-Roman traditions** prevents Armenian inscriptions from being accurately modelled (Espinosa Espinosa and Velázquez Soriano, 2021). Ontologies such as EAGLE² and FAIR Epigraphy³ rely on typologies and conceptual hierarchies that do not reflect the cultural logic or material diversity of Armenian monuments. This misalignment leads to inappropriate categorizations and forces Armenian material into external frameworks that lack cultural specificity.

Third, the highly dispersed nature of Armenian heritage further increases **the need for interoperable digital infrastructures**. With more than 80% of Armenian monuments located outside the Republic of Armenia (Tigranyan, 2023), often in regions facing conflict, neglect, or political pressure, the ability to represent inscriptions in a consistent, structured, and internationally accessible format becomes crucial for cultural preservation and scholarly continuity.

Fourth, **the absence of interoperable vocabularies** has broader implications for computational research (Tamrazyan and Hovhannisyan, 2025). Without a standardized terminology layer, the application of natural language processing, machine learning, semantic web technologies, or automated annotation workflows becomes severely limited. Establishing a SKOS vocabulary therefore functions not only as a heritage documentation effort but as an enabling infrastructure for future digital research in Armenian studies.

Finally, there is a growing recognition within the DH of the **importance of supporting underrepresented cultural traditions** and challenging the implicit Eurocentrism embedded in many digital heritage standards. Developing Armenian-specific vocabularies contributes to this broader movement by ensuring that local epistemologies and scholarly traditions are formally encoded rather than marginalized.

This article contributes to this broader effort by offering the first corpus-driven analysis of inscription-type terminology in Armenian epigraphy and by identifying the requirements for developing a culturally grounded SKOS vocabulary. Rather than importing externally defined categories, the study derives its insights from distinc-

tions that emerge within Armenian inscriptional practice and considers how these may be aligned with international standards. The data and code underlying this study are openly available⁴.

Specifically, this work contributes with:

- **A curated terminology dataset** derived from foundational Armenian epigraphic publications and produced through a dedicated OCR and text-processing workflow.
- **A terminology-aware NER model** trained to identify and classify inscription-type expressions across Armenian-language scholarship⁵.
- **A corpus-based foundation for conceptual modelling** that identifies key terminological patterns and conceptual distinctions required for further constructing a SKOS vocabulary of Armenian inscription types.

2 Data Normalization

To examine how inscription-type terminology functions across Armenian epigraphic scholarship, we develop a computational workflow designed to digitize the corpus, extract terminology, and analyze its distribution at scale. The workflow proceeds through several stages: assembling a representative set of printed publications, digitizing them into machine-readable text, preprocessing the OCR output into a clean sentence-level corpus, constructing a ground-truth NER dataset using an expert terminology list, and training transformer-based NER models to evaluate both in-domain performance and generalization to unseen books.

2.1 Adoption of the EAGLE Model

As a conceptual starting point, we adopt the seven-domain organizational structure of the EAGLE vocabularies, one of the most widely used international models for describing inscriptions in digital epigraphy. These domains—*Material*, *Execution Technique*, *Type of Inscription*, *Object Type*, *Decoration*, *Dating Criteria*, and *State of Preservation*—provide a shared descriptive template that

²<https://www.eagle-network.eu/resources/vocabularies/>

³https://ontology.inscriptiones.org/type_of_inscription

⁴<https://github.com/dhlab-epfl/dhlab-epigraphy-studies>

⁵The pipeline is designed to be transferable to other under-resourced epigraphic traditions, with adaptation mainly required for OCR, sentence segmentation, and expert terminology curation.

Book	Abbrev.	Citation
Divan Hay Vimagrut ⁷ yan 1	DHV-1	Orbeli (1965b)
Divan Hay Vimagrut ⁷ yan 2	DHV-2	Orbeli (1960)
Divan Hay Vimagrut ⁷ yan 3	DHV-3	Barkhudaryan (1967b)
Divan Hay Vimagrut ⁷ yan 4	DHV-4	Barkhudaryan (1973b)
Divan Hay Vimagrut ⁷ yan 5	DHV-5	Barkhudaryan (1982b)
Divan Hay Vimagrut ⁷ yan 6	DHV-6	Avagyan and Janpoladyan (1977b)
Divan Hay Vimagrut ⁷ yan 7	DHV-7	Grigoryan (1996)
Divan Hay Vimagrut ⁷ yan 8	DHV-8	Grigoryan (1999)
Divan Hay Vimagrut ⁷ yan 9	DHV-9	Barkhudaryan et al. (2012b)
Divan Hay Vimagrut ⁷ yan 10	DHV-10	Sargsyan et al. (2017b)
Historical Monuments of Akunk and Katnaghbyur	KAR-2014	Karapetyan (2014)
Khodjavank Monastery	KAR-2024a	Karapetyan (2024)
Northern Artsakh	KAR-2021a	Karapetyan (2021c)
Lapidary Inscriptions of Bun Aghvank	KAR-2021b	Karapetyan (2021b)
Armenian Collection of the Caucasian Museum	KAR-2004	Karapetyan (2004)
Previously Unpublished Inscriptions of Yerevan Katoghike	HAR-2019a	Harutyunyan (2019a)
Epigraphic Heritage of Armavir Province	HAR-2017	Harutyunyan (2017)
Newly Discovered Tombstones of Holy Ejmiatsin	HAR-2021	Harutyunyan and Melkonyan (2021)
Settlement of Noragavit and St. Gevorg Church	HAR-2019b	Harutyunyan (2019b)
Inscribed Artefacts of the N. Adonts Museum of Sisian	HAR-2019c	Harutyunyan (2019b)
Epigraphic Heritage of Tatev Hermitage	HAR-2023	Harutyunyan (2023b)
Natural Disasters and Celestial Phenomena in Inscriptions	HAR-2022/2023	Harutyunyan (2022, 2023a)
Melik Mansions of Artsakh and Syunik	GHU-2001	Ghulyan (2001)
Inscriptions of Armenian Settlements in India	KOR-2024	Kortoshyan (2024)
Inscriptions of Aleppo	KOR-2013	Kortoshyan (2013a)
Tsaghatskar Monastery: Historical and Archaeological Study	MEL-2024	Melkonyan (2024)
Birthplaces of Genocide Survivors in Lebanese Funerary Inscriptions	TAS-2018	(Tashjian, 2018)
Georgian State Policy and Armenian Cultural Monuments (1988-1998)	KAR-1998	(Karapetyan, 1998)
Reconstruction of Dadivank	AYV-2011	(Ayvazyan, 2011)

Table 1: Publications included in the corpus, with abbreviations used throughout the paper. Light-gray rows indicate the volumes used to construct the full training dataset for the main experiments. The earliest baseline model was trained solely on **DHV-10**. The remaining publications (non-highlighted rows) were held out as unseen material for evaluating the model’s ability to generalize and to detect potential new inscription-type terminology.

would allow Armenian inscriptions to be positioned within the existing epigraphic infrastructures.

For the present study, we focus specifically on the *Type of Inscription* domain that captures the functional category of an inscription (e.g., “commemorative”, “funerary”, “votive”, “legal”, “donor”, “construction-related”) and interacts closely with linguistic, material, and historical characteristics. To operationalize this conceptual layer for Armenian epigraphy, we first required a stable and culturally grounded vocabulary of inscription-type terminology. For this, a team of domain specialists compiled an initial inventory of terms that represent inscription types. They also used a large language model (LLM) as an assistive tool (GPT-5) to expand this inventory by generating additional variants drawn from the digitized corpus; however, because these automatically proposed terms varied in accuracy, each term candidate was manually reviewed by experts, who retained only those terms that were terminologically sound, historically attested, and conceptually dis-

tinct.

The resulting curated list comprises **41 Armenian inscription-type multiword terms**⁶. These include general descriptors such as *վիճագիր* and *վիճակալս արձանագրութիւն*; functional classes such as funerary (*տապալնագիր*), donor (*սվիրագիր*), commemorative (*հուշագիր*), legal (*կանոնագիր*), and construction-related inscriptions (*շինարարական արձանագրութիւն*); architectural subtypes specifying the placement of inscriptions on a monument, including *գավիթային*, *սյունային*, *բարավորի*, and *որմնագիր*; administrative and socio-economic categories such as tax-exemption inscriptions (*ապահարկման արձանագրութիւն*) and water-management inscriptions (*ջրօգտագործման արձանագրութիւն*); and rare or specialized types, including cryptographic inscriptions (*շրջագիր / ծածկագիր*), royal and princely inscriptions, monastic inscriptions, and manuscript-associated

⁶A term can be composed of only one term e.g. *վիճագիր* or multiple words (multiword) e.g. *վիճակալս արձանագրութիւն*

colophonic writings. We refer to this set of terms as our **expert-defined inscription type reference list** and will serve us in creating an annotated dataset for detecting new and undocumented inscription-type terms.

2.2 Corpus Construction

Because no digital corpus of Armenian epigraphy currently exists, we construct our dataset manually from authoritative printed publications. The complete set of works forming the corpus is listed in Table 1. These sources include multi-volume corpora, regional surveys, architectural studies, thematic monographs, and peer-reviewed articles, and together represent the most comprehensive and methodologically rigorous body of scholarship available for Armenian inscriptions.

Our primary source is the monumental *Divan Hay Vimagrut'yan* (DHV) series, published by the National Academy of Sciences of Armenia since the 1960s (Orbeli, 1965a,a; Barkhudaryan, 1967a, 1973a, 1982a; Barkhudaryan et al., 2012a; Avagyan and Janpoladyan, 1977a; Barkhudaryan, 1999; Sargsyan et al., 2017a). The DHV volumes remain the most detailed and methodologically consistent corpus of Armenian inscriptions, offering reliable transcriptions, functional classifications, and palaeographic descriptions. They therefore serve as the backbone of our terminology inventory.

To avoid a DHV-centric vocabulary and to reflect the broader diversity of Armenian epigraphic scholarship, we also incorporate terminology from regional corpora and thematic studies authored by leading specialists such as Karapetyan, Kortoshyan, Petrosyan, Barkhudaryan, and Harutyunyan (Karapetyan, 2021a; Kortoshyan, 2013b; Karapetyan, 1998, 2004; Kortoshyan, 2013a, 2024; Petrosyan, 2008; Harutyunyan, 2017). These works capture terminological variation across regions, generations, and disciplinary perspectives, and contribute essential conceptual depth to the corpus. The publications in Table 1 were selected according to four criteria designed to ensure representativeness, conceptual precision, and cultural fidelity.

Authorship Diversity. We include works authored across different generations, institutions, and scholarly traditions. This diversity reduces reliance on any single terminological style and captures diachronic variation in Armenian epigraphic

discourse.

Genre Balance. To reflect both stable and emerging terminology, we intentionally balance different scholarly genres. Books and monographs provide detailed historical, architectural, and linguistic analyses; non-digital corpora of inscriptions, especially the DHV volumes, offer authoritative transcriptions and typological classifications; and peer-reviewed articles and conference proceedings capture recent conceptual innovations and methodological developments.

Regional Coverage. Because more than 80% of Armenian cultural heritage lies outside the Republic of Armenia, we include publications documenting inscriptions from Armenia, Artsakh, Georgia, Turkey, Iran, India, Ukraine, Moldova, and other historically Armenian regions.

Temporal Coverage. The corpus spans publications from the 1960s to 2024, enabling us to incorporate both foundational terminology and more recent shifts in scholarly interpretation.

2.3 Digitization

After assembling the corpus, all publications were digitized and converted into machine-readable text using an OCR pipeline for printed Armenian. Although most content was preserved, the OCR output introduced structural fragmentation, punctuation errors, and inconsistent line breaks, necessitating substantial preprocessing before terminology extraction and NER annotation.⁷

2.4 Preprocessing

We, thus, build a dedicated preprocessing pipeline to restore the structural and linguistic coherence of the corpus. The raw OCR output was highly fragmented: paragraphs were split into short line segments, Armenian punctuation, especially the full stop (:), was frequently misrecognized, and the hierarchical organization of the printed editions was largely lost. These issues made sentence segmentation and downstream terminology extraction unreliable without substantial normalization.

To address this, we designed a three-stage preprocessing workflow. (1) We segmented the OCR output by page in order to preserve the structural

⁷PDFs were converted to page-level images using `pdf2image` and processed with the Calfa `hye-tesseract` OCR engine (<https://github.com/calfa-co/hye-tesseract>), which supports Mesropian and Grabar orthographies.

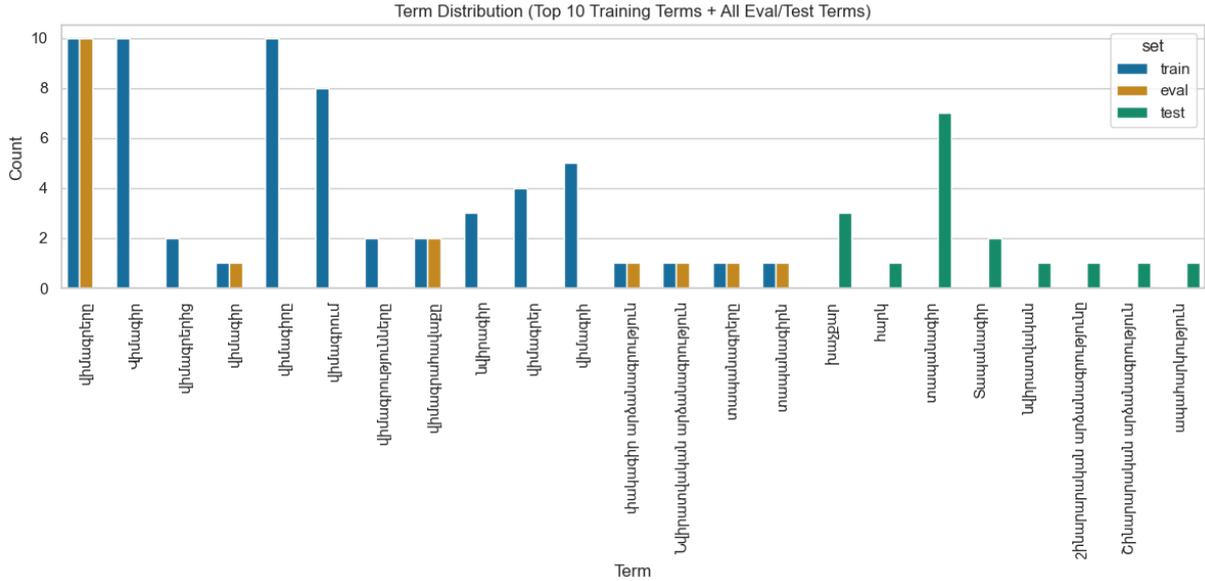


Figure 2: Distribution of terms across splits for DHV-10 data.

boundaries of the original publications. (2) Within each page, we regrouped the fragmented lines into coherent paragraphs by leveraging recurrent layout patterns characteristic of Armenian epigraphic editions that we noticed in DHV: inscription headers, line-count descriptors, uppercase inscription blocks, and commentary sections commonly introduced by the marker ծաւնթ. (“note”). This stage also included systematic correction of punctuation and the merging of artificially split sentences. (3) We segmented the reconstructed paragraphs into sentences using a custom Armenian sentence splitter, which is adapted to the Armenian full-stop “:”.

3 Experiments

With the sentence-level corpus prepared, the next step is to identify and annotate occurrences of inscription-type terminology using our expert-defined inscription type reference list. These annotations will form the supervised dataset required for training and evaluating the NER models.

Because Armenian inscription terminology is morphologically rich and appears in highly variable scholarly contexts, simple string matching is insufficient. We therefore combine rule-based detection with GPT-5’s assistance⁸. The prompt guides term extraction by focusing on predefined terms, recording each occurrence separately with

⁸Specifically, we applied GPT-5 over predefined text chunks together with the reference term list (Appendix A). GPT-5 is used for assisted annotation, while extraction is performed with reproducible fine-tuned NER models.

surrounding context for clarity. Metadata like file name and page number ensure traceability. We organized the extracted data hierarchically, with occurrences grouped under terms and terms compiled into an extraction result. The prompt emphasizes Armenian language handling to account for inflectional and orthographic variations, without relying on explicit morphological rules, ensuring robust zero-shot extraction.

Each sentence is then tokenized and labelled using the standard inside–outside–beginning (IOB) tagging scheme⁹, which marks the boundaries of multiword terms. Because Armenian exhibits rich inflection, flexible word order, and modifier-heavy expressions, the automatic tagging often produced misaligned spans. Therefore, we did a manual verification and targeted rule-based adjustments to ensure consistent token-level annotation and transformed it in the IOB format. These rules are especially adapted to Armenian, where rich inflectional morphology and the frequent use of suffixes can cause surface forms to deviate from their base forms.

We then partitioned the data into training, development, and test sets. To evaluate true generalization beyond the expert-defined inscription type reference list, the test set contains only terms absent from the training and development sets. This design prevents memorization effects and allows us to measure the model’s ability to recognize unseen

⁹[https://en.wikipedia.org/wiki/Inside-outside-beginning_\(tagging\)](https://en.wikipedia.org/wiki/Inside-outside-beginning_(tagging))

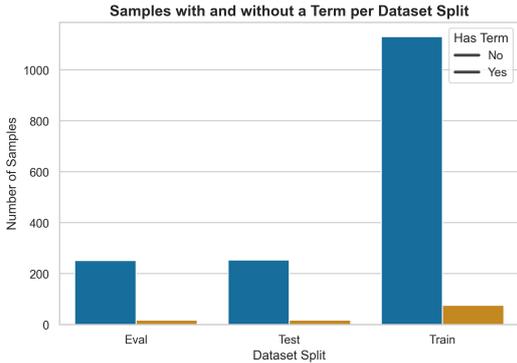


Figure 3: Distribution of sentences across dataset splits for DHV-10 data.

terminology, while the training and development splits include overlapping term types to support stable learning.

Figures 2 and 3 summarize the distribution of sentences and terms across the splits, showing a consistent 85/15 ratio of sentences with and without annotated entities and confirming the strict separation of term types between the test set and the remaining subsets. The figures also highlight the strong class imbalance inherent to the corpus, a characteristic of epigraphic scholarship that makes the task particularly challenging for NER models, which must learn to recognize sparse and morphologically variable terminology in predominantly non-annotated text.

3.1 Results

As our main baseline system, we chose **spaCy**, a neural NLP toolkit optimized for industrial use. We extend its pipeline with a custom NER component configured to recognize inscription-type entities and train it on the DHV-10 split using stochastic gradient descent with shuffled batches, periodic evaluation, and early stopping based on validation loss. For comparison, we also train three transformer-based models under identical conditions: Multilingual BERT (mBERT)¹⁰, XLM-RoBERTa¹¹, and HyeBERT¹², an Armenian-specific transformer.

As shown in Table 3, the spaCy baseline achieves moderate precision (0.50) but extremely low recall (0.06), indicating that while its few predictions are often correct, it fails to identify the vast

¹⁰<https://huggingface.co/google-bert/bert-base-multilingual-cased>

¹¹<https://huggingface.co/FacebookAI/xlm-roberta-base>

¹²<https://huggingface.co/aking11/hyebert>

majority of inscription-related entities. HyeBERT, despite being tailored to Armenian, performs similarly poorly, suggesting that limited pretraining data and a smaller architecture restrict its ability to model the highly variable descriptive language of epigraphic scholarship.

More unexpectedly, HyeBERT, despite being specifically pretrained for Armenian, exhibits similarly poor performance, with an F1 score of 0.09 and recall of only 0.07. This result suggests that language-specific pretraining alone is insufficient for this task, likely due to limited pretraining data (Armenian subset from OSCAR dataset¹³), and possibly a mismatch between HyeBERT’s training domain and the highly variable, descriptive language of epigraphic scholarship.

By contrast, the multilingual transformer models (mBERT and XLM-RoBERTa) substantially outperform both spaCy and HyeBERT, reaching F1 scores of 0.52 with recall values more than six times higher than spaCy. We attribute this performance gap to the richer contextual representations learned by multilingual transformers, which better capture Armenian’s inflectional complexity and flexible word order. In addition, cross-lingual subword sharing appears beneficial for sparse, domain-specific terminology, allowing multilingual models to generalize more effectively across diverse epigraphic contexts despite limited task-specific training data.

3.2 Scaling to the Complete Corpus

The baseline experiments on DHV-10 demonstrate that multilingual transformer models offer a clear advantage over lightweight architectures, but they also reveal an important limitation: the restricted size and lexical diversity of the initial dataset prevent the models from fully learning the range of inscription-type terminology present in Armenian epigraphic scholarship. To develop a model capable of generalizing beyond a narrow subset of terms and adapting to highly variable scholarly contexts, we decided to train on a substantially broader corpus.

To evaluate model performance under more realistic and diverse conditions, we expand the dataset to include sentence-level annotations from twenty volumes of the corpus (Table 1). This extension increases the scale of the experiment by more than an order of magnitude: the baseline setup

¹³<https://oscar-project.org/>

Split	# Sentences	# Sentences w/ Term	# Unique Terms	# Overlap w/ Train
Train	1,208	76	27	-
Development	269	17	7	7
Test	271	17	8	0

Table 2: Dataset statistics for the NER training, development, and test splits, drawn from **DHV-10**. The test set contains only unseen terms.

Model Name	Precision	Recall	F1
spaCy	0.50	0.06	0.11
mBERT	0.78	0.39	0.52
XLM-RoBERTa	0.78	0.39	0.52
HyeBERT	0.34	0.07	0.09

Table 3: Baseline performance of spaCy and transformer models on the **DHV-10** subset.

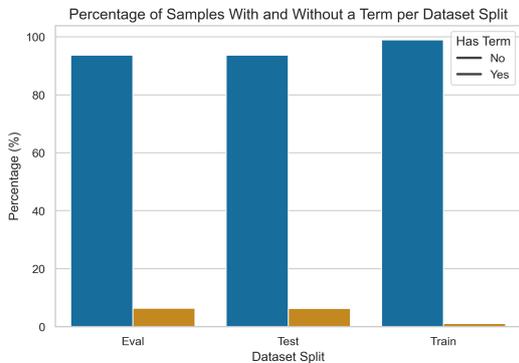


Figure 4: Distribution of sentences across dataset splits for complete data in percentages.

contained roughly 1,200 sentences, while the full dataset exceeds 56,000. The enlarged corpus introduces both a wider range of positive examples (sentences containing inscription terminology) and a substantially larger proportion of negative sentences, thereby mirroring the natural sparsity of terminology in scholarly prose. Crucially, the development and test sets remain unchanged, ensuring that all results are directly comparable across experimental settings.

Figure 4 shows the distribution of sentences across the training, development, and test splits for the complete dataset. Table 4 summarizes the key statistics of the expanded corpus.

We train the same set of models (spaCy, HyeBERT, mBERT, and XLM-RoBERTa) on this enlarged dataset. The results are presented in Table 5. We notice that, while maintaining the same trend, using a larger amount of data leads to notable improvements in transformer performance. The multilingual BERT model achieves the high-

est F1 score (0.63), with precision rising to 0.82 and recall improving to 0.50. XLM-RoBERTa also benefits from the expanded data, though to a lesser degree. These gains suggest that access to a richer and more heterogeneous set of contexts significantly enhances the models’ ability to recognize inscription terminology, especially given that such terms occur in only about one percent of sentences in the corpus. The superior performance of mBERT may reflect the alignment between its subword vocabulary and the morphological patterns of Armenian, allowing it to identify relevant terms even under highly imbalanced conditions.

Overall, these results confirm that multilingual transformer models not only outperform smaller architectures but also scale effectively with larger and more realistic training data. This motivates their use as the primary models for terminology discovery in the unseen-books experiments that follow.

4 Generalization to Unseen Books and Term Discovery

To assess generalization beyond the training corpus, we applied the best-performing NER model to a set of books excluded from all training, development, and benchmark annotation stages. The volumes were processed using the same preprocessing and sentence-segmentation pipeline, and the model was used to infer inscription-type terminology from previously unseen texts.

Figure 5 illustrates the relationship between three term sets: the *expert reference vocabulary* (54 terms), the *benchmark terms* extracted during annotation (136 terms), and the *inferred terms* predicted by the model (98 terms).

The overlaps reveal two main behaviors. First, the model reliably recognizes known inscription-type terminology in new contexts: 40 inferred terms overlap with the benchmark set, indicating robust contextual generalization across grammatical realizations and textual environments. Second, the model shows limited capacity for independent

Split	# Sentences	# Sentences w/ Term	# Unique Terms	# Overlap w/ Train
Train	51,223	530	136	-
Development	269	17	7	7
Test	271	17	8	0

Table 4: Statistics of the complete dataset.

Model Name	Precision	Recall	F1
spaCy	0.50	0.06	0.11
mBERT	0.82	0.50	0.63
XLM-RoBERTa	0.67	0.44	0.53
HyeBERT	0.50	0.06	0.10

Table 5: Performance of all models on the complete dataset.

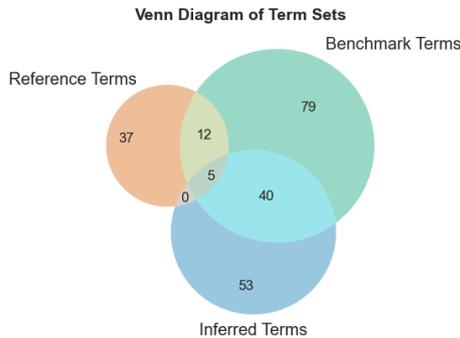


Figure 5: Venn diagram comparing reference terminology (54 terms), benchmark new terms (136 terms), and terms inferred by the model on unseen books (98 terms).

term discovery. None of the inferred terms overlap exclusively with the expert reference list, and only five terms are shared across all three sets. Furthermore, 53 inferred terms are unique to the model output but correspond to morphological variants, partial matches, or contextually ambiguous expressions rather than genuinely new inscription types.

Overall, these results indicate that while transformer-based NER models generalize well to unseen texts, they do not autonomously expand the conceptual terminology. Their primary value lies in supporting corpus exploration and expert analysis by surfacing candidate expressions and contextual variation rather than replacing expert-driven terminology formation.

5 Conclusions and Future Work

This study presented a first systematic, data-driven analysis of inscription-type terminology in Armenian epigraphic scholarship and established the

foundational resources needed for developing a culturally specific SKOS vocabulary. Through the digitization of authoritative publications, the creation of a curated terminology list, and the training of transformer-based NER models, we produced the first empirical infrastructure for identifying and examining Armenian inscription-type concepts at scale.

Our results show that automatic terminology extraction alone is insufficient for this domain. While multilingual transformer models successfully generalize to unseen texts and reliably detect attested terms, they rarely identify genuinely new concepts; most model-generated candidates are morphological variants or contextually ambiguous expressions. This highlights the necessity of expert involvement, as Armenian inscription-type categorization is deeply embedded in cultural, architectural, and functional knowledge that cannot be inferred from surface patterns alone. At the same time, the corpus reveals both categories familiar from international ontologies (such as funerary, donor, or legal inscriptions) and types highly specific to Armenian tradition (such as գավիթային, բարավորի, or ջրօգտագործման արձանագրություն), many of which are absent from EAGLE and FAIR Epigraphy vocabularies.

Eventually, we would want to continue to focus on moving from terminology extraction toward concept modelling. This involves validating terminology through contextual and distributional evidence; identifying synonymy, variation, and fine-grained distinctions; constructing hierarchical relations and selecting preferred or alternative labels; detecting inscription-type concepts expressed only implicitly through descriptive context; and aligning Armenian-specific categories with international ontologies while preserving cultural specificity. These steps will advance the project from surface-level term identification toward a structured conceptual model grounded in Armenian scholarly practice and will ultimately support the development of the first Armenian SKOS vocabulary for inscription types, enhanc-

ing the interoperability and visibility of Armenian epigraphic heritage within global digital infrastructures.

References

- S. Avagyan and H. Janpoladyan. 1977a. *Corpus Inscriptionum Armeniacarum*, volume 6.
- Suren Avagyan and Hovhannes Janpoladyan. 1977b. *Divan Hay Vimagrutyun, Prak VI*. Academy of Sciences of the Armenian SSR, Yerevan.
- Ashot Ayvazyan. 2011. *Dadivanki Verakangnumy (1997–2011 tt.)*. HChU, Yerevan.
- M. Barkhudaryan. 1999. *Aghvants Yerkir yev Dratsik: Artsakh (The Land of Aghvank and Its Neighbors: Artsakh)*. Gandzasar, Yerevan.
- S. Barkhudaryan. 1967a. *Corpus Inscriptionum Armeniacarum*, volume 3.
- S. Barkhudaryan. 1973a. *Corpus Inscriptionum Armeniacarum*, volume 4.
- S. Barkhudaryan. 1982a. *Corpus Inscriptionum Armeniacarum*, volume 5.
- S. G. Barkhudaryan, K. Ghafadaryan, and S. T. Saghumyan. 2012a. *Corpus Inscriptionum Armeniacarum*, volume 9.
- S. G. Barkhudaryan, K. Ghafadaryan, and S. T. Saghumyan. 2012b. *Divan Hay Vimagrutyun, Prak IX*. Institute of Archaeology and Ethnography, NAS RA, Yerevan.
- Sedrak Barkhudaryan. 1967b. *Divan Hay Vimagrutyun, Prak III*. Academy of Sciences of the Armenian SSR, Yerevan.
- Sedrak Barkhudaryan. 1973b. *Divan Hay Vimagrutyun, Prak IV*. Academy of Sciences of the Armenian SSR, Yerevan.
- Sedrak Barkhudaryan. 1982b. *Divan Hay Vimagrutyun, Prak V*. Academy of Sciences of the Armenian SSR, Yerevan.
- Francesco Bianchini. 2023. Looking beyond the text: Opportunities and challenges in the digitisation of sanskrit inscriptions. *Can't Touch This*.
- David Espinosa Espinosa and Isabel Velázquez Soriano. 2021. Epigraphy in the digital age: opportunities and challenges in the recording, analysis and dissemination of inscriptions.
- Artak Ghulyan. 2001. *Artsakhi yev Syuniki Melikakan Aparanknere*. HChU, Yerevan.
- Tim Greenwood. 2004. *A corpus of early medieval armenian inscriptions*. *Dumbarton Oaks Papers*, 58:27–?
- Garnik M. Grigoryan. 1996. *Divan Hay Vimagrutyun, Prak VII*. National Academy of Sciences of Armenia, Yerevan.
- Garnik M. Grigoryan. 1999. *Divan Hay Vimagrutyun, Prak VIII*. National Academy of Sciences of Armenia, Yerevan.
- Arsen Harutyunyan. 2017. Armaviri marzi vimagrakan zharangutyun usumnasirutyune. In *Metsamoryan Yntertsummer I*. Patmamshakutayin Argelots-Thangaranneri Tsarayutyun, Yerevan.
- Arsen Harutyunyan. 2019a. Antip vimagrere yerevani katoghike yekeghetsuts. In G. G. Sargsyan and A. E. Harutyunyan, editors, *Sedrak Barkhudaryan – 120. Gitakan Hodvatsneri Zhoghovatsu*. HAI, Yerevan.
- Arsen Harutyunyan. 2019b. Sisiani n. adontsi anvan patmutyan tangarani ardanagir ararkanere. *Hnagitutyun: Vem*, 11(2(66)).
- Arsen Harutyunyan. 2022. Tarerayin aghetnern u yerknayin yerevuytnere vimagrere. mas arayin: Yerkracharzh. *Hnagitutyun: Vem*, 14(1(77)).
- Arsen Harutyunyan. 2023a. [Tarerayin aghetnern u yerknayin yerevuytnere vimagrere. mas yerrord: Yerash, morekh](#). *Hnagitutyun: Vem*, 15(1(81))
- Arsen Harutyunyan. 2023b. [Tatevi mets anapati vimagrakan zharangutyune](#).
- Arsen Harutyunyan and Armine Melkonyan. 2021. S. ejmiatsni tapanageri: Ghevond vrd. pirghalemiani norahayt zhoghovatsu. *Ejmiatsin*.
- S. Karapetyan. 2021a. The armenian lapidary inscriptions of boon aghvank.
- Samvel Karapetyan. 1998. *Vratsi Petakan Kaghakakanutyune yev Hay Mshakuyti Hushardzannere (1988–1998), Book II*. Gitutyun, NAS RA, Yerevan.
- Samvel Karapetyan. 2004. *Kovkasyan Tangarani Haykakan Havaqatsun*. HChU, Yerevan.
- Samvel Karapetyan. 2014. *Akunk yev Katnaghbyur Gyugheri Patmakan Hushardzannere*. HChU, Yerevan.
- Samvel Karapetyan. 2021b. Bun aghvanki hay vimagrutyunnere.
- Samvel Karapetyan. 2021c. Hyusisayin artsakh. <https://raa-am.org/northern-artsakh/>.
- Samvel Karapetyan. 2024. *Khojivank*. HChU, Yerevan.
- Raffi Kortoshyan. 2013a. *Halepi Vimagrere, Prak 16*. RAA Publishing, Yerevan.
- Raffi Kortoshyan. 2013b. *The Inscriptions of Aleppo*, volume 16. RAA Publishing, Yerevan.

Raffi Kortoshyan. 2024. *Hndkastani Hayahots Bnakavayreri Vimagrere*. HChU, Yerevan.

Fangchao Liu, John Hindmarch, and Mona Hess. 2023. A review of the cultural heritage linked open data ontologies and models.

Husik Melkonyan. 2024. *Tsaghats Kar Vanke (Patmah-naagitakan Usumnasirutyun)*. HAI Publishing, Yerevan.

H. A. Orbeli. 1965a. *Corpus Inscriptionum Armeniacarum*, volume 1.

Hovsep A. Orbeli. 1960. *Divan Hay Vimagrutyun, Prak II*. Academy of Sciences of the Armenian SSR, Yerevan.

Hovsep A. Orbeli. 1965b. *Divan Hay Vimagrutyun, Prak I*. Academy of Sciences of the Armenian SSR, Yerevan.

Hamlet Petrosyan. 2008. *Khachkar*. Printinfo, Yerevan.

G. G. Sargsyan, A. E. Harutyunyan, and K. T. Asatryan. 2017a. *Corpus Inscriptionum Armeniacarum*, volume 10.

G. G. Sargsyan, Arsen E. Harutyunyan, and Karen T. Asatryan. 2017b. *Divan Hay Vimagrutyun, Prak X*. Institute of Archaeology and Ethnography, NAS RA, Yerevan.

Hamest Tamrazyan, Gagik Hovhannisyan, and Arman Harutyunyan. 2026. [From stone to standards: A digital heritage interoperability model for armenian epigraphy within the leiden and epidoc frameworks](#). *Heritage*, 9(1):27.

Hamest Tamrazyan and Gayane Hovhannisyan. 2024a. [Digital guardianship: Innovative strategies in preserving armenian's epigraphic legacy](#). *Heritage*, 7(5):2296–2312.

Hamest Tamrazyan and Gayane Hovhannisyan. 2024b. [Preserving endangered heritage: Integrating geonames/pleiades, armenian toponyms and regularization for cultural identity preservation in conflict zones](#). *International Journal of Humanities and Arts Computing*, 18(2):224–248.

Hamest Tamrazyan and Gayane Hovhannisyan. 2025. Cultural categorization in epigraphic heritage digitization. *Heritage*, 8(5):148.

Lori Tashjian. 2018. *Tseghaspanutyune Verapratsneri Tsndavayrere Libanani Tapanagrerum*. National Academy of Sciences of Armenia Press, Yerevan.

Armine Tigranyan. 2023. The armenian cultural heritage of artsakh. *Mechanisms for Protection In The International System For Preservation of Heritage, Vem Series*, (6).

Yingwei Yan, Kenneth Dean, Chen-Chieh Feng, Guan Thye Hue, Khee-heong Koh, Lily Kong, Chang Woei Ong, Arthur Tay, Yi-chen Wang, and Yiran Xue. 2020. Chinese temple networks in south-east asia: a webgis digital humanities platform for the collaborative study of the chinese diaspora in southeast asia. *Religions*, 11(7):334.

A Prompting

```
zero_shot_prompt = f"""
You are an expert in Armenian epigraphy.
Your task is to extract specific terms
from the provided text chunks.
For each term, you will identify its
occurrences in the text,
along with the context in which
it appears. The context should include
a few words before and after the term
to provide clarity on its usage.
```

```
Each occurrence should be documented
with the following details:
- file_name: The name of the file where
the term was found.
- page_number: The page number in the
document where the term was found.
- term: The term itself.
- context: A snippet of text
surrounding the term.
```

```
Each term can have multiple occurrences,
and each occurrence should be
recorded separately and stored in an
Occurrence object. All occurrences of
a term should be grouped under a Term
object, which includes the term, its
definition, and a list of its occurrences.
```

```
Finally, all Term objects should be
compiled into an ExtractionResult object.
```

```
The text is written in Armenian and may
require special handling. Extract all
occurrences of the following terms from
the document text.
Terms and definitions: {terms}
```

```
--- BEGIN TEXT CHUNK ---
{chunk}
--- END TEXT CHUNK ---
"""
```

Armenian AutoEpiDoc: Automated Extraction and Encoding of Armenian Inscriptions into EpiDoc TEI/XML

Hamest Tamrazyan

DHLAB, EPFL, Lausanne, Switzerland
hamest.tamrazyan@epfl.ch

Emile Cornamusaz

EPFL, Lausanne, Switzerland
emile.cornamusaz@epfl.ch

Emanuela Boros

DHLAB, EPFL, Lausanne, Switzerland

Abstract

Armenian epigraphy is extensively documented in printed scholarly corpora, yet lacks machine-readable editions that support interoperability or computational analysis. In this paper, we present *Armenian AutoEpiDoc*, a system that automatically converts expert-verified Armenian inscription records into EpiDoc-compliant TEI/XML files. Operating on curated and domain-validated data, AutoEpiDoc maps Armenian-specific metadata to EpiDoc structures through rule-based templates and schema-aware validation. The workflow significantly reduces manual encoding effort and provides a scalable path toward producing digital editions and integrating Armenian inscriptions into international epigraphic infrastructures.

1 Introduction

Armenia possesses one of the richest and most continuous inscriptional traditions in the Near East, spanning religious, commemorative, funerary, legal, and architectural contexts for over a millennium. Yet despite this abundance, no comprehensive machine-readable corpus exists, and Armenian inscriptions remain largely absent from digital epigraphic infrastructures. Recent surveys highlight the scale and geographic dispersion of the material: more than 80% of Armenian inscriptions lie outside the borders of the modern Republic of Armenia, often in regions affected by conflict or infrastructural risk (Greenwood, 2014; Tamrazyan, 2023). Major corpora, such as the multi-volume *Divan Hay Vimagrut'yan* (DHV), are authoritative but remain non-digital and unstructured, preventing the field from benefiting from interoperability, searchability, semantic modelling, and long-term preservation offered by standards such as EpiDoc¹.

¹EpiDoc is a community-developed application of the TEI Guidelines for the digital encoding, publication, and inter-

Without controlled vocabularies, structured metadata, or machine-actionable corpora, Armenian inscriptions cannot be linked to infrastructures such as EAGLE² or FAIR Epigraphy (Bianchini, 2023)³. Dominant epigraphic ontologies are largely grounded in Greco-Roman traditions and do not reflect the cultural specificity or typological diversity of Armenian inscriptions. As noted in studies of digital cultural heritage (Liu et al., 2023), such mismatches can lead to conceptual loss when local traditions are forced into externally defined categories.

This limitation is not due to a lack of scholarship. Armenian epigraphy is extensively documented in catalogues, surveys, and monographs; rather, the challenge lies in the absence of scalable mechanisms for converting expert knowledge into standard-compliant digital editions. Manual EpiDoc encoding requires both technical expertise and familiarity with Armenian epigraphic conventions, making large-scale digitization impractical (Tamrazyan and Hovhannisyan, 2024; Tamrazyan et al., 2025).

TEI⁴ provides a widely adopted XML framework for structured text encoding, with EpiDoc as its epigraphic application profile. Armenian inscriptions, however, remain largely absent from TEI/EpiDoc-based digital corpora due to the lack of scalable tools for transforming expert-curated catalogues into machine-readable editions.

To address this gap, we introduce *Armenian AutoEpiDoc*, the first system designed to automatically generate EpiDoc-compliant TEI/XML editions from expert-curated Armenian inscription records. Unlike OCR-based or text-mining approaches, AutoEpiDoc operates on inscriptions

change of ancient and medieval epigraphic and papyrological texts. <https://epidoc.stoa.org/>

²<https://www.eagle-network.eu/>

³<https://inscriptions.org/>

⁴<https://tei-c.org/>

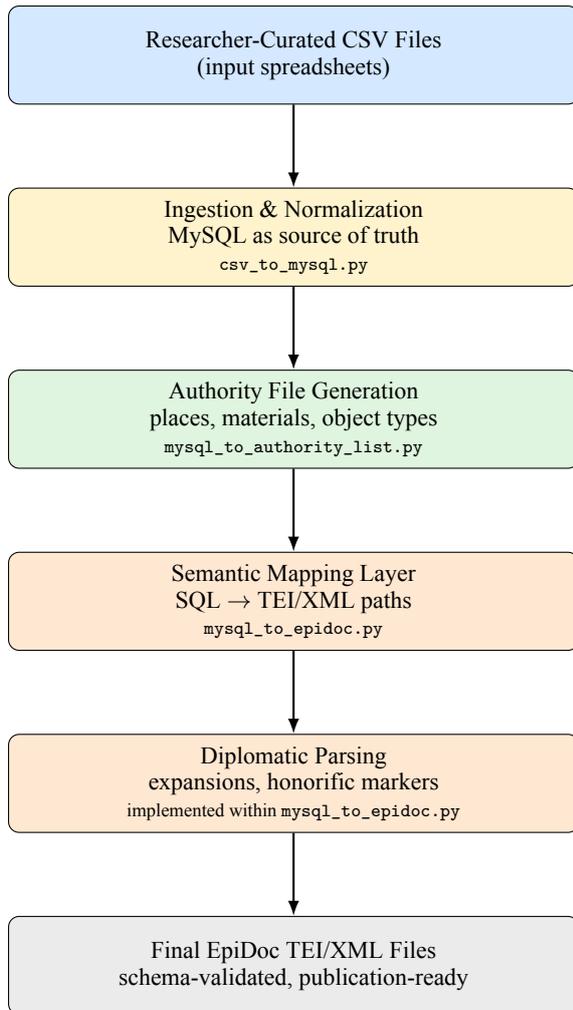


Figure 1: Color-coded architecture of the AutoEpiDoc ELT pipeline, showing Python modules and semantic processing layers.

that have already been validated by specialists and functions as a transformation layer that maps structured scholarly metadata and diplomatic transcriptions to EpiDoc elements, producing schema-valid XML. This significantly reduces the manual effort required for TEI encoding and enables the integration of Armenian epigraphic data into international digital epigraphic infrastructures and Linked Open Data (LOD) ecosystems.

In this paper, we present the design, implementation, and demonstration of *Armenian AutoEpiDoc*. All components of the system, including authority lists, generated EpiDoc files, and the complete AutoEpiDoc at <https://github.com/dhlab-epfl/autoepidoc>.

2 Methodology

Armenian AutoEpiDoc implements a semi-automated Extract–Load–Transform (ELT) pipeline designed to bridge the gap between unstructured research data and the strict hierarchical requirements of the EpiDoc standard. The architecture prioritizes data integrity and semantic interoperability through a three-stage process: (1) **Ingestion and Normalization** (Blue), (2) **Authority Management**, and (3) **XML Serialization**.

Figure 1 provides an overview of the AutoEpiDoc ELT pipeline. The color-coded layers in the figure correspond directly to these three stages:

Next, we clarify how data flows through the system and how each Python module operationalizes one component of the transformation pipeline into fully standards-compliant EpiDoc files.

2.1 Pipeline Architecture

The system utilizes an intermediate relational database (MySQL) as the primary source of truth rather than processing CSV files directly into XML. This decision allows for strict type enforcement and relational integrity checks prior to encoding.

2.1.1 Stage 1: Ingestion and Normalization

The input researcher-curated CSV files were organized into thematic datasets describing inscriptions, monuments, places, materials, and controlled vocabularies. Table 1 provides an overview of all input files and their roles in the ingestion workflow.

The ingestion module (`csv_to_mysql`) reads researcher-curated spreadsheets (`.csv`) as presented in Table 1 and it applies a heuristic normalization algorithm that sanitizes inconsistent column headers (e.g., handling variants such as “*sub-monument ID*” vs. “*sub_monument_id*”) and enforces type safety to avoid accidental numerical inference in inventory codes. The module then loads the cleaned data into a normalized Star Schema, where a central inscription table references auxiliary lookup tables for monuments, locations, materials, and object types.

2.1.2 Stage 2: Authority Management

A important requirement for modern digital epigraphy is the use of controlled vocabularies to guarantee semantic consistency. Instead of storing

CSV File	Description	Key Columns / Fields
inscriptions.csv	Primary dataset describing each inscription record; forms the core of the pipeline.	Inscription ID, monument ID, transcription, Armenian/English description, date (Armenian Era), material, object type, condition, bibliography.
monuments.csv	List of monuments and architectural structures to which inscriptions belong.	Monument ID, name (hy/en), type, village/town, region, coordinates, parent monument ID (for complexes).
locations.csv	Hierarchical geographic information; used for authority file generation.	Place ID, name (hy/en), type (village, district, region), parent place ID, contemporary state name, coordinates.
materials.csv	Controlled list of inscription materials used to generate TEI <material> vocabularies.	Material ID, Armenian label, English label, notes, AAT reference (if available).
object_types.csv	List of inscription-bearing object types used across Armenian epigraphy.	Object type ID, Armenian label, English label, definition, AAT reference (if available).
bibliography.csv	Structured citations for printed corpora, surveys, and monographs.	Bib ID, author(s), year, title, publication, pages, stable ID (e.g., Zotero).
photos.csv (optional)	Links between inscriptions and photographic documentation.	Photo ID, inscription ID, file path/URL, photographer, date, rights.

Table 1: Researcher-curated CSV (spreadsheets) files ingested by the AutoEpiDoc pipeline and their role in the MySQL-based normalization stage. Each file populates one or more relational tables used as the authoritative source before XML serialization.

metadata as unstructured strings, AutoEpiDoc generates standalone TEI-compliant Authority Files for **Places, Monuments, Materials, and Object Types**.

The authority generation module (`mysql_to_authority_list`) builds these files hierarchically. For geospatial entities, relational foreign keys (e.g., `parent_place_id`) are converted into semantic TEI nodes (<note type="relation">), thereby reconstructing administrative hierarchies such as those of the Artsakh region. The module also integrates LOD principles by aligning local terminology with international ontologies. Materials and Object Types are automatically linked to the Getty Art & Architecture Thesaurus (AAT) and to EAGLE vocabularies via a dedicated <standOff> section, ensuring interoperability with European digital infrastructures.

2.1.3 Stage 3: XML Serialization

The final stage of the pipeline converts MySQL inscription records into EpiDoc-compliant TEI/XML files. The transformation script (`mysql_to_epidoc`) uses a dynamic XPath-based mapping strategy to translate flat SQL rows into nested TEI structures, including both the <teiHeader> and the <text> body.

Diplomatic Transcription Parsing Metadata and transcription are processed separately. Armenian epigraphic transcriptions rely on scholarly diplomatic conventions that are meaningful to spe-

cialists but not directly machine-readable, such as letters functioning as numerals (e.g. date formulas), non-punctuation symbols used as numeric separators, abbreviated forms, ligatures, and overlines marking honorific or sacred names.

Transcriptions are therefore stored in a simplified ASCII notation reflecting established Armenian epigraphic practice and are subsequently parsed by a rule-based module that maps these conventions to structured EpiDoc TEI/XML. For example, abbreviated forms are encoded using <expan> with explicit <abbr> and <ex> components, numeric letters are normalized using <num>, and overlines and ligatures are represented using <hi> and <join>. This preserves palaeographic and editorial information while making Armenian diplomatic transcriptions interoperable with international digital epigraphic standards, without requiring researchers to write TEI/XML manually.

The most challenging cases for rule-based parsing involve damaged numeral sequences, ambiguous ligatures, and inconsistent scholarly transcription practices across sources. In such cases, AutoEpiDoc preserves the original transcription and flags uncertain elements for expert review rather than enforcing potentially erroneous normalization.

Chronological and Bilingual Processing To support the needs of Armenian studies, the system represents dates in both the Armenian Era and the Gregorian calendar. Original Armenian-era dating expressions are encoded in <origDate>

with a custom @calendar attribute, while computed Gregorian equivalents are provided for indexing and interoperability. All descriptive metadata is generated bilingually (Armenian/English) using xml:lang attributes to support multilingual search and display.

3 Output Examples

In this section, we present some excerpts from the EpiDoc-compliant TEI/XML files generated by AutoEpiDoc from real Armenian inscription records obtained in this work.

Example 1: TEI Header and Authority Integration. The <teiHeader> generated by AutoEpiDoc integrates descriptive and administrative metadata derived from the underlying relational schema and enriched through automatically generated authority lists. Stable URN-based identifiers for monuments, architectural components, materials, object types, techniques, and scripts are referenced via @ref attributes, while multilingual naming is supported through parallel language-specific elements. An inscription associated with the Gandzasar monastic complex is encoded as follows:

```
<msDesc xml:id="
  ms_Gandzasar_Monastic_complex_monastery">
  <msIdentifier>
    <repository ref="urn:armepic:mon:MON0002">
      <objectName xml:lang="en">Gandzasar
        Monastic complex / monastery</
        objectName>
      <objectName xml:lang="hy">Գանձասար
        վանական համալիր / վանք</objectName>
    </repository>
  </msIdentifier>
  <msPart xml:id="ms_Gavit_Narthex">
    <msIdentifier>
      <repository ref="urn:armepic:mon:
        MONPART0010">
        <objectName xml:lang="en">Gavit (
          Narthex)</objectName>
        <objectName xml:lang="hy">Գավիթ</
          objectName>
      </repository>
    </msIdentifier>
  </msPart>
  <physDesc>
    <objectDesc>
      <supportDesc>
        <support>
          <objectType ref="urn:armepic:
            objtype:OBJ0006" xml:lang="en"
            >lintel</objectType>
          <objectType xml:lang="hy">բարավոր</
            objectType>
          <material ref="urn:armepic:material:
            MAT0001" xml:lang="en">tuff</
            material>
```

```
<material xml:lang="hy">տոնիք</
  material>
  <rs type="technique" ref="urn:armepic:
    technique:TEC001" xml:lang="en">
    carved</rs>
  <rs type="technique" xml:lang="hy">
    փորագրություն</rs>
  </support>
</supportDesc>
<layoutDesc>
  <layout xml:lang="hy">Հյուսիսային մուտքի
    կիսակամար բարին,
    արտաքուստ</layout>
  <layout xml:lang="eng">On the stone
    semi-arch of the northern entrance,
    from the outside</layout>
</layoutDesc>
</objectDesc>
<handDesc>
  <handNote xml:id="hand_ART0001" scriptRef
    ="urn:armepic:script:SCR002">
    <term xml:lang="en">Bolorgits Erkatagir
      </term>
    <term xml:lang="hy">Բոլորգիտ
      երկաթագիր</term>
  </handNote>
</handDesc>
</physDesc>
</msDesc>
```

Listing 1: Example of Armenian epigraphic encoding (TEI/XML)

This example illustrates the transformation of catalog descriptions into semantically structured metadata aligned with international vocabularies.

Example 2: Diplomatic Transcription Encoding. AutoEpiDoc converts Armenian diplomatic transcription conventions into structured TEI markup while preserving palaeographic and editorial information. Line breaks, ligatures, abbreviations, expansions, and numerals are encoded explicitly, supporting both faithful representation and computational reuse.

```
<text>
  <body>
    <div type="edition">
      <ab>
        <lb n="1"/><w>:թիւ:</w>
        <w>ՉԻ
          <num value="1271">1271</num>
        </w>
        <lb n="2"/>
        <w>Կ
          <hi rend="ligature">սւ</hi>
          <hi rend="ligature">սւ</hi>
          <hi rend="ligature" xml:id="lig1">ւ</
            hi>
        </w>
        <w>
          <hi rend="ligature" xml:id="lig2">ւ</
            hi>
          <emph rend="bold">
            <expan>
              <abbr/>
              <ex>սոսնոն</ex>
```

```

        <abbr>j,</abbr>
      </expan>
    </emph>
  </w>
  <join xml:id="j1" result="ligature"
        target="#lig1 #lig2"/>
  <w>tu</w><w>Յnhwltu,</w>
</ab>
</div>
</body>
</text>

```

Listing 2: TEI/XML encoding of an Armenian inscription edition

Example 3: Chronological Representation.

Dates in Armenian epigraphic corpora are often expressed using the Armenian Era (Ա.Թ.). AutoEpiDoc encodes both the original dating expression and a normalized Gregorian equivalent, enabling chronological interoperability while preserving historical conventions.

```

<origin>
  <origPlace>
    <!-- place information -->
  </origPlace>
  <origDate>
    <date calendar="#cal_armenian" when-
      armenian="0720">ՉԻ</date>
    <date calendar="#cal_gregorian" when="1271
      ">AD 1271</date>
  </origDate>
</origin>

```

Listing 3: Encoding of original date using Armenian and Gregorian calendars

4 Conclusions

The examples presented in this paper show that AutoEpiDoc is not merely a tool for converting tabular data into XML, but a system for semantically restructuring Armenian inscription records. Rather than reproducing the flat structure of the source data, the pipeline generates hierarchical TEI/EpiDoc documents that preserve diplomatic fidelity while supporting long-term digital preservation, interoperability, and reuse.

The resulting outputs integrate multilingual metadata, Armenian-era dates with normalized Gregorian equivalents, controlled vocabularies, and explicit authority references in a fully machine-actionable form. Schema-based validation and rule-driven transformations ensure consistency and scalability across large and heterogeneous corpora.

AutoEpiDoc has been applied to a growing subset of Armenian inscriptions within the ArmEpiC

corpus and the Artsakh Epigraphy Atlas. In pilot comparisons, manual TEI/EpiDoc encoding typically required tens of minutes per inscription, whereas the system produces a first-pass EpiDoc file in seconds, requiring only targeted expert correction. Domain specialists confirmed that the generated encodings preserve Armenian epigraphic conventions while substantially reducing manual encoding effort. The system is currently in active use and continues to be iteratively validated within these projects, demonstrating its value for real epigraphic research workflows. Remaining challenges concern ambiguous ligatures and damaged numeral sequences, which are flagged for expert review.

By aligning Armenian epigraphic data with international TEI/EpiDoc standards, this work lays the foundation for the first comprehensive, standards-aligned digital corpus of Armenian inscriptions and enables interoperability with epigraphic infrastructures such as FAIR Epigraphy, EFES, and Linked Open Data environments.

References

- Francesco Bianchini. 2023. Looking beyond the text: Opportunities and challenges in the digitisation of sanskrit inscriptions. *Can't Touch This*.
- Tim Greenwood. 2014. Armenian epigraphy. In *Armenian Philology in the Modern Era*, pages 101–121. Brill.
- Fangchao Liu, John Hindmarch, and Mona Hess. 2023. A review of the cultural heritage linked open data ontologies and models.
- Hamest Tamrazyan. 2023. Digitization of the inscriptions on the monuments of armenian cultural heritage in nagorno-karabakh region. In *DH*.
- Hamest Tamrazyan and Gayane Hovhannisyan. 2024. Digital guardianship: Innovative strategies in preserving armenian’s epigraphic legacy. *Heritage*, 7(5):2296–2312.
- Hamest Tamrazyan, Gayane Hovhannisyan, and Arsen Harutyunyan. 2025. From stone to standards: A digital heritage interoperability model for armenian epigraphy within the leiden and epidoc frameworks.

Studying Expert-ese: Profiling and Classification of Domain-Specific Language Variation in Architecture with Traditional Machine Learning and LLMs

Carmen Schacht

Ruhr-University Bochum, Germany
Faculty of Philology
Department of Linguistics
carmen.schacht@rub.de

Renate Delucchi Danhier

TU Dortmund University
Department of Cultural Studies
Institute for Diversity Studies
renate.delucchi@tu-dortmund.de

Abstract

This study investigates how domain expertise shapes spontaneous oral language production, with a focus on architecture. Building on the ExpLay Corpus, which contains image descriptions by speakers with and without architectural training, we analyze linguistic variation by combining Profiling-UD and the DECAF framework. We extract a broad range of syntactic and morpho-syntactic features to build linguistic profiles for both groups and train classifiers to distinguish expert from non-expert productions. Two traditional machine learning models (logistic regression and SVM) are compared with a lightweight BiLSTM and two large language models (GliClass and LLaMA 2). While the expert and non-expert corpora diverge only subtly (pairwise Jensen–Shannon divergence (JSD)= 0.25), the BiLSTM using fastText embeddings achieves the highest F1-score (0.88), outperforming both traditional models and LLMs. This indicates that semantic representations are more predictive of domain variation than purely structural features and that smaller neural architectures generalize better on limited data. Overall, the findings provide empirical evidence that architectural expertise leaves measurable linguistic traces in spontaneous speech, supporting the Grammar of Space hypothesis.

1 Introduction

Linguistic variation reflects differences in situational, social, and cognitive contexts. This paper investigates how expertise shapes spontaneous oral language production, focusing on the architectural domain. We operationalize the concept of “expertese”: a domain-specific linguistic register analogous to “translationese” (Gellerstam, 1986; Kuniilovskaya and Corpas Pastor, 2021; Teich et al., 2016), to describe systematic differences that emerge in the language of expert communities. For the present project, the definition of register adopted is that of Argamon (2019):

"A register is described by that situational context and the linguistic features typical of the register, along with a description of how those features function specifically within that particular context of language use. That is, the linguistic features describing a register are not arbitrary, but form a complex that is useful for particular communicative purposes in a particular context."

This definition allows for the assumption that linguistic patterns emerge as shaped by their communicative occurrence—in this case, the assumption that domain-expertise shapes the linguistic phenotype of language productions within the respective domain. Previous research on register has shown that linguistic features (lexical, syntactic, and discourse-level) cluster in systematic ways depending on communicative context (Biber and Conrad, 2009). Experts often operate within shared cognitive and informational environments, leading to specialized linguistic patterns (Argamon, 2019; Degaetano-Ortlieb and Piper, 2019). This phenomenon is not limited to writing: spoken language can also exhibit register-specific variation shaped by expertise, although this has been investigated less systematically. Architecture provides a compelling domain for studying these effects, since architects undergo extensive training, use specialized terminology, and engage with spatial representations in ways that differ from laypeople (Mertins et al., 2020). These cognitive and communicative differences are expected to leave measurable linguistic traces even in spontaneous speech. This study builds on insights from information-theoretic approaches to language (Shannon, 1948; Jaeger, 2010), which view communication as the efficient transmission of information. Registers are hypothesized to optimize information flow within communities, often through conventionalized and compact

linguistic structures (Halliday, 1988/2004; Teich et al., 2016). If experts share more background knowledge, their speech may exhibit distinctive patterns of linguistic complexity and distribution. Using the ExpLay Corpus (Schacht and Delucchi Danhier, 2025), which contains spontaneous image descriptions by experts and non-experts, we conduct a multi-level analysis of linguistic variation. We apply Profiling-UD and DECAF to extract syntactic and morpho-syntactic features, and evaluate whether machine learning models can distinguish expert from non-expert speech. Methodologically, we explore to which extent the linguistic manifestation of expertise is detectable automatically. Beyond its contribution to register variation research, this work is also motivated by a broader relevance of society: in many public-facing domains, experts have trouble communicating effectively with laypeople. In architecture, where communication with clients, policymakers, citizens, and other interdisciplinary colleagues such as engineers is central to the profession, such misunderstandings can not solely be attributed to a lack of shared specialized architectural vocabulary, but may also reflect differences in how information is structured and distributed in spontaneous speech by experts and laypeople. By showing that architectural expertise leaves measurable traces in oral production beyond professional jargon, this study shifts attention from surface notions of complexity to differences in information packaging and shared knowledge assumptions. While this analysis focuses on architecture, the proposed profiling and classification approach is transferable for investigating expertise-related register variation in other domains where experts and laypeople need to communicate with each other, such as medicine or law. The project data is made available under CC BY 4.0 license¹.

2 Related Work

2.1 Linguistic Complexity and Register Variation

Linguistic complexity has been widely studied in domains such as language acquisition and second language learning (e.g., Park (2024), Lu (2010), Xia et al. (2016), Collins-Thompson (2014), Kyle (2016)). More recently, questions of linguistic complexity have also become prominent in research on machine learning and LLMs, as illustrated in Misra and Mahowald (2024). These strands of re-

search share an interest in identifying and quantifying linguistic features that may serve as indicators of structural or cognitive complexity in language use.

In order to apply linguistic complexity to the analysis of register variation—like Teich et al. (2016) or Kunilovskaya and Corpas Pastor (2021) did—this paper adopts the definition of register introduced earlier, which allows for the assumption of linguistic patterns emerging as shaped by their communicative occurrence. In the context of the present study, this assumption translates into domain-expertise shaping the linguistic characteristics of language productions within a given domain.

This reasoning leads directly to the theoretical frameworks underlying the present study. Following Schacht and Delucchi Danhier (2025), the research builds on perspectives that link domain expertise, cognition, and language production. At its core lies the theory of linguistic relativity (Whorf, 1956; Slobin, 1996), which states that language shapes human cognition and specifically attention. Empirical work has demonstrated language relativity effects in domains including color perception (Winawer et al., 2007; Roberson et al., 2000), spatial reference frames (Levinson, 2003; Majid et al., 2004), and motion events (Papafragou et al., 2008).

Analogous cognitive effects have been observed for expertise in different domains, which can influence perception and processing in ways similar to language. Classic neuroimaging research shows structural adaptations linked to spatial training in taxi drivers (Maguire et al., 2000), while other studies report perceptual and sensory-motor advantages in expert or semi-expert populations such as gamers (Ersin et al., 2022; Jiang et al., 2020). In the architectural domain specifically, eye-tracking studies reveal distinct visual attention patterns for experts compared to laypeople (Delucchi Danhier et al., 2025; Mertins et al., 2020), motivating the extension of such analyses to linguistic behavior.

2.2 Communication Efficiency, Information Theory, and Domain-Specific Conventions

As a continuation of the reasoning outlined above, one of the fundamental principles of communication in general is to achieve a smooth and ideally loss-less transfer of information while eliminating unnecessary linguistic signal—that is, to achieve efficiency—and thereby align flexibly with the communicative situation.

Mathematically speaking, communication al-

¹<https://osf.io/4hvm6/overview>.

ways takes place through a noisy channel between the sender (speaker) and the receiver (listener) of information, as described in Shannon (1948)'s Information Theory.

Communication in this sense is understood as a linear transfer of information in single communicative units (for example, in bits) and successful communication as being as loss-less as possible, striving for an ideal use of the communicative channel. This reasoning underlies the framework of the Uniform Information Density hypothesis (UIDh) (Jaeger, 2010), shifting the mathematical-engineering approach towards linguistics. The UIDh proposes that the use of the communicative channel must not exceed its capacity boundaries—neither the upper nor the lower bound—in order to avoid loss of transmitted information. Speakers therefore need to minimize divergence in the flow of information, ensuring a relatively homogeneous distribution of information across sentences.

This is achieved by choosing linguistic encoding in such a way as to keep the flow as constant as possible. From different communicative situations emerge different demands and standards, however, meaning that baseline channel capacities may vary. This leads to the assumption that different registers, for instance, will display distinct patterns of linguistic choices to achieve this balance and further that not only language but also expertise shapes cognition and subsequently linguistic production. According to Teich et al. (2021), this phenomenon manifests in the form of conventionalized linguistic codes serving to smoothen the information density profile of communication in specialized fields while still transmitting the relevant but—compared to general language—heightened amount of information characteristic to the respective field. Concrete examples of this phenomenon are studies on the language in the fields of physics (Halliday, 1988/2004), literary studies (Degaetano-Ortlieb and Piper, 2019), and scientific communication more broadly (Teich et al., 2016; Degaetano-Ortlieb and Teich, 2022), all showing conventionalized and informationally compact linguistic encoding. Similarly, shifts within the same language over time can be registered as shown in diachronic studies of scientific English (Rubino et al., 2016; Degaetano-Ortlieb and Teich, 2018; Biber et al., 2011; Biber and Gray, 2016; Juzek et al., 2020) as well as scientific German (Jakobi et al., 2024). Common structures that facilitate the condensa-

tion of information in a production take the form of for example compounding (Degaetano-Ortlieb, 2021; Gamboa et al., 2024, 2025) and metaphor (Halliday, 1988/2004; Webster, 2018), especially in highly technical or scientific language, as they condense a heightened amount of information compared to their phrasal counterparts, further underscoring the link between communicative efficiency, register variation, and domain-specific expertise.

Accordingly, the expectation of denser linguistic encoding as shown by the previous research introduced earlier is well motivated by the UIDh, as shared knowledge among a group of domain-experts can be assumed, lifting the baseline for appropriate information transmission, especially in tasks related to their common professional field where they discuss familiar information. This, in turn, allows for divergence from common structures in general language use and justifies the use of more complex, information-rich expressions, as they will not be considered more complex within their specialized surroundings and thus not spike the flow of information.

2.3 The ExpLay Corpus and Pipeline

Delucchi Danhier et al. (2025) showed that architectural experts and laypeople display distinct visual attention patterns when processing three-dimensional spatial stimuli presented in two dimensions, as revealed by eye-tracking data. In addition to these findings, Schacht and Delucchi Danhier (2025) argue that experts are likely to exhibit comparable differences when verbalizing visual perception, since linguistic production inherently involves the linearization of perceptual input (Levelt, 1989). This perspective integrates principles from Linguistic Relativity, Information Theory, and the UID hypothesis, motivating a systematic investigation of linguistic patterns in expert versus non-expert language.

The study is based on the ExpLay Corpus (Schacht and Delucchi Danhier, 2025)², that is available under a CC BY 4.0 license and contains spontaneous German image descriptions produced by speakers with and without architectural training. Participants described a set of architectural stimuli under controlled experimental conditions, resulting in a balanced dataset of expert and non-expert productions. The corpus provides matched elicitation settings and spontaneous language data, making

²<https://gitlab.ruhr-uni-bochum.de/schaccmr/explay-resource.git>.

it well suited for investigating domain-specific linguistic variation.

All data were processed using the ExpLay pipeline, which integrates automatic UD parsing, linguistic feature extraction, as well as profiling via Profiling-UD and DECAF. This infrastructure enables systematic comparisons of linguistic distributions between groups and forms the basis for the analyses presented in this paper.

3 Methodology

Several metrics and automatic analysis tools for the evaluation of linguistic complexity in various use cases exist to date, including the ExpLay resource (Schacht and Delucchi Danhier, 2025) which implements an automatic evaluation module based on the approach of Park (2024), incorporating a selected number of syntactic features automatically extracted from the annotated corpus data.

To extend the approach of ExpLay into an extensive fine-grained analysis, this paper aims to evaluate i) the divergence of the two sub-corpora, quantifying the difference between them and ii) the difference in linguistic profiles of the two sub-corpora by following the objective of Teich et al. (2016), who test if "classes have distinctive linguistic correlates and, if so, how well the classes are distinguished linguistically and which features contribute most to their distinction". Linguistic profiling has received growing attention within the last couple of years and has increasingly influenced research areas such as computational register analysis (Argamon, 2019), which investigates divergences between registers. As Profiling-UD (Brunato et al., 2020) only accepts unparsed data, the raw ExpLay data is re-parsed using the tool, to ensure a consistent analysis throughout this study. The data is then first evaluated using the DECAF framework (Müller-Eberstein et al., 2025), an automatic tool under the MIT license originally designed for the analysis and filtering of annotated datasets to create specialized training-data according to the respective research question or training objective, which also offers a module for the evaluation of divergence among corpora based on their linguistic features. We apply DECAF to the data parsed with Profiling-UD for divergence analysis. Subsequently, the data is processed with the Profiling-UD tool itself, an automatic processing tool freely available online for fine-grained linguistic profiling analysis incorporating more than 130 linguistic

features automatically extracted from the data, of which 113 are used in this study, based on their availability in the web-application of Profiling-UD (see the list in F for the sub-set used here). Those were extracted first to identify the features that contribute most strongly to the distinction between the two groups and subsequently to integrate the identified distinguishing features into the classification task, in order to test their predictive strength. The categories of features in Profiling-UD include:

1. Raw Text Properties
2. Lexical Variety
3. Morpho-syntactic Information
4. Syntactic Features

This tool is task-agnostic and explicitly multi-lingual, as it is built upon the UD-framework (de Marneffe et al., 2021) and thus supports research on language variation flexibly. Based on the profiling results, the current study follows the approach of Schacht and Delucchi Danhier (2025) and Park (2024) by conducting a Principal Component Analysis (PCA) (Jolliffe, 2002) as well as a subsequent ANOVA on the extracted profiles. These steps serve to reduce the multi-dimensional profiling results to the features that contribute most significantly to the variance in the data and to evaluate the statistical significance of these contributions. For both, the PCA and the subsequent creation of the radar charts, the data is normalized applying the standard scalar, mapping the values into a shared dimensional space. The PCA is performed with the Scikit library (Pedregosa et al., 2011) and the ANOVA using the Scipy library (Wes McKinney, 2010), both under 3-Clause BSD license.

The evaluation will focus on the following three research questions derived from the analysis methodology:

1. How much do the two sub-corpora diverge from each other?
2. Do the linguistic profiles of the domain-experts differ from the profiles of the laypeople?
3. Do the domain-experts display more complex linguistic structures than the non-experts?

In addition, two classic machine learning classifiers—a logistic regression model and a Support

Vector Machine (SVM) model—are trained for the classifications of documents into the categories of ‘expert’ and ‘non-expert’ following [Kunilovskaya and Corpas Pastor \(2021\)](#), [Teich et al. \(2016\)](#) and [Argamon \(2019\)](#), not to improve classification accuracy, but to validate the linguistic assumption of register-distinction via linguistic features. We first perform feature selection—identifying the most predictive features for our classification task of classifying expert-ese productions—using the grafting technique ([Perkins et al., 2003](#)) following [Cimino et al. \(2017\)](#) and subsequently train the classifiers. As these algorithms are feature based algorithms, the classification is used to identify the most predictive features of the two sub-corpora by virtue of the grafting feature selection returning a ranked feature importance list. These features will then be compared to the results from the PCA and ANOVA. The two classifiers are trained using the Scikit library and are set with the default values of the SVM and logistic regression. The grafting is implemented via a custom script and the resulting feature vectors are fed into the models. A five-fold cross-validation is applied to the grafting algorithm. To compare the feature-representation of the data—which can be considered a mostly structural representation capturing linguistic characteristics derivable from the linguistic surface form—with a more semantic approach, which will represent more of the meaning of the input data, a Bidirectional Long Short-Term Memory (BiLSTM) ([Hochreiter and Schmidhuber, 1997](#)) model is trained and evaluated on the data as well. For classification, we trained three BiLSTM models using random, GloVe ([Pennington et al., 2014](#)) (available under Apache-2.0 license), and fastText ([Grave et al., 2018](#)) embeddings (available under CC BY-SA 3.0 license). Models were trained for up to 15 epochs with early stopping and standard hyperparameters. The Scikit library was used for the implementation. In addition, two large language models (the Bert-based encoder model GliClass ([Stepanov et al., 2025](#)) available under Apache-2.0 license and the LLaMA-2-7b-hf model from Meta ([Touvron et al., 2023](#)), available on HuggingFace upon request under meta license) were evaluated in zero- and few-shot settings to compare the performance of large-scale semantic representations with that of smaller, feature-based architectures. The GliClass model is a variation of the GliNER architecture ([Zaratiana et al., 2024](#)) with DeBERTa ([He et al., 2020](#)) at its core which has been optimized for classification tasks. We

tested the large-v3.0 version of GliClass with 439 Million parameters and the base-v2.0-rac version, which was optimized on a Retrieval-Augmented Classification (RAC) dataset for the few-shot scenario of classification task by augmenting the training with retrieved most-similar examples of the classes which it was trained upon. To run a decoder architecture—designed for generation—against an encoder model, LLaMA 2 7b is also tested. LLaMA-2-7b-hf is a 7 Billion parameter pre-trained, autoregressive decoder language model employing an optimized transformer architecture trained on general language and not fine-tuned on instructions. Tested against the performance of the feature-based models it will compare the predictive power of both dimensions on the divergence of the two sub-corpora. Should only the simpler BiLSTM architecture prove to reliably handle the classification of the two sub-corpora, this might indicate a poor generalization on the part of the LLMs due to insufficient data. In case neither of the models perform well on the task, we can assume the semantic to be non-predictive for the classes.

3.1 Data and Data Collection

The data analyzed in this study originates from the ExpLay Corpus ([Schacht and Delucchi Danhier, 2025](#)), which consists of experimental elicitation data in German from spontaneous oral image descriptions produced by participants with and without architectural training and contains 130 productions in total. The corpus has been processed using the accompanying ExpLay-Pipeline. Annotations include morpho-syntactic and lexical parses, dependency parsing within the Universal Dependencies framework ([de Marneffe et al., 2021](#)), constituency parsing, as well as compound and coreference annotation. For the corpus, two groups of 13 participants each were selected and matched by gender, age, and multilingual status, to control for potential noise in the data. The expert-group consists of five male and eight female and the laypeople group of four male and nine female participants. The participants’ age ranges from 19 and 32 years and both groups include 12 monolingual and 1 bilingual speaker. The degree of expertise in the expert-group has to be categorized as semi-expert though, as the participants were still students of architecture at the time of data collection instead of established architects, and domain-specific effects are still expected accordingly.

4 Results

Calculation of Divergence The calculation of the JSD based on the parses from the Profiling-UD tool resulted in a divergence of 0.25, which is considered a small divergence. Thus, the differences between the two groups might be subtle, but considering the limited amount of data it still proves valuable to evaluate the contribution of features to the divergence. The top 6 features contributing most to the divergence were *Person*, *xpos*, *Verb-Form*, *PronType*, *deprel*, and *upos* (see Table 2 for their strength of contribution).

Even though the contribution of the individual features might be modest, it proves worthwhile to examine how the variation between the corpora takes shape. A closer look into the differences of distribution of the top six contributing features reveals for example variation in the use of person, with the experts using the first person more than the non-experts while the non-experts tend to prefer the third person, indicating a variation in the description of perspective (see Table 3). In addition, experts seem to produce slightly more relative pronouns compared to non-experts (see Table 4), which might indicate a variation in descriptive depth, as relative clauses add additional information to their head. Such variation might point to differences in semantics between the two sub-corpora, motivating the further investigation of linguistic patterns found within the two groups by means of a deeper structural analysis as well as semantic approaches.

Linguistic Profiling of the Sub-Corpora The linguistic profiling was conducted on the unparsed raw data and resulted in a total of 113 features out of the full set of 130 available features automatically selected and used by the parsing tool (see the list in the Appendix F for the subset used in the present study). The features are categorized into the four broader categories presented in Section 3. The syntactic features are then again grouped into sub-categories according to Brunato et al. (2020). By normalizing and subsequent pooling of the values of the individual features by group and subcategory, linguistic profiles were created in a shared dimensional space. The standard scaler from the Scikit library was applied for normalization and then the mean was calculated by group per feature. To visualize variations in the different distributions of those profiles radar charts were created, both for by-sub-category and by-broad-category, to compare

distributions. As can be derived from the combined category profiles in Figure 1, the greatest variation between the two groups can be observed in the raw text properties. Especially features like token-length or tokens per sentence stand out if we zoom into the radar chart of the raw text properties in Figure 2 in the Appendix B. Experts seem to produce overall longer tokens and longer sentences. Additionally, the type-token-ratio in chunks of 100 tokens in the lexical variety chart (see also Figure 3) is remarkable, as experts seem to produce a greater variety of different tokens than non-experts, indicating more varied semantics. Both of these findings could indicate a more specialized language in domain experts. Also, the morpho-syntactic profiles in Figure 4 in appendix B of the expert sub-group displays more varied characteristics than the profiles of the non-experts, by means of an elevated lexical density and verbal characteristics. With regard to the syntactic features, the profiling seems to be able to replicate the initial findings of Schacht and Delucchi Danhier (2025), as the experts exhibit more pronounced tree structures.

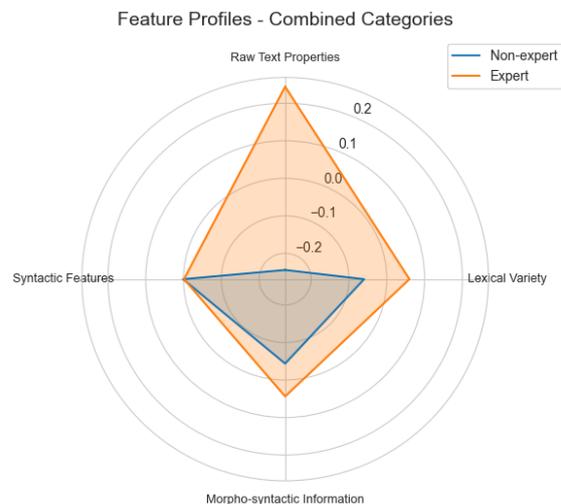


Figure 1: Radar chart of the combined categories profiles.

PCA and ANOVA To also statistically evaluate the visual findings of the linguistic profiling discussed above, a PCA and ANOVA was performed on the feature profiles of both groups. Via the PCA the principle components that reflect the relations among the variables the most can be extracted. By analyzing which variables contribute most to the first principle component (PC1) we can evaluate

Table 1: Classification reports (Micro-averaged Precision, Recall, and F1-scores) of the different classifiers across groups.

Group	Classifier	Precision	Recall	F1-score
classic ML	Logistic Regression	0.730	0.730	0.730
classic ML	SVM	0.630	0.620	0.615
BiLSTM	Random Embeddings	0.800	0.650	0.680
BiLSTM	GloVe Embeddings	0.820	0.810	0.810
BiLSTM	fastText Embeddings	0.890	0.880	0.880
GLiClass	large-v3.0/Zero-Shot	0.630	0.530	0.419
GLiClass	large-v3.0/Few-Shot	0.720	0.510	0.350
GLiClass	base-v2.0-rac/Zero-Shot	0.250	0.500	0.333
GLiClass	base-v2.0-rac/Few-Shot	0.750	0.520	0.367
LLaMA2	Zero-Shot	0.740	0.510	0.348
LLaMA2	Few-Shot	0.680	0.630	0.609

the top variables contributing to the variation in the data, as PC1 reflects the greatest explanation of variance. Table 5 shows the explained variation of the first three principal components, with PC1 explaining 15.17 % of the variation. This has to be considered a small contribution, but as in the analysis of the JSD, bound by the limited availability of data, effects are most likely modest.

Following Park (2024), a subsequent ANOVA was performed on the top five contributing variables from the PC1, which can be seen in Table 6. None of them turned out to be significant, but it is still remarkable, that four of them belong to the morpho-syntactic category (auxiliary distributions and upos distributions) and one to the syntactic category (average links length), supporting the finding of Schacht and Delucchi Danhier (2025) of the tendencies to variation in sentence structure, slight as they may be.

Feature Selection and Traditional Machine Learning Classifiers

The implemented grafting algorithm iteratively selected six features per classifier according to their predictiveness (see Table 7 for the selected features for the logistic regression model and Table 8 for the features of the SVM). For both models the most predictive feature was character per token, supporting the previously presented importance of this feature. The majority of the remaining features belong to morpho-syntactic distributions, underscoring their predictiveness, as observed in the radar charts already. The classifiers were then trained on the selected features and tested with a held-out test set of 20 percent of the original

data set. To compare the performance of all models the micro-averaged F1-scores were calculated for all of the models (see Table 1). Both models performed reasonably well, with the logistic regression outperforming the SVM achieving an F1-score of 0.73 while the SVM only achieved 0.615 (see Table 1). This indicates a relatively strong predictive power of the selected structural features and supports the structural patterns of the respective sub-corpus presented above.

Training of BiLSTM Models and Inference with LLM

To test the feature-based classifiers against the model architectures that represent a richer semantic, three different iterations of a BiLSTM were trained using random, GloVe and fastText embeddings. While GloVe captures more global relations, fastText is considered to handle out-of-vocabulary more robustly, which might be relevant in the present scenario, where a specialized domain vocabulary is assumed. Ten models were trained of each iteration and the best performing model was selected. While the model with random embeddings only achieved an F1-score of 0.68—which already outperforms the SVM—the GloVe model achieved 0.81 and the fastText 0.88 (refer to Table 1). This robust performance supports the assumption that semantics might represent the variation in the sub-corpora even more accurately than structural features, as these models are considered to represent the data’s semantics due to their context-oriented architecture. Of the LLMs, except for the LLaMA 2 in the few-shot testing, all models performed poorly indicating a poor generalization,

probably due to lack of data, as this is a comparatively small dataset. The LLaMa 2 in the few-shot setting achieved an F1-score of 0.609, but the other LLMs all scored below chance (compare Table 1). This suggests that the application of LLMs in research designs with very limited data might not be an ideal choice and the more traditional model architectures might be preferable in these scenarios.

5 Discussion

This study was conducted as a continuation of the initial analysis by [Schacht and Delucchi Danhier \(2025\)](#), extending their approach by comparing the expert and layperson sub-corpora on the basis of linguistic complexity while additionally employing the DECAF and Profiling-UD frameworks to quantify divergence between the groups. This provided a measure of overall corpus-level distance, which was subsequently complemented by a more fine-grained analysis designed to identify the features that distinguish the two groups and to construct domain-specific linguistic profiles. These profiles were then used as input for the training of traditional feature-based machine learning classifiers and compared to the performance of LLMs on the classification task, thereby combining exploratory profiling with predictive modeling.

How much do the two sub-corpora diverge from each other? In terms of quantification of the divergence between the sub-corpora by JSD, the two groups can be said to diverge at a level of 0.25. While this value suggests a small yet measurable difference, the PCA and ANOVA revealed only a small proportion of explainable variation, with no variables contributing significantly. Nonetheless, an inspection of the subtleties of the linguistic profiles revealed tendencies that are in line with the earlier findings of [Schacht and Delucchi Danhier \(2025\)](#), while also pointing towards semantic variation as a promising direction for future research.

Do the linguistic profiles of the domain-experts differ from the profiles of the laypeople? The analysis of the linguistic profiles indicates that the expert group does indeed differ from the non-expert group, albeit subtly. The most prominent differences were found in raw text features, with experts producing longer tokens and longer sentences, as well as displaying greater lexical diversity. Moreover, the structural differences already suggested in the initial analysis of [Schacht and Delucchi Dan-](#)

[hier \(2025\)](#) are corroborated by the present findings, reinforcing the assumption of distinct patterns emerging in expert language use.

Do the domain-experts display more complex linguistic structures than the non-experts? The results suggest that the expert group tends to produce slightly more complex phrasal structures compared to non-experts, thus supporting the tendencies observed in the ExpLay study. However, while syntactic complexity plays a role, the predictive characteristics of the two groups appear to be more pronounced in the domain of semantics. While the effects observed in this study are comparatively small, this is not unexpected given that the ExpLay corpus represents only a small corpus of experimental data. In contrast, many studies in register analysis are based on much larger corpora, often spanning millions of tokens, such as the work of [Teich et al. \(2016\)](#) or [Kunilovskaya and Corpas Pastor \(2021\)](#). Considering this, it is not unusual to observe only small effects on limited datasets. What is noteworthy, however, is that even under these conditions the data nevertheless displays tendencies towards register-specific linguistic patterns. Small as they may be, these tendencies indicate that there is indeed detectable variation in the language use of domain-experts within the architectural domain. The differences identified in the present study are subtle, but they mirror the initial findings of [Schacht and Delucchi Danhier \(2025\)](#). In particular, experts display a tendency towards more complex sentence and phrasal structures as well as an elevated use of content words. While some of the structural features proved less predictive in isolation, a closer examination of the individual features reveals that many of these differences ultimately point towards semantic variation. All in all, the combination of several indicators—the divergence in perspective and description depth as reflected in the JSD and the accompanying features, the elevated lexical variety, longer tokens and sentences (suggesting higher content production and potentially greater information density per token, with token length in particular functioning as a proxy for the presence of content words)—all contribute to an overall picture of semantic richness. These findings mirror those of [Schacht and Delucchi Danhier \(2025\)](#), who likewise found an elevated use of content words in the expert sub-corpus. The strong performance of the context-based BiLSTM model provides further support for

this interpretation. Since BiLSTMs are particularly effective in capturing rich semantic representations of data, their superior performance compared to traditional feature-based classifiers suggests that the predictive characteristics of the corpus are rooted more strongly in semantics than in structural features. Had the classical machine learning algorithms performed comparably or even better, this would have indicated a stronger reliance on structural predictiveness. Instead, the outperformance of BiLSTM with fastText embeddings—well suited to handling out-of-vocabulary tokens that are expected to occur in specialized, domain-specific data such as the present corpus—points clearly towards the presence of predictive semantics. Taken together, these findings closely match the characterization of expert-language as the "use of specialist terminology, nominal style, and high lexical density" of Teich et al. (2016) introduced in Section 1 and support the assumption of Teich et al. (2021) of conventionalized linguistic codes among professionals of a field and those being a cognitive result of expertise emerging from the specific communicative demands in their functional context as Argamon (2019) hypothesized. Thus, while the observed differences remain minor in scope, they nevertheless indicate that domain-specific registers can be distinguished even in relatively small datasets, thereby supporting the claims of capturing representative patterns of domain variation in the line of research of linguistic profiling. Future work could extend this approach by conducting human classification experiments to examine how reliably listeners can distinguish expert from non-expert language. In addition, information-theoretic analyses could explore how differences in linguistic structure relate to patterns of information density across groups. The present study strengthens our understanding of how domain expertise manifests linguistically and opens up future directions to investigate combined structural, semantic and information theoretic characteristics of 'expert-ese'.

6 Limitations

While this study provides insights into linguistic variation between experts and non-experts, several limitations must be acknowledged. First, the dataset is relatively small, reflecting the difficulty of collecting high-quality, naturalistic elicitation data. This constrains the statistical power of the analysis and the performance of the models. Par-

ticularly classifiers typically require substantial amounts of data to be trained and perform robustly. This also affects potential fine-tuning experiments involving pre-trained models, as the limited amount of documents will most likely not be sufficient to fine-tune a pre-trained model on the classification task. The findings of the present study must therefore be interpreted as indicative rather than exhaustive. Second, as the study relies on automatically parsed data, any annotation errors may propagate through subsequent analysis, though state-of-the-art tools were used to minimize this risk. While this issue could be alleviated by more extensive manual curation of the automatic annotations, such curation is very costly in human resources and therefore beyond the scope of this study. Finally, the study focuses exclusively on the domain of architecture, which limits the generalizability of the findings. Replicating the approach across domains and with larger datasets will be essential for testing the robustness and transferability of the observed patterns.

Acknowledgments

We are grateful to the anonymous reviewers for their helpful comments. This research is partially funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

- Shlomo Engelson Argamon. 2019. [Computational register analysis and synthesis](#). *ArXiv*, abs/1901.02543.
- Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge University Press.
- Douglas Biber and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Studies in English Language. Cambridge University Press.
- Douglas Biber, Bethany Gray, and Kornwipa Poonpon. 2011. [Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?](#) *TESOL Quarterly*, 45(1):5–35.
- Dominique Brunato, Andrea Cimino, Felice Dell’Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. [Profiling-UD: a tool for linguistic profiling of texts](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France. European Language Resources Association.

- Andrea Cimino, Martijn Wieling, Felice Dell’Orletta, Simonetta Montemagni, and Giulia Venturi. 2017. Identifying predictive features for textual genre classification: the key role of syntactic features. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, CEUR Workshop Proceedings. CLiC-it 2017 : Italian Conference on Computational Linguistics ; Conference date: 11-12-2017 Through 13-12-2017.
- Kevyn Collins-Thompson. 2014. [Computational assessment of text readability: A survey of current and future research](#). *ITL - International Journal of Applied Linguistics*, 165(2):97–135.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, 47(2):255–308.
- Stefania Degaetano-Ortlieb. 2021. [Measuring informativity: The rise of compounds as informationally dense structures in 20th century Scientific English](#), pages 291–312. John Benjamins Publishing Company.
- Stefania Degaetano-Ortlieb and Andrew Piper. 2019. [The scientization of literary study](#). In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 18–28, Minneapolis, USA. Association for Computational Linguistics.
- Stefania Degaetano-Ortlieb and Elke Teich. 2018. [Using relative entropy for detection and analysis of periods of diachronic linguistic change](#). In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33, Santa Fe, New Mexico. Association for Computational Linguistics.
- Stefania Degaetano-Ortlieb and Elke Teich. 2022. [Toward an optimal code for communication: The case of scientific english](#). *Corpus Ling.. Ling.. Theory*, 18(1):175–207.
- Renate Delucchi Danhier, Barbara Mertins, Holger Mertins, and Gerold Schneider. 2025. [Entropy as a lens: Exploring visual behavior patterns in architects](#). *Journal of Eye Movement Research*, 18(5).
- A. Ersin, H. Ceren Tezeren, N. Ozunlu Pekyavas, B. Asal, A. Atabey, A. Diri, and İ Gonen. 2022. [The relationship between reaction time and gaming time in e-sports players](#). *Kinesiology*, 54(1):36–42. Doi:10.26582/k.54.1.4.
- John Gamboa, Kristina Braun, Juhani Järvikivi, and Shanley E. M. Allen. 2025. [The distributional properties of long nominal compounds in scientific articles: an investigation based on the uniform information density hypothesis](#). *Corpus Linguistics and Linguistic Theory*, 21(1):137–171.
- John C. B. Gamboa, Leigh B. Fernandez, and Shanley E. M. Allen. 2024. [Investigating the uniform information density hypothesis with complex nominal compounds](#). *Applied Psycholinguistics*, 45(2):322–367.
- M. Gellerstam. 1986. Translationese in swedish novels translated from english. In L. Wollin and H. Lindquist, editors, *Translation studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund.
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. [Learning word vectors for 157 languages](#). In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- M. A. K. Halliday. 1988/2004. [On the language of physical science](#). In Jonathan J. Webster, editor, *The Collected Works of M. A. K. Halliday (Vol. 5)*, pages 140–158. Continuum, London and New York.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. [DeBERTa: Decoding-enhanced bert with disentangled attention](#). <https://api.semanticscholar.org/CorpusID:219531210>.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long short-term memory](#). *Neural Computation*, 9(8):1735–1780.
- T. F. Jaeger. 2010. [Redundancy and reduction: Speakers manage syntactic information density](#). *Cognitive Psychology*, 61(1):23–62.
- Deborah N. Jakobi, Thomas Kern, David R. Reich, Patrick Haller, and Lena A. Jäger. 2024. [Potec: A german naturalistic eye-tracking-while-reading corpus](#). *Preprint*, arXiv:2403.00506.
- Chunzhen Jiang, Aritra Kundu, and Mark Claypool. 2020. [Game player response times versus task dexterity and decision complexity](#). In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play, CHI PLAY ’20*, page 277–281, New York, NY, USA. Association for Computing Machinery.
- I. T. Jolliffe. 2002. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York. Doi:10.1007/b98835.
- Tom S Juzek, Marie-Pauline Krielke, and Elke Teich. 2020. [Exploring diachronic syntactic shifts with dependency length: the case of scientific English](#). In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 109–119, Barcelona, Spain (Online). Association for Computational Linguistics.
- S Kullback and R A Leibler. 1951. [On information and sufficiency](#). *The Annals of Mathematical Statistics*, 22(1):79–86.
- M. Kunilovskaya and G. Corpas Pastor. 2021. [Translationese and register variation in english-to-russian professional translation](#). In V. X. Wang, L. Lim, and

- D. Li, editors, *New Perspectives on Corpus Translation Studies. New Frontiers in Translation Studies.*, . https://doi.org/10.1007/978-981-16-4918-9_6. Springer, Singapore.
- Kristopher Kyle. 2016. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. Ph.D. thesis, Georgia State University. [Doi:10.57709/8501051](https://doi.org/10.57709/8501051).
- Willem J. M. Levelt. 1989. *Speaking: From Intention to Articulation*. The MIT Press.
- S. C. Levinson. 2003. *Space in language and cognition: Explorations in cognitive diversity*. Cambridge University Press.
- Xiaofei Lu. 2010. *Automatic analysis of syntactic complexity in second language writing*. *International Journal of Corpus Linguistics*, 15(4):474–496.
- E. A. Maguire, D. G. Gadian, I. S. Johnsrude, C. D. Good, J. Ashburner, R. S. J. Frackowiak, and C. D. Frith. 2000. Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, 97(8):4398–4403.
- A. Majid, M. Bowerman, S. Kita, D. B. Haun, and S. C. Levinson. 2004. Can language restructure cognition? the case for space. *Trends in Cognitive Sciences*, 8(3):108–114.
- H. Mertins, R. Delucchi Danhier, B. Mertins, A. Schulz, and B. Schulz. 2020. The role of expertise in the perception of architectural space. In C. Leopold, C. Robeller, and U. (Hrsg. Weber, editors, *Research Culture in Architecture*, pages 279–288. Birkhäuser, Basel.
- Kanishka Misra and Kyle Mahowald. 2024. *Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.
- Max Müller-Eberstein, Rob Van Der Goot, and Anna Rogers. 2025. *DECAF: A dynamically extensible corpus analysis framework*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 351–362, Vienna, Austria. Association for Computational Linguistics.
- A. Papafragou, J. Hulbert, and J. Trueswell. 2008. Does language guide event perception? evidence from eye movements. *Cognition*, 108(1):155–184.
- Shinjae Park. 2024. *Identifying key linguistic variables of second language speaking proficiency using principal component analysis*. *Forum for Linguistic Studies*, 6(6):623–633.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. *Glove: Global vectors for word representation*. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Simon Perkins, Kevin Lacker, and James Theiler. 2003. Grafting: Fast, incremental feature selection by gradient descent in function space. *The Journal of Machine Learning Research*, 3:1333–1356.
- D. Roberson, I. Davies, and J. Davidoff. 2000. Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, 129(3):369–398.
- Raphael Rubino, Stefania Degaetano-Ortlieb, Elke Teich, and Josef van Genabith. 2016. *Modeling diachronic change in scientific writing with information density*. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 750–761, Osaka, Japan. The COLING 2016 Organizing Committee.
- Carmen Schacht and Renate Delucchi Danhier. 2025. *ExpLay: A new corpus resource for the research on expertise as an influential factor on language production*. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 216–227, Vienna, Austria. Association for Computational Linguistics.
- Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.
- D. I. Slobin. 1996. From “thought and language” to “thinking for speaking”. In J. J. Gumperz and S. C. Levinson, editors, *Rethinking linguistic relativity*, pages 70–96. Cambridge University Press.
- Ihor Stepanov, Mykhailo Shtopko, Dmytro Vodianytskyi, Oleksandr Lukashov, Alexander Yavorskyi, and Mykyta Yaroshenko. 2025. *Gliclass: Generalist lightweight model for sequence classification tasks*. Preprint, arXiv:2508.07662.
- Elke Teich, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes, and Ekaterina Lapshinova-Koltunski. 2016. *The linguistic construal of disciplinarity: A data-mining approach using register features*. *Journal of the Association for Information Science and Technology*, 67(7):1668–1678.
- Elke Teich, Peter Fankhauser, Stefania Degaetano-Ortlieb, and Yuri Bizzoni. 2021. *Less is more/more diverse: On the communicative utility of linguistic conventionalization*. *Frontiers in Communication*, 5.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. *Llama 2: Open foundation and fine-tuned chat models*. *Preprint*, arXiv:2307.09288.

Jonathan J. Webster. 2018. *18. The Language Of Science – A Systemicfunctional Perspective*, pages 345–363. De Gruyter Mouton, Berlin, Boston.

Wes McKinney. 2010. *Data Structures for Statistical Computing in Python*. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Benjamin Lee Whorf. 1956. *Language, Thought, and Reality*. Cambridge, Ma.

J. Winawer, N. Witthoft, M. C. Frank, L. Wu, A. R. Wade, and L. Boroditsky. 2007. Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, 104(19):7780–7785.

A K Wong and M You. 1985. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Trans Pattern Anal Mach Intell*, 7(5):599–609.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. *Text readability assessment for second language learners*. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. *GLiNER: Generalist model for named entity recognition using bidirectional transformer*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

A Divergence Experiments

DECAF operates by running a high-dimensional data analysis on the indexed input data and subsequently calculating the JSD (Wong and You, 1985), which is a metric based on the Kullback-Leibler divergence (KLD) (Kullback and Leibler, 1951), which in turn is a quantification of the distance between the distributions of probabilities. Unlike KLD, however, the JSD is a symmetric measure, thus it is not important which of the distributions is being referenced. The divergence was calculated on the the values of the following feature types: *Abbr, Case, Definite, Degree, ExtPos, Foreign, Gender, Mood, NumForm, NumType, Number, Person,*

Polarity, Poss, PronType, Reflex, Style, Tense, Typo, VerbForm, Voice, deprel, upos, and xpos. Out of the full set, the top six contributing features are displayed in Table 2.

Table 2: Contribution of the top 6 variables in the JSD.

Variable	Contribution score
Person	0.070
xpos	0.050
PronType	0.050
VerbForm	0.050
deprel	0.050
upos	0.040

The contribution of the use of person in the two sub-corpora is pointed out in Table 3.

Table 3: Contribution of the use of person in both groups.

Group	First Person	Third Person
Experts	0.135	0.864
Non-experts	0.071	0.928

Table 4: Contribution of the use of pronoun types in both groups.

Pronoun Type	Experts	Non-experts
Art	0.625	0.629
Dem	0.033	0.036
Ind	0.144	0.178
Int	0.001	0.002
Neg	0.005	0.005
Prs	0.143	0.105
Rel	0.049	0.041

B Linguistic Profiling Experiments

The following four radar charts display the profiles of the two sub-groups by the four main categories of the linguistic profiling analysis. The first one is depicting the raw text property profiles (see Figure 2) and lexical variety profiles (see Figure 3).

The second radar chart shows the morpho-syntactic information profiles (see Figure 4) and syntactic features profiles (see Figure 5).



Figure 2: Radar charts of the raw text property profiles.

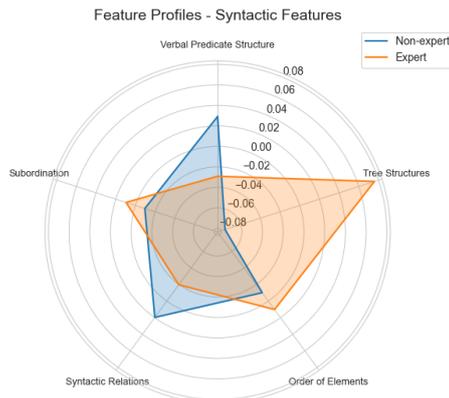


Figure 5: Radar charts of the syntactic features profiles.

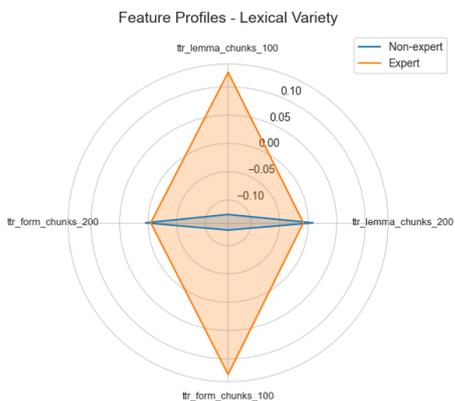


Figure 3: Radar charts of the lexical variety profiles.

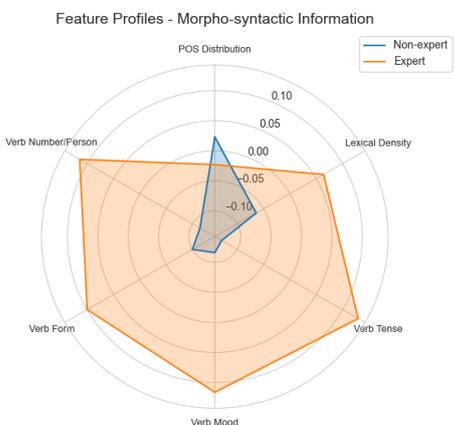


Figure 4: Radar charts of the morpho-syntactic information profiles.

C PCA Experiments

The following table show the explained variation of the PCs in the PCA (Table 5 and the results of the one-way ANOVA conducted on the contributing variables from PC1 in Table 6).

Table 5: Explained variation of the components of the PCA.

Components	Explained variation
PC1	15.170%
PC2	13.230%
PC3	9.530%

Table 6: Results of one-way ANOVA for the top 5 contributing linguistic features of PC 1.

Variable	F-statistic	p-value
avg_links_len	1.946	0.176
aux_form_dist_Fin	0.001	0.972
aux_tense_dist_Pres	0.002	0.964
aux_mood_dist_Ind	0.051	0.823
upos_dist_NOUN	0.015	0.904

D Feature Selection via Grafting

Table 7: Grafting feature selection from logistic regression coefficients and their absolute values for the selected features.

Feature	Abs. Coefficient
char_per_tok	1.663
verb_edges_dist_0	0.845
dep_dist_root	0.732
dep_dist_csubj	0.502
upos_dist_AUX	0.251
verbs_num_pers_dist_Plur+	0.138

Table 8: Grafting feature selection from SVM coefficients and their absolute values for the selected features.

Feature	Abs. Coefficient
char_per_tok	0.491
upos_dist_PART	0.453
dep_dist_csubj	0.173
dep_dist_mark	0.172
ttr_lemma_chunks_200	0.124
verbs_tense_dist_Past	0.029

E Model Evaluation

The following table shows the full classification reports from all tested models.

F Profiling Features

The following list displays the subset of features from the Profiling-UD tool used in the present study: *n_sentences*, *n_tokens*, *tokens_per_sent*, *char_per_tok*, *ttr_lemma_chunks_100*, *ttr_lemma_chunks_200*, *ttr_form_chunks_100*, *ttr_form_chunks_200*, *upos_dist_ADJ*, *upos_dist_ADP*, *upos_dist_ADV*, *upos_dist_AUX*, *upos_dist_CCONJ*, *upos_dist_DET*, *upos_dist_NOUN*, *upos_dist_NUM*, *upos_dist_PART*, *upos_dist_PRON*, *upos_dist_PROPN*, *upos_dist_PUNCT*, *upos_dist_SCONJ*, *upos_dist_VERB*, *upos_dist_X*, *lexical_density*, *verbs_tense_dist_Past*, *verbs_tense_dist_Pres*, *verbs_mood_dist_Imp*, *verbs_mood_dist_Ind*, *verbs_mood_dist_Sub*, *verbs_form_dist_Fin*, *verbs_form_dist_Inf*, *verbs_form_dist_Part*, *verbs_num_pers_dist_Plur_*, *verbs_num_pers_dist_Plur_1*, *verbs_num_pers_dist_Plur_3*, *verbs_num_pers_dist_Sing_*, *verbs_num_pers_dist_Sing_1*,

verbs_num_pers_dist_Sing_3, *aux_tense_dist_Past*, *aux_tense_dist_Pres*, *aux_mood_dist_Ind*, *aux_mood_dist_Sub*, *aux_form_dist_Fin*, *aux_form_dist_Inf*, *aux_num_pers_dist_Plur_3*, *aux_num_pers_dist_Sing_1*, *aux_num_pers_dist_Sing_3*, *verbal_head_per_sent*, *verbal_root_perc*, *avg_verb_edges*, *verb_edges_dist_0*, *verb_edges_dist_1*, *verb_edges_dist_2*, *verb_edges_dist_3*, *verb_edges_dist_4*, *verb_edges_dist_5*, *verb_edges_dist_6*, *avg_max_depth*, *avg_token_per_clause*, *avg_max_links_len*, *avg_links_len*, *max_links_len*, *avg_prepositional_chain_len*, *n_prepositional_chains*, *prep_dist_1*, *prep_dist_2*, *prep_dist_3*, *obj_pre*, *obj_post*, *subj_pre*, *subj_post*, *dep_dist_acl*, *dep_dist_advcl*, *dep_dist_advmod*, *dep_dist_amod*, *dep_dist_appos*, *dep_dist_aux*, *dep_dist_aux:pass*, *dep_dist_case*, *dep_dist_cc*, *dep_dist_ccomp*, *dep_dist_compound*, *dep_dist_compound:prt*, *dep_dist_conj*, *dep_dist_cop*, *dep_dist_csubj*, *dep_dist_dep*, *dep_dist_det*, *dep_dist_det:poss*, *dep_dist_expl*, *dep_dist_expl:pv*, *dep_dist_fixed*, *dep_dist_iobj*, *dep_dist_mark*, *dep_dist_nmod*, *dep_dist_nmod:poss*, *dep_dist_nsubj*, *dep_dist_nsubj:pass*, *dep_dist_nummod*, *dep_dist_obj*, *dep_dist_obl*, *dep_dist_parataxis*, *dep_dist_punct*, *dep_dist_root*, *dep_dist_xcomp*, *principal_proposition_dist*, *subordinate_proposition_dist*, *subordinate_post*, *subordinate_pre*, *avg_subordinate_chain_len*, *subordinate_dist_1*, *subordinate_dist_2*, *subordinate_dist_3*

CroCoSyn: A Cross-Lingual and Cross-Model Corpus of LLM-Generated Film Synopses

Louis Escoufflaire

MIT Trope Tank — Massachusetts Institute of Technology
Institute for Language and Communication — UCLouvain
escouf@mit.edu – louis.escoufflaire@uclouvain.be

Abstract

We introduce CroCoSyn, a controlled, cross-lingual and cross-model corpus of 25,920 LLM-generated film synopses in English and French. Each synopsis is generated under systematically varied conditions, including model type, temperature, genre, protagonist gender, and narrative constraints, and enriched with structured metadata capturing characters and their relationships. Comparing Mistral and Llama across different model temperature degrees, CroCoSyn enables fine-grained analysis of narrative content, style, and character representation across models and languages. The corpus supports research on gender and cultural biases and story generation evaluation, and provides a foundation for comparative studies between LLM-generated and human-written narratives.

1 Introduction

Stories are central to how societies transmit values, norms, and cultural models, from myths and novels to films and news narratives (Eliade, 1961; Bruner, 2010; Gottschall, 2012). They also encode implicit assumptions and stereotypes about gender, culture, and power (Lovatt, 2013; Casey et al., 2021).

Recent advances in large language models (LLMs) allow narrative texts to be generated at scale and on demand, supporting creative, educational, and journalistic applications (Ray, 2023; Cardon et al., 2023). However, LLM outputs reflect their training data (largely web-based and Western-centric), potentially reproducing stylistic conventions, cultural biases, and representational asymmetries (Baack, 2024).

Despite growing interest in studying LLM-generated text, resources for systematically comparing narratives across languages, models, and other parameters are lacking. To address this gap, we introduce CroCoSyn, a cross-lingual, cross-model corpus of 25,920 film synopses generated by Llama-3 and Mistral. Each synopsis is produced

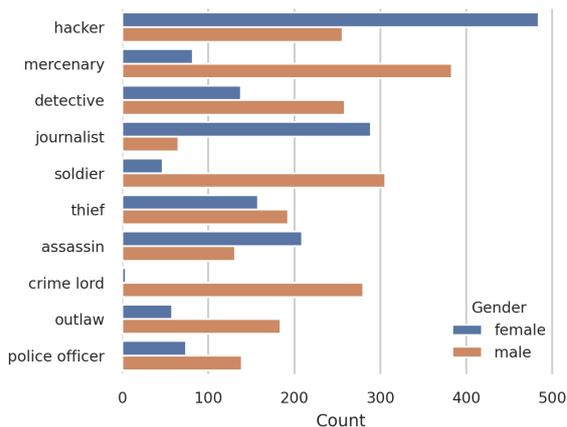


Figure 1: Most frequent character occupations by gender in the 5,192 *action* film synopses of the CroCoSyn corpus (non-binary and other genders were not included on the graph for visualization purposes).

under a fully balanced factorial design controlling language, model, temperature, genre, protagonist gender, and other narrative parameters. Beyond raw text, synopses are enriched with structured metadata capturing characters and relationships, enabling fine-grained quantitative and qualitative analyses. CroCoSyn provides a unique resource to study narrative generation, investigate gender and cultural biases, and compare outputs across models and languages. It also supports broader research on story evaluation, narrative modeling, and computational creativity. The CroCoSyn corpus will be released as an open-source resource and made freely available on GitHub.

2 Related Work

2.1 Story Generation with LLMs

Large language models such as GPT-4 and Llama-3 generate fluent and stylistically convincing long-form narratives (Achiam et al., 2023; Dubey et al., 2024), yet they often struggle with global coherence, character consistency, and overall plot

structure (Montfort and y Pérez, 2023). Researchers have highlighted the influence of prompt design and model architectures on narrative outcomes (Bender et al., 2021; Harmon and Rutman, 2023). In addition, extensive research has documented gender and cultural stereotypes in generated text, including biased associations between professions and genders, differential emotional framing, and underrepresentation of marginalized groups (Bolukbasi et al., 2016; Li and Bamman, 2021; Barroso da Silveira and Alves Lima, 2024). In narrative contexts, such biases can emerge not only lexically but also through higher-level story elements, such as character agency, role allocation, and genre-specific tropes (Rettberg, 2024). These findings highlight the need for carefully controlled narrative corpora that allow researchers to isolate the effects of individual generation variables on representational outcomes.

2.2 Corpora of Human and LLM-generated Stories

Several corpora of human-written narratives exist. ROCStories (Mostafazadeh et al., 2016) has been widely used for modeling hierarchical story generation and LLM-induced story continuation (Cavazza, 2025). Broader multilingual datasets such as StoryDB (Tikhonov et al., 2021) support cross-lingual comparative narrative research, while STORIUM (Akoury et al., 2020) provides richly annotated collaborative stories for machine-in-the-loop generation and narrative quality assessment. Datasets of LLM-generated narratives remain comparatively scarce. TinyStories (Eldan and Li, 2023) offers simple, controlled narratives to probe small model capabilities and uncover biases (Gunti and Supriya, 2025), TF1-EN-3M (Nadas et al., 2025) contains 3 million short fables generated by Llama-3, and GPT-WritingPrompts (Huang et al., 2024) consists of Reddit-formatted short stories generated by GPT-3.5. To our knowledge, no existing corpus provides LLM-generated film synopses across multiple models and languages from the same prompt, making CroCoSyn a unique resource.

3 Corpus Design

The CroCoSyn corpus was designed with the goal of enabling systematic comparison across two languages, two models, and multiple linguistic and narrative conditions. All variables leveraged in corpus generation are balanced. This approach allows

for fine-grained analysis of how generation parameters influence narrative structure, style, and content. The corpus was also designed to be sufficiently large and diverse to support both quantitative analysis and close reading.

3.1 Models

The corpus was generated using two large language models, Llama-3 (Dubey et al., 2024) and Mistral (Jiang et al., 2023). These models were selected for their widespread use, open or semi-open availability, and because they originate from companies based in different countries (the United States for Llama-3 and France for Mistral), a feature relevant to the cross-lingual and cross-cultural aspects of the corpus. We used the "small" versions of the models, respectively Llama-3-8B-Instruct and Mistral-7B. These versions were considered sufficient for generating short texts such as film synopses, while keeping resource requirements manageable. Each model generated exactly half of the corpus, with identical prompting structures. Llama-3 and Mistral were both trained on partially overlapping large-scale web and textual corpora, but the exact composition of their training datasets differs. Such differences can affect model knowledge, style, and multilingual capabilities, which can be investigated using the CroCoSyn corpus.

To investigate the effect of decoding stochasticity on narrative generation, two temperature conditions were defined: *low* and *high*. The low-temperature condition was set to 0.1 for both models, favoring rather deterministic and conservative outputs. The high-temperature condition was set to 0.8 for Llama and 1.0 for Mistral, following differences observed in both models' responses to varying levels of temperature during a preliminary research stage and aiming to produce comparable levels of variability.

3.2 Generation Variables

The corpus follows a full factorial design in which each synopsis is generated under a unique combination of controlled parameters. These parameters include language, model, model temperature, film genre, target length, writer nationality, writer gender, protagonist gender, temporal setting, and evaluative framing through prompt adjectives. All variables are evenly distributed across the corpus, ensuring that no condition is over- or under-represented. This balanced structure enables direct comparison across individual dimensions as

well as the analysis of interaction effects between them. The distributions of variables combining into 25,920 different synopses are presented below, together with a brief description of each variable.

- *language*: the language of both the prompt and the generated synopsis, English or French.
- *model*: the large language model used for generation, Mistral (Mistral-7B) or Llama (Llama-3-8B-Instruct).
- *model temperature*: the decoding temperature used during text generation, controlling the level of stochasticity. Two relative levels are defined: *low* (0.1 for both models) and *high* (0.8 for Llama and 1.0 for Mistral).
- *genre*: the intended film genre specified in the prompt. Five genres are included: crime, romance, comedy, action, and drama.
- *target length*: the approximate length requested in the prompt, set to either 250 or 350 words.
- *writer nationality*: the nationality assigned to the writer persona in the system prompt, either United States or France.
- *writer gender*: the gender assigned to the writer persona in the system prompt (male, female, or unspecified).
- *protagonist gender*: the gender of the main character of the story, as constrained by the prompt (male, female, or unspecified).
- *temporal setting*: the narrative time frame suggested in the prompt (past, present, or unspecified).
- *adjective*: an optional adjective framing the synopsis evaluatively (*good*, *great*, *compelling*, *fun*, *dark*, or unspecified).

3.3 Prompt Template

All synopses are generated using a shared prompt template, with controlled variables injected through parameterized slots. To ensure comparability across languages, English and French prompts were designed to be semantically equivalent. The prompts explicitly require that all main characters be named, facilitating character-level analysis. No additional stylistic constraints were imposed beyond the controlled variables. The English and French equivalent prompts (and system prompts) below were used to generate two of the 25,920 synopses in the corpus (two per model):

English prompt:

- Model: *Mistral*

- Temperature: *low*
- System prompt: You are a *female* script writer from *France*.
- User prompt: Write a *good* film synopsis in English (around 250 words) in the *action* genre, *set in the past*, with a *male* protagonist. All main characters should be named.

French prompt:

- Model: *Mistral*
- Temperature: *low*
- System prompt: Vous êtes *une* scénariste *française*.
- User prompt: Rédigez un synopsis de film *de qualité* en français (environ 250 mots) dans le genre *du film d'action*, *situé dans le passé*, dont le protagoniste est *une femme*. Tous les personnages principaux doivent être nommés.

3.4 Structured Data Enrichment

Each synopsis in the CroCoSyn corpus is accompanied by structured metadata capturing narrative elements such as the film title, character information, and inter-character relationships. To extract this information from each synopsis, we used a zero-shot pipeline relying on Llama-3-8B-Instruct, which was given the entry and prompted to output a valid JSON structure containing the following elements:

- *title*: the film's title.
- *characters*: a list of characters, each with a name, gender, occupation, character traits, and role (protagonist, secondary, antagonist).
- *relationships*: a list of relationships between characters, specifying source, target, and relationship type (e.g., family, romantic, enemy).

Each JSON output is automatically validated and corrected if necessary, guaranteeing that all structured metadata adheres to a consistent schema.

4 Corpus Description

Future research will be devoted to the results of a comparative analysis of character descriptions and roles in synopses generated across the different variables. Here, we present an example entry from the corpus as well as some preliminary results.

4.1 Example

Movie Title: "The Shadow of Versailles"

Synopsis:

In 18th century France, amidst the grandeur of the Palace of Versailles, a young and fearless female spy, Colette Dumont, navigates the treacherous world of espionage. Colette, a skilled thief and assassin, is recruited by the enigmatic and charismatic King Louis XV to infiltrate the inner circle of the ruthless and cunning Cardinal de Rohan. Colette's mission is to uncover the Cardinal's plans to overthrow the King and restore the power of the French monarchy to the nobility.

As she navigates the intricate web of court politics, Colette finds herself drawn to the Cardinal’s charming and cunning nephew, the Marquis de la Croix. As Colette delves deeper into the Cardinal’s plans, she must confront her own demons and confront the darkness within herself. With the help of her trusted ally, the street-smart and resourceful Pierre, Colette must use all her skills and cunning to outwit the Cardinal’s henchmen and prevent a catastrophic war.

Characters:

- *Colette Dumont* — female, spy; traits: fearless, skilled, charismatic; role: protagonist
- *King Louis XV* — male, king; traits: enigmatic, charismatic; role: secondary
- *Cardinal de Rohan* — male, cardinal; traits: ruthless, cunning; role: antagonist
- *Marquis de la Croix* — male, noble; traits: charming, cunning; role: secondary
- *Pierre* — male, unspecified; traits: resourceful; role: secondary

Relationships:

- *Colette Dumont* → *King Louis XV* — employer
- *Colette Dumont* → *Cardinal de Rohan* — enemy
- *Colette Dumont* → *Marquis de la Croix* — romantic
- *Colette Dumont* → *Pierre* — friend
- *Cardinal de Rohan* → *Marquis de la Croix* — family

This example was generated using the prompt and parameters of the English example prompt presented in section 3.3. It presents a film synopsis along with the metadata associated to it in the corpus: film title, character information (name, gender, occupation, traits, role) and relationships.

4.2 Preliminary quantitative results

While qualitative analysis of the synopses output by the models is possible, for example by examining the style and narrative structure of the example presented in the previous section, the metadata associated to each synopsis allows for quantitative analysis of the corpus across several dimensions. In this early presentation of results, we focus on trends related to the gender of characters model personas in the CroCoSyn corpus.

Figure 1 shows that some occupations are significantly more often attributed to male characters (*mercenary, detective*), while others are more often female (*hacker, journalist*). Overall, such a plot highlights gendered trends in role allocation in stories generated by the models, suggesting that certain occupations are more stereotypically associated with one gender than the other.

Figure 2 suggests that attributing a gender persona to the LLM tends to influence its generation towards stereotyped outputs: words that emphasize conflict and crime (*redemption, deadly, under-*

world) are more common in “male-written” stories, and they tend to include more male characters (as shown by the male names appearing). Female persona-associated words include a higher frequency of female character names, and words related to relationships and emotions (*mother, secret, closer*).

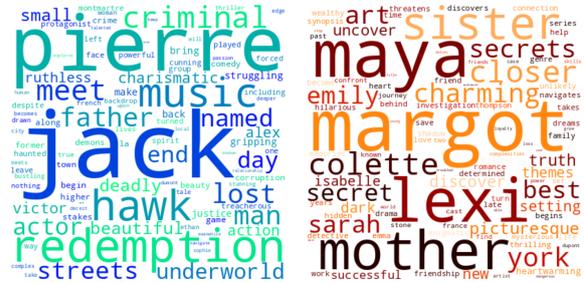


Figure 2: WordClouds of words most distinctive to all synopses generated by LLMs with male- (left) and female- (right) writer personas, as specified in the system prompt. Word importance was computed using log-odds ratios of normalized word frequencies among the top 200 words, after removing stopwords and non-informative tokens. Word size reflects the strength of association with each persona.

5 Conclusion

The CroCoSyn corpus provides a large, controlled, and cross-lingual dataset of LLM-generated film synopses, enriched with structured narrative metadata. This resource enables systematic investigation of how the choice of a specific model, language, temperature, and prompt variables influence narrative content, style, and character representation. Beyond bias and representational studies, the corpus supports applications in narrative modeling, story generation evaluation, and computational creativity research. It opens avenues for both quantitative and qualitative analyses, providing a foundation for future work on understanding and improving the sociocultural and structural properties of LLM outputs.

Looking ahead, we plan to extend the corpus to additional languages, models, and narrative variables, further broadening the scope for cross-lingual and cross-model analyses. A first line of research we aim to pursue with the corpus is a systematic comparison between LLM-generated and human-written synopses, to better understand potential representation imbalance in machine-generated storytelling.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. [Storium: A dataset and evaluation platform for machine-in-the-loop story generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Story generation dataset and platform.
- Stefan Baack. 2024. A critical analysis of the largest source for generative ai training data: Common crawl. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2199–2208.
- Julia Barroso da Silveira and Ellen Alves Lima. 2024. Racial biases in ais and gemini’s inability to write narratives about black people. *Emerging Media*, 2(2):277–287.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.
- Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to home-maker? debiasing word embeddings. *Advances in neural information processing systems*, 29.
- Jerome Bruner. 2010. Narrative, culture, and mind. *Telling Stories: Language, Narrative, and Social Life*, 46.
- Peter Cardon, Carolin Fleischmann, Jolanta Aritz, Minna Logemann, and Jeanette Heidewald. 2023. The challenges and opportunities of ai-assisted writing: Developing ai literacy for the ai age. *Business and Professional Communication Quarterly*, 86(3):257–295.
- Kennedy Casey, Kylee Novick, and Stella F Lourenco. 2021. Sixty years of gender representation in children’s books: Conditions associated with overrepresentation of male versus female protagonists. *Plos one*, 16(12):e0260566.
- Marc Cavazza. 2025. Large language models preserve semantic isotopies in story continuations. *arXiv preprint arXiv:2510.04400*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- Ronen Eldan and Yuanzhi Li. 2023. Tinstories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.
- Mircea Eliade. 1961. *Mythes, rêves et mystères*. Gallimard.
- Jonathan Gottschall. 2012. *The Storytelling Animal: How Stories Make Us Human*. Houghton Mifflin Harcourt.
- Geethika Gunti and M Supriya. 2025. Uncovering hidden narratives: Discovering and classifying archetypes in tiny stories. In *2025 International Conference on Advanced Computing Technologies (ICoACT)*, pages 1–6. IEEE.
- Sarah Harmon and Sophia Rutman. 2023. Prompt engineering for narrative choice generation. In *International Conference on Interactive Digital Storytelling*, pages 208–225. Springer.
- Xi Yu Huang, Krishnapriya Vishnubhotla, and Frank Rudzicz. 2024. The gpt-writingprompts dataset: A comparative analysis of character portrayal in short stories. *arXiv preprint arXiv:2406.16767*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Lucy Li and David Bamman. 2021. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the 3rd Workshop on Narrative Understanding*, pages 48–55.
- Helen Lovatt. 2013. *The Epic Gaze: Vision, Gender and Narrative in Ancient Epic*. Cambridge University Press.
- Nick Montfort and Rafael Pérez y Pérez. 2023. Computational models for understanding narrative. *Revista de Comunicação e Linguagens*, 58:97–117.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. [A corpus and cloze evaluation for deeper understanding of commonsense stories](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, CA. Association for Computational Linguistics.
- Mihai Nadas, Laura Diosan, Andrei Piscoran, and Andreea Tomescu. 2025. Tf1-en-3m: Three million synthetic moral fables for training small, open language models. *arXiv preprint arXiv:2504.20605*.

Partha Pratim Ray. 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154.

Jill Walker Rettberg. 2024. How generative ai endangers cultural narratives. *Issues in Science and Technology*, 40(2):77–79.

Alexey Tikhonov, Igor Samenko, and Ivan P. Yamshchikov. 2021. [Storydb: Broad multi-language narrative dataset](#). *Computing Research Repository*. ArXiv:2109.14396.

Identity Without Action: Rethinking Collective Action Models in Disinformation Research

Lorella Viola

Vrije Universiteit Amsterdam / De Boelelaan 1105

1081 HV

Amsterdam, The Netherlands

l.viola@vu.nl

Abstract

Despite the rapid growth of disinformation research, the fundamental reasons behind user engagement with such content remain poorly understood. Recently, several scholars have suggested that researchers should study engagement with disinformation as a form of collective action (CA). Drawing on Social Identity Theory (SIT) and the Social Identity Model of Collective Action (SIMCA), this study empirically verifies this assumption by testing it across two distinct linguistic communities, English and Spanish. Specifically, it investigates whether mobilizing CA language functions as a uniform predictor of engagement, or if engagement is primarily driven by community specific identity dynamics. The experiment analysed a bilingual corpus of 4,035 X (formerly Twitter) posts associated with conspiracy theory and disinformation-related hashtags (e.g., #Agenda2030, #TheGreatReset). Using a mixed-methods approach combining BERTopic for narrative discovery, non-parametric statistical testing and Random Forest Regressor, we disentangled the effects of language presence from community behaviour. The results reveal that the Spanish community exhibits a higher baseline engagement compared to the English community indicating that engagement is primarily driven by macro-level community norms (i.e., identity) rather than micro-level linguistic triggers. We argue that rather than treating mobilizing language as a uniform predictor of engagement, future application of SIMCA in disinformation research should account for these identity-based baseline differences.

1 Introduction

Online disinformation, understood here as the deliberate dissemination of false or misleading content with the potential to cause public harm (Tucker et al., 2018), has surged in recent years, particularly in the wake of the COVID-19 crisis. Widely acknowledged as a major threat to public and indi-

vidual safety, substantial research has examined its structure (Van Prooijen and Douglas, 2018; Van Prooijen and Van Vugt, 2018), spread (Bonnie et al., 2021), impact (Simms et al., 2020; Stabile et al., 2019; Chen et al., 2020), and content (Wiggins, 2023; Demata et al., 2022; Fallis, 2009). Scholars have also investigated the role of conspiracy theories and disinformation in shaping public perceptions and decision-making (Chen et al., 2021; Yagi et al., 2024), as well as the influence of network structures and user interactions in amplifying disinformation (Quintana et al., 2022; Gunaratne et al., 2019). More recently, attention has turned to understanding the deeper mechanisms through which disinformation persuades individuals to accept unlikely or false narratives, giving more importance to the cognitive and identity-based factors that may explain why individuals engage with disinformation (Reddi et al., 2023; Bastick, 2021; Butter and Knight, 2020).

Despite the rapid growth of this field, however, the fundamental reasons behind user engagement with disinformation remain poorly understood. Some researchers argue that this gap stems from contradictory and fragmented findings (Birchall and Knight, 2022; Kirchner and Reuter, 2020), while others point to the overly functionalist approach to disinformation, treating it as merely ‘the opposite of true’ and overlooking its cognitive and subjective dimensions (Viola, 2025b; Reddi et al., 2023; Bastick, 2021). This often leads to disinformation consumers being dismissed as irrational or paranoid actors and to counter-measures being mostly inefficient (Alava et al., 2017; Conway, 2017; Johnson, 2018; Mølmen and Ravndal, 2021; Reicher and Haslam, 2016). This scholarship also contends that engagement with disinformation is not merely an act of passive belief but an active discursive process, where individuals construct and negotiate their cultural identities. It further argues that interacting with disinformation, through

shares, comments, reactions, and reposts, constitutes a discursive practice that equally encodes collective agency. This behaviour would be shaped by perceived injustice, disillusionment with mainstream media, and the affordances of alternative information ecosystems (Wintterlin et al., 2023).

These arguments are supported by two key considerations. First, much of the current research on disinformation remains disproportionately focused on a small subset of industrialized democracies, particularly the United States and the United Kingdom (Bajaj, 2024). A study by Seo & Faris (Seo and Faris, 2021) found that 62.8% of empirical studies published in communication journals between 2015 and 2020 relied on U.S.-based data (p. 1166). Scholars such as Bajaj (Bajaj, 2024) highlight that disinformation is not a universal phenomenon and that this geographical bias distorts our understanding of its cultural dimensions. Consequently, mitigation efforts that ignore these cultural dynamics risk being ineffective.

Second, disinformation has been linked to citizens' decrease of trust in mainstream media and other sources of authoritative information (MacFarquhar, 2016; Lewis and Marwick, 2017; Allcott and Gentzkow, 2017), mistrust of establishment political figures and institutions and increased acceptance of, or indeed support for, fringe, anti-establishment or radical actors and movements (Beauchamp, 2019; Amlinger, 2022; Reichardt, 2022). Research on the 2020 health crisis, for example, has demonstrated that COVID-19 disinformation motivated individuals to protest by offering a sense of agency and empowerment (Reichardt, 2022; Amlinger, 2022; Birchall and Knight, 2022). Thus, according to this view, even unlikely or improbable disinformation narratives succeed in mobilizing individuals by fostering a belief in their capacity to effect change. In this sense, online participation, e.g., expressing dissent, signing petitions, or sharing content, would function as a low-cost form of collective action (CA) (Brunsting and Postmes, 2002). Through social media, users would be able to signal group membership, express opposition to elites, and reinforce a shared identity, transforming engagement with disinformation into a performative act of resistance. Online activities such as sharing or liking would in this way offer an easy and effective way for people to express dissent with others or to demonstrate their belonging to a group.

Building on this literature, the present study in-

vestigates engagement with disinformation on social media through the lens of social identity and CA, while explicitly questioning whether these frameworks adequately capture platform-mediated interaction. Rather than assuming that observable engagement reflects intentional mobilization or participatory efficacy, the study examines whether CA language functions as a universal engagement booster, or if its apparent effect is a byproduct of the higher baseline activity inherent to specific linguistic communities. To this end, we analyse a bilingual corpus of 4,085 English and Spanish posts from X associated with disinformation and conspiracy theory narratives. By combining BERTopic for narrative discovery with non-parametric statistical testing (Mann-Whitney U) and Random Forest Regressor to predict engagement levels, we assesses whether call-to-action (CTA) vocabulary provides a consistent engagement effect (i.e., likes, reposts, quotes, replies) across groups, or if engagement is primarily driven by identity-based community behaviour. This distinction is central in refining how CA models are applied to social networks and reassessing the validity of engagement metrics as proxies for real-world mobilization connected to disinformation.

2 User engagement, disinformation, and collective action

The scholarly literature on user engagement in online and social media contexts has approached this phenomenon through various conceptual frameworks and methodologies. Engagement is often conceptualized as user-initiated actions that contribute to value co-creation, as proposed by Brodie et al. (Brodie et al., 2013). This broad definition underscores the interplay between behavioural, cognitive, and emotional dimensions of engagement, emphasizing the need to explore its motivations and nuances. Shao (Shao, 2009) categorized user interaction into three primary behaviours: consumption (viewing and reading), participation (interacting with content), and production (creating and uploading content). Following this framework, researchers have examined how engagement manifests across platforms, particularly differentiating between active participation and passive consumption. On Facebook and YouTube, active engagement involves actions such as liking, commenting, and sharing, whereas passive engagement consists of clicking, watching, or hovering over content

(Kaur et al., 2019; Khan, 2017). On X, active engagement further includes reposting, quoting, and following (Chen, 2011). Studies also highlight the prevalence of passive users (often called ‘lurkers’), who primarily consume content without actively engaging, comprising up to 90% of users in many online communities (Nonnecke and Preece, 1999; Preece et al., 2004). This contrast between active contributors and passive consumers underscores the need to understand what motivates users to actively engage with content, particularly disinformation.

Due to the urgency of the topic, recent research has therefore explored the drivers of engagement with disinformation and fake news on social media. Emotionally charged content has been identified as one of the strongest amplifiers of engagement, with sensationalized headlines and narratives strategically crafted to trigger emotional responses such as fear, anger, and anticipation, thus encouraging interaction and dissemination (Horner et al., 2023). Additionally, visual elements seem to play a significant role in enhancing credibility and audience response (Cao et al., 2020; Viola, 2025a). Features such as clickbait, emotionally charged language, and references to specific individuals, organizations, or events further heighten emotional resonance, thereby boosting engagement (Ali et al., 2023). Other factors influencing the likelihood of sharing disinformation would include fear of missing out, source credibility, information quality, cognitive overload, and social media fatigue (Kumar et al., 2020; Islam et al., 2020). These content strategies would capitalize on psychological triggers to grab attention, manipulate perceptions, and enhance virality. This study builds upon these findings and integrates the Social Identity Theory (SIT) (Tajfel and Turner, 2004) with the SIMCA model (Turner, 1991; Tajfel and Turner, 2004; van Zomeren et al., 2008) to provide a theoretical lens for understanding engagement with disinformation.

3 Methodology

This study integrates SIT (Tajfel and Turner, 2004) with the SIMCA model (van Zomeren et al., 2008) to provide a theoretical lens for testing engagement with disinformation. Rather than accepting these models as given, we use them to formulate competing hypotheses regarding the drivers of user engagement. Engagement is operationalized as the cumulative sum of likes, shares, reposts, quotes, and replies associated with each post, as found in

the literature¹ (Chen, 2011).

SIT posits that individuals derive a significant portion of their self-concept from their perceived membership in social groups. In the context of this study, the “in-group” is operationalised as the linguistic community (i.e., Anglosphere vs. Hispanosphere). The author acknowledges that language is not a one-to-one substitute for cultural identity (Edwards, 2009). At the same time, however, in the context of transnational disinformation particularly on social media, language barriers often form the primary perimeter of information ecosystems. Evidence shows that language is not merely symbolic but also structural: it organizes who interacts with whom and, by extension, what information circulates within which audiences. With regard to Twitter specifically, Hale (2014) demonstrates that its connectivity is strongly stratified by language, with interaction patterns clustering within language communities and comparatively fewer ties crossing language boundaries; multilingual users can bridge communities, but they do not erase the underlying segmentation. Similarly, Eleta and Golbeck’s work on multilingual Twitter (2014) finds that most discourse remains anchored within language-specific public confirming that interaction on platforms frequently aligns with language-defined communities. Crucially, SIT suggests that different groups may develop distinct norms of interaction. Therefore, high engagement levels within a specific linguistic community may not necessarily reflect the quality of the content, but rather a community-specific baseline of expressive responding (Winterlin et al., 2023). This framework helps explain why distinct engagement cultures might emerge independent of specific mobilizing cues, helping us to test the valence of SIT for disinformation research.

The SIMCA framework on the other hand posits that collective action is driven by three psychological predictors: social identity, perceived injustice, and participatory efficacy (van Zomeren et al., 2008). Applied to disinformation, SIMCA implies that narratives framing mainstream institutions as corrupt (Injustice) or urging users to “wake up” (Efficacy) should trigger a transition from passive consumption to active distribution. If SIMCA is universally applicable to social media disinforma-

¹While high engagement metrics can indicate both support and heated opposition, from the SIMCA theoretical perspective, both reactions indicate that the content successfully triggered an engagement response, regardless of valence.

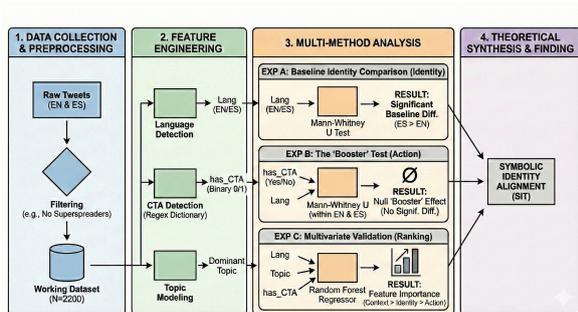


Figure 1: Research design: methodology overview.

tion, the presence of CA language should act as an engagement booster regardless of the language spoken, due to its explicit signalling of injustice and efficacy. By contrast, if engagement is symbolic, interaction may be driven primarily by the user’s cultural baseline, with mobilizing language having little to no additive effect. We test this hypothesis using non-parametric statistical testing (Mann-Whitney U). To further validate the relative impact of mobilizing language compared to identity and contextual factors, we trained a Random Forest Regressor to predict engagement levels.

To identify CTA expressions without relying on a priori assumptions, we adopted a two-step ‘Human-in-the-Loop’ approach. We first applied Embedding-Based Topic Modelling using BERTopic (Grootendorst, 2022) to identify latent narrative clusters, the most driving topics, and a post excerpt representative of each topic. For each topic, we then qualitatively inspected the top 10 representative terms associated with mobilization. From these terms, we extracted a lexicon of explicit CTA expressions such as ‘Resist the reset’, ‘donot-comply’), power-laden language (e.g., us vs. them frames and agency attribution such as ‘Say no to digital IDs’, ‘Rise up!’) and thematic organisation (e.g., New World Order, WEF Puppets). A post was coded as having CTA expressions only if it contained language from this manually validated lexicon. This ensured that the classification was interpretable and not subject to the noise of probabilistic topic modelling. Additionally, the author applied Critical Discourse Analysis (CDA) (Dijk, 1985, 1997) on a randomised sample of 200 posts (100 posts per language) to extract further expressions. The list of CTA expressions is provided in Table 4. A visual representation of the workflow is provided in Figure 1.

4 Data-set

Data retrieval was conducted through targeted queries that extracted posts containing specific hashtags, including #Agenda2030, #The GreatReset, #NewWorldOrder, #wefpuppets. These hashtags have been found in the literature as typically associated with disinformation, fake news, and conspiracy theories (Laquière, 2025; Christensen and Au, 2023; Sa’ad Abdullahi and Pindiga, 2023). The full list of the seed hashtags is provided in the Appendix. Additionally, the data-set was enriched with information about content diffusion and filtered for posts drawn from sources known for low credibility, as identified in the Iffy+ list (Golding, 2025), which catalogues 2,042 outlets flagged by professional fact-checkers as non-reliable sources, a method commonly employed in prior work (DeVerna et al., 2024; Yang et al., 2021)².

The data-set was later pseudonymised to remove any identifiable references and it can be provided upon request to the author. The working data-set covers 394 days from 1 August 2022 to 30 August 2023. It contains 4,035 posts in two languages almost identically distributed (2,000 in English - EN, 2,035 in Spanish - ES) thus making the two sets highly comparable. The data-set also includes several attributes such as the post texts, the hashtags, likes, replies, reposts, shares, and quotes count. A detailed description is given in Table 7 in the Appendix.

5 Analysis

5.1 Superspreaders and bot activity

First we remove superspreaders and bot activity from the dataset. Superspreaders were identified using a multi-dimensional engagement and network-activity criterion, operationalized according to three criteria: per-user engagement metrics, conversational out-degree, and thresholding (DeVerna et al., 2024). Per-user engagement was calculated by computing three summary statistics on repost volume and three on reply volume for each user in the data-set as explained below:

- **Repost sum** (rt_sum_i): the total number of times user i ’s posts were reposted.

²The author acknowledges that while this strategy allows for broad coverage, it cannot account for the fact that not all the posts in the data-set contain links to external content and that individual articles from a low-credibility outlet may still be accurate.

- **Repost mean** (rt_mean_i): the average number of reposts per post for user i .
- **Repost max** (rt_max_i): the highest repost count received by any single post of user i .
- **Reply sum** (rp_sum_i): the total number of replies received by user i 's posts.
- **Reply mean** (rp_mean_i): the average number of replies per post for user i .
- **Reply max** (rp_max_i): the maximum number of replies received by any single post of user i .

These six metrics captured both total volume (sum), typical activity level (mean), and individual peaks (max). Conversational out-degree was computed to assess how broadly each user initiated or contributed to discussions. This was operationalized by constructing a directed reply network in which each edge represented a reply from one user to another. A user's out-degree thus indicated the number of distinct users they replied to. Finally, the top five percent of users on each metric were identified independently to define the superspreader threshold. A user was designated a superspreader if they ranked in the top 5 percentile on any of the five measures (aggregate reposts, peak reposts, aggregate replies, peak replies, or out-degree). This approach ensured that both consistently active amplifiers and users responsible for individual cascades could be captured. The results stratified per language are displayed in Table 1 below.

Most notable insights include that among the identified superspreaders, English users ($N = 248$) tend to reply to more distinct users (median out-degree = 3) than Spanish users (median = 1). Spanish superspreaders receive slightly more replies on average (mean rp -sum = 7.57 vs. 6.73) and exhibit higher reply peaks (mean rp -x = 5.50 vs. 4.17). Repost behaviour is comparable across languages: mean repost peak (20) and sum (44 EN vs. 36 ES) show only modest differences. Based on the metrics, a total number of 1,915 superspreaders was identified (940 for English and 975 for Spanish). This means that in the English sub-set, superspreaders account for about 43% of posts, whereas in the Spanish sub-set, superspreaders contribute around 48%.

The remaining accounts were further inspected for potential bot activity using *BotometerLite* ([Observatory on Social Media](#)). *BotometerLite* produces scores derived from historical Twitter data collected prior to 31 May 2023, thereby aligning

with the data-set's temporal scope and estimating the likelihood that a user was a bot at the time of observation. As none of the accounts was scored as a potential bot, no further filtering was applied to the data-set which finally included 1,140 posts in English and 1,060 in Spanish, making the two sets once again highly comparable.

5.2 Statistical analysis of language and engagement

The analysis now examines possible engagement differences between the two communities that can indicate that audience responses vary significantly according to community identity factors. To address the non-normal distribution of engagement metrics in social media data (Shapiro-Wilk $p < .001$), we employed the non-parametric Mann-Whitney U test. The results indicated a statistically significant difference in engagement between the language groups ($U = 558123.0, p < .001$). Although the median engagement for both groups was 0 (reflecting the long-tail nature of the data after removing superspreaders), the Spanish subset exhibited a higher mean engagement ($M = 1.87$) compared to the English subset ($M = 1.37$). The effect size was small ($r = 0.076$), suggesting that while language plays a statistically significant role in engagement patterns, it is likely one of multiple contributing factors. To ensure robustness, we repeated this analysis on the full dataset including superspreaders (Total $N=4,035$). The difference remained statistically significant ($U = 1,869,946, p < .001, r = 0.081$), confirming that the observed cross-linguistic difference is not an artifact of outlier exclusion. The results of both tests are displayed in Table 2. In the next section, the analysis tests the effect of CTA expressions in driving engagement with disinformation for both groups.

5.3 Collective action

This part of the analysis now investigates the role of CA language in user engagement with disinformation posts. The tested hypothesis is that perceived injustice strengthens motivation for collective action, which in an online environment would translate into engaging with and sharing disinformation as an act of defiance and regained agency. First, we identified CTA expressions combining the statistical (topic modelling) and linguistic (CDA) methodology as explained in 3.

Language	Statistic	out-degree	rp-max	rp-sum	rt-max	rt-sum
EN	50%	3.00	1.00	2.00	2.00	3.00
	75%	5.00	2.00	3.00	10.00	15.25
	Total	248.00	248.00	248.00	248.00	248.00
	Max	45.00	131.00	219.00	808.00	2495.00
	Mean	4.25	4.17	6.73	20.61	44.07
	Min	0.00	0.00	0.00	0.00	0.00
	Std	7.00	13.16	22.98	80.74	216.37
ES	50%	1.00	1.00	2.00	2.50	4.00
	75%	4.00	2.00	3.00	10.00	13.75
	Total	286.00	286.00	286.00	286.00	286.00
	Max	48.00	354.00	440.00	808.00	2495.00
	Mean	3.20	5.50	7.57	19.75	35.78
	Min	0.00	0.00	0.00	0.00	0.00
	Std	5.42	25.39	33.42	75.23	191.84

Table 1: Summary statistics (50th %, 75th %, total count, maximum, mean, minimum, and standard deviation) of superspreader metrics by language (EN vs. ES).

Group	N	Mean	Mdn	<i>U</i> -Stat	<i>p</i>	<i>r</i>
Dataset: Filtered (No Superspreaders)						
EN	1,140	1.37	0.0	558,123	< .001	.08
ES	1,060	1.87	0.0			
Dataset: Full (With Superspreaders)						
EN	2,000	8.72	1.0	1,869,946	< .001	.08
ES	2,035	14.19	1.0			

Table 2: Comparison of engagement metrics between English (EN) and Spanish (ES) groups with and without superspreaders

5.3.1 Topic modelling

Topic modelling was used to identify posts containing mobilising language and call-to-action expressions. Specifically, we used BERTopic with a multilingual SentenceTransformer (Grootendorst, 2022) as in the literature, it is found to outperform traditional LDA models on multilingual, short texts. Feature extraction was performed by applying a bag-of-words with unigrams and bigrams, and $\text{max_df} = 0.85$ and $\text{min_df} = 0.02$ to remove overly common/rare terms. Preliminary runs with unconstrained topic discovery produced a large number of fine-grained clusters, many of which overlapped semantically. Setting $\text{nr_topics}=10$ allowed for a more coherent, high-level representation of discourse themes while retaining sufficient diversity to capture major narrative patterns across languages. This choice also ensured comparability across analyses and facilitated qualitative interpretation. The model returned per-document topic probabilities, per-topic term rankings, and most representative document per topic, enabling both quantitative summaries and qualitative interpretation across languages crucial for the extraction of

CTA expressions.

Several topics identified by the model displayed strong mobilising language, encouraging resistance, participation, or direct action against perceived threats linked to Agenda2030 and related governance narratives. Representative posts (Table 3) show imperatives such as ‘wake up’, ‘join’ and ‘resist’, reflecting how conspiracy-linked discourses often blend moral urgency with collective calls to action. These patterns suggest that in disinformation discourse, mobilising language is often present as shared narratives of opposition and control.

Additionally, further expressions were extracted applying CDA (Dijk, 1985, 1997) on a sample of 200 posts (100 posts per language). The full list of expressions (91 in English and 82 in Spanish) is provided in Table 4.

6 Call-to-Action and user engagement

To test the hypothesis that mobilizing language functions as an engagement booster (as predicted by SIMCA), we now compared the engagement levels of posts containing CTA expressions against those without, within each language community. Results are displayed in Table 5. Contrary to theoretical expectations, no significant engagement premium was found for mobilizing content in either group. In the English data-set, posts with CTA expressions did not elicit significantly higher engagement than non-CTA posts ($U = 57,446, p = .39$). Similarly, in the Spanish data-set the presence of mobilizing language had no significant impact on engagement levels ($U = 23,054, p = .56$). These findings suggest that in social media disinformation ecosystems, the mere presence of explicit ‘calls

Topic ID	Count	Top Representative Terms	Representative Example (excerpt)
0	1689	agenda2030, nwo, video, com, world, order, people, global, reset, truth	"Immortalized cell lines used in lab-grown meat... global elites want to control the food supply."
1	362	chile, españa, onu, méxico, argentina, política, gobierno, países, sociedad, latinoamérica	"Y ahora @GiorgioJackson, @gabrielboric... hablan de la ONU y la Agenda2030 en Latinoamérica."
2	292	climatescam, climate, agenda2030, globalwarming, wef, co2, fake, scam, science, propaganda	"Desmintiendo el calentamiento global bufones de la ONU... #ClimateScam."
3	46	canada, canadians, trudeau, agenda2030, climate, policy, protest, covid, freedom, rights	"Canadians need to support Alberta Oil & Gas — Trudeau is destroying our economy for Agenda2030."
4	37	vaccine, covid, depopulation, populationcontrol, health, agenda, wef, control, elite, world	"Vaccines are part of the depopulation agenda — open your eyes."
5	26	climatechange, greenagenda, sdgs, un, sustainability, policy, leaders, summit, goals, development	"The UN Climate Summit pushes Agenda2030 — same green agenda with new branding."
6	19	esg, economy, finance, corporations, capitalism, elites, governance, power, wef, global	"ESG is just a corporate version of Agenda2030 — economic control by the few."
7	16	energy, oil, gas, renewables, canada, europe, policy, crisis, cost, transition	"Energy transition in Europe is a scam — higher costs and less freedom."
8	14	digitalid, cbdc, surveillance, government, privacy, citizens, control, technology, freedom, rights	"Digital ID and CBDCs mean full surveillance — say no to control systems."
9	12	conspiracy, hoax, lies, fake, propaganda, media, agenda2030, truth, narrative, misinformation	"The media lies again — Agenda2030 is propaganda to hide the real plan."

Table 3: Topics identified by BERTopic with document count, most representative keywords, and a representative post excerpt.

EN	ES
wake, resist, join, sign, take action, act, must, do not comply, resistthereset, stop, share, help, demand, make, stand, stand up, speak out, raise your voice, get involved, take a stand, participate, march, protest, mobilize, defend, spread the word, volunteer, organize, boycott, take part, engage, rally, fight back, join the movement, act now, donot-comply	despierta, resiste, únete, firma la petición, actúa ahora, comparte, movilízate, haz algo, noalaagenda2030, noalplan, detente, toma, actúa, firma, ayuda, exige, haz, mantente, levántate, alza la voz, participa, actúa ya, protesta, lucha, defiende, propaga el mensaje, voluntario, organiza, boicotea, toma parte, únete al movimiento, comprométete, alístate, actúa hoy, únete ahora, haz tu voz escuchada, súmate, movilízate hoy

Table 4: Call-to-Action Expressions by Language

to fight’ or ‘wake up’ does not trigger increased user interaction, reinforcing the interpretation that engagement is driven by community norms, e.g., identity rather than mobilizing appeals, i.e., action.

7 Multivariate Analysis: Drivers of Engagement

To further investigate the drivers of engagement, we trained a Random Forest Regressor using Language (Identity), Dominant Topic (Context), and Collective Action (Mobilization) as features. The model yielded a negative R^2 score (-0.003), indicating that textual and linguistic features alone cannot predict the magnitude of user engagement. This null result is however theoretically significant: it demonstrates that engagement in disinformation ecosystems is not a mechanical response to spe-

Lang	CTA	N	Mdn	U	p	r
EN	Yes	110	0.0	57,446	.39	-.01
	No	1030	0.0			
ES	Yes	46	0.0	23,054	.56	.01
	No	1014	1.0			

Table 5: Comparisons of engagement for posts with and without CTA expressions.

Feature	Category	Importance
Dominant Topic	Context	43.8%
Language (EN/ES)	Identity	37.4%
Collective Action	Mobilization	18.8%

Table 6: Feature Importance scores from the Random Forest Regressor predicting user engagement.

cific trigger words or topics, but likely a stochastic process driven by algorithmic amplification and network dynamics (Gonzalez-Bailon et al., 2011). At the same time, the Feature Importance analysis (Table 6 and Figure 2) provides insight into the relative weight of these weak signals. Dominant Topic (43.8%) and Language (37.4%) largely outperformed CTA expressions (18.8%). This ranking confirms that even among the weak textual predictors available, the context (Topic) and community (Language) carry twice as much weight as the mobilizing rhetoric (CTA).

8 Discussion

This study provided empirical evidence that challenges the universalist application of the SIMCA model to disinformation on social media. By disentangling the effects of community identity from

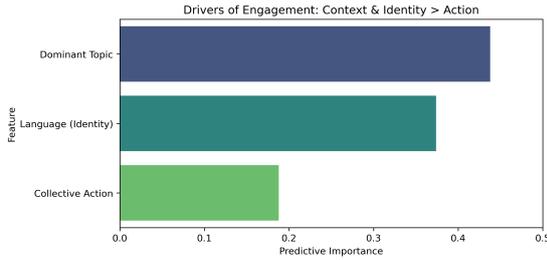


Figure 2: Feature importance ranking from the Random Forest model predicting user engagement.

mobilizing language (CTA expressions), our analysis revealed that language and cultural identity remain a primary driver of engagement. Cross-linguistic comparison between the two language communities showed a significantly higher baseline of activity in the Spanish-speaking community compared to the Anglosphere, a difference that persisted even after excluding superspreaders. This finding supports the Social Identity Theory perspective that online communities exhibit distinct engagement behaviours heavily structured by language (Hale, 2014).

Second, and most critically, our results demonstrate that mobilizing rhetoric fails to translate into measurable action such as explicit endorsement. Contrary to the predictions of the SIMCA model, the presence of explicit CTA language provided no significant engagement effect in either language group. This null result was further corroborated by multivariate analysis, where the CA feature proved to be the weakest predictor of engagement (18.8%), far outweighed by narrative context (43.8%) and language identity (37.4%).

These findings fundamentally reframe how we should interpret engagement metrics in disinformation research. The fact that users do not engage more with posts demanding them to act, resist, or wake up suggests that their interaction is symbolic rather than functional. In other words, users likely engage explicitly with a disinformation post to signal their alignment with the group and opposition to the mainstream (identity performance), rather than exclusively signalling collective resistance.

These results encourage us to rethink current applications of collective action models to social media as they may overestimate the mobilizing nature of engagement. Digital interactions in this context should be viewed as low-cost expressive behaviours, that is forms of identity signalling that satisfy affective needs for belonging without nec-

essarily implying a commitment to collective mobilization. Finally, by documenting the significant baseline disparity between Spanish and English communities, this study challenges the field’s reliance on Anglocentric data. It demonstrates that engagement is a culturally situated behaviour, proving that theoretical models built solely on English datasets cannot be generalized to other linguistic ecosystems without local validation.

9 Conclusion

This study set out to test the validity of the Social Identity Model of Collective Action (SIMCA) to social media disinformation. The findings suggest that engagement with disinformation functions primarily as symbolic identity alignment rather than functional collective action. Although mobilizing language is pervasive in disinformation discourse, its presence does not trigger the behavioural response predicted by collective action models. Instead, users appear to engage with content to signal group belonging and opposition to the mainstream, regardless of whether the content explicitly demands action.

By refining the application of SIT and SIMCA to social media, this study highlights the need to reconsider how engagement metrics are interpreted in disinformation research. While collective action models remain valuable for understanding how identities and grievances form, their explanatory power diminishes when mapped directly onto platform metrics. We argue that high engagement levels should be interpreted as signals of community cohesion and future research should therefore explore how cultural identity, emotional valence, narrative framing, and platform affordances shape identity-driven engagement with disinformation to develop more effective analytical frameworks and intervention strategies against the spread of disinformation online.

10 Limitations

While this study provides valuable insights, several limitations must be acknowledged. The data-set consists of posts from X, which may not be representative of broader social media engagement patterns on Facebook, YouTube, or TikTok. Future studies could investigate how different platforms’ algorithmic amplification patterns alter engagement dynamics. Due to lack of resources, the study primarily analysed explicit collective action

expressions from a sample of the data-set but did not account for *all* the mobilization strategies in the data-set. Further research should incorporate larger semantic and qualitative analysis to assess a larger range of such linguistic strategies.

Declaration on Generative AI

During the preparation of this work, the author used Gemini 3 in order to: Grammar and spelling check. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

- S raphin Alava, Divina Frau-Meigs, Ghayda Hassan, Hasna Hussein, and Yuanyuan Wei. 2017. *Youth and violent extremism on social media: violent extremism on social media: mapping the research*. United Nations Educational, Scientific, and Cultural Organization, Paris. OCLC: 1089113298.
- Maged Ali, Lucas Moreira Gomes, Nahed Azab, Jo o Gabriel de Moraes Souza, M. Karim Sorour, and Herbert Kimura. 2023. *Panic buying and fake news in urban vs. rural England: A case study of twitter during COVID-19*. *Technological Forecasting and Social Change*, 193:122598.
- Hunt Allcott and Matthew Gentzkow. 2017. *Social Media and Fake News in the 2016 Election*. *Journal of Economic Perspectives*, 31(2):211–236.
- Carolin Amlinger. 2022. *Gekr nkte Freiheit: Aspekte des libert ren Autoritarismus*, erste auflage, originalausgabe edition. Suhrkamp, Berlin.
- Shelly Ghai Bajaj. 2024. *Digital Disinformation Threats and Ethnocultural Diasporas*. In Gitanjali Adlakha-Hutcheon and Candyce Kelshall, editors, *(In)Security: Identifying the Invisible Disruptors of Security*, pages 53–65. Springer Nature Switzerland, Cham.
- Zach Bastick. 2021. *Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation*. *Computers in Human Behavior*, 116:106633.
- Zack Beauchamp. 2019. *Social media is rotting democracy from within*. *Vox*.
- Clare Birchall and Peter Knight. 2022. *Conspiracy Theories in the Time of Covid-19*, 1 edition. Routledge, London.
- Erika Bonnevie, Allison Gallegos-Jeffrey, Jaclyn Goldberg, Brian Byrd, and Joseph Smyser. 2021. *Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic*. *Journal of communication in healthcare*, 14(1):12–19. ISBN: 1753-8068 Publisher: Taylor & Francis.
- Roderick J. Brodie, Ana Ilic, Biljana Juric, and Linda Hollebeek. 2013. *Consumer engagement in a virtual brand community: An exploratory analysis*. *Journal of Business Research*, 66(1):105–114.
- Suzanne Brunsting and Tom Postmes. 2002. *Social Movement Participation in the Digital Age: Predicting Offline and Online Collective Action*. *Small Group Research*, 33(5):525–554. Publisher: SAGE Publications Inc.
- Michael Butter and Peter Knight, editors. 2020. *Routledge handbook of conspiracy theories*. Routledge, Abingdon, Oxon ; New York, NY.
- Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. 2020. *Exploring the Role of Visual Content in Fake News Detection*. In Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu, editors, *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, pages 141–161. Springer International Publishing, Cham.
- Gina Masullo Chen. 2011. *Tweet this: A uses and gratifications perspective on how active Twitter use gratifies a need to connect with others*. *Computers in Human Behavior*, 27(2):755–762.
- Li Chen, Qi Ling, Tingjia Cao, and Ke Han. 2020. *Mis-labeled, fragmented, and conspiracy-driven: a content analysis of the social media discourse about the HPV vaccine in China*. *Asian Journal of Communication*, 30(6):450–469.
- Li Chen, Yafei Zhang, Rachel Young, Xianwei Wu, and Ge Zhu. 2021. *Effects of Vaccine-Related Conspiracy Theories on Chinese Young Adults' Perceptions of the HPV Vaccine: An Experimental Study*. *Health Communication*, 36(11):1343–1353.
- Michael Christensen and Ashli Au. 2023. *The great reset and the cultural boundaries of conspiracy theory*. *International Journal of Communication*, 17:19–19.
- Maura Conway. 2017. *Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research*. *Studies in Conflict & Terrorism*, 40(1):77–98. Publisher: Routledge _eprint: <https://doi.org/10.1080/1057610X.2016.1157408>.
- Massimiliano Demata, Virginia Zorzi, and Angela Zottola, editors. 2022. *Conspiracy theory discourses*. Number volume 98 in *Discourse approaches to politics, society and culture*. John Benjamins Publishing Company, Amsterdam ; Philadelphia.
- Matthew R. DeVerna, Rachith Aiyappa, Diogo Pacheco, John Bryden, and Filippo Menczer. 2024. *Identifying and characterizing superspreaders of low-credibility content on Twitter*. *PLOS ONE*, 19(5):e0302201.
- Teun A. van Dijk, editor. 1985. *Discourse and Communication: New Approaches to the Analysis of Mass Media Discourse and Communication*. DE GRUYTER.

- Teun Adrianus van Dijk. 1997. *Discourse studies: a multidisciplinary introduction*. 2, 2,. Sage. OCLC: 634170329.
- John Edwards. 2009. *Language and Identity: An introduction*, 1 edition. Cambridge University Press.
- Irene Eleta and Jennifer Golbeck. 2014. Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior*, 41:424–432.
- Don Fallis. 2009. *A Conceptual Analysis of Disinformation*.
- Barrett Golding. 2025. *Iffy Index of Unreliable Sources*.
- Sandra Gonzalez-Bailon, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. 2011. The Dynamics of Protest Recruitment through an Online Network. *Scientific Reports*, 1(1):197. ArXiv:1111.5595 [physics].
- Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Keith Gunaratne, Eric A. Coomes, and Hourmazed Haghbayan. 2019. Temporal trends in anti-vaccine discourse on Twitter. *Vaccine*, 37(35):4867–4871.
- Scott A. Hale. 2014. Global connectivity and multilinguals in the Twitter network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 833–842, Toronto Ontario Canada. ACM.
- Christy Galletta Horner, Dennis Galletta, Jennifer Crawford, and Abhijeet Shirsat. 2023. Emotions: The Unexplored Fuel of Fake News on Social Media. In *Fake News on the Internet*. Routledge. Num Pages: 28.
- A. K. M. Najmul Islam, Samuli Laato, Shamim Talukder, and Erkki Sutinen. 2020. Misinformation sharing and social media fatigue during COVID-19: An affordance and cognitive load perspective. *Technological Forecasting and Social Change*, 159:120201.
- Jessica Johnson. 2018. The Self-Radicalization of White Men: “Fake News” and the Affective Networking of Paranoia. *Communication, Culture and Critique*, 11(1):100–115.
- Wandeep Kaur, Vimala Balakrishnan, Omer Rana, and Ajantha Sinniah. 2019. Liking, sharing, commenting and reacting on Facebook: User behaviors’ impact on sentiment intensity. *Telematics and Informatics*, 39:25–36.
- M. Laeeq Khan. 2017. Social media engagement: What motivates user participation and consumption on YouTube? *Computers in Human Behavior*, 66:236–247.
- Jan Kirchner and Christian Reuter. 2020. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–27.
- Gaurav Kumar, Rishabh Joshi, Jaspreet Singh, and Promod Yenigalla. 2020. AMUSED: A Multi-Stream Vector Representation Method for Use in Natural Dialogue. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 750–758, Marseille, France. European Language Resources Association.
- Ugo Laquière. 2025. #LGBTpropaganda #GenderTheory #Wokism: Expanding and blurring the boundaries of francophone anti-gender discourse propagated on Twitter. *Politikon: The IAPSS Journal of Political Science*, 59(1):88–114.
- Becca Lewis and Alice E. Marwick. 2017. *Media Manipulation and Disinformation Online*. Technical report, Data & Society. Publisher: Data & Society Research Institute.
- Neil MacFarquhar. 2016. *A Powerful Russian Weapon: The Spread of False Stories - The New York Times*.
- Guri Nordtorp Mølmen and Jacob Aasland Ravnald. 2021. Mechanisms of online radicalisation: how the internet affects the radicalisation of extreme-right lone actor terrorists. *Behavioral Sciences of Terrorism and Political Aggression*, 0(0):1–25. Publisher: Routledge _eprint: <https://doi.org/10.1080/19434472.2021.1993302>.
- Blair Nonnecke and Jennifer Preece. 1999. Shedding light on lurkers in online communities. *Ethnographic studies in real and virtual environments: Inhabited information spaces and connected communities*, Edinburgh, 123128.
- Observatory on Social Media. *Botometer X*.
- Jenny Preece, Blair Nonnecke, and Dorine Andrews. 2004. The top five reasons for lurking: improving community experiences for everyone. *Computers in human behavior*, 20(2):201–223. Publisher: Elsevier.
- Ignacio Ojea Quintana, Ritsaart Reimann, Marc Cheong, Mark Alfano, and Colin Klein. 2022. Polarization and trust in the evolution of vaccine discourse on Twitter during COVID-19. *PLOS ONE*, 17(12):e0277292. Publisher: Public Library of Science.
- Madhavi Reddi, Rachel Kuo, and Daniel Kreiss. 2023. Identity propaganda: Racial narratives and disinformation. *New Media & Society*, 25(8):2201–2218. Publisher: SAGE Publications.
- Sven Reichardt, editor. 2022. *Die Misstrauensgemeinschaft der "Querdenker": die Corona-Protteste aus*

- kultur- und sozialwissenschaftlicher Perspektive*, Sonderausgabe für die bundeszentrale für politische bildung edition. Number Band 10857 in Schriftenreihe. Bundeszentrale für Politische Bildung, Bonn.
- Stephen D. Reicher and Alexander S. Haslam. 2016. *Fueling Terror: How Extremists Are Made*. *Scientific American*.
- Bashir Sa’ad Abdullahi and Habeeb Idris Pindiga. 2023. *Tracking the Diffusion of Disinformation on the SDGs Across Social Media Platforms*. In Jan Servaes and Muhammad Jameel Yusha’u, editors, *SDG18 Communication for All, Volume 2: Regional Perspectives and Special Cases*, pages 145–174. Springer International Publishing, Cham.
- Hyunjin Seo and Robert Faris. 2021. *Comparative Approaches to Mis/Disinformation Introduction*. *International Journal of Communication*, 15(0):8. Number: 0.
- Guosong Shao. 2009. *Understanding the appeal of user-generated media: a uses and gratification perspective*. *Internet Research*, 19(1):7–25. Publisher: Emerald Group Publishing Limited.
- Kate T. Simms, Sharon J. B. Hanley, Megan A. Smith, Adam Keane, and Karen Canfell. 2020. *Impact of HPV vaccine hesitancy on cervical cancer in Japan: a modelling study*. *The Lancet. Public Health*, 5(4):e223–e234.
- Bonnie Stabile, Aubrey Grant, Hemant Purohit, and Kelsey Harris. 2019. *Sex, Lies, and Stereotypes: Gendered Implications of Fake News for Women in Politics*. *Public Integrity*, 21(5):491–502. Publisher: Routledge _eprint: <https://doi.org/10.1080/10999922.2019.1626695>.
- Henri Tajfel and John C. Turner. 2004. *The Social Identity Theory of Intergroup Behavior*. Political psychology: Key readings. Psychology Press, New York, NY, US. Pages: 293.
- Joshua A. Tucker, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature*.
- John C. Turner. 1991. *Social influence*. Social influence. Thomson Brooks/Cole Publishing Co, Belmont, CA, US. Pages: xvi, 206.
- Jan-Willem Van Prooijen and Mark Van Vugt. 2018. *Conspiracy Theories: Evolved Functions and Psychological Mechanisms*. *Perspectives on Psychological Science*, 13(6):770–788.
- Jan-Willem Van Prooijen and Karen M. Douglas. 2018. *Belief in conspiracy theories: Basic principles of an emerging research domain*. *European Journal of Social Psychology*, 48(7):897–908.
- Martijn van Zomeren, Tom Postmes, and Russell Spears. 2008. *Toward an integrative social identity model of collective action: A quantitative research synthesis of three socio-psychological perspectives*. *Psychological Bulletin*, 134(4):504–535. Place: US Publisher: American Psychological Association.
- Lorella Viola. 2025a. *What about Language? A Multilingual Behavioural Study of User Engagement with Disinformation on X: 5th Workshop on Reducing Online Misinformation through Credible Information Retrieval, ROMCIR 2025*. *ROMCIR 2025 Reducing Online Misinformation through Credible Information Retrieval 2025*, pages 54–69. Publisher: CEUR-WS.
- Lorella Viola. 2025b. *‘Barren lesbians plotting sterilization’: gender stereotypes and prejudices in health disinformation narratives, a cross-cultural analysis of social media of the HPV vaccine*. In Catherine Tebaldi, Alistair Plum, and Christoph Purschke, editors, *Conspiracy as Genre: Narrative, Power and Circulation*. Bloomsbury Academic, London.
- Bradley Wiggins. 2023. *‘Nothing Can Stop What’s Coming’: An analysis of the conspiracy theory discourse on 4chan’s /Pol board*. *Discourse & Society*, 34(3):381–398. Publisher: SAGE Publications Ltd.
- Florian Winterlin, Tim Schatto-Eckrodt, Lena Frischlich, Svenja Boberg, Felix Reer, and Thorsten Quandt. 2023. *“It’s us against them up there”: Spreading online disinformation as populist collective action*. *Computers in Human Behavior*, 146:107784.
- Asami Yagi, Yutaka Ueda, and Tadashi Kimura. 2024. *HPV Vaccine Issues in Japan: A review of our attempts to promote the HPV vaccine and to provide effective evaluation of the problem through social-medical and behavioral-economic perspectives*. *Vaccine*, 42(22):125859.
- Kai-Cheng Yang, Francesco Pierri, Pik-Mai Hui, David Axelrod, Christopher Torres-Lugo, John Bryden, and Filippo Menczer. 2021. *The COVID-19 Infodemic: Twitter versus Facebook*. *Big Data & Society*, 8(1):20539517211013861.

A Appendix

Table 7: data-set description

Value	Count
Likes	26,007
Quotes	1,281
Replies	3,329
Reposts	15,693
Users	2,091
Mentions	1448
Hashtags	13,691

B Appendix: Seed Hashtags

Table 8: Frequency of seed hashtags in the dataset. The list is sorted by frequency.

Hashtag	Count	Hashtag	Count	Hashtag	Count
#Agenda2030	2,950	#chemtrails	44	#klausschwab	23
#agenda2030	336	#BillGates	42	#WEFpuppet	22
#WEF	291	#SocialCreditSystem	42	#noalaagenda2030	22
#ClimateScam	147	#PureBlood	40	#malditaagenda2030	22
#GreatReset	144	#WEFpuppets	38	#Fauci	22
#NWO	121	#billgates	34	#depopulation	21
#NewWorldOrder	112	#Globalist	32	#NetZero	21
#Agenda2030.	90	#ClimateCult	32	#DictaduraSanitaria	20
#AGENDA2030	82	#SDGs	30	#Soros	20
#NOM	76	#TheGreatReset	29	#CrimesAgainstHumanity	20
#nwo	68	#Bilderberg	28	#Agenda2030?	20
#Agenda2030,	66	#Repentinitis	28	#CBDCs	19
#Agenda21	66	#NuevoOrdenMundial	27	#vaccineinjuries	18
#Plandemia	65	#VaccineGenocide	27	#CIA	18
#DigitalID	61	#FBI	26	#agenda21	18
#ODS	58	#wef	26	#FueraONU	17
#KlausSchwab	57	#GeorgeSoros	26	#WorldEconomicForum	17
#DiedSuddenly	56	#vaccine	25	#Nuremberg2	16
#15minutecities	55	#plandemia	25	#Chemtrails	16
#CBDC	50	#ESG	25	#15MinuteCities	15
#DepopulationAgenda	50	#HunterBiden	24	#Jarnac	15
#climatescam	48	#WEF2030Agenda	24	#VaccineSideEffects	14
#cbdc	45	#repentinitis	23	#CambioClimatico	14

Weakly Supervised Named Entity Recognition for Historical Texts

Marco Sorbi^{1,2}, Laurent Moccozet², Stephane Marchand-Maillet²

¹Research Institute for Statistics and Information Science, University of Geneva, Switzerland

²Centre Universitaire d’Informatique, University of Geneva, Switzerland

Correspondence: Marco.Sorbi@unige.ch

Abstract

Named Entity Recognition has emerged as a critical task in natural language processing, particularly for extracting meaningful information from unstructured text. Although traditional approaches rely heavily on large annotated datasets, recent advances have explored weak supervision techniques to address the limitations of resource-intensive annotation processes. Historical texts provide unique challenges to this task because of their linguistic peculiarities, and several approaches exist to address texts of this domain in a supervised way, but they involve lengthy manual annotations of the documents of interest by domain experts. To address this issue, this paper explores how recent weakly supervised NER techniques can be adapted to historical texts, analyzing their suitability for this domain. The experiments show that domain-specific architectures can be effectively trained on low-resource corpora with weak supervision over a small set of entity labels. Using only 10% of the annotations, the performance of these architectures remains above 80% of the supervised quality in terms of F1-Score.

1 Introduction

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing to identify and classify named entities from unstructured text into predefined categories such as people, organizations, and locations, and serves as an important step in the structuring of textual data (Lauriola et al., 2021). General purpose NER methods range from traditional rule-based approaches to modern deep learning architectures, mainly transformer-based (Kerghel et al., 2024). Various domains exist where ad hoc methods have been studied to address domain-specific peculiarities, such as biomedicine (Lauriola et al., 2021) and history (Ehrmann et al., 2023).

Historical documents present distinct challenges compared to modern texts, ranging from an archaic

language with non-standardized orthography (Santini et al., 2025), to degraded scripts and complex document layouts (Ehrmann et al., 2020). The lack of large annotated corpora in this low-resource context makes it difficult to design and train specific NER architectures (Novotny et al., 2023) that are required for effective NER in historical documents, with a significant impact on digital humanities and cultural heritage activities.

When manually annotated data are scarce, weakly supervised NER (WS-NER) has emerged to address the lack of training data. One common WS-NER framework is distantly-supervised NER (DS-NER), which takes advantage of knowledge bases or dictionaries to automatically generate labels to use as supervision signals (Fang et al., 2021). However, DS-NER faces significant challenges due to noisy and incomplete labels caused by limited dictionary coverage and imperfect distant annotation (Zhou et al., 2022). Previous works have studied multiple approaches to DS-NER, which we will present in Section 2, which have the potential to progress NER in the historical domain given its noise context and lack of extensive labeled datasets.

This paper aims to study WS-NER in the historical domain, exploiting existing DS-NER techniques for modern texts, to evaluate their effectiveness in the domain, and use them to adapt architectures from supervised Historical NER studies to the WS-NER task¹. This work is motivated by the feasibility of creating a partial domain-specific knowledge base relative to labeling a complete domain-specific document to train a supervised NER system. For this purpose, we identify two research questions to address:

- **RQ1:** Varying the size of the dictionaries, how do DS-NER techniques that are designed for modern texts perform on historical documents with domain-specific dictionaries? How do

¹ Code is available at www.github.com/msorbi/hwsner

they compare to domain-specific supervised NER methods in WS-NER and fully supervised NER settings?

- **RQ2:** Can we exploit these DS-NER techniques to adapt domain-specific supervised NER architectures to WS-NER?

The remainder of this paper is structured as follows. In Section 2, we discuss and organize a selection of related works concerning DS-NER and supervised NER in historical documents. Then, Section 3 presents the datasets used in this work to evaluate the models. Section 4 describes the experiments carried out and presents an overview of the metrics used for evaluation in this work. The results are presented in Section 5 and discussed in Section 6, where the research questions are also answered. Finally, Section 7 presents conclusions and suggestions for future work.

2 Related Work

The positioning of this paper is at the intersection of weak supervision and the historical domain of Named Entity Recognition. The presentation of an overview of related work on the two topics is therefore appropriate.

2.1 Distantly-supervised Named Entity Recognition

DS-NER is a WS-NER framework that relies on the availability of dictionaries or knowledge bases instead of labeled sequences as training supervision signal. It avoids the need for large-scale human annotation, but introduces noise due to incomplete and inaccurate annotations (Zhang et al., 2021). Early research sees the development of AutoNER (Shang et al., 2018), which introduced a Tie-or-Break tagging scheme with a Fuzzy-LSTM-CRF architecture designed to be robust to noisy distant supervision. We divide subsequent work into two main branches: Positive-Unlabeled Learning (PUL) and Self-Supervised Learning (SSL).

Positive-Unlabeled Learning It formulates DS-NER as training on positively labeled data, created using the knowledge base, and unlabeled data. Contributions using this approach include:

- AdaPU (Peng et al., 2019), which firstly introduced PUL in DS-NER, designing a training algorithm that can unbiasedly estimate the task loss as if there were fully labeled data.

- Conf-MPU (Zhou et al., 2022), which extended PUL in DS-NER to a multi-class setting and introduced confidence-based risk estimation.
- CuPUL (Li et al., 2025), which adds curriculum learning to PUL to stabilize training and weaken the impact of noise.

Self-Supervised Learning Used as a training stage, it allows reducing the impact of noisy labels to works like:

- BOND (Liang et al., 2020), which first adapts Pre-trained Language Models to noisy distant labels, and then applies teacher-student self-training for refinement.
- SCDL (Zhang et al., 2021), which trains two teacher-student networks to jointly denoise labels.
- CENSOR (Si et al., 2024), which introduces uncertainty-aware teacher learning, to reduce reliance on miscalibrated high-confidence labels, and student-student label sharing, to mitigate error propagation.

Other approaches There are other recent works using different approaches, for example:

- SANTA (Si et al., 2023), addressing inaccurate and incomplete annotations separately, with a memory-smoothed focal loss and a noise-tolerant loss.
- MProto (Wu et al., 2023), employing a prototype network to capture intra-class variance, and formulating token-prototype assignment as an optimal transport problem.

2.2 Historical Named Entity Recognition

In the domain of historical documents, NER becomes substantially more difficult, as it faces multiple challenges due to the linguistic characteristics of these texts, which the survey by Ehrmann et al. (2023) identifies in language dynamics, noisy input, and lack of resources. Analyzing methodologies from rule-based to deep learning, the survey emphasizes the need for models robust to multilingualism, spelling variation, and sparse data. When dealing with digitized documents, non-standard layout and Optical Character Recognition (OCR) errors add a layer of noise that can significantly degrade performance (Kettunen et al., 2017).

We can distinguish the application of different techniques for NER in this domain.

Rule-based Early research mainly employs rule-based systems tailored to documents of interest, using lexical heuristics, gazetteers, and token-level rules, in order to extract people and places from British parliamentary records (Grover et al., 2008), for example. This approach is transparent, but typically struggles to face the challenge of integrating variability in the rules when increasing the corpus size.

Machine learning It enables the use of more flexible feature-based techniques that commonly rely on Hidden Markov Models, Support Vector Machines, Conditional Random Fields (CRF), and decision trees (Yadav and Bethard, 2018). In historical context, Torres Aguilar et al. (2016) applied a CRF on latin charters, obtaining good results based on a large number of features. These methods reduce the effect of variability, but they require a large number of hand designed features.

Deep learning These systems, often relying on sequence labeling and transformer-based models such as Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) for automated feature learning, and CRF for classification, have further reduced the effect of variability. Among deep learning works in the historical domain,

- Boros et al. (2020) propose a hierarchical transformer stack for historical German and French datasets, specifically to mitigate OCR noise in digitized historical texts,
- Blouin et al. (2021) investigate transfer learning from modern to historical domains, studying annotation effort, domain mismatch, and pre-training data selection for historical NER,
- Torres Aguilar (2022) creates a human-labeled dataset of multilingual medieval charters, and addresses multilingual NER by combining stacked embeddings with BERT-based models fine-tuned on the dataset,
- Novotny et al. (2023) create a small dataset with sentences containing known entities from late medieval European texts, and propose a bootstrapping annotation pipeline to build larger corpora of the texts.

However, these methods rely on heavily annotated datasets to learn features, which are rarely available.

Few-shot learning It exploits Large Language Models (LLMs) through prompting in order to label the historical texts without relying on annotated corpora. Among these works in the historical domain,

- Hiltmann et al. (2025) prompt LLMs with historical context, outperforming established frameworks on historical texts,
- Zhang and Colavizza (2025) use few-shot learning of LLMs on the HIPE dataset (Ehrmann et al., 2020).

This work focuses on deep learning systems to reduce the amount of annotations required to train them. Specifically, we study the applicability of weak supervision techniques to train deep learning models on the historical NER, studying the impact of the amount of annotated data on the quality of results. Even if we can consider the approach of Novotny et al. (2023) to belong to the WS-NER paradigm, to the knowledge of the authors, there is no extensive study of WS-NER or DS-NER in the historical domain.

3 Dataset

Our experiments are based on the dataset introduced by Torres Aguilar (2022), consisting of a human-labeled medieval NER dataset whose ancient French charters come from:

- Corpus de la Bourgogne du Moyen Âge (“Corpus of Burgundy in the Middle Ages”, CBMA) - Cartulary of the city of Arbois (Magnani, 2020): a municipal cartulary commissioned in 1384, containing public issues such as military services and war costs, taxes and customs, or lawsuits in court.
- Diplomata Belgica (“Belgian Diplomatic Sources”, CDBE – de Hemptinne, Thérèse and Deploige, Jeroen and Kupper, Jean-Louis and Prevenier, Walter, 2015): a database published by the Belgian Royal Historical Commission, containing French charters dated the 13th century, containing legal actions concerning individuals, corporations, and private affairs.
- HOME History of Medieval Europe (HOME) - Alcar (Stutzmann et al., 2021): it contains cartularies dated between the 12th and 14th centuries, reporting donations, exchanges, and other legal acts.

	CBMA	CDBE	HOME
Train	38658	235643	114640
Validation	896	18510	18510
Test	8133	56081	18554

Table 1: Number of tokens per each dataset split.

	CBMA	CDBE	HOME
PERS	652	4118	925
LOC	347	2662	922

Table 2: Number of distinct training entities for each type and source, before sampling.

The dataset contains annotations for entities in the Person (PERS) and Location (LOC) classes, and the sizes for each source are reported in Table 1.

Separately for each source, to adapt the dataset to a WS-NER setting, we extracted all the tagged entities from the train split, and randomly sampled the entities to make the dictionaries for each type. Then, we used the sampled dictionaries to create the noisy labels for the train splits of the mentioned sources via string matching. Table 2 shows the dimensions of each dictionary. The validation and test splits remained unchanged.

4 Experimental Setup

To address RQ1, we tested DS-NER techniques belonging to the Positive-Unlabeled and Self-Supervised Learning frameworks, plus the two specific works mentioned in Section 2 and the Stacked Embeddings supervised architecture (Torres Aguilar, 2022), and compared their performances on historical French texts in two settings:

- Fully supervised NER, using the original human-labeled train splits, to analyze the best supervision condition, and
- WS-NER, varying the sampling size for the dictionaries according to Section 3, to see its impact on the performances. We used sampling sizes from 20% to 100%, increasing by 20%, that are commonly used in previous works (Zhou et al., 2022; Wu et al., 2023), and 10% to test with increased noise conditions,

training every method on the train split of each setting and sample size independently. To better summarize the results, we will refer with “small

dictionaries” to the 10% and 20% samples, with “medium dictionaries” to the 40% and 60% samples, and with “large dictionaries” to the 80% and 100% samples.

Then, to address RQ2, we used the best performing DS-NER technique to train

1. XLM-RoBERTa base (Conneau et al., 2020), a large multilingual masked language model, and
2. Stacked Embeddings (Torres Aguilar, 2022), a domain-specific architecture for supervised NER in medieval texts,

on the WS-NER task, using the same settings as in the first part, and checked whether DS-NER can generalize their performance from the supervised task.

Evaluation We now discuss the evaluation metrics used to assess NER performance in historical documents. According to the works presented in Section 2, the typical evaluation metric for NER is the F1-score, usually presented alongside precision and recall to show their impact. Novotny et al. (2023) use the $F_{0.25}$ -score, which significantly gives more weight to precision than to recall, but otherwise the two metrics are averaged in a balanced way, that is, using the F1-score.

In WS-NER, training supervision is based on incomplete labels, which are characterized by high precision but low recall (Zhou et al., 2022). This effect is amplified in historical documents by orthographic variations that complicate entity identification. As DS-NER training algorithms aim to improve the recall at the expense of precision, we continue to use the common F1-score as a balanced measure, giving importance to the recall metric, too. We compute these metrics using `seqeval` (Nakayama, 2018) default token-level mode, in order to grant partial scores to entities that are correctly identified and classified, even if their boundary is not correct, and mitigate the problem of annotation granularity (Ehrmann et al., 2020).

If multiple classes are present, as in our case with PERS and LOC, micro averaging is commonly used for global measures (Shang et al., 2018; Torres Aguilar, 2022) to ensure that frequent classes have an appropriate influence on the final results.

5 Results

We analyze here the test performances, in terms of micro-averaged F1-Score, Precision, and Recall,

obtained by the various methods under fully supervised NER and different weak supervision settings. For each technique, we micro-average measures among classes and then, for each setting, we average among datasets and report them in percentage in Figure 1. Detailed measures are reported in Appendix A. For comparison, Stacked Embeddings (StEmb), a domain-specific architecture for supervised NER in medieval texts (Torres Aguilar, 2022), is also reported along with dictionary matching and the CuPUL DS-NER technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings as underlying classifier. We can notice that:

- Despite having a precision among the best in every setting, if we consider recall and F1-score, the Stacked Embeddings architecture is the best performer with full supervision and large dictionaries, but its performances drop quickly when the sample size is decreased.
- If trained under full supervision, the best DS-NER techniques are BOND (Liang et al., 2020) and SANTA (Si et al., 2023), with an F1-score close to the baseline supervised Stacked Embeddings architecture.
- SANTA maintains good recall with large dictionaries, but if we consider the F1-score, BOND remains the best DS-NER technique with large and medium dictionaries, overcoming the supervised baseline as well for medium dictionaries, and is the best DS-NER technique in every setting in terms of precision.
- With small dictionaries, the best technique in terms of F1-score is CuPUL (Li et al., 2025), which has the best recall even with medium dictionaries.
- Considering all three metrics, CuPUL is the least affected by the noise introduced with distant labels, with a performance that only slightly changes on the different settings. Conf-MPU (Zhou et al., 2022), which is the other PUL-based technique, and MProto (Wu et al., 2023) also have stable but lower performances.
- Except for the simpler BOND, the SSL techniques obtained generally unstable performances across settings, especially in terms of recall and F1-score.

Because CuPUL is the best performing technique in low supervision settings, and the least affected by the noise of distant labels, we use it to train (1) XLM-RoBERTa (Conneau et al., 2020) and (2) Stacked Embeddings (Torres Aguilar, 2022) on the WS-NER task. This is done by replacing the original RoBERTa classifier in CuPUL. Despite that they perform slightly worse than the baseline with full supervision and large dictionaries, we can see once again in Figure 1 that CuPUL is little affected by the noise of distant labels. Although XLM-RoBERTa (1) still has some noise-related problems, especially with small dictionaries, CuPUL allows the ad-hoc classifier (2) to have performances in the WS-NER settings that are significantly closer to that of the fully supervised task for which it has been designed.

6 Discussion

The results of our experiments show that weak supervision can achieve F1-Scores within 20% of supervised Named Entity Recognition performance on historical texts using only 10% of the annotations, making them particularly appealing for this context where large annotated datasets are often unavailable. To mitigate the linguistic challenges that characterize NER in historical texts, the experiments demonstrate that DS-NER techniques can effectively adapt domain-specific NER models to low-resource historical corpora, to achieve performance levels that are close to fully supervised methods even with additional noise introduced with distant labels. The amount of noise in the distant labels, controlled via the dictionary size, has a significant impact on the final performance, causing noticeable drops in most of the techniques. Nevertheless, CuPUL and MProto have shown good robustness, maintaining consistent performance across different dictionary sizes, and revealing particularly suitable for this scenario. Considering these results, we can now answer the research questions formulated in Section 1.

6.1 RQ1: Validity of modern text DS-NER techniques for ancient texts

The results reported in Figure 1 show that most DS-NER techniques suffer from performance drops when they are applied to NER in ancient texts, especially in low supervision settings with small dictionaries. CuPUL (Li et al., 2025), BOND (Liang et al., 2020), and SANTA (Si et al., 2023), belong-

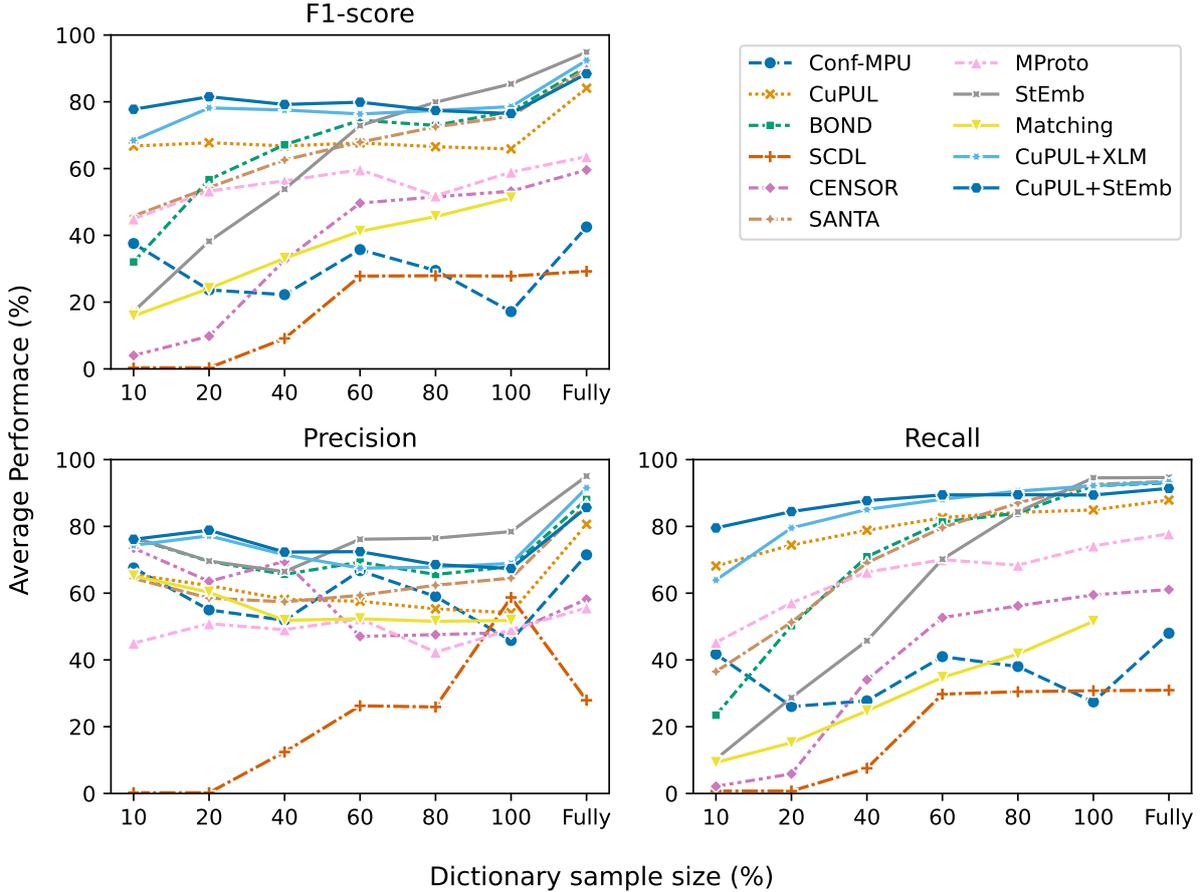


Figure 1: Test results of Fully supervised NER and Weakly supervised NER varying the dictionary sample size. The Stacked Embeddings (StEmb) supervised architecture, dictionary matching, and the CuPUL DS-NER technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings as underlying classifier are reported using solid lines.

ing to three different categories as identified in Section 2, are able to achieve good performance with large dictionaries, but still suffer substantial drops when switching to smaller dictionaries, especially the latter two techniques. CuPUL is the most robust to the noise of low supervision, thanks to a high recall as opposed to the other techniques.

6.2 RQ2: Adaptation of domain-specific supervised NER architectures using modern text DS-NER

Based on the answer to RQ1, we selected CuPUL to carry out further experiments on the adaptability to the WS-NER task in historical texts. Figure 1 shows that, especially for the Stacked Embeddings architecture (Torres Aguilar, 2022), the performance in all settings is comparable and closer to the supervised performance of the architecture, remaining above 80% of the supervised F1-Score. This indicates that CuPUL enables the supervised NER architecture to generalize well to weak supervision.

The relatively low capacity of Stacked Embeddings, which is based on a BiLSTM-CRF instead of a transformer, may help in this, as it is, in principle, less prone to overfitting and easier to generalize.

7 Conclusion and Future Work

This study shows that weak supervision can effectively be used to train domain-specific models to achieve NER performance levels, on historical texts with little annotations, that are close to fully supervised methods. This is particularly important for the historical domain, where large annotated datasets are often unavailable.

Further assessment of the robustness of this framework and improvement of its performance are the two main areas of focus for future work. Additional analysis could involve examining the adaptability of the framework under various sources of noise that are characteristic of the domain, such as the use of external historical knowledge bases (Uckelman; Wrisley, 2018) to build annotation dictionary-

ies and its application to historical documents transcribed with OCR (Ehrmann et al., 2020). These investigations will determine the framework’s applicability when manual transcription of texts or compilation of entity lists is impractical or unfeasible. Additionally, similar analysis can be conducted for other information extraction tasks, including Entity linking and Relationship extraction. In order to increase the NER performance, a major development of the technique would be the integration of active learning, which offers the dual benefit of assisting the model in resolving difficult instances and exploiting domain expertise from historians and linguists where it is most needed and appropriate.

By reducing the need for large annotated datasets, these techniques support the adaptation of NER tools to a wider range of historical low-resource text collections, enabling more comprehensive analysis of historical documents (Bouillon et al., 2024), with significant implications for the fields of digital humanities and cultural heritage.

Limitations

The empirical analysis of this work is restricted to manually transcribed medieval French charter collections with two entity types (persons and locations), which limits the generalizability of the findings to other historical languages and periods, document genres, larger sets of entity categories, and OCR-transcribed documents. In order to avoid the quality of knowledge bases impacting the results, weak supervision is performed via randomly sampled dictionaries constructed from the training splits; manually curated or external bases, which could be biased towards some subset of entities, may exhibit different noise profiles and cause a different model behavior. Moreover, a fully annotated, albeit small, validation set is kept for model evaluation and hyperparameter tuning. Finally, the evaluation relies on standard NER metrics and does not include human analysis of downstream impact on digital humanities tasks.

Ethical Considerations

Historical corpora reflect the social, cultural, and political biases of the periods and institutions that produced them. The medieval charters considered in this work primarily document legal and administrative activities of municipal and ecclesiastical authorities, emphasizing the interests of elites. As a result, NER models trained on these sources could

reproduce existing biases about people and places, and the usage of models and techniques designed for modern languages or pretrained on modern texts, which is often necessary due to the inherent low volume of available historical data, could amplify these biases when facing challenges specific to the domain, such as linguistic variation and orthographic inconsistency. Moreover, learning from genre-specific conventions like formulaic expressions may capture patterns that are characteristic of the genre and which could not generalize to other genres and time periods.

The systems proposed in this work are intended as assistive tools for historical research, and their outputs should be interpreted in collaboration with domain experts who can assess biases in both the sources and the models.

Acknowledgments

The Swiss National Science Foundation funds this work under grant number 215733 “Une édition sémantique et multilingue en ligne des registres du Conseil de Genève (1545-1550)”. The project website is accessible at <https://www.unige.ch/registresconseilge/en>. We express our gratitude to Christophe Chazalon, who helped process the texts thanks to his historical domain expertise, H el ene de Ribaupierre and Gilles Falquet for their knowledge and insights about Information Extraction and Semantic Web, and the anonymous reviewers and editors for their comments and suggestions.

References

- Baptiste Blouin, Benoit Favre, Jeremy Auguste, and Christian Henriot. 2021. [Transferring modern named entity recognition to the historical domain: How to take the step?](#) In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 152–162, NIT Silchar, India. NLP Association of India (NLP AI).
- Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adri an Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. 2020. [Alleviating digitization errors in named entity recognition for historical documents](#). In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 431–441, Online. Association for Computational Linguistics.
- Pierrette Bouillon, Christophe Chazalon, Sandra Coram-Mekkey, Gilles Falquet, Johanna Gerlach, Stephane Marchand-Maillet, Laurent Moccozet, Jonathan Mutal, Raphael Rubino, and Marco Sorbi. 2024. [RC-num: A semantic and multilingual online edition of](#)

- the geneva council registers from 1545 to 1550. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 21–22, Sheffield, UK. European Association for Machine Translation (EAMT).
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- de Hemptinne, Thérèse and Deploige, Jeroen and Kupper, Jean-Louis and Prevenier, Walter, editor. 2015. *Diplomata Belgica: les sources diplomatiques des Pays-Bas méridionaux au Moyen Âge. The Diplomatic Sources from the Medieval Southern Low Countries*. Commission royale d’Histoire/Koninklijke Commissie voor Geschiedenis.
- Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. [Named entity recognition and classification in historical documents: A survey](#). *ACM Comput. Surv.*, 56(2).
- Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. [Overview of clef hipe 2020: Named entity recognition and linking on historical newspapers](#). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*, page 288–310, Berlin, Heidelberg. Springer-Verlag.
- Zheng Fang, Yanan Cao, Tai Li, Ruipeng Jia, Fang Fang, Yanmin Shang, and Yuhai Lu. 2021. [TEBNER: Domain specific named entity recognition with type expanded boundary-aware network](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 198–207, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. 2008. [Named entity recognition for digitised historical texts](#). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- Torsten Hiltmann, Martin Dröge, Nicole Dresselhaus, Till Grallert, Melanie Althage, Paul Bayer, Sophie Eckenstaler, Koray Mendi, Jascha Marijn Schmitz, Philipp Schneider, Wiebke Sczeponik, and Anica Skibba. 2025. [NER4all or Context is All You Need: Using LLMs for low-effort, high-performance NER on historical texts. A humanities informed approach](#). *Preprint*, arXiv:2502.04351.
- Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. [Recent advances in named entity recognition: A comprehensive survey and comparative study](#). *Preprint*, arXiv:2401.10825.
- Kimmo Kettunen, Eetu Mäkelä, Teemu Ruokolainen, Juha Kuokkala, and Laura Löfberg. 2017. [Old content and modern tools - searching named entities in a finnish ocred historical newspaper collection 1771-1910](#). *Digital Humanities Quarterly*, 11.
- Ivano Lauriola, Fabio Aioli, Alberto Lavelli, and Fabio Rinaldi. 2021. [Learning adaptive representations for entity recognition in the biomedical domain](#). *Journal of Biomedical Semantics*, 12(1):10.
- Yuepei Li, Kang Zhou, Qiao Qiao, Qing Wang, and Qi Li. 2025. [Re-examine distantly supervised NER: A new benchmark and a simple approach](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10940–10959, Abu Dhabi, UAE. Association for Computational Linguistics.
- Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. [Bond: Bert-assisted open-domain named entity recognition with distant supervision](#). In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD ’20*, page 1054–1064, New York, NY, USA. Association for Computing Machinery.
- Eliana Magnani. 2020. *Des chartae au Corpus: La plateforme des CBMA - Chartae/Corpus Burgundiae Medii Aevi*, volume 27 of *Atelier de Recherche Sur Les Textes Médiévaux*, pages 57–67. Brepols Publishers.
- Hiroki Nakayama. 2018. [seqeval: A python framework for sequence labeling evaluation](#). Software available from <https://github.com/chakki-works/seqeval>.
- Vit Novotny, Kristina Luger, Michal Štefánik, Tereza Vrabcova, and Ales Horak. 2023. [People and places of historical Europe: Bootstrapping annotation pipeline and a new corpus of named entities in late medieval texts](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14104–14113, Toronto, Canada. Association for Computational Linguistics.
- Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. [Distantly supervised named entity recognition using positive-unlabeled learning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2409–2419, Florence, Italy. Association for Computational Linguistics.
- Cristian Santini, Laura Melosi, and Emanuele Frontoni. 2025. [Named entity recognition in historical italian: The case of giacomo leopardi’s zibaldone](#). *Preprint*, arXiv:2505.20113.
- Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. [Learning named entity](#)

- tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064, Brussels, Belgium. Association for Computational Linguistics.
- Shuzheng Si, Zefan Cai, Shuang Zeng, Guoqiang Feng, Jiaying Lin, and Baobao Chang. 2023. **SANTA: Separate strategies for inaccurate and incomplete annotation noise in distantly-supervised named entity recognition**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3883–3896, Toronto, Canada. Association for Computational Linguistics.
- Shuzheng Si, Helan Hu, Haozhe Zhao, Shuang Zeng, Kaikai An, Zefan Cai, and Baobao Chang. 2024. **Improving the robustness of distantly-supervised named entity recognition via uncertainty-aware teacher learning and student-student collaborative learning**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5533–5546, Bangkok, Thailand. Association for Computational Linguistics.
- Dominique Stutzmann, Sergio Torres Aguilar, and Paul Chaffenet. 2021. **HOME-Alcar: Aligned and Annotated Cartularies**. Type: dataset.
- Sergio Torres Aguilar. 2022. **Multilingual named entity recognition for medieval charters using stacked embeddings and bert-based models**. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 119–128, Marseille, France. European Language Resources Association.
- Sergio Torres Aguilar, Xavier Tannier, and Pierre Chastang. 2016. **Named entity recognition applied on a data base of Medieval Latin charters. The case of chartae burgundiae**. In *3rd International Workshop on Computational History (HistoInformatics 2016)*, Krakow, Poland.
- Sara L Uckelman. **DRAFT: Names in the 1292 census of Paris**.
- David Joseph Wrisley. 2018. **The literary geographies of christine de pizan**. MLA.
- Shuhui Wu, Yongliang Shen, Zeqi Tan, Wenqi Ren, Jietian Guo, Shiliang Pu, and Weiming Lu. 2023. **MProto: Multi-prototype network with denoised optimal transport for distantly supervised named entity recognition**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2361–2374, Singapore. Association for Computational Linguistics.
- Vikas Yadav and Steven Bethard. 2018. **A survey on recent advances in named entity recognition from deep learning models**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Shibingfeng Zhang and Giovanni Colavizza. 2025. **Named entity recognition of historical text via large language model**. *Preprint*, arXiv:2508.18090.
- Xinghua Zhang, Bowen Yu, Tingwen Liu, Zhenyu Zhang, Jiawei Sheng, Xue Mengge, and Hongbo Xu. 2021. **Improving distantly-supervised named entity recognition with self-collaborative denoising learning**. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1518–1529, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Kang Zhou, Yuepei Li, and Qi Li. 2022. **Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7198–7211, Dublin, Ireland. Association for Computational Linguistics.

A Detailed Results

Tables 3 to 5 report the micro-averaged F1-Score, Precision, and Recall obtained by the different techniques, averaged on the test split of each source, in the cases of fully supervised NER and varying distant supervision settings. These measures are used to draw Figure 1.

Tables 6 to 8 report the micro-averaged F1-Score, Precision, and Recall obtained by the different techniques on the test split of each source, in the cases of fully supervised NER and varying distant supervision settings.

B Hyperparameter tuning

For RQ1, we used the default hyperparameters for every technique, computing prior class probabilities from the validation set when required. For RQ2, according to Li et al. (2025) and using grid searches, we made a first tuning step for the voters’ hyperparameters, and a second step for the curriculum training hyperparameters, as we used CuPUL with Stacked embeddings (Torres Aguilar, 2022) and XLM-RoBERTa (Conneau et al., 2020) voters.

B.1 Voters’ hyperparameter tuning

The tuned hyperparameters for the voters are the train epochs, drop negative, loss type, and m. The learning rate is set to $1e-3$, and other hyperparameters are left at their default values. Tables 9 to 11 show the validation performances of Stacked Embeddings and XLM-RoBERTa voters with every set of hyperparameters. We select the set of hyperparameters based on the mean and standard deviation of the performances and on the performance with

Sample (%)	10	20	40	60	80	100	Fully	MEAN
Conf-MPU	38 ± 27	24 ± 33	22 ± 31	36 ± 27	29 ± 22	17 ± 24	43 ± 31	30 ± 8
CuPUL	67 ± 1	68 ± 0	67 ± 3	68 ± 5	67 ± 5	66 ± 5	84 ± 7	69 ± 6
BOND	32 ± 18	57 ± 13	67 ± 5	74 ± 5	73 ± 5	77 ± 11	90 ± 4	67 ± 16
SCDL	0 ± 0	0 ± 0	9 ± 12	28 ± 39	28 ± 39	28 ± 38	29 ± 40	18 ± 12
CENSOR	4 ± 5	10 ± 7	33 ± 32	50 ± 36	52 ± 37	53 ± 38	60 ± 42	37 ± 19
SANTA	46 ± 0	54 ± 3	63 ± 3	68 ± 6	72 ± 4	76 ± 5	90 ± 4	67 ± 13
MProto	45 ± 6	53 ± 3	56 ± 9	60 ± 10	52 ± 15	59 ± 8	64 ± 18	56 ± 5
StEmb	17 ± 11	38 ± 20	54 ± 6	73 ± 8	80 ± 3	85 ± 7	95 ± 1	63 ± 24
Matching	16 ± 9	24 ± 7	33 ± 4	41 ± 2	46 ± 3	51 ± 5		
CuPUL+XLM	68 ± 6	78 ± 3	78 ± 2	76 ± 5	77 ± 5	79 ± 7	92 ± 3	78 ± 6
CuPUL+StEmb	78 ± 3	82 ± 3	79 ± 1	80 ± 4	77 ± 6	76 ± 7	88 ± 4	80 ± 3

Table 3: Test results in terms of F1-Score of Fully supervised NER and Weakly supervised NER varying the dictionary sample size. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results among DS-NER techniques for each setting are reported in bold. For reference, the Stacked Embeddings (StEmb) supervised architecture and the pure dictionary matching are reported, too. The bottom rows report performances of the CuPUL technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings, respectively, as underlying classification model.

low supervision settings. The selected hyperparameters are:

- Stacked Embeddings voter hyperparameters:
 - train epochs (E) = 15, drop negative (Neg) = 0.1, loss type = MPN, m = 20
- XLM-RoBERTa voter hyperparameters:
 - train epochs (E) = 5, drop negative (Neg) = 0.5, loss type = MPN-CE, m = 20

B.2 Curriculum training hyperparameter tuning

Tables 12 and 13 show the validation performances of the CuPUL curriculum train stage with Stacked Embeddings and XLM-RoBERTa, respectively, with every set of hyperparameters. We select the set of hyperparameters based on the mean and standard deviation of the performances and on the performance with low supervision settings. The selected hyperparameters are:

- CuPUL with Stacked Embeddings:
 - train epochs (E) = 20, loss type = MPN
- CuPUL with XLM-RoBERTa:
 - train epochs (E) = 10, train sub-epochs (sE) = 1, loss type = Conf-MPU-CE, learning rate (LR): 1e−5

C Online Resources

The source code is available online at <https://github.com/msorbi/hwsner>

Sample (%)	10	20	40	60	80	100	Fully	MEAN
Conf-MPU	68 ± 23	55 ± 41	52 ± 41	67 ± 24	59 ± 29	46 ± 41	71 ± 21	60 ± 8
CuPUL	66 ± 2	62 ± 2	58 ± 4	58 ± 6	55 ± 6	54 ± 6	81 ± 7	62 ± 8
BOND	76 ± 18	70 ± 8	66 ± 5	69 ± 6	66 ± 10	68 ± 16	88 ± 6	72 ± 7
SCDL	0 ± 0	0 ± 0	12 ± 17	26 ± 37	26 ± 36	59 ± 42	28 ± 39	22 ± 18
CENSOR	73 ± 33	64 ± 28	70 ± 24	47 ± 34	48 ± 34	48 ± 34	58 ± 41	58 ± 10
SANTA	64 ± 10	59 ± 8	57 ± 5	59 ± 7	62 ± 5	65 ± 8	87 ± 5	65 ± 9
MProto	45 ± 7	51 ± 7	49 ± 7	52 ± 10	42 ± 14	49 ± 8	56 ± 21	49 ± 4
StEmb	77 ± 4	70 ± 9	66 ± 1	76 ± 8	76 ± 7	78 ± 11	95 ± 1	77 ± 8
Matching	65 ± 17	60 ± 9	52 ± 7	52 ± 10	52 ± 9	52 ± 9		
CuPUL+XLM	74 ± 3	77 ± 4	71 ± 4	67 ± 6	68 ± 7	69 ± 10	92 ± 3	74 ± 7
CuPUL+StEmb	76 ± 2	79 ± 4	72 ± 3	72 ± 6	69 ± 8	67 ± 10	86 ± 5	74 ± 5

Table 4: Test results in terms of Precision of Fully supervised NER and Weakly supervised NER varying the dictionary sample size. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results among DS-NER techniques for each setting are reported in bold. For reference, the Stacked Embeddings (StEmb) supervised architecture and the pure dictionary matching are reported, too. The bottom rows report performances of the CuPUL technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings, respectively, as underlying classification model.

Sample (%)	10	20	40	60	80	100	Fully	MEAN
Conf-MPU	42 ± 30	26 ± 37	28 ± 39	41 ± 34	38 ± 32	27 ± 39	48 ± 35	36 ± 7
CuPUL	68 ± 3	74 ± 2	79 ± 6	83 ± 6	84 ± 5	85 ± 5	88 ± 6	80 ± 6
BOND	23 ± 16	50 ± 19	71 ± 14	81 ± 8	84 ± 5	92 ± 1	93 ± 2	71 ± 22
SCDL	1 ± 1	1 ± 1	8 ± 9	30 ± 41	30 ± 41	31 ± 42	31 ± 42	19 ± 13
CENSOR	2 ± 3	6 ± 4	34 ± 38	53 ± 39	56 ± 40	59 ± 42	61 ± 43	39 ± 22
SANTA	37 ± 4	51 ± 4	69 ± 3	80 ± 6	87 ± 4	93 ± 2	94 ± 3	73 ± 19
MProto	45 ± 5	57 ± 5	66 ± 14	70 ± 12	68 ± 15	74 ± 10	78 ± 9	66 ± 10
StEmb	10 ± 7	29 ± 19	46 ± 8	70 ± 8	84 ± 3	95 ± 1	95 ± 1	61 ± 29
Matching	9 ± 6	15 ± 5	25 ± 4	35 ± 2	42 ± 3	52 ± 2		
CuPUL+XLM	64 ± 9	80 ± 5	85 ± 3	88 ± 3	91 ± 2	92 ± 2	93 ± 2	85 ± 9
CuPUL+StEmb	80 ± 5	84 ± 2	88 ± 1	89 ± 1	89 ± 1	89 ± 2	91 ± 3	87 ± 4

Table 5: Test results in terms of Recall of Fully supervised NER and Weakly supervised NER varying the dictionary sample size. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results among DS-NER techniques for each setting are reported in bold. For reference, the Stacked Embeddings (StEmb) supervised architecture and the pure dictionary matching are reported, too. The bottom rows report performances of the CuPUL technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings, respectively, as underlying classification model.

Sample (%)	10	20	40	60	80	100	Fully	MEAN
Conf-MPU	55	0	0	43	37	0	71	30 ± 26
CuPUL	66	67	62	63	63	64	75	66 ± 4
BOND	29	43	60	70	75	86	90	65 ± 19
SCDL	1	1	1	1	1	1	1	1 ± 0
CENSOR	0	0	0	0	0	0	0	0 ± 0
SANTA	46	53	59	59	68	79	85	64 ± 12
MProto	37	49	47	46	31	47	39	42 ± 6
StEmb	32	66	59	83	83	92	93	73 ± 19
Matching	27	33	38	41	44	51		
CuPUL+XLM	77	82	80	75	77	82	89	80 ± 4
CuPUL+StEmb	80	81	80	84	80	80	83	81 ± 1
Conf-MPU	0	71	67	64	51	51	56	52 ± 21
CuPUL	68	68	69	65	63	61	89	69 ± 8
BOND	55	74	72	73	66	61	86	69 ± 9
SCDL	0	0	26	82	82	82	86	51 ± 35
CENSOR	1	12	75	85	85	76	91	61 ± 33
SANTA	45	59	64	71	72	68	94	68 ± 13
MProto	51	57	69	64	62	64	83	64 ± 9
StEmb	6	22	57	66	76	76	96	57 ± 28
Matching	5	17	33	39	43	45		
CuPUL+XLM	65	74	75	72	71	69	93	74 ± 8
CuPUL+StEmb	80	78	78	74	69	67	91	77 ± 7
Conf-MPU	57	0	0	0	0	0	0	8 ± 19
CuPUL	66	68	69	75	73	73	88	73 ± 6
BOND	12	53	70	81	78	85	95	68 ± 24
SCDL	0	0	0	0	0	0	0	0 ± 0
CENSOR	11	17	22	64	69	83	87	51 ± 28
SANTA	46	51	65	74	78	80	91	69 ± 14
MProto	47	53	53	70	63	66	69	60 ± 8
StEmb	14	27	46	70	81	88	95	60 ± 27
Matching	15	23	29	44	50	57		
CuPUL+XLM	64	78	78	83	84	85	95	81 ± 8
CuPUL+StEmb	73	86	80	82	83	83	91	82 ± 5

Table 6: Test results in terms of F1-Score of Fully supervised NER and Weakly supervised NER varying the dictionary sample size on each of the three datasets (CBMA, CDBE, HOME), respectively. Measures are micro-averaged among classes. The MEAN column reports the performance averaged across settings with its standard deviation. The best results among DS-NER techniques for each setting are reported in bold. For reference, the Stacked Embeddings (StEmb) supervised architecture and the pure dictionary matching are reported, too. The bottom rows report performances of the CuPUL technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings, respectively, as underlying classification model.

Sample (%)	10	20	40	60	80	100	Fully	MEAN
Conf-MPU	52	0	0	48	39	0	63	29 ± 24
CuPUL	64	62	54	55	53	54	70	59 ± 6
BOND	57	59	65	69	73	80	89	70 ± 10
SCDL	1	1	1	1	1	1	1	1 ± 0
CENSOR	100	100	100	0	0	0	0	43 ± 46
SANTA	51	50	52	50	58	70	80	59 ± 10
MProto	35	42	41	40	23	39	28	35 ± 6
StEmb	74	80	68	85	83	89	93	82 ± 8
Matching	74	66	58	52	52	53		
CuPUL+XLM	77	78	73	66	68	74	87	75 ± 6
CuPUL+StEmb	76	76	75	78	72	73	79	76 ± 2
Conf-MPU	100	65	56	52	38	37	51	57 ± 18
CuPUL	64	60	57	52	49	46	86	59 ± 12
BOND	73	72	60	62	52	46	80	63 ± 11
SCDL	0	0	37	78	77	75	83	50 ± 32
CENSOR	28	32	67	78	79	67	90	63 ± 21
SANTA	74	68	57	60	59	53	92	66 ± 12
MProto	49	54	58	52	49	51	77	56 ± 9
StEmb	74	57	66	66	66	64	96	70 ± 11
Matching	41	48	42	41	41	40		
CuPUL+XLM	70	72	66	61	59	55	93	68 ± 11
CuPUL+StEmb	79	76	69	63	57	54	90	70 ± 11
Conf-MPU	50	100	100	100	100	100	100	93 ± 16
CuPUL	69	65	63	66	63	62	85	68 ± 7
BOND	100	78	72	77	71	78	95	82 ± 10
SCDL	0	0	0	0	0	100	0	14 ± 33
CENSOR	93	59	42	63	64	78	85	69 ± 15
SANTA	68	57	63	67	70	70	88	69 ± 8
MProto	51	57	48	65	56	58	62	57 ± 5
StEmb	83	72	66	78	80	82	96	79 ± 8
Matching	80	67	55	65	62	62		
CuPUL+XLM	76	82	76	75	76	78	94	80 ± 6
CuPUL+StEmb	73	85	74	76	76	75	88	78 ± 5

Table 7: Test results in terms of Precision of Fully supervised NER and Weakly supervised NER varying the dictionary sample size on each of the three datasets (CBMA, CDBE, HOME), respectively. Measures are micro-averaged among classes. The MEAN column reports the performance averaged across settings with its standard deviation. The best results among DS-NER techniques for each setting are reported in bold. For reference, the Stacked Embeddings (StEmb) supervised architecture and the pure dictionary matching are reported, too. The bottom rows report performances of the CuPUL technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings, respectively, as underlying classification model.

Sample (%)	10	20	40	60	80	100	Fully	MEAN
Conf-MPU	59	0	0	39	35	0	81	31 ± 28
CuPUL	68	74	73	74	77	78	80	75 ± 3
BOND	20	33	55	70	77	92	92	63 ± 24
SCDL	2	2	2	2	3	3	3	2 ± 0
CENSOR	0	0	0	0	0	0	0	0 ± 0
SANTA	42	55	67	71	82	90	90	71 ± 15
MProto	40	60	53	54	49	61	67	55 ± 8
StEmb	20	56	52	82	83	95	93	69 ± 24
Matching	17	22	28	34	39	49		
CuPUL+XLM	76	86	87	85	88	92	91	86 ± 4
CuPUL+StEmb	84	86	86	90	88	88	87	87 ± 2
Conf-MPU	0	78	83	84	79	82	63	67 ± 26
CuPUL	72	78	87	86	88	88	92	84 ± 6
BOND	45	77	90	88	89	91	92	82 ± 15
SCDL	0	0	21	87	89	90	90	54 ± 38
CENSOR	0	8	87	92	93	89	93	66 ± 37
SANTA	33	52	73	86	92	95	96	75 ± 21
MProto	53	61	85	81	84	85	90	77 ± 12
StEmb	3	13	50	66	88	94	96	59 ± 33
Matching	3	10	27	37	45	53		
CuPUL+XLM	61	77	87	87	90	90	94	84 ± 10
CuPUL+StEmb	82	81	90	89	89	88	93	87 ± 4
Conf-MPU	66	0	0	0	0	0	0	9 ± 22
CuPUL	64	72	77	87	87	89	91	81 ± 9
BOND	6	40	68	85	86	94	95	68 ± 29
SCDL	0	0	0	0	0	0	0	0 ± 0
CENSOR	6	10	15	66	76	90	90	50 ± 33
SANTA	34	46	67	82	87	93	94	72 ± 20
MProto	43	50	60	75	72	77	77	65 ± 12
StEmb	8	17	35	63	81	94	95	56 ± 32
Matching	8	14	19	33	42	53		
CuPUL+XLM	55	75	81	92	94	95	96	84 ± 13
CuPUL+StEmb	73	86	87	89	92	92	94	87 ± 6

Table 8: Test results in terms of Recall of Fully supervised NER and Weakly supervised NER varying the dictionary sample size on each of the three datasets (CBMA, CDBE, HOME), respectively. Measures are micro-averaged among classes. The MEAN column reports the performance averaged across settings with its standard deviation. The best results among DS-NER techniques for each setting are reported in bold. For reference, the Stacked Embeddings (StEmb) supervised architecture and the pure dictionary matching are reported, too. The bottom rows report performances of the CuPUL technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings, respectively, as underlying classification model.

E	Neg	Loss	m	10	20	40	60	80	100	Fully	MEAN
10	0.1	MPN	10	53 ± 37	54 ± 38	55 ± 39	54 ± 38	83 ± 5	82 ± 6	92 ± 2	68 ± 15
10	0.1	MPN	20	76 ± 5	82 ± 5	81 ± 3	81 ± 3	80 ± 6	80 ± 7	90 ± 2	81 ± 4
10	0.1	MPN-CE	10	75 ± 1	79 ± 2	83 ± 3	83 ± 3	84 ± 6	83 ± 7	93 ± 2	83 ± 5
10	0.1	MPN-CE	20	73 ± 5	79 ± 3	81 ± 2	82 ± 3	81 ± 6	80 ± 8	92 ± 1	81 ± 5
10	0.3	MPN	10	54 ± 38	53 ± 37	55 ± 39	54 ± 38	83 ± 5	82 ± 7	92 ± 3	67 ± 15
10	0.3	MPN	20	74 ± 9	81 ± 6	81 ± 4	80 ± 3	80 ± 5	78 ± 7	90 ± 3	80 ± 4
10	0.3	MPN-CE	10	75 ± 2	79 ± 1	83 ± 3	83 ± 3	83 ± 5	83 ± 7	93 ± 1	83 ± 5
10	0.3	MPN-CE	20	73 ± 4	80 ± 2	82 ± 2	82 ± 3	81 ± 6	81 ± 8	92 ± 1	81 ± 5
15	0.1	MPN	10	53 ± 37	55 ± 39	87 ± 2	86 ± 3	85 ± 5	83 ± 6	92 ± 2	77 ± 14
15	0.1	MPN	20	80 ± 2	81 ± 7	84 ± 2	81 ± 2	79 ± 6	80 ± 8	91 ± 1	82 ± 3
15	0.1	MPN-CE	10	69 ± 5	75 ± 4	81 ± 3	85 ± 5	84 ± 7	84 ± 8	94 ± 0	82 ± 7
15	0.1	MPN-CE	20	71 ± 5	75 ± 6	79 ± 0	83 ± 4	82 ± 6	82 ± 8	92 ± 1	81 ± 6
15	0.3	MPN	10	51 ± 36	53 ± 37	87 ± 2	85 ± 3	85 ± 5	85 ± 7	92 ± 2	77 ± 15
15	0.3	MPN	20	78 ± 2	81 ± 6	83 ± 1	82 ± 3	79 ± 6	81 ± 7	90 ± 1	82 ± 3
15	0.3	MPN-CE	10	69 ± 4	75 ± 2	81 ± 3	85 ± 5	84 ± 7	84 ± 8	94 ± 1	82 ± 7
15	0.3	MPN-CE	20	70 ± 5	75 ± 6	81 ± 2	83 ± 4	82 ± 6	81 ± 9	92 ± 1	81 ± 6
20	0.1	MPN	10	52 ± 37	53 ± 38	84 ± 3	86 ± 3	86 ± 5	84 ± 6	93 ± 1	77 ± 15
20	0.1	MPN	20	77 ± 2	81 ± 7	84 ± 2	83 ± 3	82 ± 6	82 ± 8	91 ± 1	83 ± 3
20	0.1	MPN-CE	10	67 ± 6	71 ± 6	80 ± 3	84 ± 5	84 ± 6	85 ± 9	94 ± 1	81 ± 8
20	0.1	MPN-CE	20	68 ± 7	73 ± 6	79 ± 2	83 ± 5	82 ± 6	83 ± 9	94 ± 1	80 ± 7
20	0.3	MPN	10	51 ± 36	53 ± 38	87 ± 3	86 ± 3	86 ± 5	85 ± 7	92 ± 1	77 ± 15
20	0.3	MPN	20	79 ± 0	81 ± 7	82 ± 1	83 ± 4	81 ± 6	81 ± 8	91 ± 2	83 ± 3
20	0.3	MPN-CE	10	67 ± 5	72 ± 5	81 ± 5	84 ± 6	84 ± 6	85 ± 9	94 ± 1	81 ± 8
20	0.3	MPN-CE	20	68 ± 7	73 ± 6	80 ± 2	83 ± 5	82 ± 7	83 ± 9	94 ± 1	81 ± 7

Table 9: Validation results in terms of F1-Score of Stacked Embeddings voters. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results for each setting are reported in bold. Selected hyperparameters are in bold.

E	Neg	Loss	m	10	20	40	60	80	100	Fully	MEAN
10	0.1	MPN	10	40 ± 8	68 ± 1	81 ± 3	87 ± 4	90 ± 3	92 ± 1	92 ± 3	78 ± 16
10	0.1	MPN	20	39 ± 4	69 ± 3	79 ± 5	84 ± 7	90 ± 3	89 ± 4	92 ± 4	77 ± 16
10	0.1	MPN-CE	10	25 ± 5	40 ± 7	56 ± 8	77 ± 5	87 ± 2	88 ± 3	94 ± 3	67 ± 23
10	0.1	MPN-CE	20	32 ± 11	44 ± 5	60 ± 9	77 ± 2	87 ± 0	91 ± 0	93 ± 4	69 ± 21
10	0.1	MPU	10	69 ± 11	55 ± 19	47 ± 1	45 ± 2	43 ± 3	36 ± 7	42 ± 4	48 ± 9
10	0.1	MPU	20	65 ± 17	50 ± 10	56 ± 15	52 ± 14	39 ± 1	39 ± 1	48 ± 8	50 ± 8
10	0.1	MPU-CE	10	64 ± 9	60 ± 16	61 ± 7	54 ± 13	51 ± 12	46 ± 12	40 ± 16	54 ± 7
10	0.1	MPU-CE	20	58 ± 13	55 ± 13	54 ± 10	45 ± 15	41 ± 10	36 ± 12	43 ± 9	47 ± 7
10	0.3	MPN	10	31 ± 7	69 ± 3	80 ± 3	86 ± 6	90 ± 3	90 ± 2	93 ± 2	77 ± 19
10	0.3	MPN	20	35 ± 9	67 ± 0	81 ± 4	86 ± 5	88 ± 5	89 ± 4	91 ± 3	77 ± 17
10	0.3	MPN-CE	10	27 ± 7	38 ± 6	59 ± 6	77 ± 4	86 ± 2	88 ± 2	94 ± 2	67 ± 23
10	0.3	MPN-CE	20	32 ± 11	44 ± 5	62 ± 6	78 ± 3	87 ± 0	91 ± 0	94 ± 3	70 ± 21
10	0.3	MPU	10	70 ± 11	52 ± 14	46 ± 3	47 ± 0	47 ± 11	48 ± 9	44 ± 3	50 ± 8
10	0.3	MPU	20	63 ± 20	57 ± 17	53 ± 15	55 ± 13	38 ± 0	37 ± 2	41 ± 1	49 ± 9
10	0.3	MPU-CE	10	65 ± 9	59 ± 13	64 ± 7	55 ± 14	52 ± 11	38 ± 19	51 ± 9	55 ± 8
10	0.3	MPU-CE	20	58 ± 11	59 ± 12	57 ± 8	45 ± 15	40 ± 12	34 ± 15	56 ± 14	50 ± 9
10	0.5	MPN	10	36 ± 11	67 ± 3	80 ± 3	85 ± 2	90 ± 2	92 ± 1	93 ± 3	78 ± 18
10	0.5	MPN	20	40 ± 8	71 ± 5	82 ± 7	85 ± 6	90 ± 3	88 ± 3	93 ± 3	78 ± 16
10	0.5	MPN-CE	10	26 ± 6	39 ± 5	61 ± 5	78 ± 3	86 ± 2	87 ± 3	94 ± 2	67 ± 23
10	0.5	MPN-CE	20	33 ± 11	44 ± 3	63 ± 5	77 ± 2	85 ± 2	91 ± 1	95 ± 2	70 ± 21
10	0.5	MPU	10	69 ± 12	61 ± 21	43 ± 2	44 ± 1	44 ± 1	48 ± 8	45 ± 2	51 ± 9
10	0.5	MPU	20	63 ± 17	56 ± 19	57 ± 13	44 ± 8	41 ± 3	35 ± 6	45 ± 7	49 ± 9
10	0.5	MPU-CE	10	63 ± 9	61 ± 12	62 ± 11	52 ± 13	48 ± 11	39 ± 18	55 ± 21	54 ± 8
10	0.5	MPU-CE	20	56 ± 15	55 ± 16	54 ± 11	44 ± 12	42 ± 13	31 ± 18	51 ± 15	48 ± 8

Table 10: Validation results in terms of F1-Score of XML-RoBERTa voters – Part I. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results for each setting are reported in bold.

E	Neg	Loss	m	10	20	40	60	80	100	Fully	MEAN
5	0.1	MPN	10	55 ± 2	74 ± 14	82 ± 7	86 ± 6	91 ± 2	90 ± 2	93 ± 3	81 ± 12
5	0.1	MPN	20	55 ± 2	81 ± 6	77 ± 11	78 ± 8	80 ± 11	82 ± 8	89 ± 5	78 ± 9
5	0.1	MPN-CE	10	59 ± 9	75 ± 1	77 ± 2	82 ± 2	87 ± 2	87 ± 3	93 ± 3	80 ± 10
5	0.1	MPN-CE	20	64 ± 9	75 ± 5	75 ± 7	81 ± 6	83 ± 4	85 ± 5	92 ± 4	79 ± 8
5	0.1	MPU	10	67 ± 12	62 ± 8	64 ± 10	49 ± 3	44 ± 1	43 ± 6	36 ± 7	52 ± 11
5	0.1	MPU	20	67 ± 15	65 ± 14	59 ± 11	57 ± 8	45 ± 8	43 ± 9	37 ± 7	53 ± 10
5	0.1	MPU-CE	10	60 ± 11	62 ± 14	56 ± 17	51 ± 9	40 ± 14	41 ± 2	34 ± 0	49 ± 9
5	0.1	MPU-CE	20	56 ± 13	55 ± 16	53 ± 14	47 ± 8	41 ± 6	28 ± 3	32 ± 1	45 ± 10
5	0.3	MPN	10	59 ± 2	77 ± 10	82 ± 8	86 ± 5	90 ± 3	90 ± 2	92 ± 2	82 ± 10
5	0.3	MPN	20	59 ± 5	82 ± 8	76 ± 11	79 ± 9	80 ± 12	82 ± 8	89 ± 4	78 ± 8
5	0.3	MPN-CE	10	59 ± 9	76 ± 2	78 ± 1	84 ± 2	86 ± 1	87 ± 3	92 ± 3	80 ± 9
5	0.3	MPN-CE	20	63 ± 9	76 ± 6	75 ± 7	81 ± 5	83 ± 5	85 ± 5	91 ± 5	79 ± 8
5	0.3	MPU	10	67 ± 13	54 ± 2	49 ± 8	45 ± 5	42 ± 4	41 ± 9	43 ± 10	49 ± 8
5	0.3	MPU	20	67 ± 15	64 ± 16	60 ± 12	48 ± 3	47 ± 3	43 ± 10	41 ± 5	53 ± 9
5	0.3	MPU-CE	10	61 ± 12	61 ± 16	57 ± 16	54 ± 10	45 ± 16	40 ± 8	34 ± 4	50 ± 9
5	0.3	MPU-CE	20	58 ± 12	56 ± 16	52 ± 14	46 ± 9	42 ± 7	26 ± 1	32 ± 0	44 ± 10
5	0.5	MPN	10	48 ± 10	76 ± 11	81 ± 8	86 ± 5	92 ± 3	91 ± 1	92 ± 2	81 ± 14
5	0.5	MPN	20	57 ± 1	82 ± 7	77 ± 11	79 ± 9	80 ± 12	82 ± 8	89 ± 5	78 ± 9
5	0.5	MPN-CE	10	60 ± 8	76 ± 2	76 ± 2	82 ± 3	86 ± 2	87 ± 3	93 ± 4	80 ± 9
5	0.5	MPN-CE	20	65 ± 8	76 ± 6	75 ± 7	81 ± 5	82 ± 4	85 ± 4	92 ± 4	80 ± 7
5	0.5	MPU	10	66 ± 15	54 ± 1	52 ± 3	48 ± 4	42 ± 4	39 ± 7	35 ± 2	48 ± 9
5	0.5	MPU	20	70 ± 18	67 ± 11	61 ± 12	46 ± 2	46 ± 4	43 ± 9	41 ± 6	53 ± 10
5	0.5	MPU-CE	10	59 ± 12	60 ± 13	57 ± 17	53 ± 11	51 ± 11	38 ± 11	42 ± 10	51 ± 7
5	0.5	MPU-CE	20	59 ± 13	56 ± 14	53 ± 13	49 ± 9	41 ± 9	32 ± 6	46 ± 12	48 ± 8

Table 11: Validation results in terms of F1-Score of XML-RoBERTa voters – Part II. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results for each setting are reported in bold. Selected hyperparameters are in bold.

E	Loss	10	20	40	60	80	100	Fully	MEAN
15	Conf-MPU-CE	77 ± 1	80 ± 6	82 ± 3	81 ± 4	79 ± 7	79 ± 9	90 ± 1	81 ± 4
15	Conf-MPU	81 ± 1	83 ± 5	81 ± 3	83 ± 4	80 ± 8	78 ± 8	89 ± 3	82 ± 3
15	MPN-CE	72 ± 5	77 ± 3	80 ± 2	82 ± 4	82 ± 7	80 ± 8	92 ± 1	81 ± 5
15	MPN	78 ± 2	81 ± 5	82 ± 2	82 ± 3	82 ± 7	80 ± 7	90 ± 3	82 ± 3
15	MPU-CE	64 ± 12	66 ± 12	59 ± 14	51 ± 14	44 ± 14	29 ± 18	43 ± 21	51 ± 12
15	MPU	59 ± 26	66 ± 23	62 ± 23	60 ± 20	58 ± 17	47 ± 20	58 ± 33	59 ± 5
20	Conf-MPU-CE	77 ± 2	78 ± 8	81 ± 2	81 ± 4	81 ± 7	80 ± 9	91 ± 1	81 ± 4
20	Conf-MPU	77 ± 2	82 ± 6	83 ± 2	82 ± 5	80 ± 6	79 ± 8	90 ± 3	82 ± 4
20	MPN-CE	69 ± 5	74 ± 6	80 ± 1	83 ± 5	82 ± 6	83 ± 9	94 ± 1	80 ± 7
20	MPN	80 ± 0	83 ± 6	85 ± 3	84 ± 4	82 ± 6	82 ± 7	90 ± 3	84 ± 3
20	MPU-CE	63 ± 11	63 ± 17	58 ± 15	51 ± 12	43 ± 15	27 ± 18	38 ± 19	49 ± 12
20	MPU	65 ± 18	64 ± 26	74 ± 9	68 ± 10	63 ± 11	60 ± 12	61 ± 29	65 ± 4
25	Conf-MPU-CE	76 ± 3	81 ± 5	82 ± 3	82 ± 4	82 ± 7	81 ± 9	91 ± 1	82 ± 4
25	Conf-MPU	79 ± 2	82 ± 6	84 ± 3	82 ± 5	79 ± 7	80 ± 8	90 ± 3	82 ± 3
25	MPN-CE	70 ± 5	75 ± 4	78 ± 1	82 ± 4	83 ± 6	83 ± 9	93 ± 1	81 ± 6
25	MPN	77 ± 3	84 ± 6	84 ± 2	84 ± 3	83 ± 7	82 ± 8	91 ± 2	84 ± 3
25	MPU-CE	65 ± 12	61 ± 17	58 ± 16	50 ± 15	44 ± 16	27 ± 18	36 ± 20	49 ± 12
25	MPU	70 ± 13	67 ± 23	76 ± 6	68 ± 8	65 ± 9	61 ± 9	63 ± 26	67 ± 4

Table 12: Validation results in terms of F1-Score of CuPUL curriculum train with Stacked Embeddings. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results for each setting are reported in bold. Selected hyperparameters are in bold.

E	sE	Loss	LR	10	20	40	60	80	100	Fully	MEAN
1	1	Conf-MPU	1	28 ± 24	36 ± 31	40 ± 27	42 ± 33	43 ± 37	42 ± 38	44 ± 40	39 ± 5
1	1	Conf-MPU	3	51 ± 29	50 ± 33	57 ± 28	57 ± 32	58 ± 30	53 ± 35	56 ± 36	55 ± 3
1	1	Conf-MPU-CE	1	29 ± 25	43 ± 33	39 ± 28	36 ± 34	38 ± 35	39 ± 34	44 ± 41	38 ± 4
1	1	Conf-MPU-CE	3	67 ± 15	68 ± 13	61 ± 17	60 ± 18	60 ± 17	66 ± 14	74 ± 17	65 ± 5
1	1	MPU-CE	1	27 ± 23	37 ± 31	38 ± 26	31 ± 28	30 ± 27	26 ± 23	33 ± 30	32 ± 4
1	1	MPU-CE	3	59 ± 15	57 ± 6	46 ± 7	53 ± 10	37 ± 5	34 ± 6	27 ± 12	45 ± 11
1	2	MPU-CE	1	47 ± 17	53 ± 11	48 ± 16	46 ± 13	45 ± 10	43 ± 5	35 ± 5	45 ± 5
1	2	MPU-CE	3	53 ± 12	61 ± 6	54 ± 9	52 ± 11	35 ± 7	42 ± 7	30 ± 10	47 ± 10
5	1	Conf-MPU	1	70 ± 8	76 ± 9	69 ± 14	73 ± 14	74 ± 17	79 ± 12	79 ± 14	74 ± 3
5	1	Conf-MPU	3	69 ± 0	80 ± 4	79 ± 7	86 ± 3	83 ± 7	89 ± 1	92 ± 3	82 ± 7
5	1	Conf-MPU-CE	1	69 ± 5	78 ± 10	74 ± 11	76 ± 7	76 ± 8	80 ± 5	91 ± 3	78 ± 6
5	1	Conf-MPU-CE	3	70 ± 7	80 ± 2	81 ± 1	84 ± 3	88 ± 1	87 ± 1	93 ± 2	83 ± 6
5	1	MPU-CE	1	56 ± 9	55 ± 11	52 ± 14	49 ± 13	46 ± 12	39 ± 7	39 ± 13	48 ± 6
5	1	MPU-CE	3	52 ± 12	67 ± 3	57 ± 8	57 ± 1	47 ± 4	50 ± 5	42 ± 11	53 ± 7
5	2	Conf-MPU	1	66 ± 9	79 ± 4	81 ± 3	88 ± 1	85 ± 7	90 ± 2	92 ± 3	83 ± 8
5	2	Conf-MPU	3	61 ± 3	84 ± 0	80 ± 6	87 ± 3	89 ± 1	91 ± 1	94 ± 2	84 ± 9
5	2	Conf-MPU-CE	1	73 ± 14	81 ± 3	81 ± 4	78 ± 8	83 ± 5	88 ± 0	93 ± 3	83 ± 6
5	2	Conf-MPU-CE	3	64 ± 14	77 ± 3	81 ± 4	81 ± 7	84 ± 5	90 ± 0	94 ± 3	82 ± 8
5	2	MPU-CE	1	65 ± 7	63 ± 3	56 ± 4	53 ± 5	49 ± 10	44 ± 6	57 ± 12	55 ± 7
5	2	MPU-CE	3	63 ± 3	60 ± 13	50 ± 17	66 ± 5	59 ± 10	44 ± 21	65 ± 21	58 ± 7
10	1	Conf-MPU	1	70 ± 5	79 ± 5	80 ± 5	85 ± 4	84 ± 6	88 ± 3	93 ± 2	83 ± 6
10	1	Conf-MPU	3	65 ± 10	78 ± 2	82 ± 4	89 ± 0	87 ± 4	89 ± 2	92 ± 3	83 ± 8
10	1	Conf-MPU-CE	1	73 ± 11	82 ± 2	81 ± 4	80 ± 8	85 ± 4	89 ± 0	92 ± 5	83 ± 5
10	1	Conf-MPU-CE	3	62 ± 16	75 ± 6	86 ± 0	83 ± 5	86 ± 2	90 ± 0	94 ± 3	82 ± 9
10	1	MPU-CE	1	62 ± 8	63 ± 11	54 ± 8	51 ± 10	51 ± 6	47 ± 3	60 ± 13	55 ± 5
10	1	MPU-CE	3	57 ± 7	63 ± 17	52 ± 20	51 ± 20	60 ± 1	51 ± 11	56 ± 31	56 ± 4
10	2	Conf-MPU	1	61 ± 8	83 ± 2	85 ± 1	86 ± 3	91 ± 1	90 ± 1	92 ± 4	84 ± 9
10	2	Conf-MPU	3	54 ± 9	83 ± 2	86 ± 0	90 ± 0	92 ± 1	91 ± 1	91 ± 4	84 ± 12
10	2	Conf-MPU-CE	1	67 ± 17	77 ± 3	78 ± 7	79 ± 8	85 ± 4	91 ± 1	95 ± 2	82 ± 8
10	2	Conf-MPU-CE	3	57 ± 17	80 ± 3	82 ± 1	80 ± 8	87 ± 1	90 ± 0	95 ± 2	82 ± 11
10	2	MPU-CE	1	60 ± 13	55 ± 9	62 ± 3	54 ± 11	55 ± 9	46 ± 6	67 ± 19	57 ± 6
10	2	MPU-CE	3	67 ± 8	62 ± 14	67 ± 7	63 ± 15	72 ± 5	59 ± 17	63 ± 14	65 ± 4

Table 13: Validation results in terms of F1-Score of CuPUL curriculum train with XLM-RoBERTa. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results for each setting are reported in bold. Selected hyperparameters are in bold. Learning rate is in the scale of $1e-5$

Invisible Speakers? Gender Disparity in German AI Discourse and Its Reflection in Language Models

Milena Belosevic

Bielefeld University

Faculty of Linguistics and Literary Studies

milena.belosevic@uni-bielefeld.de

Abstract

This paper investigates how language models (LMs) reproduce the existing gender disparity found in German media discourse about artificial intelligence (AI). Building on a human-annotated corpus of quotations from German media discourse on AI, we first quantify the frequency with which male and female speakers are directly cited across domains and speaker roles. We then train LLäMmlein (Pfister et al., 2025), a state-of-the-art German-only language model, GBERT, and a logistic regression model using only the quoted text as input and without providing any gender cues to classify the quotation as originating from a male or female speaker. By comparing model predictions with corpus-based gold labels, we find that male voices dominate both the corpus and the model predictions. Balancing the data mitigates but does not fully eliminate this disparity, indicating that the strong male-default tendency of transformer models cannot be explained by corpus skew alone, but also by their priors from pretraining. The study contributes to the interpretability of language models' output for DH-related tasks, adaptation of NLP tools to domain-specific humanities corpora, and knowledge modelling in the humanities.

1 Introduction

Current language models are trained on large corpora in which male speakers are typically more visible than female speakers (Brennen et al., 2019), particularly in domains stereotypically associated with men, where male speakers are more frequently cited than female speakers. We hypothesise that such models replicate patterns from their training data and tend to assign the male label more regularly than the female label. This model behaviour can be tested in cases where gender is not explicitly marked in the text. However, to our knowledge, this kind of model performance has not been investigated on German-only language models.

In this paper, we draw on a gender classification task as a diagnostic probe. Therefore, the goal of the paper is not to build a "gender predictor" in which the model learns gender from linguistic cues in the quotes, nor to create a system that accurately infers the "true" gender of quoted speakers, but to test how the training distribution (including bias) affects the model's behaviour and how pre-existing biases in language models surface when they are applied to quotations. By removing overt gender markers and evaluating on a balanced test set, any systematic male–female asymmetry reflects learned associations in the model and/or in the training distribution rather than explicit cues. This makes the setup meaningful for DH use cases where such models may be used to (semi-)automatically tag speaker metadata in corpora.

Specifically, we focus on two research questions: (1) How are gender and social roles distributed in German AI discourse?, and (2) Do model predictions reproduce the empirical gender imbalance in our dataset? We compile a set of direct quotations from the German media discourse about AI, for a stratified subset in which the speaker's gender can be identified from external sources, and we compare the corpus-based gender distribution of cited speakers with the gender predictions produced by the fine-tuned German-only model LLäMmlein (Pfister et al., 2025). We compare its performance with GBERT and logistic regression.

After providing a theoretical background (Section 2) and methods (Section 3), we present results in Section 4 and discuss their implications in Section 5.

2 Related work

We use *disparity* as the main term for the corpus and model outputs because our primary measurements are descriptive and observable (Barocas

et al., 2023): (1) the unequal representation of male vs. female quoted speakers in the dataset, and (2) asymmetric error rates of models. We do not claim to identify the underlying causal mechanisms (e.g., intentional discrimination, editorial practices, or broader societal inequities) and therefore avoid the more normatively loaded label *bias* for the empirical counts themselves. At the same time, we connect our findings to the broader bias-in-NLP literature (Bartl et al., 2025; Gallegos et al., 2024) by treating these disparities as downstream risks. When used as annotation assistants, models can systematically under-tag female speakers.

The topic has received growing attention with the advancement of language models (Jenny et al., 2024; Ho et al., 2025) and use mainly English data (Yang et al., 2025). Recent studies focus on detecting gender bias in literary texts or song lyrics, using English data and word embeddings or topic modelling methods (Chen et al., 2025). Another line of research is centred around identifying gender disparities in LLM-generated responses (Rhue et al., 2024; Wan et al., 2023; Fulgu and Capraro, 2024), creating alignment datasets (Zhang et al., 2025) or proposing (LLM-based) methods for identifying gender imbalances in training data (Derner et al., 2025). In German, other types of bias, such as epistemological bias in parliamentary debates, were investigated in Rehbein et al. 2024. Brennen et al. 2019 examine which AI experts are granted authority to shape narratives in the media over 30 years in the US and the UK, noting that the ten most-cited AI scholars account for 70% of the media space in both countries. They also find a strong bias against female experts in their dataset. Nguyen and Hekman 2022 found that reporting about AI is dominated by a small group of ‘AI alpha journalists’ who account for a disproportionate number of articles. Ryazanov et al. 2024 observe a significant increase in mentions of government agencies and people in leadership positions (via the Leadership frame), suggesting that industrial leaders and governments play a greater role in AI discourse post-ChatGPT.

Quotation detection is only part of our methodology in Section 3, and we do not aim to develop a new approach to this technique. However, it is noteworthy that many studies address methods for automatic quotation identification and attribution, but usually do not focus on gender asymmetries (e.g., Petersen-Frey and Biemann, 2024b; Brunner, 2013; Tekir et al., 2023; Petersen-Frey and

Biemann, 2024a) and use literary dialogues as a dataset (Van Cranenburgh and Van Den Berg 2023, Underwood et al. 2018).

3 Methodology

Data collection and preprocessing. We queried the subcorpus Webmonitor from the Digitales Wörterbuch der deutschen Sprache (DWDS)¹ for the token *KI* (künstliche Intelligenz ‘artificial intelligence’) for the period 01.01.2025-01.12.2025. The Webmonitor subcorpus (Nolda et al., 2023) is a daily-updated collection of German-language web sources. It is therefore particularly suitable for observing current linguistic trends. It comprises online texts, blogs, and news articles on various topics (politics, sports, lifestyle). Given the high number of hits (131,094) for this period, we restricted the search results to one month (31.10.2025-30.11.2025) to make human annotation manageable. This restriction yields 15,003 hits. Based on the search results, we constructed a corpus of articles that explicitly use this term. Henceforth, we refer to this corpus as the AI Corpus. Note that we define AI discourse as quotations from news and web articles whose main topic is AI. Many quotations do not explicitly mention AI or use the term, but they address its consequences, risks, governance, or related economic and social issues. We exported hit lists with source URLs from DWDS. Since DWDS does not support full-text retrieval, we crawled and cleaned full-text articles using Trafilatura (Barbaresi, 2021). We retrieved 6,059 full articles. Importantly, DWDS was used only to export hit lists including article URLs, not to retrieve full texts. Full texts were downloaded from the linked publisher pages and extracted with Trafilatura based on the exported URL lists.

Quotes subset and preprocessing. We automatically extracted candidate quotations and speaker candidates from the full articles using a combined regex–spaCy heuristic. First, we identified all text segments enclosed in German quotation marks (,...““, “...“, »...«) in each article and retained a context window of ± 200 characters around each segment. On these context snippets, we applied the German spaCy model (de_core_news_md) to obtain tokenisation, POS tags, lemmata and dependency parses. We then searched each snippet for a reporting verb whose lemma belongs to a manually

¹<https://www.dwds.de/>

compiled list of communication verbs and whose grammatical subject (dependency labels sb/nsubj) refers to a speaker candidate. Our reporting-verb list is intentionally conservative: we prioritise precision in speaker attribution over exhaustive coverage, and accept lower recall because candidates are subsequently sampled and manually filtered for relevance.

As potential speaker candidates, we treated (i) proper names, (ii) personal pronouns and (iii) noun subjects whose dependency subtree contains person-like elements, capturing patterns like *die Sprecherin von Bitkom, Anja Müller* 'the speaker of Bitkom, Anna Müller'. For each quotation, we stored a boolean flag indicating whether such a reporting-verb–subject configuration was present, as well as the surface form of the subject as a *speaker_candidate*. Only quotations with at least one speaker candidate near a reporting verb were retained as candidates for subsequent manual annotation. The quotations coming from the same speaker were not merged. In this way, we extracted 7,142 quotations.

Quotations that were clearly not relevant to our research question (e.g., quoted dialogues from book and song titles, or ironical quotes like KI "understands") were manually excluded by the author (3174 cases). This cleaned subset, comprising 3968 quotations, was provided to human annotators. We refer to this subset as the Quotation Corpus.

For all modelling experiments, we used only the quotation text as input and systematically removed explicit gender cues referring to the speaker (e.g. names, pronouns, or gendered titles). Thus, the models were exposed to quotations that, as far as could be automatically and manually detected, contained no direct gender information about the quoted person. Indirect correlations between wording and gender (e.g. topic or style differences) may remain. We did not alter non-speaker-related gendered language inside the quote.

Human annotation. Two annotators first examined the quotations from the corpus for their relevance to AI discourse, providing a binary decision for each quotation (relevant or irrelevant) based on annotation guidelines². Annotators saw the quotation together with the surrounding reporting clause and ± 1 sentence of context. Because of cognitive load and potential inconsistencies (annotators may

read different amounts of text), we retain full articles only for complex cases during the adjudication step. The agreement between annotators was substantial (Cohen's $\kappa = 0.72$). Cases of disagreement were resolved through majority vote performed by the author. Of 3968 quotations, 3195 were annotated as relevant, and 773 were labelled as irrelevant after adjudication.

The author annotated these 3195 quotations for the perceived gender of the quoted speaker (male, female, or neutral/unknown). Because gender was not identifiable in all quotations and the annotators were not provided with the full texts, the speaker's gender was annotated by the author based on context or, in ambiguous cases, on the full article. To estimate inter-annotator reliability, 200 of the 3,195 quotations with clearly identifiable speaker gender were randomly sampled and independently labelled by a second annotator, yielding Cohen's $\kappa = 0.98$ (almost perfect agreement with the author's labels). After adjudication, these quotations remained in the corpus and are included in the final dataset used for modelling. We treat gender as a property of how speakers are represented in the media, not as a claim about their identity, and explicitly acknowledge the limitations of binary gender labels. Quotations annotated as *neutral/unknown* (277 of 3,195, e.g., multiple speakers or only a surname mentioned in the full text) were excluded, leaving 2,918 quotations for modelling.

The two annotators who assessed the relevance of quotations for AI discourse also annotated the 2,918 quotations for *social_role* (company_executive, expert, researcher, politician, journalist, spokesperson, worker, user, artist, cleric, activist, celebrity) and *domain* (work, everyday use, IT domain, finance, culture and arts, data_privacy, education, health, cybersecurity, military) according to the annotation guidelines. Domain labels were derived from outlet section metadata (via DWDS hit lists) and harmonised into a unified scheme. Source-role labels follow the taxonomy proposed by Asr et al. 2021b. The author resolved disagreements to obtain a gold standard. Inter-annotator agreement was very high, with Cohen's $\kappa = 0.982$ for social roles and $\kappa = 0.922$ for domains, providing a robust basis for the subsequent modelling experiments. The *speaker_candidate* field was automatically identified (see above). The author reviewed these candidates and, where necessary, drew on the full article to provide information about the speaker's job or role.

²Annotation guidelines and code are available at: <https://osf.io/m5xqw/overview>

Baselines. As a trivial reference point, we include a majority baseline that always predicts the majority class (“male”) observed in the annotated dataset. This baseline reflects the global skew in the data. Evaluating all models against this baseline on the balanced test set allows us to see whether they learn anything beyond reproducing the overall male dominance. As a standard supervised NLP baseline, we train a logistic regression classifier on TF-IDF representations of the quotation text. Unlike pretrained language models, this classifier has no external-world knowledge and no pretraining corpus. Any systematic preference for male or female labels would therefore directly reflect the labelled AI data. Next, we fine-tune the GBERT base (deepset/gbert-base, Chan et al. 2020) transformer model as a binary sequence classifier with labels *male* and *female*. We select GBERT as a strong German encoder baseline because it is a widely used BERT-style model pretrained on large German corpora (news, web text, Wikipedia), thereby providing a well-established reference point for German text classification.

Main model. We use the LLäMmlein 1B (LSX-UniWue/LLaMmlein_1B, Pfister et al. 2025), an open German-only decoder-only language model trained from scratch on German data. In contrast to GBERT’s encoder architecture, LLäMmlein is a generative transformer that is typically used for text generation. We adapt it to our classification task by fine-tuning it with QLoRA, a memory-efficient adapter method well-suited to small datasets (Detmers et al., 2023). By comparing GBERT and LLäMmlein, we can test whether the observed male-default tendency persists across different transformer types. The TF-IDF/logistic baseline serves as a non-pretrained control. The key hyperparameters for GBERT-N/B, LLäMmlein-N/B, and the TF-IDF/logistic regression baseline are provided in Appendix Table 14. We also explored several prompting-based approaches (zero and three-shot prompting with the unfine-tuned LLäMmlein-1B model), but all proved methodologically unreliable. For all models, the inputs are de-gendered quotations, without broader article context. We excluded quotations with fewer than four tokens (191 of the 2,918 manually annotated quotations) from modelling experiments, as they provide too little linguistic material for gender classification. From the 2,727 remaining quotations, we first drew a fixed, held-out test set of 171 quotations (86

female, 85 male) by stratified random sampling on gender only. All models are evaluated on this same balanced test set. The remaining 2,556 quotations form the pool for training and development. In the natural condition (-N), we randomly split this pool into 2,044 training instances and 512 development instances, preserving the observed gender skew. In the balanced condition (-B), we downsample the non-test items only by gender to create a 50/50 balanced training set of 796 quotations and a separate balanced development set of 171 quotations. Thus, -N and -B are two alternative ways of partitioning the same 2,556 non-test items. In both cases, the balanced test set is never used for downsampling or model selection. Alternative strategies such as class-weighted loss functions were considered but not pursued, as our primary aim was to contrast a “natural” gender distribution with an explicitly balanced training regime in a transparent and easily interpretable way. Exploring more fine-grained debiasing techniques (e.g. class weights, focal loss) is left for future work. The -N/-B design was applied only to the two transformer models to assess how they reflect the gender disparity in our dataset and how changes in the training distribution affect their performance. Logistic regression is trained and tuned on the same balanced train/validation splits as the -B transformer models and evaluated on the same balanced test set.

Evaluation metrics include overall accuracy, macro-averaged F1 score, per-gender precision and recall, and the proportion of quotations predicted as male versus female. The qualitative error analysis focuses on misclassifications across social roles and domains.

4 Results

Descriptive statistics. Of 2,918 manually annotated quotations, 79% are attributed to male speakers (2,304 instances) and only 21% to female speakers (614 cases), indicating a substantial gender disparity in the dataset and supporting results obtained on English data (Asr et al., 2021a).

The distribution of social roles is highly uneven (see Appendix Table 8). Half of all quotations are attributed to company executives (1,459 instances), followed by experts (15%), researchers (11%) and politicians (7.8%); all other roles (artist, user, worker, spokesperson, journalist, cleric, activist, celebrity) each account for less than 4% of the corpus. In almost all roles, male speakers pre-

dominate: for example, 85.3% of company executive quotations and 77.9% of expert quotations are attributed to men, and men almost exclusively hold roles such as activist (95% male). Female speakers dominate only in a few comparatively small categories, such as journalists (64% female), and are closer to parity among spokespersons (44.8% female), users (36.4% female), and workers (38.6% female, all row-wise percentages). Because several of these roles have low absolute counts, we treat such patterns as descriptive tendencies rather than robust statistical differences. Human annotations serve as a gold standard to assess how different models reproduce this gender disparity.

The two largest domains are work (31.2% of all quotations) and everyday use (29.8%), followed by IT (17.6%). Other domains, such as finance, education, culture, or health, account for much smaller shares (Appendix Table 11). Within almost all domains, male speakers clearly predominate. For example, 84.3% of quotations in the work domain and 86.9% in the IT domain are attributed to men, with similarly high male shares in finance (76.4%), data_privacy (81.3%) and especially cybersecurity (94.1%). Female speakers constitute a higher relative proportion only in a few areas, most notably education (49.5% female, essentially parity), culture (35.3% female), and health (36.1% female). Column-wise percentages show that male quotations are concentrated in work, everyday use and IT. In contrast, female quotations are more dispersed and relatively more frequent in everyday use, education, culture and health. As with social roles, some of these domain-specific patterns involve small absolute numbers and should be interpreted cautiously, but overall they reinforce the picture of an AI discourse in which male voices dominate across most topics. Tables 1 and 2 summarise the gender distribution across most frequent social roles and domains (for a more detailed breakdown, see Appendix Tables 9,10,12, and 13).

Finally, we examined gender based differences in quotation length. Overall, quotations in the corpus are relatively short (median length 16 tokens, IQR 9–26), with a long tail of longer quotations (up to 188 tokens). Female quotations have a mean length of 19.8 tokens (median 15), whereas male quotations have a mean length of 20.4 tokens (median 16). These minor differences suggest that the

³Aggregated category comprising artist, user, worker, spokesperson, journalist, cleric, activist, and celebrity.

Social role	<i>N</i>	Female <i>n</i> (%)	Male <i>n</i> (%)
company ex.	1459	215 (14.7%)	1244 (85.3%)
expert	439	97 (22.1%)	342 (77.9%)
researcher	320	89 (27.8%)	231 (72.2%)
politician	228	56 (24.6%)	172 (75.4%)
other ³	472	157 (33.3%)	315 (66.7%)
Total	2918	614 (21.0%)	2304 (79.0%)

Table 1: Gender distribution across major social roles in the Quotation Corpus (row-wise percentages: for each social role, they indicate the proportion of female vs. male quotations within that role).

Domain	<i>N</i>	Female <i>n</i> (%)	Male <i>n</i> (%)
work	909	143 (15.7%)	766 (84.3%)
everyday use	869	216 (24.9%)	653 (75.1%)
IT	514	67 (13.0%)	447 (87.0%)
other ⁴	626	188 (30.0%)	438 (70.0%)
Total	2918	614 (21.0%)	2304 (79.0%)

Table 2: Gender distribution across major domains in the Quotation Corpus (row-wise percentages).

strong gender imbalance observed in our data is not an artefact of men being quoted at much greater length.

Majority baseline. On the balanced test set, this baseline reaches an accuracy of 0.50 but completely fails to recognise female quotations: male instances are classified correctly in 100% of cases, while all 86 female quotations are misclassified as male (F1 = 0.00 for female; confusion matrix: 85 male→male, 86 female→male). This behaviour mirrors a pure “male-by-default” strategy and provides a lower bound against which we can compare more sophisticated models.

Logistic regression. On the balanced test set, this model provides a comparatively neutral benchmark: accuracy and macro-F1 are 0.57, with almost identical performance for male and female quotations (F1 ≈ 0.57 for both classes). Compared with the majority baseline, the TF-IDF model moves beyond reproducing the global skew and instead offers a relatively balanced treatment of male and female speakers (37 female→male vs. 36 male→female misclassifications on the balanced test set), providing a useful classical benchmark for transformer- and LLM-based models.

⁴Aggregated category comprising finance, culture, data_privacy, education, health, cybersecurity and military.

As shown in Appendix Tables 16 and 15, within individual roles and domains, row-wise proportions fluctuate (e.g., a slight dominance of male→female errors for company executive and expert, and more female→male errors in finance and culture), but these patterns are based on small absolute counts and remain relatively balanced overall. For example, the model incorrectly predicts that a quotation by a female researcher in finance is authored by a male. However, it also predicts a female speaker for a quotation by a male expert in the IT domain, which contrasts with stereotypical associations with typical male and female professions and expertise (see Table 3⁵). This error profile supports our interpretation of the TF-IDF model as a comparatively “neutral” baseline on the balanced data, against which the more strongly gender-skewed behaviour of GBERT and LLäMmleIn can be contrasted.

Quote	Gold	Pred	Role / Domain
It appears that investors are taking profits on these AI-related stocks.	female	male	finance / expert
AI is trained, not programmed.	male	female	expert / IT

Table 3: Logistic regression misclassification examples.

GBERT. Under natural training (-N), GBERT exhibits a strong male-default behaviour on the balanced test set: a good performance on the validation set (val. accuracy ≈ 0.81 , macro-F1 ≈ 0.53 after three epochs) changes completely when we evaluate on the balanced test set, where overall accuracy drops to 0.49 (chance level), with an F1-score of 0.65 for male but only 0.06 for female. The confusion matrix shows that GBERT correctly identifies 81 of 85 male quotations (recall 0.95) but misclassifies 83 of 86 female quotations as male (recall 0.03). In total, it assigns the male label to 164 of 171 test instances (96%) despite the 50/50 class balance. Taken together, GBERT-N largely reproduces the training skew and heavily underrepresents female speakers in quote-only classification.

The qualitative error analysis shows highly asymmetric error patterns. All misclassified male quotes cluster around the use of AI in domains stereotypically associated with women, such as children,

⁵In all tables, examples are translated from German, glosses are omitted for space reasons.

parenting, and education (see Table 4). Similarly, GBERT-N frequently misclassifies female roles stereotypically associated with men, such as female company executives or experts (see Appendix Table 16 for a quantitative summary). A similar pattern appears at the domain level. Female speakers in the core AI domains, such as IT, are almost exclusively misclassified as male. Only in everyday use do we observe a small number of male→female errors at all. In contrast to the comparatively symmetric errors of the logistic regression baseline, these results show that GBERT-N does not merely make random errors on the balanced test set. Instead, it exhibits a male default, which is particularly evident among female speakers in high-status roles and in central AI domains.

Quote	Gold	Pred	Role / Domain
The models reproduce these mental patterns and reinforce them.	female	male	researcher / IT
AI cannot replace the relationship with friends or parents.	male	female	expert / use

Table 4: GBERT-N misclassification examples.

Balanced fine-tuning of GBERT (GBERT-B) improves over the naturally trained model, but a male-default tendency remains. After three epochs, validation accuracy stabilises at ≈ 0.59 (macro-F1 = 0.59). On the balanced test set, GBERT-B achieves an accuracy of 0.56 and a macro-F1 of 0.54. However, the confusion matrix shows a clear asymmetry: 67/85 male quotations are classified correctly (recall 0.79) but only 29/86 female quotations (recall 0.34), with female→male errors still more than three times as frequent as male→female (57 vs. 18).

Qualitative examples (Table 5) illustrate that misclassified male examples are related to domains stereotypically associated with women (e.g., feelings), and vice versa: topics such as robotics are wrongly associated with male speakers (see examples in Table 5). Across social roles, female company executives, researchers, and experts are much more often misclassified as male than vice versa (see Appendix Table 15 for a quantitative summary). Overall, balancing mitigates but does not fully remove the model’s inherited gender dis-

parity.

Quote	Gold	Pred	Role / Domain
When robots are equipped with AI, AI gets a body.	female	male	journalist / IT
AI does not feel sad when it experiences 'pain'.	male	female	comp. exec. / IT

Table 5: GBERT-B misclassification examples.

LLäMmlein (natural skew). Under natural training (-N), LLäMmlein also exhibits an evident male-default tendency on the balanced test set. After three epochs, it reaches an accuracy of 0.55 and a macro-F1 of 0.47. Class-wise, it performs well on male quotations (F1 = 0.68, recall = 0.94), but much worse on female quotations (F1 = 0.27, recall = 0.16). The confusion matrix shows that it correctly identifies 80 of 85 male quotes, but misclassifies 72 of 86 female quotes as male, with only 14 female quotes correctly recognised.

The error analysis again reveals a strongly asymmetric pattern. Especially female company executives are systematically “pulled” towards the male label (see examples in Table 6). Across domains, we observe the same disparity: female speakers in IT and finance are overwhelmingly misclassified as male, with male-to-female errors occurring only sporadically in everyday use and work (see Appendix Table 16 for a quantitative summary). Compared with the more symmetric error profile of the logistic regression baseline, LLäMmlein-N thus behaves very similarly to GBERT-N.

Quote	Gold	Pred	Role / Domain
The algorithm itself plays an important role (...).	female	male	expert / use
AI chatbots pose a serious threat to our children (...).	male	female	politician / use

Table 6: LLäMmlein-N misclassification examples.

LLäMmlein (balanced fine-tuning). Balanced fine-tuning of LLäMmlein (LLäMmlein-B) leads to a much more symmetric behaviour than the naturally trained variant. On the balanced test set, the model achieves an accuracy of 0.57 and a macro-F1 of 0.57, with performance similar for male (F1

= 0.58, recall = 0.61) and female quotations (F1 = 0.55, recall = 0.52). Overall, despite a slight tendency to overpredict male labels, balancing the training data substantially reduces the strong male default observed in LLäMmlein-N.

However, domain-linked gender associations remain. For example, the model assumes a female speaker for educational contexts and a male speaker in the health domain (see Table 7). The errors are almost evenly distributed (41 female→male and 33 male→female). The symmetry in social roles is most evident among company executives, where female and male speakers are misclassified as the opposite gender at equal rates. In other roles, however, female researchers and experts are still more often “pulled” towards the male label, whereas some smaller categories (e.g., journalists) show only female→male errors. Across domains, LLäMmlein-B behaves more evenly: in IT and culture, errors are roughly balanced, and in everyday use male→female errors even outnumber female→male ones, whereas in finance, health, and military, all misclassifications still go from female to male (see Appendix Tables 16 and 15 for a quantitative summary).

Quote	Gold	Pred	Role / Domain
With AI, we bring precision to the surgery room.	female	male	politician/health
This is how we combine digital innovation with real pedagogical impact.	male	female	expert/education

Table 7: LLäMmlein-B misclassification examples.

Comparing GBERT and LLäMmlein shows that per-gender gaps (Δ = male – female for F1 and recall) are large under -N: GBERT-N yields $\Delta F1 = 0.59$ and $\Delta Recall = 0.92$, and LLäMmlein-N $\Delta F1 = 0.41$ and $\Delta Recall = 0.78$. Under -B, LLäMmlein becomes close to symmetric ($\Delta F1 = 0.04$, $\Delta Recall = 0.09$), whereas GBERT still shows a notable gap ($\Delta F1 = 0.21$, $\Delta Recall = 0.45$). A simple non-parametric bootstrap over the 171 test instances (B=10,000; percentile 95% CI) confirms that the -N gaps are far larger than sampling noise (e.g., GBERT-N $\Delta F1$ CI [0.49, 0.67]; LLäMmlein-N [0.30, 0.53]), while the balanced LLäMmlein-B gap is compatible with zero (CI includes 0).

Exploratory tests find no statistically significant domain dependence of error direction: only LLäMmlein-B shows a modest omnibus associ-

ation (permutation $p=0.024$), with finance being suggestive (Fisher $p=0.006$; FDR $q=0.057$). For social roles, an omnibus permutation test indicates an association only for GBERT-N ($p=0.026$); male→female errors are rare and confined to few roles, and no role remains significant after FDR.

5 Conclusions

This paper investigated the extent to which German language models reproduce the existing gender disparities found in German media discussions about AI. The approach accounts for the need to contextualise fairness criteria used to evaluate biases in LLM systems (Anthis et al., 2025) by focusing on a specific context (public AI discourse).

On a balanced test set (50% male / 50% female), neutral behaviour would correspond to roughly symmetric performance across genders. Instead, we observe that both GBERT-N and LLäMmlein-N systematically overpredict male speakers when female speakers are equally represented in the gold data. Therefore, gender disparity in their predictions cannot be reduced to the annotated data alone but is also shaped by their pretraining and architecture. This result is also supported by the observation that a logistic regression baseline performs on par with, or slightly better than, GBERT and LLäMmlein on the balanced test set and exhibits a highly symmetric treatment of male and female quotations. The value of more complex architectures in our study lies, therefore, less in raw performance and more in their ability to reveal the interaction between training distributions (natural vs. balanced) and gender disparities in predictions. Hence, if DH researchers use GBERT or LLäMmlein as an annotation assistant for speaker gender in AI corpora, they risk systematic under-tagging of female speakers, especially in high-status roles. To mitigate skew, they should monitor per-group errors and rely on human validation.

In addition to the mitigation strategies (see e.g., Ferrara 2023), we recommend raising researchers' awareness of such biases by cultivating healthy distrust (Paaßen et al., 2025) toward model outputs.

Retraining on a balanced subset of the same material improves the performance of both models, yet in different ways: GBERT-B becomes a stronger classifier overall but still systematically prefers the male label, while LLäMmlein-B not only gains in macro-F1 but also moves much closer to a symmetric treatment of male and female speakers, with

comparable recall and more balanced error patterns. Therefore, model choice and training regime matter for downstream DH analyses, such as authorship attribution (Rybicki, 2025). From a DH perspective, this also suggests that simple data balancing cannot completely eliminate gender disparities in model outputs, and that different German architectures respond differently to such corrective interventions, which is highly relevant for the use of these models in the study of public discourse. Additionally, the comparison between a natural and a balanced training illustrates that fine-tuning is not a neutral operation: depending on how we curate training data, we can obtain models with different bias profiles, which is highly relevant for DH scenarios where such models might be used as annotation assistants.

Overall, this paper makes three contributions: (1) a quote corpus on German AI media discourse annotated with gender, social roles, and domains (2) a diagnostic probe setup for gender prediction, enabling interpretable analysis of model priors, and (3) empirical evidence that transformer-based models can exhibit a male-default tendency that persists under balancing, with implications for using LMs as annotation tools in DH pipelines.

Limitations and future work

The study has several limitations. First, the analysis is based on a one-month snapshot of texts from a single dataset. As a result, the observed gender distributions and model behaviours reflect this specific period and text type. They should not be interpreted as long-term trends in German media as a whole. Future work will extend the time span and include additional domains to examine temporal variation and generalizability beyond AI discourse.

While we treat gender as the target variable, a complementary line of work could investigate implicit gender disparity, i.e., whether models treat male and female speakers differently on other tasks such as credibility assessment, stance detection or topic classification. This would involve using gender only as a group attribute and analysing differences in model behaviour across gender groups. Furthermore, reliance on binary gender labels may limit the generalizability of our findings to non-binary or gender-diverse speakers. Next, our extraction heuristic is intentionally conservative (high precision for speaker attribution) and does not aim to capture all quotations. As estimating overall re-

call would require full manual annotation of quotes, we leave it for future work.

Finally, since we trained and evaluated the models on quotations without their local context and speaker-related gender markers, future work should examine model behaviour by including (de-gendered) local context.

References

- Jacy Reese Anthis, Kristian Lum, Michael Ekstrand, Avi Feller, and Chenhao Tan. 2025. [The impossibility of fair LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 105–120, Vienna, Austria. Association for Computational Linguistics.
- Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vagrant Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021a. [The gender gap tracker: Using natural language processing to measure gender bias in media](#). *PLOS ONE*, 16(1):e0245533.
- Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vasundhara Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021b. [The gender gap tracker: Using natural language processing to measure gender bias in media](#). *PLoS One*, 16(1):e0245533.
- Adrien Barbaresi. 2021. [Trafalatura: A web scraping library and command-line tool for text discovery and extraction](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.
- Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.
- Marion Bartl, Abhishek Mandal, Susan Leavy, and Suzanne Little. 2025. [Gender bias in natural language processing and computer vision: A comparative survey](#). *ACM Computing Surveys*, 57(6).
- J Brennen, Anne Schulz, Philip Howard, and R Nielsen. 2019. [Industry, experts, or industry experts? Academic sourcing in news coverage of AI](#). *RIJS Factsheets*.
- Annellen Brunner. 2013. [Automatic recognition of speech, thought, and writing representation in german narrative texts](#). *Literary and Linguistic Computing*, 28(4):563–575.
- Branden Chan, Stefan Schweter, and Timo Möller. 2020. [German’s next language model](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Danqing Chen, Adithi Satish, Rasul Khanbayov, Carolin Schuster, and Georg Groh. 2025. [Tuning into bias: A computational study of gender bias in song lyrics](#). In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 117–129, Albuquerque, New Mexico. Association for Computational Linguistics.
- Erik Derner, Sara Sansalvador De La Fuente, Yoan Gutierrez, Paloma Moreda Pozo, and Nuria M Oliver. 2025. [Leveraging large language models to measure gender representation bias in gendered language corpora](#). In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 468–483, Vienna, Austria. Association for Computational Linguistics.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized LLMs](#). *Preprint*, arXiv:2305.14314.
- Emilio Ferrara. 2023. [Should ChatGPT be biased? Challenges and risks of bias in large language models](#). *First Monday*.
- Raluca Alexandra Fulgu and Valerio Capraro. 2024. [Surprising gender biases in GPT](#). *Computers in Human Behavior Reports*, 16:100533.
- Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. [Bias and fairness in large language models: A survey](#). *Computational Linguistics*, 50(3):1097–1179.
- Jerlyn Q.H. Ho, Andree Hartanto, Andrew Koh, and Nadyanna M. Majeed. 2025. [Gender biases within artificial intelligence and chatgpt: Evidence, sources of biases and solutions](#). *Computers in Human Behavior: Artificial Humans*, 4:100145.
- David F. Jenny, Yann Billeter, Bernhard Schölkopf, and Zhijing Jin. 2024. [Exploring the jungle of bias: Political bias attribution in language models via dependency analysis](#). In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 152–178, Miami, Florida, USA. Association for Computational Linguistics.
- Dennis Nguyen and Erik Hekman. 2022. [The news framing of artificial intelligence: a critical exploration of how media discourses make sense of automation](#). *AI & Society*.
- Andreas Nolda, Adrien Barbaresi, and Alexander Geyken. 2023. *Korpora für die lexikographische Beschreibung diatopischer Variation in der deutschen Standardsprache*, pages 29–52. De Gruyter, Berlin, Boston.

- Benjamin Paaßen, Suzana Alpsancar, Tobias Matzner, and Ingrid Scharlau. 2025. [Healthy distrust in AI systems](#).
- Fynn Petersen-Frey and Chris Biemann. 2024a. [Dataset of quotation attribution in German news articles](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4412–4422, Torino, Italia. ELRA and ICCL.
- Fynn Petersen-Frey and Chris Biemann. 2024b. Fine-grained quotation detection and attribution in German news articles. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 196–208, Vienna, Austria. Association for Computational Linguistics.
- Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. [LLäMlein: Transparent, compact and competitive German-only language models from scratch](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.
- Ines Rehbein, Josef Ruppenhofer, Annelen Brunner, and Simone Paolo Ponzetto. 2024. Out of the mouths of MPs: Speaker attribution in parliamentary debates. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12553–12563, Torino, Italia. ELRA and ICCL.
- Lauren Rhue, Sofie Goethals, and Arun Sundararajan. 2024. [Evaluating LLMs for gender disparities in notable persons](#).
- Igor Ryazanov, Carl Öhman, and Johanna Björklund. 2024. How ChatGPT changed the media’s narratives on AI: A semi-automated narrative analysis through frame semantics. *Minds & Machines*, 35(1).
- Jan Rybicki. 2025. [Can machine translation of literary texts fool stylometry?](#) *Digital Scholarship in the Humanities*, 40(1):268–276.
- Selma Tekir, Aybüke Güzel, Samet Tenekeci, and Bekir Haman. 2023. [Quote detection: A new task and dataset for NLP](#). In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 21–27, Dubrovnik, Croatia. Association for Computational Linguistics.
- Ted Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in English-language fiction. *Journal of Cultural Analytics*.
- Andreas Van Cranenburgh and Frank Van Den Berg. 2023. [Direct speech quote attribution for Dutch literature](#). In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 45–62, Dubrovnik, Croatia. Association for Computational Linguistics.
- Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. “Kelly is a warm person, Joseph is a role model”: Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.
- Jinrui Yang, Xudong Han, and Timothy Baldwin. 2025. Demographics and democracy: Benchmarking LLMs’ gender bias and political leaning in European parliament. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 416–439, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.
- Tao Zhang, Ziqian Zeng, YuxiangXiao YuxiangXiao, Huiping Zhuang, Cen Chen, James R. Foulds, and Shimei Pan. 2025. [GenderAlign: An alignment dataset for mitigating gender bias in large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11293–11311, Vienna, Austria. Association for Computational Linguistics.

A Appendix

A.1 Overall distribution of social roles

Social role	Count	Percent of all quotes
company_executive	1459	50.00%
expert	439	15.04%
researcher	320	10.97%
politician	228	7.81%
artist	115	3.94%
user	88	3.02%
worker	70	2.40%
spokesperson	67	2.30%
journalist	50	1.71%
cleric	47	1.61%
activist	20	0.69%
celebrity	15	0.51%
Total	2918	100.00%

Table 8: Distribution of social roles in the annotated AI Quotation Corpus.

A.2 Gender distribution by social role (row-wise percentages)

Social role	<i>N</i>	Female <i>n</i>	Male <i>n</i>	Female %	Male %
company_executive	1459	215	1244	14.74%	85.26%
expert	439	97	342	22.10%	77.90%
researcher	320	89	231	27.81%	72.19%
politician	228	56	172	24.56%	75.44%
artist	115	31	84	26.96%	73.04%
user	88	32	56	36.36%	63.64%
worker	70	27	43	38.57%	61.43%
spokesperson	67	30	37	44.78%	55.22%
journalist	50	32	18	64.00%	36.00%
cleric	47	0	47	0.00%	100.00%
activist	20	1	19	5.00%	95.00%
celebrity	15	4	11	26.67%	73.33%

Table 9: Gender distribution by social role in the annotated AI Quotation Corpus (row-wise percentages).

A.3 Gender distribution by social role (column-wise percentages)

Social role	Fem. % of all fem.	Male % of all males
company_executive	35.02%	53.99%
expert	15.80%	14.84%
researcher	14.50%	10.03%
politician	9.12%	7.47%
artist	5.05%	3.65%
journalist	5.21%	0.78%
user	5.21%	2.43%
worker	4.40%	1.87%
spokesperson	4.89%	1.61%
celebrity	0.65%	0.48%
activist	0.16%	0.82%
cleric	0.00%	2.04%

Table 10: Gender distribution by social role in the annotated AI Quotation Corpus (column-wise percentages).

A.4 Overall domain distribution

Domain	Count	Percent of all quotes
work	909	31.15%
everyday use	869	29.78%
IT	514	17.61%
finance	144	4.93%
culture	116	3.98%
data_privacy	112	3.84%
education	101	3.46%
health	86	2.95%
cybersecurity	34	1.17%
military	33	1.13%
Total	2918	100.00%

Table 11: Distribution of domains in the annotated AI Quotation Corpus.

A.5 Gender distribution by domain (row-wise percentages)

Domain	<i>N</i>	Female <i>n</i>	Male <i>n</i>	Female %	Male %
work	909	143	766	15.73%	84.27%
everyday use	869	216	653	24.86%	75.14%
IT	514	67	447	13.04%	86.96%
finance	144	34	110	23.61%	76.39%
culture	116	41	75	35.34%	64.66%
data_privacy	112	21	91	18.75%	81.25%
education	101	50	51	49.50%	50.50%
health	86	31	55	36.05%	63.95%
cybersecurity	34	2	32	5.88%	94.12%
military	33	9	24	27.27%	72.73%

Table 12: Gender distribution by domain in the annotated AI Quotation Corpus (row-wise percentages).

A.6 Gender distribution by domain (column-wise percentages)

Domain	Female % of all females	Male % of all males
work	23.29%	33.25%
everyday use	35.18%	28.34%
IT	10.91%	19.40%
finance	5.54%	4.77%
culture	6.68%	3.26%
data_privacy	3.42%	3.95%
education	8.14%	2.21%
health	5.05%	2.39%
cybersecurity	0.33%	1.39%
military	1.47%	1.04%

Table 13: Distribution of domains within each gender (column-wise percentages). Percentages indicate the distribution of each gender across domains.

A.7 Key hyperparameters and configurations for all models

Model	Variants	LR	Epochs	Batch _{train}	Max seq
GBERT	-N, -B	2×10^{-5}	3	16	128
LLäMmlein	-N, -B	2×10^{-4}	3	4	128
TF-IDF+LogReg	-B only	–	–	–	128 ^a
Majority	–	–	–	–	–

Table 14: Key hyperparameters for all models. GBERT-N/B and LLäMmlein-N/B share the same optimisation settings within each architecture and differ only in the training distribution (natural vs. downsampled balanced).

A.8 Misclassifications by social role and model

Social role	LogReg		GBERT-N		GBERT-B		LLäMmlein-N		LLäMmlein-B	
	F→M	M→F	F→M	M→F	F→M	M→F	F→M	M→F	F→M	M→F
company_executive	11	15	30	0	22	7	25	1	16	16
expert	5	8	9	2	9	2	8	2	5	3
researcher	5	3	15	0	11	2	12	0	8	1
politician	6	4	10	1	6	2	11	1	7	5
journalist	3	0	5	0	3	0	4	0	2	0
spokesperson	2	0	3	0	2	0	3	0	2	0
user	2	1	5	0	2	0	3	0	1	1
artist	2	2	5	0	2	2	5	0	2	4
worker	1	0	1	0	0	0	1	0	0	0
activist	0	0	0	0	0	0	0	0	0	1

Table 15: Misclassifications by social role and model. F→M = female quotation predicted as male; M→F = male quotation predicted as female. Empty cells indicate zero errors for that combination.

A.9 Misclassifications by domain and model

Domain	LogReg		GBERT-N		GBERT-B		LLäMmlein-N		LLäMmlein-B	
	F→M	M→F	F→M	M→F	F→M	M→F	F→M	M→F	F→M	M→F
IT	3	6	8	0	7	2	5	0	4	4
finance	5	2	11	0	9	0	11	0	8	0
work	7	8	22	0	16	5	20	1	13	9
everyday use	15	17	27	4	12	9	22	4	7	14
education	0	1	2	0	2	1	2	0	1	2
culture	2	1	4	0	3	1	4	0	2	2
health	3	1	5	0	4	0	5	0	3	0
military	2	0	3	0	3	0	3	0	2	0
data_privacy	0	0	1	0	1	0	1	0	1	2

Table 16: Misclassifications by domain and model. F→M = female quotation predicted as male; M→F = male quotation predicted as female. Empty cells indicate zero errors for that combination.

GlobLingDiv: A global dataset linking linguistic diversity and digital support to reveal landscapes with under-resourced languages for NLP

Katharina Zeh

Department of German Studies
University of Vienna
Vienna, Austria
katharina.zeh@univie.ac.at

Hannes Essfors

Faculty of Informatics
TU Wien
Vienna, Austria

Juliane Benson

Department of German Studies
University of Vienna
Vienna, Austria

Lale Tüver

Department of History
University of Vienna
Vienna, Austria

Andreas Baumann

Department of German Studies
University of Vienna
Vienna, Austria

Hannes Fellner

Department of
Comparative Literature
and Language Studies
University of Vienna
Vienna, Austria

Abstract

Linguistic diversity is increasingly under pressure globally and is becoming ever more relevant in digital contexts, where many languages remain structurally under-resourced, limiting access to language technologies and inhibiting equitable NLP development. To support linguistic diversity, publicly available data are needed that capture both the number of languages spoken and the distribution of speakers across them. We introduce GlobLingDiv, a database that uses country-level speaker distributions to derive language richness and entropy-based diversity measures, alongside a population-weighted digital language support measure. Applying these metrics globally, we examine the association between linguistic diversity and digital support conditions. The results reveal a substantial imbalance: highly diverse linguistic landscapes show comparatively low digital support, underscoring the need for more inclusive NLP environments.

1 Introduction

Linguistic diversity refers to the variety of spoken languages in a given community and the balance of speakers among those languages (Grin and Fürst, 2022). In recent years, linguistic diversity has decreased considerably, and languages are under threat of extinction. According to Ethnologue, a total of 7,164 languages are estimated worldwide, and by the end of the century, nearly half of them could disappear (Hutson et al., 2024; Kandler and Unger, 2023). Beyond cultural significance, linguistic diversity is increasingly relevant in digital contexts, where many lan-

guages remain structurally under-resourced and digitally marginal (Benson et al., 2025; Blasi et al., 2022). Recent initiatives have begun to systematically examine inequalities in digital language support—often focusing on specific regions or subsets of languages, such as the European Language Equality initiative (Gaspari et al., 2022; Grützner-Zahn and Rehm, 2022), yet comparable global-scale, speaker-based diversity estimates remain scarce.

To protect linguistic diversity worldwide and to work towards more inclusive digital and NLP environments, it is important not only to study the factors affecting it but also to provide reliable diversity estimates to begin with. However, a significant challenge is the lack of a comprehensive and publicly available dataset that quantifies linguistic diversity through both the number of languages and the distribution of speakers across them. An effective method of quantifying diversity is Shannon entropy (Shannon, 1948), which captures both richness and evenness based on speaker counts (see Section 3). Although entropy-based measures are increasingly used in linguistic diversity research (Grin and Fürst, 2022), they have been applied mainly to smaller regions or individual countries.

To address this gap, we have constructed a comprehensive database, GlobLingDiv, that quantifies linguistic diversity globally using Shannon entropy. The dataset draws on the Joshua Project’s ethnolinguistic data (Joshua Project, 2025), as well as information from Ethnologue (Eberhard et al., 2022).¹

¹We acknowledge that Joshua Project and Ethnologue have missionary origins and are subject to ethical debate. Our use

Continent	langs	fams	spk %	spk M
Africa	2180	45	17.7	1345
Americas	1141	89	13.4	1019
Asia	2315	57	58.8	4469
Europe	414	25	9.6	727
Oceania	1348	77	0.5	405
Global	6745	214	100.0	7601

Table 1: Dataset summary aggregated by continent. *langs*: number of languages; *fams*: language families; *spk %*: global speakers share; *spk M*: speakers in Millions.

In addition, we provide a speaker-weighted digital support measure based on the Digital Language Support scale (Simons et al., 2022), enabling linguistic diversity to be considered alongside the degree of digital representation. As an application, we examine country-level correlations between diversity and digital support conditions.

Our contribution is twofold: (1) we introduce a publicly available dataset that quantifies linguistic diversity using richness and entropy-based measures and adds a abundance-weighted digital language support indicator at the country level; and (2) we demonstrate the value of combining these measures by examining global patterns and the association between linguistic diversity and digital support, offering a basis for identifying regions with under-resourced linguistic settings that we suggest future NLP research to focus on.

2 Constructing the dataset

Since our aim is to provide a fine-grained measure of country-level linguistic diversity on a global scale, we constructed a dataset containing language distributions based on speaker numbers per country. As our objective was to approximate overall linguistic landscapes rather than to capture individual acquisition patterns, we did not differentiate between L1 and L2 speakers. Data processing and integration were carried out in Python.²

First, the Joshua Project provides detailed demographic information on people groups within each country, including spoken language and population size. To address the fact that multiple people

of their data is limited to quantitative information and does not imply endorsement. We encourage future work to integrate additional linguistic data sources to broaden perspectives and reduce reliance on any single provider.

²Data and code are made available at <https://gitlab.tuwien.ac.at/digilingdiv/globlingdiv>

groups might speak the same language, the populations were in such cases aggregated to produce total speaker counts per language per country. Second, information from Ethnologue (Eberhard et al., 2022) was also integrated into the final dataset. In some cases, speaker counts were embedded in text fields, requiring the use of regular expressions to extract numeric values.

The data were aligned so that each row represents a unique language–country pair. When speaker counts were available from more than one source, the minimum value was selected to avoid overestimation of speaker counts. Entries were merged using ISO language and country codes (ISO, 2007, 2013), enabling integration with other datasets. Speaker counts were transformed to fractions (i.e., probabilities) by using total speaker counts per country as the basis for normalization. That is, for each country, all entries, representing the distribution of languages in the country’s linguistic landscape, sum to 1.

The final dataset comprises 6,745 languages and 7,600,502,492 speakers in 239 countries, with a total of 12,249 language–country pairs (see Table 1), plus country-level totals. Approximately 95% of the linguistic richness estimated by Ethnologue is covered. GlobLingDiv is structured as follows: we provide five CSV tables containing: 1) a country-language-probability triplet. 2) country-level diversity metrics across three columns: richness, H , $exp(H)$; 3) a look-up table for each country with four columns: country ISO-code, country name, continent, abundance-weighted digital support score, total speaker count; 4) a look-up table for all languages included, consisting of two columns: language ISO-code and language name.

The limited coverage in our dataset can partly be attributed to the deliberate exclusion of extinct languages, which are included in Ethnologue. Additionally, inherent uncertainty associated with language data contributes to coverage issues: there is no universally accepted distinction between languages and dialects, such that closely related varieties may be treated as separate languages in some sources but as dialects in others, which often results in conflicting figures regarding both the number of languages and their speaker populations (Jolad and Agarwal, n.d.). Speaker estimates also rely heavily on national census data (Boissonneault et al., 2025), which have been criticized for political bias (Duchêne and Humbert, 2018), thereby reducing their ability to capture the true underlying language

ISO	Lang.	H	exp(H)
PG	845	4.57	96.65
CM	289	3.98	53.59
CA	240	2.13	8.41
VU	115	4.00	54.34
DE	94	1.60	4.94
CH	45	1.75	5.74
AT	43	1.05	2.85
GL	3	0.62	1.86

Table 2: An excerpted GlobLingDiv table showing richness (Lang.), Shannon entropy (H), and Exponent Shannon ($exp(H)$) for selected countries. Cameroon appears diversier than Vanuatu according to richness, but Vanuatu is more diverse once relative abundance is considered. Likewise, Germany has nearly twice the richness of Switzerland, yet Switzerland is more diverse according to $exp(H)$ due to a more even speaker distribution.

and speaker distribution. However, addressing potential political biases in census-based speaker estimates at the country-level lies beyond the scope of the present study. Overall, the dataset spans 214 language families and 153 isolates, covering most of the global linguistic and phylogenetic diversity currently documented (Hammarström et al., 2024).

3 Entropy-based linguistic diversity

Linguistic diversity can be understood and measured in multiple ways. One of the most common approaches is language richness, i.e., the total number of languages within a given area. Other measures focus on linguistic abundance—how languages are distributed within a population. A third category, phylogenetic diversity, considers evolutionary relationships (Gavin et al., 2013). While phylogenetic approaches yield valuable insights, this study focuses on abundance, that is, the distribution of speakers across languages. To capture this, we compute Shannon entropy H of the probability distribution of speakers across all languages spoken in a country. Originating from information theory, the exponent of Shannon entropy $exp(H)$ is employed in biodiversity research to measure the ‘effective number of species’ and is here interpreted as the ‘effective number of languages’ (it is equivalent to the ‘Hill number of order 1’; Tuomisto, 2010). It is a measure of complexity, and its values represent the level of uncertainty in the language distribution: higher (exp-)entropy indicates a more even (diverse) distribution of speakers across languages, while lower (exp-)entropy

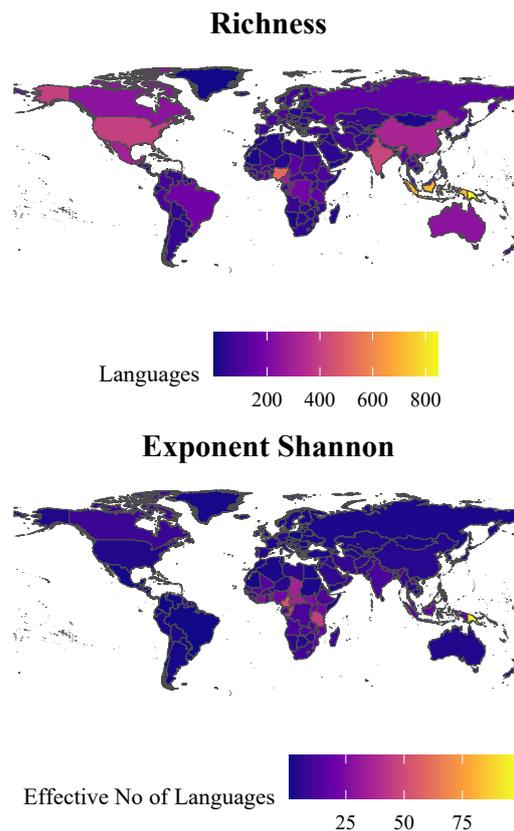


Figure 1: Global linguistic diversity visualized using both richness (language counts) and $exp(H)$, representing the effective number of equally abundant languages per country. Higher values reflect both greater balance and richness in a country’s linguistic landscape.

suggests a rather skewed distribution with a small set of dominant languages. In our context, $exp(H)$ can be interpreted as the number of languages in a given region that one would encounter if all of them were equally frequent. This approach offers a more nuanced perspective than simple language counts (Grin and Fürst, 2022).

While both measures in Figure 1 reveal substantial cross-national variation, $exp(H)$ highlights countries with more balanced speaker distributions as diversity hotspots. Cameroon and Vanuatu, ranked 8th and 20th by richness (289 and 115 languages), appear as the second and third most diverse countries once relative abundance is considered (53.6 and 54.3 effective languages), while Papua New Guinea is unequivocally perceived as the global hotspot of linguistic diversity (richness = 289, $exp(H)$ = 96.65). $exp(H)$ thus enables a relative abundance aware comparison of linguistic diversity across countries, which can—and should—be used in addition to richness, both of

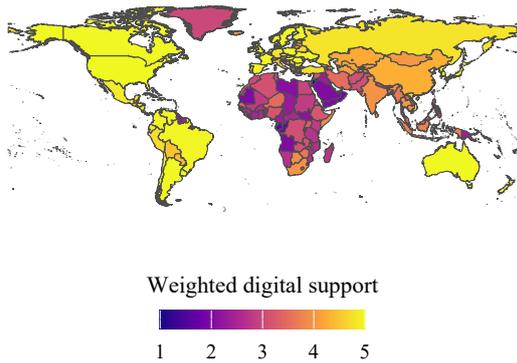


Figure 2: Country-level weighted digital language support. Higher values reflect digital representation aligned with national language distributions, while lower scores indicate limited support; values are comparatively low across much of Africa.

which can be derived using GlobLingDiv.

4 Country-level digital support

Several attempts have been made to quantify the digital vitality of languages (Gibson, 2015; Kornai, 2015). One of the most recent measures is the Digital Language Support (DLS) scale provided by Ethnologue (Simons et al., 2022). The scale captures the presence of languages across 216 digital tools and platforms and groups them into five ordered categories that reflect differing degrees of digital support (Still, Emerging, Ascending, Vital, Thriving). For our study, we recoded these categories to numeric values from 1 to 5. To obtain a country-level value, we combined these language-specific scores with the speaker information from our dataset and calculated a population-weighted average per country. In this way, languages with larger speaker populations contribute more strongly to the country score than those with fewer speakers. The resulting values provide a single, interpretable measure of how well the linguistic landscape of a country is supported in digital environments.

Countries such as Australia, the United States and the Netherlands show some of the highest digital support values (all 4.99), reflecting strong digital support for the languages spoken by the majority of the population. In contrast, several sub-Saharan African countries score toward the lower end of the scale (e.g., Republic of the Congo: 1.77). Vanuatu, previously highlighted as one of the most linguistically diverse countries, also displays a relatively low digital support value (1.66). These results (see

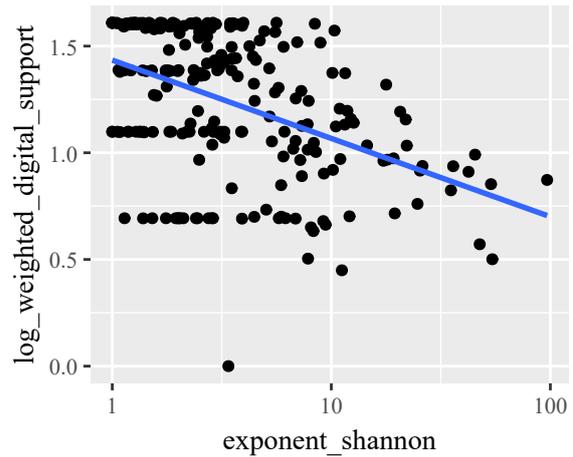


Figure 3: Scatter plot showing the association between $\exp(H)$ (the x-axis has been scaled using \log_{10}) and country-level digital language support (\log_e -transformed). Higher linguistic diversity is generally associated with lower digital support.

Figure 2), illustrate that linguistic diversity and digital support do not necessarily align: countries with rich linguistic landscapes may have comparatively weak digital representation.

5 Correlation analysis

To examine the relationship between linguistic diversity and digital language support, we conducted a country-level correlation analysis using $\exp(H)$ and the country-level digital support values (log-transformed to account for skewness). The analysis was performed using Spearman’s rank correlation coefficient due to the skewed distribution of $\exp(H)$ (Kendall and Gibbons, 1990).

The result shows a statistically significant moderate to strong negative effect between the two measures ($r = -.47, p < .001$). This pattern indicates that countries with higher linguistic diversity tend to have lower levels of digital support, whereas countries with lower levels of linguistic diversity are more likely to exhibit higher digital support values.

While correlation does not imply causation, the strength and direction of the association highlight a substantial global imbalance: linguistic diversity and digital presence are unevenly distributed across countries. Indeed, a potential causal link between the two domains could go in both directions: highly diverse linguistic landscapes appear more vulnerable to digital underrepresentation, suggesting that linguistic diversity is not necessarily mirrored in

the digital sphere (Mikami and Kodama, 2012); or it could be that lack of digital language resources further promotes biased language distributions in the non-digital domain.

6 Conclusion

To conclude, counteracting threats to linguistic diversity requires a clearer understanding of the factors that shape global linguistic landscapes, which in turn depends on extensive and reliable data (Bromham et al., 2022). The dataset presented here supports this effort by providing entropy-based diversity measures and a country-level indicator of digital language support. These measures reveal substantial cross-national variation in linguistic diversity and digital representation and show that the two do not necessarily align (Bella et al., 2023; Simons et al., 2022). Correlation patterns further suggest that linguistically diverse contexts may face disproportionate limitations in digital environments, pointing to structural imbalances.

At the same time, the results should be interpreted with care. Country-level aggregation inevitably obscures sub-national patterns, and the precision of speaker counts varies across sources and regions. Because the entropy-based diversity measures are derived from these speaker distributions, their interpretation is sensitive to such uncertainties. Future extensions could further draw on complementary data sources where available, such as national census data, to refine speaker estimates and language classifications. Future work could also complement the entropy-based results with qualitative validation, for example, by drawing on expert knowledge of regional and country-specific linguistic situations. In addition, future work could relate these patterns to those identified by existing initiatives and projects addressing inequalities in digital language support. Digital Language Support values offer a useful proxy, but do not capture all dimensions of digital presence or language use, particularly in informal or platform-specific settings (Gibson, 2015). The correlation analysis reflects association rather than causation. These limitations point to opportunities for future refinement, including sub-national modelling, multivariate approaches, and the integration of additional digital indicators and temporal data.

Beyond linguistic research, the dataset offers value for HSS and NLP. By providing structured and comparable measures, it can support quanti-

tative analyses and help identify linguistic setups with limited digital support (Anastasopoulos et al., 2020; Blasi et al., 2022). The resource will be made publicly available in CSV format to support reuse and integration with other datasets.

Acknowledgments

This research was funded by WWTF (grant number ICT23-012).

References

- Antonios Anastasopoulos, Christopher Cox, Hilaria Cruz, and Graham Neubig. 2020. Endangered languages meet modern nlp. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 39–45, Online. International Committee on Computational Linguistics.
- Gábor Bella, Paula Helm, Gertraud Koch, and Fausto Giunchiglia. 2023. [Towards bridging the digital language divide](#). *Preprint*, arXiv:2307.13405.
- Juliane Benson, Katharina Zeh, Hannes Essfors, Hannes Fellner, Julia Neidhardt, and Andreas Baumann. 2025. [Linguistic diversity and digitalization: An ambivalent relationship](#). In *Digital Humanism*, number 16319 in Lecture Notes in Computer Science, pages 358–365. Springer Nature.
- Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. [Systematic Inequalities in Language Technology Performance across the World’s Languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.
- Michaël Boissonneault, Adam Tallman, Volker Gast, and Simon J Greenhill. 2025. Projected speaker numbers and dormancy risks of canada’s indigenous languages. *Royal Society Open Science*, 12(2):241091.
- Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua. 2022. Global predictors of language endangerment and the future of linguistic diversity. *Nature ecology & evolution*, 6(2):163–173.
- Alexandre Duchêne and Philippe N Humbert. 2018. Surveying languages: the art of governing speakers with numbers. *International Journal of the Sociology of Language*, 2018(252):1–20.
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. *Ethnologue: Languages of the World*, 25 edition. Dallas: SIL International, Dallas, TX.
- Federico Gaspari, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne, and Andy Way. 2022. Introducing the digital language equality

- metric: Technological factors. In *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022), co-located with LREC 2022*, pages 1–12, Marseille, France.
- Michael C Gavin, Carlos A Botero, Claire Bower, Robert K Colwell, Michael Dunn, Robert R Dunn, Russell D Gray, Kathryn R Kirby, Joe McCarter, Adam Powell, and 1 others. 2013. Toward a mechanistic understanding of linguistic diversity. *BioScience*, 63(7):524–535.
- Megan L. Gibson. 2015. A framework for measuring the presence of minority languages in cyberspace. In *Linguistic and Cultural Diversity in Cyberspace. Proceedings of the 3rd International Conference*, pages 61–70, Moscow, Russia. Interregional Library Cooperation Centre.
- François Grin and Guillaume Fürst. 2022. Measuring linguistic diversity: A multi-level metric. *Social indicators research*, 164(2):601–621.
- Annika Grützner-Zahn and Georg Rehm. 2022. Introducing the digital language equality metric: Contextual factors. In *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022), co-located with LREC 2022*, pages 13–26, Marseille, France.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. *Glottolog 5.1*. <https://glottolog.org>. Max Planck Institute for Evolutionary Anthropology, Leipzig.
- James Hutson, Pace Ellsworth, and Matt Ellsworth. 2024. Preserving linguistic diversity in the digital age: a scalable model for cultural heritage continuity. *Journal of Contemporary Language Research*, 3(1).
- ISO. 2007. *ISO 639-3:2007 – Codes for the Representation of Names of Languages – Part 3: Alpha-3 Code for Comprehensive Coverage of Languages*. International Organization for Standardization, Geneva, Switzerland. [Accessed: 2025-04-08].
- ISO. 2013. *Codes for the Representation of Names of Countries and Their Subdivisions – Part 1: Country Codes (ISO 3166-1)*. International Organization for Standardization, Geneva, Switzerland.
- Sreenivasan Jolad and Anubha Agarwal. n.d. *Mapping india’s language and mother tongue diversity and its exclusion in the indian census*. <https://osf.io/sjxc6>. OSF Preprints.
- Joshua Project. 2025. *Joshua project: People groups of the world*. Accessed: 2025-07-15.
- Anne Kandler and Roman Unger. 2023. Modeling language shift. In *Diffusive spreading in nature, technology and society*, pages 365–387. Springer.
- Maurice G Kendall and Jean Dickinson Gibbons. 1990. *Rank correlation methods*, 5. ed. edition. A Charles Griffin title. Arnold, London [u.a.].
- András Kornai. 2015. A new method of language vitality assessment. In *Linguistic and Cultural Diversity in Cyberspace. Proceedings of the 3rd International Conference*, pages 132–138, Moscow, Russia. Interregional Library Cooperation Centre.
- Yoshiki Mikami and Shigeaki Kodama. 2012. *Measuring linguistic diversity on the web*. In Laurent Vannini and Hervé Le Crosnier, editors, *Net.lang (version en anglais): Toward the Multilingual Cyberspace*, pages 118–139. C&F Éditions. Chapter in edited volume on multilingual cyberspace.
- Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.
- Gary F. Simons, Abbey L. Thomas, and Chad K. White. 2022. *Assessing digital language support on a global scale*. Preprint, arXiv:2209.13515.
- Hanna Tuomisto. 2010. A diversity of beta diversities: Straightening up a concept gone awry. part 1. defining beta diversity as a function of alpha and gamma diversity. *Ecography*, 33(1):2–22.

A Appendix

This appendix summarizes the quantitative measures used to capture linguistic diversity and digital language support at the country level.

A.1 Shannon Entropy

Shannon entropy captures the diversity of languages spoken within a country, taking into account both richness and evenness.

$$H = - \sum_{i=1}^N p_i \log p_i \quad (1)$$

where p_i is the proportion of speakers of language i , N the total number of languages in the country, and where \log refers to the natural logarithm (Shannon, 1948).

A.2 Exponentiated Shannon Entropy

To improve interpretability, Shannon entropy is exponentiated, yielding the Hill number of order $q = 1$, which corresponds to the effective number of equally abundant languages:

$$\exp(H). \quad (2)$$

This measure reflects the number of languages that would be present if all languages in a country were spoken by equal proportions of speakers.

A.3 Country-Level Weighted Digital Language Support

Country-level digital language support is computed as a speaker-weighted average of language-specific digital support scores:

$$D = \frac{\sum_i s_i n_i}{\sum_i n_i}, \quad (3)$$

where s_i denotes the digital support score associated with language i , and n_i the number of speakers of language i in the country. This measure represents the average level of digital language support experienced by speakers within a country.

For analyses sensitive to skewness, this measure is log-transformed prior to statistical modeling and visualization.

LLMs Got Rhyme? Hybrid Phonological Filtering for Greek Poetry Rhyme Detection and Generation

Stergios Chatzikyriakidis

Department of Philology
University of Crete
stergios.chatzikyriakidis@uoc.gr

Anastasia Natsina

Department of Philology
University of Crete
natsina@uoc.gr

Abstract

Large Language Models (LLMs), even though exhibiting multiple capabilities on many NLP tasks, struggle with phonologically-grounded phenomena like rhyme detection and generation. When one moves to lower-resource languages such as Modern Greek, this is even more evident. In this paper, we present a hybrid neural-symbolic system that combines LLMs with deterministic phonological algorithms to achieve accurate rhyme identification and generation. We implement a comprehensive taxonomy of Greek rhyme types and employ an agentic generation pipeline with phonological verification. We use multiple prompting strategies (zero-shot, few-shot, Chain-of-Thought, and RAG-augmented) across several LLMs including Claude 3.7 and 4.5, GPT-4o, Gemini 2.0 and open-weight models like Llama 3.1 8B and 70B and Mistral Large. Results reveal a significant reasoning gap: while native-like models (Claude 3.7) perform intuitively (40% accuracy in identification), reasoning-heavy models (Claude 4.5) achieve state-of-the-art performance (54%) only when prompted with Chain-of-Thought. Most critically, pure LLM generation fails significantly (under 4% valid poems), while our hybrid verification loop restores performance to 73.1%. Along with the system presented, we further release a corpus of 40,000+ rhymes, derived from the *Anemoskala* and *Interwar Poetry* corpora, to support future research.¹

1 Introduction

Rhyme is a significant feature in many poetic traditions. In Modern Greek, rhyme (ομοιοκαταληξία) has been systematically employed from the Cretan Renaissance (14th–17th centuries) all the way to contemporary poetry and popular music, including hip-hop and rap (Topintzi et al., 2019).

¹The material needed to run the experiments and verify the results of this paper can be found here: https://github.com/StergiosCha/Greek_Rhyme

LLMs, unsurprisingly given their prevalent text-based training, exhibit notable weaknesses in phonologically-aware reasoning. LLMs process text at the token level, which poorly aligns with phonological units like syllables, stress patterns, and rhyme domains.

We tackle the issue of rhyme identification and generation using a hybrid Neural-Symbolic architecture that combines the generative and reasoning capabilities of LLMs with deterministic phonological algorithms. Our contributions are:

1. A dataset of 40k rhymes aggregated from existing Modern Greek corpora, based on the phonological taxonomy found in Topintzi et al. (2019).
2. A hybrid proposer-verifier identification pipeline where LLM predictions are grounded by deterministic phonological verification.
3. An agentic generation pipeline with a Generate-Verify-Refine loop that raises the validity of generated poems from nearly 0% to 73.1%.
4. A multi-model evaluation across proprietary (Claude, GPT-4o) and open-weight models, with analysis of prompting strategies including RAG augmentation.

2 Background

2.1 Rhyme in Modern Greek Poetry

Rhyme in Modern Greek poetry dates to the medieval period, appearing in works like those of Stefanos Sachlikis (14th c.), and flourishing in Cretan Renaissance masterpieces such as Kornaros’s *Erotokritos*. The phenomenon spans the Heptanesian Romanticism of Solomos, the Athenian school, Parnassian and Symbolist movements, and continues, albeit more selectively, in modern and contemporary poetry (Κοχόλης, 1993).

Greek rhyme exhibits several distinctive characteristics:

Stress-based Classification Greek is a stress-accent language where rhyme domains are defined relative to the stressed syllable. More particularly, Greek is subject to the three-syllable rule according to which accent must fall in one of the last three syllables of the word (broadly defined, includes also cases of what we call phonological word, e.g. a word plus a weak object pronoun). We follow [Topintzi et al. \(2019\)](#) and use the following categories:

- **Masculine (M)**: Stress is on the final syllable (oxytone), e.g., καρδιά / φωτιά [/'kaɾ'dja / fo'tɕa/].
- **Feminine (F2)**: Stress is on the penultimate syllable (paroxytone), e.g., ελπίδα / πατρίδα [/'el'piða / pa'triða/].
- **Proparoxytone (F3)**: Stress is on the antepenultimate syllable, e.g., στόματα / σώματα [/'stomata / 'somata/].

Greek poetry further employs several other rhyme types that are independent of stress position:

Rich Rhyme (RICH): In this type, the onset consonant(s) of the stressed syllable must match. This is further distinguished into Total Rich (TR) rhyme where complete onset matching is at play, or Partial Rich (PR) with partial matching:

- TR-S (singleton): καλά / μαλά [/'ka'la / ma'la/]
- PR-C1 (first consonant): στόματα / σώματα [/'stomata / 'somata/]

Identical Pre-rhyme Vowel (IDV): The vowel that precedes the stressed syllable must match:

- ξανθή / γραφή [/'ksan'thi / gra'fi/] (the pre-stress vowel /a/ matches).

Mosaic (MOS): The rhyme domain spans across more than one word:

- όνομά της / ομπάτης [/'onoma tis / o'batis/]

Imperfect (IMP): Partial phonetic matching with systematic variation:

- IMP-V: Vowel differs (χάνετε / γίνετε [/'xanete / 'jinete/])

- IMP-C: Consonant differs (ξαφνίζει / τεχνίτη [/'ksaf'nizi / te'xniði/])
- IMP-0F: Final consonant-zero alternation (πιστοί / χαρείς [/'pi'sti / xa'ris/])

2.2 Computational Approaches to Rhyme

Early approaches on rhyme identification used classic unsupervised machine learning ([Reddy and Knight, 2011](#)) or probabilistic models using phoneme frequencies for rhyme detection in rap music ([Hirjee and Brown, 2010](#)). More recent supervised methods based on neural networks achieve higher accuracies. For example, [Haider and Kuhn \(2018\)](#) achieve 97% accuracy via a single Siamese Recurrent Network model trained in German, English, and French, using no explicit phonetic features.

In poetry generation, we find early statistical machine translation for Classical Chinese quatrain generation by [He et al. \(2012\)](#), where they in effect treat each line as a kind of translation of the previous line. The Hafez system ([Ghazvininejad et al., 2016](#)) is a hybrid system that puts together finite-state acceptors that encode metrical and rhyme constraints with RNNs for English sonnet generation. [Lau et al. \(2018\)](#) develop Deep-speare, which is a joint neural model for Shakespearean sonnets. The outputs of this system have been shown to be to a large extent indistinguishable from human verse in crowd evaluations. More recent work has moved towards modeling rhyme as continuous similarity rather than categorical classes (e.g., [Nagy, 2022](#), for Latin), and unsupervised detection through tools like RhymeTagger ([Plecháč, 2018](#)). Moving on to the LLM era, we find byte-level transformers such as ByGPT5 ([Belouadi and Eger, 2023](#)), as well as synthetic-data approaches like GPoET ([Popescu-Belis et al., 2023](#)).

For Greek specifically, the foundational computational work is that of [Topintzi et al. \(2019\)](#), who established the categorical taxonomy for Greek rhyme and developed an initial *Greek Rhyme* (GrRh) database. While their work provided the theoretical framework and the original corpus curation, our study introduces several critical advancements: (i) the first full algorithmic implementation of the Greek-specific G2P and rhyming rules in a deterministic Python environment; (ii) a large-scale reproduction and cleaning of the corpus, using our verifier to identify and correct systematic extraction errors in the original GrRh, thereby recover-

Poet	Rhyme Pairs
Palamas	20,620
Tellos Agras	6,119
Valaoritis	4,202
Solomos	2,518
Karyotakis	1,692
Cavafy	1,585
Fotos Giomyllis	1,565
Kostas Ouranis	792
Napoleon Lapathiotis	495
Romos Filiras	484
Mitsos Papanikolaou	404
Kalvos	100
Total	40,576

Table 1: Corpus composition by poet.

ing over 40,000 high-quality rhyme pairs; and (iii) the integration of these rules into an agentic LLM framework for generation and "proposer-verifier" identification.

3 Dataset

We construct our dataset by using two primary resources for Modern Greek poetry. The first source is the Anemoskala archive from the Centre for the Greek Language (KEG - Κέντρο Ελληνικής Γλώσσας), which provides extensive digitized collections of major poets. The second source is the Interwar Poetry Dataset, an open-access dataset created by Dr Natsina and Professor Chatzikyriakidis with the help of undergraduate students at the Philology Department, University of Crete². This corpus comprises over 600 poems by interwar Greek poets. We apply phonological filtering to extract rhyme pairs from both corpora and then merge them. The corpus statistics are shown in Table 1.

4 System Architecture

We implement a hybrid, neural-symbolic architecture that combines LLMs with symbolic phonological rules (Figure 1). The aim is to tackle the inherent "blindness" of sub-token processing by delegating the phonological validation to a symbolic engine.

4.1 Phonological Engine

The basis of the phonological symbolic processor/verifier is based on (Topintzi et al., 2019). Its purpose is exactly to handle Greek-specific rhyme analysis:

²Dataset available at: https://github.com/StergiosChatzikyriakidis/Modern_Greek_Literature/tree/v1.

- **Syllabification.** A number of Greek syllabification rules are implemented. The algorithm is designed to identify syllable boundaries based on Greek-specific phonotactic constraints.
- **Stress Detection.** We use the accent marks found in Greek orthography (acute, grave, circumflex, dieresis, diaeresis, etc.) in order to identify stress. In case of words with clitics (notably weak object pronouns, e.g., κάλεσέ με), we have a mechanism to handle stress domain extension.
- **Rhyme Domain Extraction.** The domain of the rhyme extends from the stressed vowel until the end of the phonological phrase. In the case of Mosaic rhymes, it can span multiple orthographic words.
- **Phonetic Transcription.** We convert Greek orthography to a phonetic representation.

4.2 Rhyme Classification Module

We develop a classification module based on the phonological module. It compares rhyme domain pairs and assigns labels. More specifically it checks:

1. **Position Classification:** Here it determines where the stress falls and classifies as M/F2/F3.
2. **Perfect Match Check:** In this part, the post-stress material is checked to see if it is an exact match.
3. **Feature Detection:** This checks for RICH rhyme (onset matching), IDV rhyme (pre-stress vowel) and/or Mosaic (MOS) rhyme (word boundary crossing).
4. **Imperfection Analysis:** In case the rhyme is not perfect, classify accordingly (IMP-V, IMP-C, IMP-OF, IMP-OM).

The output result is a compound label such as F2-TR-S-IDV (Feminine-2, Total Rich Singleton, Identical pre-rhyme Vowel).

4.3 LLM Integration and Prompting Strategies

We evaluate the performance of several state-of-the-art LLMs (Claude 3.7/4.5, GPT-4o, Gemini 2.0) using a progressive series of prompting strategies: Zero-Shot Structured, Chain-of-Thought (CoT), and RAG-Augmented.

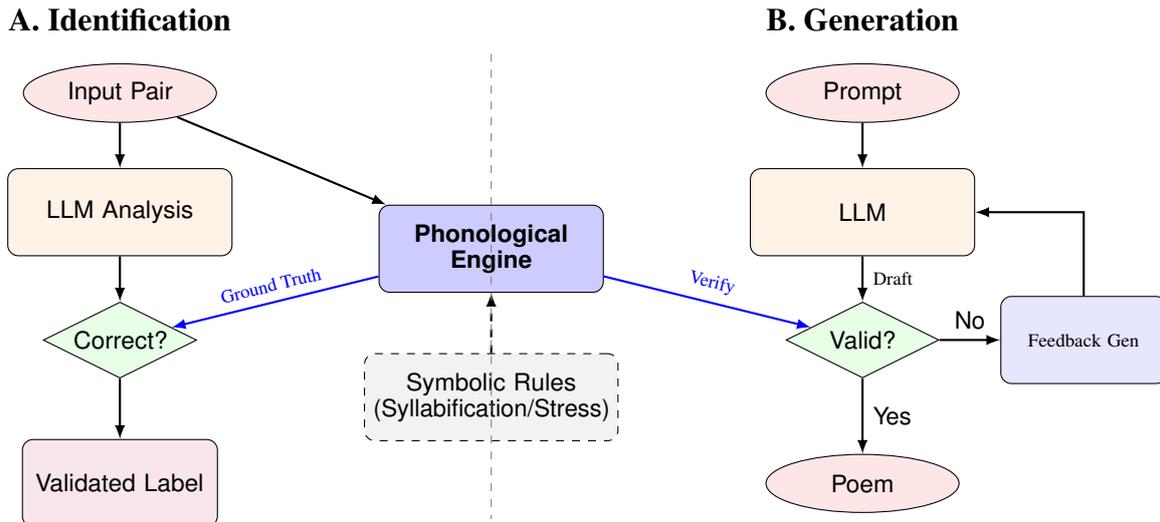


Figure 1: Hybrid system architecture. **Left:** Identification combines LLM predictions with Engine-generated ground truth for validation. **Right:** Generation uses the Engine to verify and refine LLM outputs.

Hybrid Combination Mechanism For the identification task, our system utilizes the LLM as a *proposer* and the symbolic engine as a *verifier*. In the "Verified Detection" mode, the system first obtains an initial classification from the LLM. We note here that the phonological ground truth for the validation is provided by the symbolic engine. The LLM receives this as prompt and is asked to *self-reflect* and compare its initial hypothesis with the deterministic rules and correct its final output, if needed. This ensures that the linguistic reasoning of the transformer is grounded in the deterministic accuracy of the phonological engine.

RAG Implementation Details Our RAG component (*Rhyme-RAG*) retrieves relevant examples from our cleaned 40,000-rhyme corpus. The retrieval we use is feature-based in the following sense: the system extracts phonological features (e.g., "F3", "RICH", "MOSAIC") from the user query via keyword markers and retrieves the top-15 most relevant rhyme pairs that exhibit these exact patterns. This provides the model with high-precision examples of the specific, often obscure, Greek rhyme types it is tasked to generate or identify.

4.4 Agentic Generation Pipeline

For rhyme generation, we use a Generate-Verify-Refine loop. The agent first generates a draft stanza based on a theme, requested rhyme pattern and number of lines. The symbolic part of the system is then tasked with identifying final words for every line and verifying them using symbolic rules. If

Algorithm 1 Generate-Verify-Refine Loop

```

1: Input: Theme, rhyme_type, features, num_lines
2: Output: Phonologically valid poem
3: attempts  $\leftarrow$  0
4: while attempts < 15 do
5:   poem  $\leftarrow$  LLM.generate(prompt)
6:   errors  $\leftarrow$  verify_rhymes(poem)
7:   if errors =  $\emptyset$  then
8:     return poem
9:   else
10:    feedback  $\leftarrow$  format_errors(errors)
11:    prompt  $\leftarrow$  update_prompt(feedback)
12:    attempts  $\leftarrow$  attempts + 1
13:   end if
14: end while
15: return poem with warning

```

a rhyme fails, the verifier generates a precise error message, which is sent back to the LLM for correction.

Verifier Feedback Strings Concrete examples of these feedback strings include:

- Stress mismatch: Expected F3, found F2 for 'αναβιώνει' / 'απλώνει'.
- Identical words 'αγαπώ' / 'αγαπώ' (COPY) are not allowed unless requested.
- RICH requested but 'σκληρό' / 'θησαυρό' is PURE.

This iterative feedback continues for up to 15 attempts.

The verifier's task is to check each rhyme pair using the phonological rules and provide feedback in case the rhyme does not comply with the rules.

5 Experimental Setup

We use a test set of 40 poems for rhyme identification. This produces a total of 160 test cases in total (2 strategies \times 2 RAG configs \times 40 poems). The set has a balanced distribution across rhyme types (13 Masculine, 16 Feminine (F2), 11 Proparoxytone (F3)), as well as comprehensive coverage of rhyme features including more rare types of rhyme (21 PURE, 10 RICH, 10 IDV, 6 IMPERFECT, 5 MO-SAIC). We went for a balance that was big enough to make meaningful claims, while, at the same time, maintaining evaluation feasibility across 8 models and 4 configurations (1,280 total API calls).

For generation, we evaluate the system on 26 test cases with specified rhyme constraints (e.g., “Write a 4-line poem with F3-RICH rhyme on theme: love”). To calculate the “Generation Validity (%)” metric, we evaluate the LLM output against the requested rhyme scheme. A poem is considered valid only if **all** pairs in the requested scheme (e.g., A-A and B-B in a quatrain) are verified as correct matches by the symbolic engine. If a model generates three lines of monorhyme (AAA) when AABB was requested, it is penalized for the structural mismatch. Each test runs twice: once with our verification loop (Generate-Verify-Refine) and once without (pure LLM generation).

6 Results

6.1 Rhyme Identification Results

In Table 2, we see the results across all prompting configurations.

Table 3 breaks down performance by rhyme type, revealing systematic biases.

The results across all configurations (Table 2) show that Claude 4.5 reaches the highest accuracy with 53.8%, but this number comes with a caveat: it requires both CoT and RAG to get there. GPT-4o goes from 7.7% to 50.0% once CoT is switched on, the model knows more than it lets on, but needs to be walked through the reasoning step by step. Claude 3.7 sits around 38-42% regardless of what you throw at it.

The picture we get from open models is rather poor. Llama 3.1 70B tops out at 23.1% and CoT makes things worse for Llama 3.3 (down to 7.7%). This is not just about size. If it were, CoT should help the 70B models the way it helps GPT-4o. It does not, which means the Greek phonetic representations in these models are simply not there to reason over.

Model size helps, the 70B Llama shows better performance than the 8B easily, but it does not close the gap with the proprietary systems.

Analyzing performance by rhyme type (Table 3) exposes systematic biases prevalent across the LLM landscape. Proparoxytone (F3) rhymes represent a clear ceiling for all models, with accuracy dropping significantly compared to Masculine (M) rhymes. This performance decay is likely due to the increased distance between the stress and the word boundary, which multiplies the potential for syllabification and stress-mapping errors during sub-token processing.

Most critically, complex features like RICH or MOSAIC are almost entirely invisible to current transformers (see Appendix for full statistics); MO-SAIC detection achieved 0% accuracy across almost all configurations. This reinforces our central argument: LLMs, while creatively powerful, lack the precision required for dealing with rhyme identification and generation.

While our proposed system architecture (Section 4) includes a verified detection mode where the LLM can reflect on symbolic ground truth, the benchmarks reported here measure the models’ performance in unaided, CoT, and RAG configurations. This experimental design allows us to isolate the phonological reasoning gap and measure the inherent difficulty of the Greek rhyming taxonomy for neural architectures.

To understand the models’ capabilities and failure modes, we categorized the 9 test cases into thematic observations (detailed outputs in Appendix A):

- 1. Phonological vs. Lexical Conflation:** Many models struggled to distinguish phonemically identical but orthographically distinct pairs (homonyms) from simple repetition (COPY rhyme). For example, in the pair *κρίνοι/κρίνει* (lilies / it judges), which are perfect homophones, Claude 4.5 falsely flagged a ‘COPY’ rhyme, failing to recognize the distinct lemmata despite the sound identity.
- 2. The Proparoxytone Distance Problem:** As seen in the quantitative results, proparoxytone rhymes like *παράπνοο/άπνοο* caused significant structural failures. Models often failed to map the three-syllable distance between the stress and the word boundary, frequently misidentifying these as paroxytone (F2) or simple masculine rhymes.

Model	Structured	Structured+RAG	CoT	CoT+RAG
<i>Proprietary Models</i>				
Claude 4.5	53.8%	26.9%	46.2%	53.8%
Claude 3.7	38.5%	42.3%	42.3%	30.8%
GPT-4o	7.7%	7.7%	50.0%	26.9%
Gemini 2.0	23.1%	42.3%	19.2%	11.5%
<i>Open-Weight Models</i>				
Mistral Large	11.5%	15.4%	26.9%	11.5%
Llama 3.1 70B	23.1%	19.2%	11.5%	15.4%
Llama 3.3 70B	23.1%	23.1%	7.7%	7.7%
Llama 3.1 8B	7.7%	15.4%	7.7%	3.8%

Table 2: Rhyme identification accuracy (%) across configurations. Bold means best configuration per model.

Model	M (n=32)	F2 (n=40)	F3 (n=32)
Claude 4.5	65.6%	47.5%	21.9%
Claude 3.7	37.5%	52.5%	21.9%
GPT-4o	31.2%	25.0%	12.5%
Gemini 2.0	43.8%	20.0%	9.4%
Mistral Large	25.0%	17.5%	6.2%
Llama 3.1 70B	12.5%	32.5%	3.1%
Llama 3.3 70B	25.0%	12.5%	9.4%
Llama 3.1 8B	6.2%	12.5%	6.2%

Table 3: Identification accuracy by rhyme type. All models struggle most with F3 (proparoxytone) rhymes.

Model	No Verify	With Verify	Improvement
Claude 4.5	0.0%	34.6%	+34.6%
Claude 3.7	3.8%	73.1%	+69.3%
GPT-4o	0.0%	42.3%	+42.3%

Table 4: Generation validity (% of perfectly valid poems) with and without phonological verification loop.

- 3. Visual vs. Phonetic Identity:** A core challenge for sub-token transformers is getting tricked by orthography. Most models failed on the pair $\sigma\rho\theta\acute{o}\zeta/\phi\omega\zeta$ due to the visual mismatch, yet reasoning-heavy configurations (Claude 4.5 CoT+RAG) were able to parse the underlying /os/ identity, suggesting that step-by-step reasoning can partially overcome the visual bias of the tokenizer.
- 4. Morphological Variation:** Test cases involving archaic forms (e.g., the dative plural $\sigma\upsilon\rho\alpha\nu\acute{o}\iota\zeta$) proved particularly challenging.

6.2 Rhyme Generation Results

Table 5 shows generation validity broken down by feature complexity.

The verification loop enhances rhyme validity across all models. Claude 3.7 achieves the highest verified generation rate (73.1% valid poems),

Feature Type	n	Claude 3.7	Claude 4.5	GPT-4o
BASIC (no features)	6	100.0%	83.3%	90.9%
IMPERFECT	2	50.0%	50.0%	33.3%
IDV+PURE	2	50.0%	0.0%	66.7%
IDV+RICH	2	50.0%	0.0%	0.0%
IDV+IMPERFECT	2	100.0%	0.0%	100.0%
IDV+MOSAIC+PURE	2	0.0%	0.0%	0.0%

Table 5: Generation validity (%) with verification, by feature complexity. BASIC rhymes (no features) achieve highest success. Complex multi-feature combinations (e.g., IDV+MOSAIC+PURE) fail even with verification. We note that the small sample sizes per feature category (n=2 for most non-BASIC types) limit the reliability of these breakdowns. As such, they are presented to illustrate qualitative trends rather than robust estimates.

demonstrating that the hybrid approach can be used in order to compensate for LLM phonological weaknesses. Pure LLM generation fails catastrophically (0-4% validity), confirming that deterministic verification is essential for constrained poetry generation, at least for the models used.

6.3 Qualitative Analysis of Generation

Three main types of generation errors (full output traces are provided in Appendix B) are generally corrected by the loop:

- 1. Correcting the stress pattern (F3 vs F2):** When asked for F3 (proparoxytone) rhymes, models are often incorrectly defaulting to F2 (paroxytone). *Example:* GPT-4o initially generated $\alpha\nu\alpha\beta\acute{\iota}\omega\nu\epsilon\iota / \alpha\pi\lambda\acute{\omega}\nu\epsilon\iota$ (F2). The verifier returned "Stress mismatch: Expected F3, found F2". The model corrected this to a valid F3 rhyme in the subsequent iteration.
- 2. Feature Precision (Rich vs Pure):** LLMs often treat rhyme types loosely. When M-PURE was requested, Claude 3.7 generated $\sigma\kappa\lambda\eta\rho\acute{o} / \theta\eta\sigma\alpha\upsilon\rho\acute{o}$ (which is RICH, sharing

the /r/ onset). After the intervention of the loop, the model successfully refined the output to a strict PURE rhyme, demonstrating the system’s ability to enforce precise stylistic constraints.

3. **Non-rhymes:** In some cases, Claude 4.5 proposed non-rhyming outputs, e.g. pairs like *αγαπώ / ξέρω*. The verification loop noted these invalid pairs (“No rhyme”), and forced the model to regenerate valid phonological matches.

7 Discussion

7.1 Why LLMs Struggle with Rhyme

One issue is potentially tied to the nature of LLM tokenizers (e.g., BPE, WordPiece), i.e. the fact that they segment text based on statistical co-occurrence and not linguistically motivated units. For example, a word like *παράπονο* (/paˈrapono/) might be tokenized as [πα, ρά, πο, νο] in an ideal case, but often appears as [παρ, άπ, ονο] depending on the vocabulary, stripping the model of the ability to map the stress position (F3) relative to the final syllable.

Unlike orthographical systems where text maps close to 1:1 to sound, Greek’s retention of historical orthography, i.e. a system of orthography that has been kept essentially the same for millennia and, thus, has not followed the changes in the language, features many-to-one mappings (e.g., the sound /i/ can be spelled as *ι, η, υ, ει, οι, υι*). LLMs trained primarily on text often rely on visual similarity rather than phonetic identity. This explains why models fail on *χρίνοι/χρίνει* (visual mismatch but phonetic match) while hallucinating rhymes for *ορθός/ψώς* (visual mismatch and phonetic mismatch).

Finally, note that Greek rhyme is strictly defined by stress position (M, F2, F3). Since LLMs lack an explicit prosodic module, they are prone to fail when attempting to distinguish minimal pairs that differ only in stress (e.g., *νόμος* vs *νομός*).

7.2 Creativity vs. Hallucination

A number of words that the LLM proposes are hallucinations, in the sense that these words are not existing words in Modern Greek. However, the analysis of these words shows a very nuanced trade-off between semantic grounding and poetic creativity. Two categories of invented vocabulary are observed.

Some of the hallucinations can be very well-taken as poetic neologism. For example, Claude 4.5 gives us *αιωνημένα* (eternal-ized) and *θάραμα* (courage-thing), which are not only valid phonotactically, but also are meaningful enough neologisms (for example *αιωνημένα* can be an adjective that means something related to eternity, while *θάραμα* can be seen as a collage of *θάρρος* (courage) and *χάραμα* (dawn). A striking example is *πυραφί* (fire-colored), which appears to be constructed by analogy to *χρυσάφί* (gold-colored), demonstrating a deep (if unauthorized) grasp of Greek morphology. Similarly, *αγέρη* appears as a valid homophone of *αγέρι* (breeze), suggesting a spelling variation rather than a failure. Phonetic failures, on the other hand, produce words that are phonotactically correct but lack semantic transparency. Examples include *τσιγκλο* and *σκάμπουνε*. It is significant that no phonotactic violations occur in these cases, pointing to the models’ robust internal representation of Greek syllable structure. We find a distinct behavioral difference between models: GPT-4o is the safest model, producing almost zero nonsense but also fewer neologisms. In contrast, Claude 4.5 is the boldest, with a high rate of invention, the majority of which are plausible neologisms rather than nonsense. This suggests that what is often penalized as hallucination in factual tasks can serve as a proxy for creativity in poetic tasks.

Crucially, this higher level of invention coincides with stronger performance in the verification loop. It appears that the strict constraints imposed by the verifier push capable models to neologize in order to conform to the poetic form, inventing new words when standard vocabulary fails to meet the phonological requirements.

8 Future Work

The natural next step for such a system is to go beyond rhyme and implement metrical verification. Greek poetry relies heavily on stress-timed constituent meters (e.g., iambic 15-syllable verse) and systems that could handle both rhyme and metrical structure would be a very interesting research direction to take.

Furthermore, we plan to explore Reinforcement Learning from Phonological Feedback (RLPF). The idea is that instead of a simple rejection sampling loop at inference time, the signals from our deterministic verifier could be used as a reward function to fine-tune a smaller model (e.g., Llama

8B), potentially internalizing phonological constraints directly into the model’s weights.

9 Conclusion

In this work, we presented a hybrid neuro-symbolic system for Modern Greek poetry identification and generation. We showed that while LLMs seem to possess latent creative capabilities, they struggle with the precision needed in order to deal with rhyme detection and generation. We combined a deterministic phonological engine with an agentic generation loop, and achieved a significant improvement in generation validity, raising success rates from a baseline of under 4% to 73.1%. Furthermore, our identification benchmarks revealed a reasoning gap, in which only the most advanced reasoning models (using Chain-of-Thought) could compete with symbolic verification. We hope that our released codebase and the curated corpus of 40,000+ Modern Greek rhymes will serve as foundational resources for future research into phonologically-aware NLP.

Limitations

There are a number of limitations to this work. The rhyme taxonomy we use is categorical, inherited from [Topintzi et al. \(2019\)](#). There is recent work arguing that rhyme is better treated as a gradient phenomenon ([Nagy, 2022](#)), and our deterministic engine has nothing to say about degrees of rhyme, i.e. a pair either matches or it does not. The corpus, while large by Greek standards, remains small next to what is available for English or German. On the generation side, passing the verifier means the rhymes are structurally correct, but says nothing about whether the poem is any good; phonological validity and poetic quality are orthogonal. Finally, every failed verification triggers a new LLM call, which means the loop can be substantially slower than ordinary single-pass generation.

Acknowledgments

We gratefully acknowledge the Centre for the Greek Language (Κέντρο Ελληνικής Γλώσσας) for providing access to the Anemoskala corpus and granting permission to derive our rhyming dataset from their digital resources. The original corpus is available through the Portal for the Greek Language (www.greek-language.gr).

References

- Jonas Belouadi and Steffen Eger. 2023. [ByGPT5: End-to-end style-conditioned poetry generation with token-free language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7364–7381, Toronto, Canada. Association for Computational Linguistics.
- Marjan Ghazvininejad, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191.
- Thomas Haider and Jonas Kuhn. 2018. Supervised rhyme detection with siamese recurrent networks. In *Proceedings of the Workshop on Stylistic Variation*, pages 81–86.
- Jing He, Ming Zhou, and Long Jiang. 2012. Generating chinese classical poems with statistical machine translation models. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1650–1656.
- Hussein Hirjee and Daniel Brown. 2010. Automatic detection of rhyme in rap music. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*, pages 395–400.
- Jey Han Lau, Trevor Cohn, Timothy Baldwin, Julian Brooke, and Adam Hammond. 2018. [Deep-speare: A joint neural model of poetic language, meter and rhyme](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1948–1958, Melbourne, Australia. Association for Computational Linguistics.
- Ben Nagy. 2022. Rhyme in classical latin poetry: stylistic or stochastic? *Digital Scholarship in the Humanities*, 37(4):1097–1118.
- Petr Plecháč. 2018. A collocation-driven method of discovering rhymes (in czech, english, and french poetry). In *Taming the Corpus: From Inflection and Lexis to Interpretation*, pages 79–95. Springer.
- Andrei Popescu-Belis, Àlex R. Atrio, Bastien Bernath, Etienne Boisson, Teo Ferrari, Xavier Theimer-Lienhard, and Giorgos Vernikos. 2023. [GPoeT: a language model trained for rhyme generation on synthetic data](#). In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 10–20, Dubrovnik, Croatia. Association for Computational Linguistics.
- Sravana Reddy and Kevin Knight. 2011. Unsupervised discovery of rhyme schemes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 77–82.

Nina Topintzi, Konstantinos Avdelidis, and Thomai Valkanou. 2019. Quantifying greek rhyme. In *Selected Papers from the 23rd International Symposium on Theoretical and Applied Linguistics*, pages 429–447. School of English, Aristotle University of Thessaloniki.

Ξενοφών Κοκόλης. 1993. Η ομοιοκαταληξία: Τύποι και λειτουργικές διαστάσεις. Στιγμή, Athens.

A Detailed Qualitative Results

This appendix presents the full raw outputs for the 9 representative test cases discussed in the Qualitative Analysis. For each case, we show the predicted Rhyme Type and Features for all 8 models across 4 configurations (Structured vs CoT, No RAG vs RAG).

Legend: ✓ = Correct Type & Features, ⚡ = Correct Type but Wrong Features, ✗ = Incorrect Type.

A.1 1. Baseline Success (M) (Keyword: 'απαιτώ')

Poem: Έτσι από σένα περιμένω κι απαιτώ. / της Τραγωδίας τον Λόγο τον λαμπρό —
True Label: M ['PURE']

Model	Struct (No RAG)	Struct (RAG)	CoT (No RAG)	CoT (RAG)
Claude 4.5	⚡ M	✓ M [PURE]	✗ S	⚡ M
Claude 3.7	✗ MISS	✗ MISS	⚡ M	✗ I
GPT-4o	✗ MISS	⚡ M [IMPERFECT]	⚡ M	✗ S
Gemini 2.0	⚡ M [IMPERFECT]	⚡ M [IMPERFECT]	✗ MISS	✗ MISS
Llama 70B	⚡ M [IDV, RICH]	✗ F2 [PURE]	✗ F2	✗ +
Llama 3.3	✗ S	⚡ M [IMPERFECT]	✗ MISS	✗ OF
Mistral	⚡ M [IMPERFECT]	✗ STRESS	✗ S	✗ IN
Llama 8B	✗ S	✗ MISS	✗ WORKERJOE	✗ ANN

A.2 2. Baseline Success (F2) (Keyword: 'κρίνοι')

Poem: Από ρουμπίνια ρόδα, από μαργαριτάρια κρίνοι, / από αμεθύστους μενεξέδες. Ως αυτός τα κρίνει,
True Label: F2 ['RICH']

Model	Struct (No RAG)	Struct (RAG)	CoT (No RAG)	CoT (RAG)
Claude 4.5	⚡ F2 [COPY, RICH]	✗ BETWEEN	✗ S	⚡ F2
Claude 3.7	⚡ F2 [COPY]	⚡ F2 [COPY]	✗ M	✗ TR [CC]
GPT-4o	✗ S	✗ S	⚡ F2	✗ MASCULINE
Gemini 2.0	✗ MISS	✗ MISS	✗ S	✗ MISS
Llama 70B	⚡ F2 [2W, C2, IDV, RICH]	✓ F2 [RICH]	✗ S	✗ TR
Llama 3.3	✗ MISS	⚡ F2 [PURE]	✗ M	✗ MISS
Mistral	✗ STRESS	⚡ F2 [CC, RICH]	✗ S	✗ 2
Llama 8B	✗ S	✗ RECEIVE	✗ CLIENT	✗ M

A.3 3. Structural Failure (F3) (Keyword: 'παράπονο')

Poem: αρχίζει το μωρό ένα παράπονο, / που ήρθε σ' έναν κόσμο τόσο άπονο.
True Label: F3 ['PURE']

Model	Struct (No RAG)	Struct (RAG)	CoT (No RAG)	CoT (RAG)
Claude 4.5	✗ MISS	✗ IN	✗ S	✗ NOT
Claude 3.7	✗ MISS	⚡ F3 [MOSAIC]	⚡ F3	⚡ F3 [IMPERFECT]
GPT-4o	✗ MISS	✗ MISS	✗ MATCH	⚡ F3 [IMPERFECT, RICH]
Gemini 2.0	✗ MISS	⚡ F3 [IDV, RICH]	✗ MISS	✗ F2
Llama 70B	✗ MISS	✗ F2 [PURE]	✗ F2	✗ M
Llama 3.3	✗ M [C1, RICH]	✗ MISS	✗ S	✗ MISS
Mistral	✗ STRESS	✗ F2 [IDV, PURE]	✗ S	✗ MISS
Llama 8B	✗ K	✗ F2 [IDV, RICH]	✗ S	✗ ZIGSPACE

A.4 4. Mosaic Failure (Keyword: 'λυγμέ')

Poem: του αθηναϊκού εσύ χινόπαρου λυγμέ, / ψιχάλα κυνηγάρα, που έβρεχες για με;
True Label: M ['MOSAIC']

Model	Struct (No RAG)	Struct (RAG)	CoT (No RAG)	CoT (RAG)
Claude 4.5	⚡ M [COPY]	⚡ M [COPY]	⚡ M [COPY]	⚡ M [MOSAIC, PURE]
Claude 3.7	⚡ M [IMPERFECT]	✓ M [MOSAIC]	✗ S	⚡ M [PURE]
GPT-4o	✗ S	✗ MISS	⚡ M [COPY]	⚡ M
Gemini 2.0	✗ MISS	⚡ M [IMPERFECT]	✗ MISS	⚡ M
Llama 70B	✗ S	✗ MISS	✗ S	✗ NO
Llama 3.3	✗ GIVEN	⚡ M [IMPERFECT]	✗ S	✗ MISS
Mistral	✗ STRESS	✗ STRESS	⚡ M [IMPERFECT, RICH]	⚡ M [IMPERFECT]
Llama 8B	⚡ M	✗ MISS	✗ SATURDAY	✗ S

A.5 5. Feature Hallucination (Keyword: 'αφρός')

Poem: Γιασεμιά, και κοράκια. Και των άσπρων ο αφρός / και του μαύρου η φοβέρα πάντα εντός μου κι εμπρός.

True Label: M ['PURE']

Model	Struct (No RAG)	Struct (RAG)	CoT (No RAG)	CoT (RAG)
Claude 4.5	☞ M [IMPERFECT]	☞ M [C1, RICH]	☞ M	✓ M [PURE]
Claude 3.7	✗ S	☞ M [RICH]	☞ M	✗ IN
GPT-4o	✗ S	✗ IN	☞ M	✗ IN
Gemini 2.0	☞ M [C2, RICH]	✓ M [PURE]	☞ M	✗ S
Llama 70B	☞ M [IMPERFECT]	✗ MISS	✗ [IA]	✗ F2
Llama 3.3	✗ S	✗ OF	✗ IN	☞ M
Mistral	✗ STRESS	✗ MISS	☞ M [F2, IMPERFECT]	☞ M [IMPERFECT]
Llama 8B	✗ MISS	✗ F2 [MOSAIC]	✗ STRING	✗ COMPLETE

A.6 6. Archaic Failure (Keyword: 'Ελληνίς')

Poem: Την εγέννησεν εις δήμος, μία πόλις Ελληνίς, / αλλ' ευθύς εκείνη έπτη, κι έστησεν εν ουρανοίς

True Label: M ['RICH']

Model	Struct (No RAG)	Struct (RAG)	CoT (No RAG)	CoT (RAG)
Claude 4.5	✗ MISS	✗ S	☞ M	☞ M [PURE]
Claude 3.7	✗ F2 [RICH]	☞ M [PURE]	✗ S	☞ M
GPT-4o	✗ OF	✗ OF	☞ M [IDENTICAL]	✗ S
Gemini 2.0	☞ M [IMPERFECT]	✗ MISS	☞ M	✗ MISS
Llama 70B	✗ ING	✗ MISS	☞ M	✗ NO
Llama 3.3	✓ M [RICH]	✗ F2 [PURE]	✗ BOTH	✗ OF
Mistral	✗ MISS	✗ F2 [C1, IMPERFECT, RICH]	✗ S	✗ MISS
Llama 8B	✗ IRSECURITY	✗ DAV	✗ CHOOSING	✗ MISS

A.7 7. Imperfect Detection (Keyword: 'γρήγορο')

Poem: άόριστη, με διάβα γρήγορο, / Στου καφενειού την είσοδο

True Label: F3 ['IMP', 'C', 'IMPERFECT']

Model	Struct (No RAG)	Struct (RAG)	CoT (No RAG)	CoT (RAG)
Claude 4.5	✗ S	✗ S	☞ F3 [IMPERFECT]	✗ IMPERFECT
Claude 3.7	☞ F3 [C1, IMPERFECT, RICH]	✗ S	✗ F2	✗ I
GPT-4o	☞ F3 [IMPERFECT]	✗ F2 [IMPERFECT]	✗ F2 [F2]	✗ S
Gemini 2.0	✗ MISS	☞ F3 [IMPERFECT]	✗ MISS	✗ F2
Llama 70B	✗ MISS	✗ F2 [IMPERFECT]	✗ M [IMPERFECT]	✗ S
Llama 3.3	✗ MISS	✗ M [IMPERFECT]	✗ MISS	✗ MISS
Mistral	✗ STRESS	✗ STRESS	✗ F2 [IMPERFECT]	✗ STRESS
Llama 8B	✗ F2 [2W, IDV]	✗ MISS	✗ D	✗ S

A.8 8. Proper Noun Failure (Keyword: 'νιότα')

Poem: Πολεμιστή, τα γαληνά σου νιότα / Όπου έπεσες, κλωνάρια κι απ' του Ευρώτα

True Label: F2 ['PURE']

Model	Struct (No RAG)	Struct (RAG)	CoT (No RAG)	CoT (RAG)
Claude 4.5	☞ F2	✗ STEP	☞ F2	☞ F2
Claude 3.7	☞ F2 [IMPERFECT]	✗ S	✗ F3	✗ I
GPT-4o	✗ S	✗ IN	☞ F2	✗ M [IDV, PURE]
Gemini 2.0	✗ MISS	✗ MISS	☞ F2	☞ F2
Llama 70B	✗ MISS	✗ MISS	☞ F2	☞ F2
Llama 3.3	✗ MISS	✗ THE	✗ SINCE	✗ BUT
Mistral	✗ STRESS	✗ MISS	✗ S	✗ 3
Llama 8B	✗ S	✗ F3 [2W, CC, IDV, RICH]	✗ THE	✗ F3

A.9 9. Visual vs Phonetic (Keyword: 'ορθός')

Poem: άξαφνα το παράθυρο και στάθηκα ορθός, / τις μυρωδιές, τα χρώματα και το ιλαρό το φως.

True Label: M ['MOSAIC', 'IDV']

Model	Struct (No RAG)	Struct (RAG)	CoT (No RAG)	CoT (RAG)
Claude 4.5	✖ M [COPY]	✖ M [IMPERFECT]	✖ M [IDV, IMPERFECT]	✖ M [PURE]
Claude 3.7	✖ MISS	✖ M [IMPERFECT]	✖ I	✖ IN
GPT-4o	✖ S	✖ S	✖ MISS	✖ M [PURE]
Gemini 2.0	✖ MISS	✖ M [IMPERFECT]	✖ MISS	✖ MISS
Llama 70B	✖ S	✖ MISS	✖ S	✖ M [IMPERFECT]
Llama 3.3	✖ M [IMPERFECT]	✖ M [IMPERFECT]	✖ MISS	✖ IN
Mistral	✖ MISS	✖ M [PURE]	✖ S	✖ ONSET
Llama 8B	✖ MISS	✖ THE	✖ BETWEEN	✖ S

B Appendix: Detailed Generation Traces

We present full interaction traces for the three classes of generation errors discussed in the Qualitative Analysis. These examples demonstrate how the hybrid verification loop intercepts compliant but incorrect outputs (Stress, Feature Precision) and hallucinations, guiding the model to phonologically valid results.

B.1 1. Stress Pattern Correction (GPT-4o)

Test Case: Theme "Sky" (ουρανός), Constraint: **F3 (Proparoxytone) Rhyme**.

The model initially defaults to the more common F2 pattern.

Initial Draft (Pure LLM):

Στο χέρι σου το μυστικό αναβιώνει,
και μέσα στα σκιερά πελάγη απλώνει.

Verifier Feedback: Stress mismatch: Expected F3, found F2 for 'αναβιώνει'/'απλώνει'.

Refined Output (Hybrid):

Ω θεέ μου, στον ουρανό να ψιθυρίσουμε,
τα όνειρά μας να πετάξουν να αγγίξουμε.

Analysis: The corrected rhymes (ψιθυρίσουμε / αγγίξουμε) correctly carry stress on the antepenultimate syllable.

B.2 2. Feature Precision (Claude 3.7)

Test Case: Theme "Light" (φως), Constraint: **M-PURE Rhyme**.

The model generates a valid rhyme that violates the "Pure" constraint (by being "Rich").

Initial Draft (Pure LLM):

Θα λάμπει πάντα φωτεινό στον κόσμο το σκληρό,
γλυκαίνει κάθε σκοτεινή γωνιά σαν θησαυρό.

Verifier Feedback: PURE requested but 'σκληρό'/'θησαυρό' is RICH (Onset: /r/).

Refined Output (Hybrid):

Το φως σαν κύμα της χαράς ξυπνά
τα σκοτεινά μου όνειρα φωτίζει καθαρά

Analysis: The refined pair (ξυπνά / καθαρά) has distinct onsets (/p/ vs /th/), satisfying the PURE constraint.

B.3 3. Hallucination Correction (Claude 4.5)

Test Case: Theme "Love" (αγάπη), Constraint: **M Rhyme**.

The model hallucinates rhymes with no phonetic overlap.

Initial Draft (Pure LLM):

Η καρδιά μου χτυπά δυνατά για σένα που αγαπώ,
κάθε στιγμή που περνά χωρίς εσένα είναι πικρή, το ξέρω.

Verifier Feedback: No rhyme: 'αγαπώ' / 'ξέρω'.

Refined Output (Hybrid):

Μέσα στην καρδιά μου κρύβεται αγαπώ,
ένα όνειρο γλυκό που με κρατώ,

Analysis: The model successfully regenerates a valid rhyme pair (αγαπώ / κρατώ).

C Appendix: Granular Feature Detection Statistics

This section provides the detailed identification accuracy for specific rhyme features (RICH, MOSAIC, IDV, IMPERFECT). These features represent the most challenging phonological patterns for current LLMs.

Model	PURE (n=68)	RICH (n=24)	MOSAIC (n=4)	IDV (n=24)	IMP (n=8)
Mistral Large	4.4%	20.8%	0.0%	12.5%	12.5%
Llama 3.1 70B	5.9%	16.7%	0.0%	12.5%	37.5%
Claude 3.7	8.8%	12.5%	0.0%	0.0%	25.0%
Claude 4.5	5.9%	8.3%	0.0%	4.2%	25.0%
Llama 3.3 70B	2.9%	8.3%	0.0%	0.0%	37.5%
Gemini 2.0	5.9%	0.0%	0.0%	0.0%	37.5%
GPT-4o	5.9%	0.0%	0.0%	0.0%	25.0%

Table 6: Feature detection accuracy (%) by individual feature type across all models.

D Representative Prompt Template

We provide the system prompt used for the **Zero-Shot Structured** identification task. This template establishes the phonological taxonomy and the structured output format as it was provided to the LLMs.

You are a Greek poetry rhyme analyzer. Your task is to identify rhyme patterns in Modern Greek poetry.

RHYME TAXONOMY:

1. Position-based Classification:

- M (Masculine): Rhyme from final stressed vowel to line end
- F2 (Feminine-2): Rhyme from penultimate stressed vowel to line end
- F3 (Feminine-3)³: Rhyme from antepenultimate stressed vowel to line end

2. Complexity Features:

- **RICH (onset consonants match):**
 - TR-S: Total rich with singleton onset (καλά - ξαλά)
 - TR-CC: Total rich with complex onset (αυγή - ναυγή)
 - PR-C1: Partial rich, first consonant matches (στόματα - σώματα)
 - PR-C2: Partial rich, second consonant matches (φοβερίζουν - τρίζουν)
- **IDV (Pre-rhyme vowel identity):** Vowel before stressed syllable matches.
- **MOS (Mosaic):** Rhyme domain crosses word boundaries.
- **IMP (Imperfect):**
 - IMP-V: Stressed vowel differs (χάνετε - γίνετε)
 - IMP-C: Consonants differ (ξαφνίζει - τεχνίτη)
 - IMP-0F: Final consonant-zero alternation
 - IMP-0M: Medial consonant-zero alternation
- **COPY:** Complete word/phrase repetition.

ANALYSIS PROCEDURE: 1. Identify line-final stress position. 2. Extract rhyme domain (from stressed vowel rightward). 3. Compare with other lines (scan 4-line window by default). 4. Classify position (M/F2/F3) first. 5. Then identify features (RICH, IDV, MOS, IMP, COPY). 6. Use phonetic similarity, NOT just orthography.

POEM TO ANALYZE: *[Input Text]*

³The prompt retains the original experimental term. In the paper body, we use the label “F3” (proparoxytone rhyme) following Topintzi et al. (2019).

Style as Signature: Profile-Based Authorship Verification of Mihai Eminescu’s Journalistic Corpus

Ioana-Roxana Boriceanu¹, Liviu P. Dinu^{1,2}

¹Faculty of Mathematics and Computer Science

²Human Language Technologies Research Center

University of Bucharest, Romania

ioana-roxana.boriceanu@s.unibuc.ro, ldinu@fmi.unibuc.ro

Abstract

Authorship verification aims to assess whether a questioned text is stylistically compatible with an author’s known writings, a task that is particularly challenging in historical corpora with partial ground truth. We address this problem in the context of Mihai Eminescu’s journalistic corpus, a historically grounded collection comprising published articles, manuscripts, and texts of uncertain authorship. Using a profile-based framework with character n-grams and function words, we examine how stylistic compatibility behaves across different profile construction settings and temporal splits. The results show that character trigram profiles consistently accept verified texts while producing a small and stable set of rejections among disputed items, whereas function word profiles show near complete acceptance across the corpus. A qualitative analysis shows that rejected texts exhibit meaningful differences in discourse structure and communicative purpose. These findings illustrate how authorship verification can support literary scholarship through stable signals for close reading.

1 Introduction

Mihai Eminescu (1850-1889) is widely regarded as the most influential poet of Romania and one of the central figures of its cultural history. Often described as a literary genius, Eminescu played a decisive role in shaping modern Romanian poetry and literary language. He is best known for the originality of his poetic work, which is commonly associated with European Romanticism and shows strong influences from German literature and philosophy. Eminescu’s literary work addresses themes such as love, nature, time, history, and the human condition, often combining lyrical expression with philosophical reflection. Drawing on folklore, mythology, and classical sources, his poetry has had a lasting influence on Romanian literature and remains central to literary scholarship (Iorga, 1969).

Eminescu’s career spanned several distinct stages, ranging from early work in theatre as a prompter, to periods of study in Vienna and Berlin, and later to administrative and editorial positions such as library director and journalist. These experiences shaped both his intellectual outlook and his writing. Beyond his literary activity, Eminescu also played a significant role in the development of Romanian journalism. His journalistic work, produced between 1870 and 1889, forms a large and influential corpus that reflects both the intellectual climate of the period and his distinctive authorial voice. His journalistic texts address a wide range of topics, including politics, education, culture, economics, and social reform, offering a sustained commentary on the tensions produced by the rapid modernisation of society (Mocanu, 2020).

An important aspect of this journalistic corpus is that a substantial portion of the articles were published without signature, while others appeared under pseudonyms such as *Fantasio* or *Varro*. Eminescu rarely signed his work explicitly, and for most of his journalistic career he avoided attaching his name to published texts (Mocanu, 2020). Although editorial scholarship has clarified the authorship of many articles, a small group of texts remains of disputed authorship, making this body of journalistic writing a particularly suitable case for authorship verification (AV).

Authorship verification (AV) (Stamatatos, 2016) is a task in natural language processing that aims to determine whether a given text was written by a particular author. In contrast to authorship attribution (AA) (Stamatatos, 2009), which assigns a text to one author from a predefined set of candidates, AV examines whether a questioned document is stylistically compatible with an author’s known body of work. This problem arises in a variety of contexts, including the resolution of literary authorship disputes (Juola, 2013), the authentication of historical manuscripts (Tuccinardi, 2017), and forensic text

analysis (Juola, 2021).

In the case of Mihai Eminescu, his writings have been extensively studied from a literary perspective, with particular attention to their cultural, political, and ideological context. His writing style has also been the subject of early quantitative and statistical analyses, primarily focused on his poetic corpus (Marcus, 1970; Seche, 1974; Irimia, 2009). By contrast, systematic studies of authorship and stylistic variation in his journalistic corpus have received limited attention. To address this gap, we apply authorship verification techniques to Mihai Eminescu’s journalistic corpus in order to support literary scholarship in the analysis of texts of uncertain provenance. Our contributions can be summarised as follows:

- We present a detailed case study of profile-based authorship verification on a historically grounded Romanian journalistic corpus with verified and uncertain texts.
- We show that rank-based character trigram profiles achieve a stable balance between acceptance of verified texts while retaining structured variation among disputed items, whereas function word profiles tend to over-accept in this setting.
- We demonstrate that verification rejections are interpretable and motivate closer reading of differences in discourse structure and communicative purpose, supporting the use of profile-based authorship verification as an exploratory analytical tool and not only as a purely classificatory mechanism.

2 Related Work

A variety of approaches have been explored for authorship verification (AV). Early work includes the unmasking method (Koppel et al., 2007), which assesses authorship by progressively removing discriminative features and observing changes in classification performance. Another line of research relies on profile-based methods (Potha and Stamatatos, 2014), where an author’s known texts are aggregated into a single stylometric representation and compared with questioned documents using distance based measures. Compression based methods (Halvani et al., 2017) avoid explicit feature engineering and rely on similarity scores derived from text compression. Another widely used approach is the impostors method (Koppel and Winter, 2014),

which evaluates authorship by contrasting a questioned text against large sets of distractor documents using randomly selected feature subsets.

More recent studies have employed language model based techniques for authorship verification (Kim et al., 2025; Hung et al., 2023). While these approaches often achieve strong empirical performance, their reliance on opaque representations makes them less suitable for settings where interpretability and transparency are required, such as literary scholarship and forensic analysis. In contrast, we focus on a transparent profile-based setting where decisions are based on explicit ranked feature profiles and where outcomes can be interpreted in relation to textual structure.

3 Data and Methods

3.1 Mihai Eminescu’s feuillets

Our dataset consists of Mihai Eminescu’s journalistic writings as published in the *Opere* edition of the Romanian Academy (1989). This corpus is organized into five volumes, each corresponding to a distinct period of his journalistic activity and containing three categories of texts: (1) published articles (PUB) attributed to Eminescu in contemporary newspapers, (2) manuscript texts (MAN) preserved in his own handwriting, and (3) texts of disputed authorship, referred to as dubia (DUB) in this work. The distribution of texts across volumes is summarized in Table 1.

Volume	PUB	MAN	DUB
Vol. 9 (1870-1877)	510	52	12
Vol. 10 (1877-1878)	264	14	44
Vol. 11 (1880)	323	11	7
Vol. 12 (1881)	356	31	14
Vol. 13 (1882-1883, 1888-1889)	242	96	26
Total	1695	204	103

Table 1: Distribution of texts across the five journalistic volumes

The corpus exhibits substantial variation in document length, with texts ranging from 10 to 12,767 words and a median length of 536 words. Manuscript texts tend to be shorter, with a median of 119 words, whereas published articles and texts of disputed authorship are generally longer. During preprocessing, we removed metadata such as headers, editorial notes, and footnotes where present.

We applied standard text normalization, including lowercasing and the removal of artefacts introduced by optical character recognition (OCR)

in the transcription of manuscript materials. In particular, lowercasing was applied to reduce sparsity and to avoid mixing up stylistic signals with editorial or contextual variation in capitalization, particularly for proper names, titles, and institutional references, which are unevenly distributed across journalistic texts. Original punctuation and spacing were preserved, as these features are relevant for the character n-gram representations used in our experiments.

3.2 Methods

We adopt a profile-based authorship verification framework inspired by [Potha and Stamatatos, 2014](#) and integrate the rank distance proposed by [Dinu, 2003](#), which has been shown to be effective in computational stylistics ([Popescu and Dinu, 2008](#); [Dinu et al., 2012](#)). Texts with verified authorship, including both published articles and manuscript materials, are used to construct the author profile, while texts of disputed authorship are treated as the evaluation set. Each questioned text is represented in the same feature space as the profile, and authorship is assessed by computing its distance from the author profile.

In this study, we treat style operationally as recurring patterns in low-level textual features, following standard practice in stylometry. We consider two types of features: function words (FW) and character n-grams (CNG), both extracted directly from the normalized text. Character n-grams are widely used in stylometric analysis because they capture local orthographic and morphological regularities while remaining robust for short and medium-length documents. Function words are generally assumed to be processed non-consciously by authors and are therefore well suited for capturing stable stylistic patterns that are less influenced by topic or content ([Chung and Pennebaker, 2011](#)).

To measure stylistic similarity, we primarily employ the rank-based distance proposed in [Dinu, 2003](#):

$$D_{\text{rank}}(P, Q) = \sum_{i=1}^n |r_P(f_i) - r_Q(f_i)|, \quad (1)$$

where f_i denotes the features under comparison, and $r_P(f_i)$ and $r_Q(f_i)$ represent their ranks in the author profile P and the questioned text Q , respectively. In cases of equal feature frequencies, tied features are assigned the average of the corresponding rank positions, following the standard formulation of rank distance.

For comparison, we also evaluate the asymmetric frequency-based CNG dissimilarity function introduced in [Potha and Stamatatos, 2014](#), which serves as a strong profile-based baseline in our experiments:

$$D_{\text{freq}}(Q, P) = \sum_{g \in Q} \left(\frac{2(f_Q(g) - f_P(g))}{f_Q(g) + f_P(g)} \right)^2, \quad (2)$$

where Q denotes the questioned text, P the author profile, and $f_Q(g)$ and $f_P(g)$ are the normalised frequencies of the character n-gram g in Q and P , respectively.

For the CNG representation, we retain the top L_{AUTHOR} trigrams as the author vocabulary. For each document, we also extract its top L_{DOC} trigrams. In our experiments, we set $L_{\text{AUTHOR}} = 1500$ and $L_{\text{DOC}} = 1000$, corresponding to the number of most frequent character trigrams retained for the author profile and for each questioned document, respectively.

The function word list used in our experiments, shown in [Table 2](#), includes frequent words that mainly serve grammatical and discursive functions. These include conjunctions, prepositions, auxiliary verbs, pronouns, determiners, and discourse particles. The selection is grounded in prior work on Romanian computational stylistics and authorship analysis ([Dinu et al., 2012](#)) and was slightly adjusted using word-frequency analyses of Eminescu’s literary corpus ([Seche, 1974](#)), as well as the most frequent items observed in the constructed author profiles.

și în să se cu o la nu a ce mai din pe un că ca mă fi care era lui fără ne pentru el ar dar îl tot am mi însă într cum când toate al după până decât ei nici numai dacă eu avea fost le sau spre unde unei atunci mea prin ai atât au chiar cine iar noi sunt acum ale are asta cel fie fiind peste această cele face fiecare nimeni încă între aceasta aceea acest aceasta acestei avut ceea cât da făcut noastră poate acestui alte celor cineva către lor unui altă ați dintre doar foarte unor vă aceste astfel avem aveți cei ci deci este suntem va vom vor de cari

Table 2: The function words used in our experiments

Distances are computed by comparing the rank distributions of the author profile and the questioned document over a shared feature space defined as the union of their respective vocabularies. For each text of uncertain authorship, we compute its distance to the author profile using both feature types and both distance formulations. The acceptance threshold is derived from the distribution of

distances obtained for texts with verified authorship. We define the threshold as

$$T = \mu + k\sigma, \quad (3)$$

where μ and σ denote the mean and standard deviation of the distances for genuine texts. We set $k = 2$ as a conservative choice that retains the vast majority of verified texts under approximately normal distance variation, while still allowing informative rejections among uncertain items. A questioned text is considered stylistically compatible with the author profile if its distance satisfies $d \leq T$. We deliberately avoid optimizing the threshold for maximal separation, as our objective is not classification performance but the identification of stable and interpretable deviations within an author’s stylistic space.

Our analysis also considers continuous distance distributions across splits and volumes, and we evaluate threshold stability under resampling. In this sense, the threshold does not define a sharp stylistic boundary, but provides an interpretable reference point for identifying consistent outliers.

Within this framework, the evaluation focuses on internal stylistic structure in a historically grounded corpus under partial ground truth. The primary analyses are therefore restricted to verified and disputed Eminescu texts, using verified material to estimate the expected range of authorial variation. This design targets the stability and interpretability of deviations inside a single authorial space.

4 Experiments and Results

4.1 Experiments

The experiments are designed to answer two questions: whether verification behaviour is stable across reasonable parameter choices, and whether observed rejections correspond to interpretable textual differences. We evaluate two profile construction settings. In the first setting, the author profile is constructed from all published texts (PUB), while thresholds are estimated from the manuscript texts (MAN), which provide an independent set of verified writings. In the second setting, the profile is constructed from a random subset of the published texts (1356 documents, 80%) together with all manuscript texts, and thresholds are estimated using the held-out published texts (the remaining 20%, 339 documents). This second setting serves as a robustness check in which profile composition

and threshold estimation are varied using disjoint subsets of the published texts.

Our primary distance measure is the rank-based distance (Dinu, 2003), which has been used in computational stylistics as an interpretable method for comparing ranked feature profiles. For comparison, we also report results obtained with the asymmetric frequency-based CNG dissimilarity function proposed by Potha and Stamatatos, 2014. This allows us to assess whether verification behaviour depends on the specific distance formulation or remains consistent across different measures.

Parameter choices for CNG based profiles

We investigate character n -gram representations with $n \in \{3, 4, 5\}$ under varying profile size configurations. In this context, profile size refers to the number of distinct n -gram types retained after sorting by frequency. For each setting, we construct an author profile from the published texts (PUB) and evaluate acceptance rates for manuscript texts (MAN) and dubia (DUB) using the rank-based distance. Parameter selection is guided by two criteria: high acceptance of verified manuscript texts and avoidance of configurations that yield uniform acceptance across text categories, including those of uncertain authorship. In practice, we treat MAN acceptance rates above 95% as acceptable.

Detailed results for all parameter combinations are reported in Table 3. Each row corresponds to a specific combination of author profile size (L_{AUTHOR}) and questioned document profile size (L_{DOC}), while columns report results for character 3-grams, 4-grams, and 5-grams, respectively. Values are reported as pairs of percentages in the form (MAN%, DUB%), indicating the proportion of texts classified as stylistically compatible with the Eminescu profile under the corresponding setting.

Across the parameter grid, character trigram profiles change gradually with profile size and yield a stable set of rejected texts across a wide range of profile sizes. While larger n -gram (4-grams and 5-grams) representations can produce stronger separation between MAN and DUB texts, they also alter the composition of the rejected set, leading to rejections that are less consistent across parameter settings. In the present study, we prioritise stability and interpretability of verification outcomes over maximal discrimination. Trigram representations maintain high acceptance of verified manuscript texts while producing a limited and interpretable

L_{AUTHOR}	L_{DOC}	3-grams	4-grams	5-grams
1000	800	(95.6, 91.3)	(99.0, 97.1)	(100.0, 100.0)
1000	1000	(96.6, 93.2)	(98.5, 95.1)	(100.0, 100.0)
1500	800	(98.0, 98.1)	(97.5, 95.1)	(99.5, 99.0)
1500	1000	(96.6, 94.2)	(97.1, 94.2)	(99.5, 99.0)
1500	1500	(92.6, 85.4)	(97.5, 91.3)	(98.5, 98.1)
2000	800	(99.0, 100.0)	(97.1, 93.2)	(99.5, 99.0)
2000	1000	(98.5, 98.1)	(96.6, 94.2)	(99.5, 98.1)
2000	1500	(97.5, 93.2)	(95.6, 89.3)	(98.5, 95.1)
2000	2000	(95.6, 92.2)	(95.6, 84.5)	(95.6, 92.2)
3000	800	(100.0, 100.0)	(97.1, 93.2)	(99.0, 97.1)
3000	1000	(100.0, 100.0)	(95.6, 92.2)	(99.0, 98.1)
3000	1500	(100.0, 100.0)	(95.6, 89.3)	(98.0, 95.1)
3000	2000	(100.0, 100.0)	(94.1, 83.5)	(96.6, 91.3)
3000	3000	(100.0, 100.0)	(91.7, 77.7)	(91.7, 79.6)

Table 3: Acceptance rates (MAN%, DUB%) across n-gram lengths and profile sizes.

set of rejections among uncertain items. For this reason, we adopt character trigrams for the remaining experiments.

For each value of n , we vary the number of retained character n-grams for the author profile ($L_{\text{AUTHOR}} \in \{1000, 1500, 2000, 3000\}$) and for each questioned document ($L_{\text{DOC}} \in \{800, 1000, 1500, 2000, 3000\}$), subject to $L_{\text{DOC}} \leq L_{\text{AUTHOR}}$. As a compromise between robustness and discriminative capacity, we select $L_{\text{AUTHOR}} = 1500$ and $L_{\text{DOC}} = 1000$ for the trigram representation. This configuration satisfies the MAN acceptance criterion while preserving variation in DUB acceptance and yields a stable and interpretable set of rejected texts, making it suitable for the remaining experiments.

Stability analyses

We evaluate the robustness of the proposed verification framework by testing whether its behaviour remains stable under variation in data splits, temporal coverage, and threshold estimation. These analyses are intended to assess whether verification outcomes reflect consistent stylistic properties of the corpus and not artefacts of a particular sampling or configuration.

We first perform five-fold cross-validation on the manuscript set (MAN). Manuscript texts provide an independently verified set of writings that are stylistically close to the published articles, while remaining separate from the texts used to construct the author profile in our primary setting. In each fold, the acceptance threshold is estimated from four fifths of the MAN texts and evaluated on the held-out subset. Acceptance remains consistently high for both feature types (CNG: 97.6–100%; FW:

95.0–100%), indicating that the manuscript texts form a stylistically coherent set under the proposed distance based criteria.

Next, we examine the distribution of distances across journalistic volumes and text categories (PUB, MAN, DUB) to assess potential period specific effects. For each split and volume specific subset, we compute descriptive statistics including mean, standard deviation, and range. For both CNG and FW representations, distance distributions are comparable across volumes, with no abrupt shifts that would suggest strong temporal drift within the verified material.

To further probe temporal sensitivity, we construct separate author profiles for earlier and later phases of Eminescu’s journalistic activity and evaluate manuscript texts from the opposite period. We define the *early* period as Vols. 9–10 (1870–1878) and the *late* period as Vols. 12–13 (1881–1883, 1888–1889), excluding Vol. 11 (1880) to increase temporal separation. Distance distributions remain similar across early and late configurations for both feature sets, suggesting that verification outcomes are not driven by a narrow temporal slice of the verified corpus.

Finally, we assess the stability of the decision threshold under sampling variation by bootstrapping the manuscript set used for threshold estimation and recomputing the threshold over 200 trials. For both feature representations, the resulting bootstrap distributions exhibit limited variance, indicating that the acceptance criterion is stable under resampling.

Method	Profile	Threshold	PUB held-out (339)	MAN (204)	DUB (103)
CNG rank	PUB	MAN	–	198	97
FW rank	PUB	MAN	–	204	103
CNG rank	80% PUB + MAN	20% PUB	332	–	100
FW rank	80% PUB + MAN	20% PUB	338	–	102
CNG freq	PUB	MAN	–	204	103

Table 4: Number of accepted texts across feature types and profile configurations

Title	Volume	Words	Summary	Rejected
Fata mamei Ango și Giroflé-Girofla	Vol. 9	567	theatre commentary with moral reflection	CNG
Nefericitul X	Vol. 9	376	humorous anecdotal sketch	CNG
Ziarele din Viena	Vol. 10	475	biographical news report on Francisc Schuselka	CNG
Asaltul Angelescu	Vol. 10	649	political–military critique	CNG
Din istoria calului	Vol. 10	555	historical essay on horses in antiquity	CNG
O serată literară. Despot Vodă	Vol. 10	183	cultural news report on a literary salon event	FW
Printr-o indiscrețiune	Vol. 12	350	diplomatic news brief containing a French telegram	CNG

Table 5: DUB texts rejected by at least one rank-based profile

4.2 Results

Table 4 reports, for each feature type and profile construction setting, the number of texts accepted as stylistically compatible with the author profile. The numbers in parentheses in the column headers indicate the number of texts evaluated in each category for the corresponding setting: manuscript texts (MAN) and held-out published texts (PUB) used for threshold estimation, depending on the profile construction strategy, and the full set of disputed texts (DUB). Across all feature types and profile construction settings, only 7 of the 103 disputed texts are rejected at least once, while the remaining 96 are accepted under all feature types and profile construction settings.

We observe that a small number of published or manuscript texts also fall marginally outside their respective acceptance thresholds. Such deviations are expected, as thresholds are derived statistically from empirical distance distributions and are not intended to include every individual verified text. Importantly, these cases are rare and do not indicate systematic instability.

For rank-based CNG representations, acceptance remains high for verified texts while avoiding complete acceptance of DUB texts. Although most disputed texts are accepted, the observed rejections indicate that the method maintains discriminative sensitivity instead of converging toward uniform acceptance. In the PUB profile with MAN threshold setting, CNG rank accepts 97 of 103 DUB texts (94%), while in the second profile setting it accepts 100 of 103 (97%). The three DUB texts rejected in the second setting are a subset of the six rejected in

the first. This balance aligns with the goals of authorship verification, which aims to accommodate natural stylistic variation while remaining sensitive to meaningful deviations.

In contrast, the function word based rank distance yields uniformly high acceptance across all text categories, including DUB texts. Under this feature representation, the second profile setting accepts 102 of the 103 disputed texts, while the first setting accepts all of them. This behaviour suggests that, in the present setting, function word distributions might be insufficiently discriminative. Two interacting factors contribute to this effect. First, the newspaper register limits how much function word use can vary across texts, leading to very similar rank profiles throughout the corpus. Second, short texts introduce some local rank noise, but because the threshold is estimated from comparable verified texts, this variation is treated as normal rather than leading to rejections. As a result, distance values are tightly clustered and acceptance is high.

Finally, the frequency-based CNG dissimilarity function, used here as a baseline, produces a highly permissive profile that accepts all DUB texts. This confirms that the rank-based formulation provides a stricter and more informative decision criterion in this authorship verification setting.

Results analysis

Table 5 lists the DUB texts rejected by the stricter rank-based criterion, together with their volume of publication, length, brief content summaries, and the feature representation under which rejection occurs. To understand what drives these rejections,

we examine this small subset qualitatively and consider factors known to affect profile-based stylistic distances.

A close reading of Eminescu's verified journalistic texts reveals a recurrent structural tendency. Even when prompted by concrete events or documents, his articles rarely remain at the level of description. Reporting is typically followed by interpretive expansion, causal explanation, or normative critique, often extending the immediate topic into broader reflections on institutions, social dynamics, or political principles. Several rejected DUB texts, by contrast, remain confined to the informational layer and do not develop this progression from occasion to diagnosis.

Some rejected texts illustrate this contrast particularly clearly. The humorous sketch *Nefericitul X* consists of a sequence of loosely connected anecdotes built around a generic figure, without developing a sustained argument or interpretive trajectory. *Din istoria calului*, a historical essay on the horse, similarly presents a factual overview, compiling information without developing a broader interpretive or argumentative claim. Both are rejected by both CNG profiles and differ from the dominant mode of the verified corpus, where historical or biographical material usually serves as the basis for polemical or diagnostic argumentation.

The remaining rejected DUB texts also converge in discourse organisation. They rely on episodic or self-contained forms of exposition, without developing a sustained argumentative trajectory. Meaning is produced through a sequence of scenes, illustrative examples, or isolated incidents, without the progressive accumulation of causal or institutional analysis that characterises much of Eminescu's verified journalism. This pattern appears across genres, from *Ziarele din Viena*, a biographical news report, to the pedagogical theatre commentary *Fata mamei Ango și Giroflé-Girofla*, and the political military commentary *Asaltul Angelescu*, which addresses a concrete incident without extending it into a broader institutional critique.

Other rejected cases highlight more specific sensitivities of the feature representations. The only text rejected under the function word profile, *O serată literară. Despot Vodă*, is an extremely short cultural notice reporting on a literary salon event. Approximately one third of the text consists of a list of personal names. In such a short document, this concentration substantially alters the relative distribution of function words, producing an atyp-

ical profile. Notably, this text is not rejected by the character n-gram based profile, suggesting that CNG features are less sensitive to this specific distributional effect.

Another illustrative case is *Printr-o indiscrețiune*, which consists of a diplomatic news report reproducing and commenting on telegraphic exchanges between European political figures. A substantial portion of the article is made up of verbatim quotations in French, reflecting diplomatic conventions of the period. This text is rejected under both CNG based profile configurations. However, a close qualitative reading suggests stylistic compatibility with Eminescu's journalistic writing in the Romanian framing passages. To probe this further, we repeated the analysis after removing the French-language segment. Under this condition, the remaining Romanian text was no longer rejected by the CNG profile. This indicates that the rejection is primarily driven by the presence of extended foreign-language quotations, which substantially alter character n-gram distributions, with the surrounding text remaining stylistically compatible.

Taken together, the rejected DUB texts display recurring characteristics that help explain their distance from the author profile. These properties are relevant for authorship analysis because they intersect with recurring aspects of Eminescu's journalistic practice, which typically integrates reporting with interpretive or argumentative development. The verification outcomes therefore bear on authorial compatibility, while remaining sensitive to the possibility that register or compositional factors may also contribute to observed deviations.

From a methodological perspective, these findings suggest that profile-based authorship verification should not be evaluated solely in terms of aggregate acceptance rates. Instead, the explanatory value of rejections and their alignment with known discourse and genre properties are important. In this respect, the present study illustrates how verification outcomes function as analytic signals that prompt closer textual examination, not as categorical judgments about authorship.

5 Conclusion

This paper investigated profile-based authorship verification in a historically grounded journalistic corpus, focusing on the writings of Mihai Eminescu. Using rank-based distance measures over

character n-grams and function words, we showed that character trigram profiles provide stable verification behaviour across different profile constructions, temporal splits, and threshold estimation procedures, while remaining tolerant of natural stylistic variation within verified texts.

The proposed framework does not aim to maximize discrimination, but to preserve stylistic continuity while identifying meaningful deviations. The rank-based profile representation yields decisions that are stable under resampling and can be related to concrete textual properties, making them suitable for interpretive analysis. Our qualitative examination shows that texts rejected by the model systematically differ from the verified corpus in discourse structure, communicative purpose, and textual composition, with differences that are not attributable to random variation. These findings suggest that profile-based authorship verification is well suited for exploratory analysis of stylistic boundaries within an author's body of work.

More broadly, this study argues that in historical and literary contexts with partial ground truth, authorship verification methods should be evaluated not only by acceptance rates, but by the stability, interpretability, and coherence of their outcomes. Our approach offers these properties and is therefore well suited to corpus-driven literary scholarship in settings with partial and uncertain ground truth. Future work may extend this approach to comparative author profiles, mixed author corpora, or other languages, and explore how verification signals interact with genre, register, and editorial practices in diachronic corpora.

Limitations

We focus on a single author and a single historical corpus, so generalization to other authors, languages, and genres remains to be established. Evaluation on comparable journalistic texts by other authors would provide an explicit estimate of false acceptance under domain-matched negative controls. Document length variation may affect rank stability, especially for shorter texts. While rank-based distances are less sensitive to absolute frequency differences than frequency-based measures, a more systematic assessment using controlled text segmentation would further clarify the influence of document length on verification behaviour. Finally, the method assesses stylistic compatibility, and acceptance and rejection should be interpreted

as analytic signals, not as definitive authorship judgments.

Acknowledgments

This work was supported by Ministry of Research, Innovation and Digitization, CNCS-UEFISCDI, project SIROLA, number PN-IV-P1-PCE-20231701, within PNCDI IV and by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization, and Financial Instruments Program, 2021-2027, MySMIS no. 334906.

References

- Cindy Chung and James Pennebaker. 2011. The psychological functions of function words. In *Social communication*, pages 343–359. Psychology Press.
- Liviu P Dinu. 2003. On the classification and aggregation of hierarchies with different constitutive elements. *Fundamenta Informaticae*, 55(1):39–50.
- Liviu P Dinu, Vlad Niculae, and Octavia-Maria Şulea. 2012. Pastiche detection based on stopword rankings: exposing impersonators of a romanian writer. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 72–77.
- Oren Halvani, Christian Winter, and Lukas Graner. 2017. Authorship verification based on compression models. *arXiv preprint arXiv:1706.00516*.
- Chia-Yu Hung, Zhiqiang Hu, Yujia Hu, and Roy Lee. 2023. Who wrote it and why? prompting large-language models for authorship verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14078–14084.
- Nicolae Iorga. 1969. *Studii literare. Scriitori români*, volume 1–2. Lyceum, Bucureşti.
- Dumitru Irimia, editor. 2009. *Dicţionarul limbajului poetic eminescian. Concordanţele prozei antume*. Editura Universităţii din Bucureşti, Bucureşti.
- Patrick Juola. 2013. How a computer program helped show jk rowling write a cuckoo's calling. *Scientific American*, 20.
- Patrick Juola. 2021. Verifying authorship for forensic purposes: A computational protocol and its validation. *Forensic Science International*, 325:110824.
- Junghwan Kim, Haotian Zhang, and David Jurgens. 2025. Leveraging multilingual training for authorship representation: Enhancing generalization across languages and domains. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34855–34880.

- Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8(6).
- Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187.
- Solomon Marcus. 1970. *Poetica matematică*. Editura Academiei, București.
- Mihaela Mocanu. 2020. Eminescu’s journalistic works—editing approaches and reading patterns. *Philobiblon*, 25(1):43–61.
- Marius Popescu and Liviu P Dinu. 2008. Rank distance as a stylistic similarity. In *Coling 2008: Companion volume: Posters*, pages 91–94.
- Nektaria Potha and Efstathios Stamatatos. 2014. A profile-based method for authorship verification. In *Hellenic Conference on Artificial Intelligence*, pages 313–326. Springer.
- Luiza Seche. 1974. *Lexicul artistic eminescian în lumină statistică*. Editura Academiei R.S.R., București.
- Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.
- Efstathios Stamatatos. 2016. Authorship verification: A review of recent advances. *Res. Comput. Sci.*, 123(1):9–25.
- Enrico Tuccinardi. 2017. An application of a profile-based method for authorship verification: Investigating the authenticity of pliny the younger’s letter to trajan concerning the christians. *Digital Scholarship in the Humanities*, 32(2):435–447.

Measuring Social Integration Through Participation: Categorizing Organizations and Leisure Activities in the Displaced Karelians Interview Archive using LLMs

Joonatan Laato¹, Veera Schroderus², Jenna Kanerva¹,
Jenni Kauppi², Virpi Lummaa², and Filip Ginter^{1,3}

¹TurkuNLP, University of Turku, Finland

²Department of Biology, University of Turku, Finland

³ELLIS Institute Finland

{joonatan.m.laato, veera.i.schroderus,
jmybl, jejakau, virpi.lummaa, figint}@utu.fi

Abstract

Digitized historical archives make it possible to study everyday social life on a large scale, but the information extracted directly from text often does not directly allow one to answer the research questions posed by historians or sociologists in a quantitative manner. We address this problem in a large collection of Finnish World War II Karelian evacuee family interviews. Prior work extracted more than 350K mentions of leisure time activities and organizational memberships from these interviews, yielding 71K unique activity and organization names—far too many to analyze directly.

We develop a categorization framework that captures key aspects of participation (the kind of activity/organization, how social it typically is, how regularly it happens, and how physically demanding it is). We annotate a gold-standard set to allow for a reliable evaluation, and then test whether large language models can apply the same schema at scale. Using a simple voting approach across multiple model runs, we find that an open-weight LLM can closely match expert judgments. Finally, we apply the method to label the 350K entities, producing a structured resource for downstream studies of social integration and related outcomes.

1 Introduction

Digitizing large collections of historical records, combined with the use of large language models (LLMs), has opened new possibilities for analyzing cultural and social patterns at a large scale. LLMs make it straightforward to automate tasks such as extracting entities and other structured information from massive archival collections. In many cases, however, entity extraction alone is not sufficient; additional steps—such as grounding and grouping

entities—are often needed to support the underlying research.

In this study, we examine a unique historical corpus: *Siirtokarjalaisten tie* (*The Path of Displaced Karelians*), a large interview collection documenting World War II evacuees who were permanently relocated from the Karelia region to different parts of Finland. The archive contains testimonies from approximately 160,000 individuals, offering rare coverage of everyday social life at scale. The interviews are brief (see Figure 1), but they typically include information about organization membership and leisure activities. Organizations include a wide variety of different clubs, choirs, work unions, boards, and other similar organizations, while leisure activities primarily refer to hobbies, sports, outdoors activities, and similar. Together, these in many cases allow inferring the degree of social integration of the individuals, and, if extracted across the whole data, of the displaced population at large. Extracting this information is the primary focus of this study.

In prior work, we used LLMs to extract the organization membership and leisure activity mentions from the interviews (Laato et al., 2024). The resulting dataset contains 354,302 extracted entity mentions from 89,339 interviews. However, without further grounding and interpretation, these entities are not directly usable as a proxy for social integration. For instance, the popular “Martha” club of homemakers focusing on spreading household advice appears under multiple free-form variants (e.g., “Marthas” vs. “Martha Club”), and historically meaningful groups may be described with widely differing names across regions and dialects. Critically, mere mentions of organizations and leisure activities lack metadata that would help estimate the degree of social integration implied by a given activity.

This motivates a central step between extraction and domain analysis: normalization and categorization. For sociological questions—especially those focused on social integration and its links to health and longevity—researchers need a representation of participation that is comparable across individuals and communities. Achieving this requires mapping historical mentions to a consistent, interpretable classification framework that reflects culturally specific forms of Finnish social life.

To support this downstream analysis, we evaluate whether LLMs can assist with this structuring step and how closely they can match human expert judgment. In collaboration with sociologists (domain experts), we develop a classification framework designed to capture the dimensions of social participation most relevant to studying integration patterns, and we assess both human reliability and model performance within that framework.

We address three primary questions:

1. Is it feasible to represent the multi-dimensional nature of historical Finnish social organizations and leisure activities in a way that still yields reliable human consensus?
2. If so, to what extent can LLMs replicate domain-expert categorizations of culturally specific historical entities?
3. Can iterative and automated prompt engineering further improve model performance on this task?

2 Data

Our primary source “Siirtokarjalaisten tie” (The Path of Displaced Karelians) is a 4-volume book series documenting Finnish citizens permanently displaced from Eastern Karelia following the 1939-1944 war with the USSR. The volumes contain 89,339 family interviews conducted in 1968–1970 by approximately 300 trained interviewers, representing approximately 160,000 adults from the roughly 420,000 total refugees. All interviews were originally conducted in Finnish. The books were digitized through OCR by [Loehr et al. \(2017\)](#).

The interviews combine standardized demographic sections with free-text information about daily life, including social organization membership and hobbies. Figure 1 illustrates an example interview showing the information typically captured. Previously, we used OpenAI’s GPT-4-turbo LLM to extract all mentions of social entities from these texts ([Laato et al., 2024](#)), resulting in 192,388

Primary person: KALLE VIRTANEN

Spouse: Anna Inkeri Nieminen

Source text: Managing director, born 22. 7. -29 In Antrea. Spouse, Anna Inkeri née Nieminen, shop assist., born 21. 11. -29 Uusikirkko. Children: Jari Tapio -57 Riihimäki, Jouni Juhani -58 Riihimäki, Matti Mikael -61 Loppi. Residence, in Karelia: Antrea, Kaltove-denmaankylä 39, 42-44. Other residences: Tammela -42, Ikaalinen 44-, Janakkala, Loppi, Sajaniemi 47-54, Loppi, Pilpala 54-58, Riihimäki, Loppi kk. 60-. Kalle Virtanen lives with his family in their house. Managing director Virtanen is a member in Karelians of Loppi Reg. Assn. Tax Committee and Sajaniemi’s Deer Lads Regn. Assn. His hobbies include hunting and fishing. Mrs. is a member in Karelians of Loppi Reg. Assn. and serves as the secretary of the Loppi Kuparsaari Marttas. Her hobbies include handicrafts and gardening.

Figure 1: A single interview example with entities pertaining to the husband (yellow) and wife (blue) highlighted. Translated from Finnish, preserving typographic features of the original data. Note: Names and dates have been changed.

Top 10 Hobbies	
handicrafts (<i>käsityöt</i>)	— 31,780
fishing (<i>kalastus</i>)	— 20,222
literature (<i>kirjallisuus</i>)	— 13,128
gardening (<i>puutarhanhoito</i>)	— 10,519
reading (<i>lukeminen</i>)	— 9,324
hunting (<i>metsästys</i>)	— 6,287
skiing (<i>hiihto</i>)	— 5,912
outdoor activities (<i>ulkoilu</i>)	— 5,597
sports (<i>urheilu</i>)	— 4,997
handicrafts [partitive case] (<i>käsitöitä</i>)	— 4,708
Top 10 Organizations	
Karelian Society (<i>Karjalaseura</i>)	— 6,610
Martha Association (<i>Marttayhdistys</i>)	— 5,725
youth association (<i>nuorisoseura</i>)	— 2,545
Farmers’ Association (<i>Maamiessseura</i>)	— 2,476
Marthas (<i>Martat</i>)	— 1,967
Agricultural Women (<i>Maatalousnaiset</i>)	— 1,957
Lotta Svärd (<i>Lotta Svärd</i>)	— 1,930
sports club (<i>urheiluseura</i>)	— 1,252
church choir (<i>kirkkokuoro</i>)	— 1,230
elementary school (<i>kansakoulu</i>)	— 1,212

Table 1: Top 10 most frequently mentioned hobbies and organizations. Finnish terms in italics.

mentions of hobbies and 161,914 mentions of organizations, extracted with 88.8% F-score, close to human performance at the task. These mentions comprise a total of 71,874 unique entities (7,674 hobbies and 64,200 organizations), as common activities were mentioned by many individuals.

The top 10 most frequently mentioned entities (Table 1) already show very active lifestyle and organizational involvement of the Karelian popula-

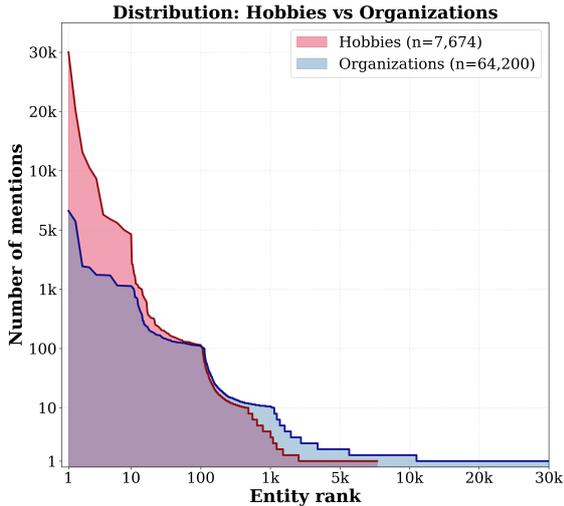


Figure 2: Distribution of hobby and organization mentions across entity ranks. Hobbies (red) show concentrated distribution in the first few mentions, with some 7.6k unique entities, while organizations (blue) show a long tail extending beyond 60k unique entities.

tion in post-war Finland. Common hobbies center on traditional rural activities such as handicrafts, fishing, and gardening. Meanwhile organizations include Karelian cultural societies, women’s associations, and agricultural groups.

As illustrated in Figure 2, the extracted hobbies and organizations display unsurprisingly different distribution. Hobbies include only 7,674 unique entities, where the top 200 account for 92.6% of all mentions (192,388 total). In contrast, organizations display a long-tail distribution with 64,200 unique entities, where the top 200 capture only 49.0% of all mentions (161,914 total).

2.1 Entity Normalization Challenges

Pattern matching reveals variations referring to effectively the same activity. The term “mart*” matches 156 unique organization strings (out of 64,200 total unique organizations) including: Martha Association (*Marttayhdistys*), Marthas (*Martat*), marthas (*martat*), Martha Society (*Marttaseura*), Martha Club (*Marttakerho*). The same root appears in 40 unique hobby strings (out of 7,674 total unique hobbies) including: Martha work (*marttatyö*), Martha activities (*marttatoiminta*). These all for the most part refer to the same popular homemaker organization and its activities.

The pattern “karjala*” (Karelia) yields 590 unique organizational strings referring to various societies established by the Karelians, often to pro-

mote their interests and maintain connections, and 94 unique hobby strings predominantly describing Karelia-specific food preparation: e.g. making Karelian pastries (*karjalanpiirakoiden valmistus*).

Handicraft variations (“käsit*”) show 78 unique hobby strings: handicrafts (*käsityöt*), handicrafts-partitive (*käsitöitä*), making handicrafts (*käsitöiden tekeminen*), handicraft making (*käsityöiden tekeminen*). The pattern also matches 22 unique organizational strings: handicraft club (*käsityökerho*).

Additionally, many organizations with entirely different names belong to the same functional category. For example, various sports clubs while having unique names should nevertheless be grouped together for the analysis.

2.2 Pre-processing: Hierarchy Extraction

We pre-process the organizations using an LLM to separate out hierarchies of organizations and individual roles: given an entity like “Chairman of the Administrative Board of the Karelian Society,” we prompt an LLM (Llama-3.3) to separate the individual’s role as well as the underlying hierarchy. In this case we would extract “Chairman” as a role and “Administrative Board” as a hierarchical element. The remaining entity is then interpreted as the actual social organization, here “Karelian Society”. This reduced the unique social organization count from 64,200 to 52,982, extracting 2,147 different roles and 709 unique hierarchies.

3 Related Work

Historical and cultural data presents unique challenges for entity normalization and clustering: (1) classifying entities into meaningful groups, (2) normalizing variant names, and (3) capturing semantic relationships through multi-dimensional categorization.

Previously, these have been approached using distributional representations such as word2vec and fastText (Mikolov et al., 2013; Bojanowski et al., 2017) to capture semantic similarities through vector clustering (Baroni et al., 2014). Transformer-based models like BERT (Devlin et al., 2019) advanced multi-label classification through contextualized representations, but require substantial domain-specific labeled data for training. More recently, LLMs have enabled clustering entities based on contextual embeddings. Huang and He (2025) transform text clustering into classification by having the LLM generate potential labels, merge

semantically similar labels, and assign entities accordingly. This label-merging approach captures underlying data structure without requiring any fine-tuning. Viswanathan et al. (2024) demonstrated few-shot clustering through LLM-generated pairwise constraints and entity canonicalization with minimal examples. Recent work on entity matching (Huang and Zhao, 2024) highlights how LLMs implement semantic equivalence beyond syntactic similarity, though challenges remain for culturally-specific entities and abbreviated names. However, multi-label classification with LLMs presents distinct challenges. Ma et al. (2025) found that LLMs suppress multiple labels during generation despite applicable categories. Lan et al. (2024) addressed this through multi-label contrastive learning. Tabatabaei et al. (2025) demonstrated successful industrial-scale deployment with custom taxonomies for hierarchical multi-label classification. Across these tasks, careful prompt engineering significantly improves LLM’s classification accuracy, with automated methods like APE (Zhou et al., 2023) and OPRO (Yang et al., 2024) emerging to generate and iteratively optimize prompts.

4 Methods

While recent work shows that LLMs can cluster semantically (Huang and He, 2025) and normalize entities (Viswanathan et al., 2024), these approaches focus on single-label classification or binary similarity judgments. However, as demonstrated in our data analysis, the diversity of historical social organizations resists discrete categorization. Attempting to single-label these entities would require hundreds of narrowly-defined categories. Instead, a multi-label approach that captures specific dimensions of each entity proves more appropriate. Furthermore, rather than forcing data into pre-existing taxonomies (as is commonly used in NER), our approach develops classification dimensions directly towards sociological research questions about social integration patterns.

4.1 A Classification Framework for Social Integration Research

The design of this framework is directly guided by the main research objective of sociologists: characterizing the social integration of Karelian refugees after relocation. Combined with further connected data on this population, such as health and mortality information, this characterization of social integra-

tion can be used to answer questions about how different levels of integration influenced mortality outcomes. Furthermore, interactions of social integration level can be measured relative to numerous factors, such as age, residential history, distance from relatives, and whether the relocation occurred individually or with the community. The dataset’s scale and demographic uniformity provide a rare opportunity to isolate the effects of post-migration social environments on individual outcomes.

While this data, and approach, will eventually be utilized broadly across different fields, initially, the main focus will be on investigating the health effects of different kinds of social integration. Consequently, key questions and category types of the annotation framework were constructed based on well-known health factors, such as physical activity (Iwasaki et al., 2001; Rhodes et al., 2017), sociality (Holt-Lunstad et al., 2010; Hansdottir et al., 2022), and mental stimulation (Yates et al., 2016; Hansdottir et al., 2022; McQuade and O’Sullivan, 2024). The frequency of activity is relevant to health in both physical activity (Iwasaki et al., 2001; Rhodes et al., 2017) and mentally stimulating leisure activities (Weziak-Białowolska et al., 2023). For non-solitary activity, frequency and group size additionally reflect the social intensity of participation.

These considerations led to the design of an annotation schema with four main aspects for each activity/organization, formulated into four questions: **Q1 - Categories:** The general type(s) of organization/activity; **Q2 – Group Size:** How many people typically participate? **Q3 – Regularity:** How often do they meet/engage? **Q4 – Physical Intensity:** What level of physical activity is involved? Each question has predefined answer options, detailed in Table 2, with special options for ambiguous cases: “Cannot be determined” applied when the judgment is impossible to make based on the information given, and “Data error” applied when the entity is not, in fact, a leisure-time activity or organization, but rather an extraction error.

To support culturally informed annotation, a questionnaire was distributed through the Karjalan Liitto ry Facebook page, where Karelian migrants and their descendants provided ratings and descriptions for approximately 60 common ambiguous entities. While these responses were not directly incorporated into the current annotation framework, they represent a valuable reference point in the refinement of the guidelines.

Q1 – Categories Professional/Work Cultural/Traditional Religious/Spiritual Sports/Physical activity Creative/Artistic Educational/Academic Social welfare Administrative Political General social group Nature-related Health-related Property/ownership Animal-husbandry Cooking Military-related Special Hobbies Non-definable Data error	Q2 – Group Size Alone Small group Large group Non-definable Data error Q3 – Regularity Regular Occasional Continuous Event-based Non-definable Data error Q4 – Physical Intensity Intense Continuous Light Stationary Non-definable Data error
---	--

Table 2: Questionnaire schema showing all options for each question in our final round of annotations.

4.2 Annotation Process

Four annotators, two domain experts and two machine learning specialists, annotated five iterative rounds of 50 entities, split evenly between hobbies and organizations. After each round the schema and guidelines were refined. For example, annotators should distinguish between organizations and their administrative bodies, separate “Administrative” (governance) from “Political” (party activities), and select “Cannot be determined” rather than guess when uncertain. Once a sufficient agreement had been reached, a set of 200 entities was annotated by all four annotators, and split into a development/evaluation set of 50 entities, and a final test set of 150 entities.

We measured inter-annotator agreement using pairwise Cohen’s Kappa across both evaluation (n=50) and test (n=150) sets (See Appendix Table 8 for full results). The pairwise Kappa score shows moderate-to-substantial average agreement of 0.704 (evaluation set) and 0.690 (test set). Annotator agreement is also a proxy for the difficulty of making the judgements, and the reliability of the resulting annotations. We therefore also inspect the agreement on a per-question basis. Here Q2 (group size) reached the highest agreement at 0.772/0.781 (eval/test), while Q4 (physical activity) showed the lowest, yet still moderate agreement at 0.678/0.579.

The final ground truth was established using majority vote across the four annotators (Snow et al., 2008; Sheng et al., 2008). Due to the multi-label nature of the task, a label was selected when at least

two annotators selected it (2-out-of-4 threshold). In evaluation, human annotators are scored against leave-one-out majority, excluding their own labels (3vs1), while LLM predictions will be evaluated against the full four-annotator consensus (4vs1).

In terms of F-score on the consensus annotation labels pooled across all four questions, the annotators achieved on average 78.6% F-score (test, n=150) and 80.7% F-score (evaluation, n=50) against the leave-one-out majority vote (Appendix Table 8).

4.3 LLM Annotation

With minimal editing, the annotation guidelines were crafted into the LLM prompt (see Appendix A.1 for the full prompt). We evaluated six state-of-the-art open-weight models to determine which could best replicate human performance. Additionally, using the best open-weight model, we investigated variants such as prompting all four questions simultaneously in one prompt, as compared to prompting each question individually in separate prompts, so as to establish whether the models benefit from holistic entity understanding versus focused attention on individual classification dimensions.

4.4 Models

We evaluated the following six state-of-the-art open-weight models: Qwen2.5-72B-Instruct, a 72B dense instruction model trained on 18T tokens with RLHF supporting 131k-token context (Yang et al., 2025b); Qwen3-235B-A22B, a 235B-parameter MoE (22B active) trained on 36T tokens with native chain-of-thought capabilities (Yang et al., 2025a); DeepSeek-R1-Distill-Llama-70B, a 70B student model distilled from the 671B DeepSeek-R1 teacher with RL-driven chain-of-thought (Guo et al., 2025); Meta-Llama-3.3-70B-Instruct, a December 2024 refresh with 70B parameters and 128k context (Grattafiori et al., 2024); Llama-4-Scout-17B-16E, a 17B-active/16-expert MoE enabling 10M-token contexts (Meta-AI, 2025); Mistral-Large-Instruct-2407, a 123B flagship model with 128k context (Mistral-AI, 2024). Notably, neither Qwen3 nor DeepSeek-R1, despite their chain-of-thought capabilities, outperformed Mistral-Large on this task.

The overall results of the models in terms of F-score on the evaluation set (n=50) are summarized in Table 3, demonstrating substantial differences

Model	Avg F1
Qwen2.5-72B-Instruct	65.0
Qwen3-235B-A22B	68.2
DeepSeek-R1-Distill-Llama-70B	62.5
Meta-Llama-3.3-70B-Instruct	73.0
Llama-4-Scout-17B-16E	70.4
Mistral-Large-Instruct-2407	74.6
Mistral-Large-Instruct-2407 [†]	73.6

[†]Same model, but prompted one question at a time rather than all four questions for each entity.

Table 3: Average F-scores of each model on the evaluation set (n=50).

between individual models.

4.5 Prompt Optimization

Among the six models evaluated, Mistral-Large achieved the highest performance with 74.6% average F-score (Table 3). To test whether automated prompt engineering could further improve this baseline, two optimizer models (OpenAI’s o3 and Anthropic’s Claude 4 Opus) were used to generate optimized prompt variants when being informed about the errors observed in the annotation carried out with the original prompt.

To this end, the 50-entity evaluation set was divided into 5 splits (20 items for optimization, 30 items for evaluation), with each entity appearing in exactly two optimization sets and three evaluation sets. For each split, both optimizer models produced two prompt variants based on error feedback, yielding 20 optimized prompts total. Only 7 showed improvements on the held-out data, and these gains were individually relatively minor, on average 4pp F-score (when averaging only data splits with an improvement, i.e. an overestimate of any actual gain). This suggests the iterative human annotation process had already produced instructions clear enough for the model to approximate human judgment.

5 Results

The entire framework was evaluated against the final ground truth test set of 150 entities, split evenly between hobbies and organizations. The results are summarized in Table 4, demonstrating that human annotators clearly outperform the LLM (72.7% vs. 78.6% F-score), however the difference is not even across the four annotation questions. From the 20 automatically-optimized prompts, only one out-

Method	Schema	Q1	Q2	Q3	Q4	Avg
Orig. prompt	Full	77.0	85.4	73.6	54.6	72.7
Best prompt	Full	78.2	85.8	71.7	55.9	72.9
7 prompt ens.	Coarse	80.0	87.7	76.8	62.7	76.8
7× orig. ens.	Coarse	76.9	87.4	79.9	60.0	76.1
7 prompt ens.	Full	80.0	85.0	71.4	58.9	73.7
7× orig. ens.	Full	76.9	85.9	75.3	55.0	73.3
Human	Full	82.5	88.6	73.8	69.3	78.6
Human	Coarse	82.5	90.1	81.2	73.4	81.8

Table 4: Performance of single-prompt methods, ensemble methods, and human annotators on test set (n=150) in full and simplified coarse label schema. All values are F-score.

performed the original human-engineered prompt, achieving 72.9% F-score compared to 72.7% F-score for the original, a marginal gain of 0.2 percentage points.

We further evaluated two ensemble approaches: a majority vote across the 7 optimized prompts which have previously shown an improvement on the evaluation set, and for comparison a majority vote across 7 runs of the original prompt (leveraging the fact that the LLMs sample in generation, and the output therefore is not deterministic). As seen in Table 4, the 7-prompt ensemble achieved 73.7% F-score (93.8% of human baseline), compared to 73.3% for the repeated original prompt. Simply running the same prompt multiple times showed minor improvement (73.3% vs 72.7%), suggesting that ensemble voting can partially compensate for the stochastic nature of LLM outputs. The most notable improvement came on Q1 (Categories), where the optimized ensemble achieved 80.0% F-score versus the repeated original’s 76.9%. However, the optimized ensemble performed worse on Q3 (Frequency), suggesting that different prompts may have complementary strengths and weaknesses.

5.1 Coarse Label Schema: Improving Performance Through Reduced Granularity

Considering that the downstream use of the labeled data requires a maximal attainable accuracy, we also experimented with reduced label granularity, producing a coarse schema by treating the following labels as equivalent: **Q2**: “Small group” = “Large group”; **Q3**: “Occasional” = “Event-based”; **Q4**: “Intense” = “Continuous” = “Light”. This simplification essentially reduces Q2 to the choice of “solitary vs social,” Q3 to the choice of “regular vs rare,” and Q4 to the choice “active vs stationary,”

while keeping Q1’s full categorical granularity.

With these coarse labels, the optimized ensemble achieved 76.8% F1 score, representing 93.9% of human baseline. This represents a substantial 3.1 percentage point improvement over the same ensemble with full label granularity (73.7%). The gains were most pronounced in Q2 (Group Size), where collapsing “Small group” and “Large group” into a single “Social” category improved performance from 85.0% to 87.7%.

5.2 Processing the Complete Dataset

Based on the results outlined above, the full dataset of 60,647 entities was processed with the 7-prompt ensemble method, resulting in total of 424,529 LLM queries. The run was carried out on a cluster system using the vLLM library for efficient LLM generation. The system achieved 2219.38 tokens/s input and 54.05 tokens/s output. Despite identical formatting instructions, the prompts varied in reliability: the best required only 1.1 attempts per entity on average, while the most problematic needed 1.9 attempts (maximum 5 attempts per entity). Most failures stemmed from malformed JSON or empty responses; others involved incorrect category choices or hallucinated words. After majority voting across all 7 prompts, we obtained the final LLM annotation of all entities.

For Q1 (Categories), the most common labels were Administrative (19.5%, n=11,844), Professional/Work-related (17.9%, n=10,830), and General Social Group (12.8%, n=7,744), reflecting the formal organizational structure of post-war Finnish society. Multi-label classification proved essential for Q1: 29.2% of entities (17,717) received multiple category labels, with all 18 categories appearing in multi-label combinations (Table 5). Administrative entities most frequently combined with Professional/Work-related (1,654 cases), Educational/Academic (1,583), and Social Welfare (1,072). In contrast, Q2–Q4 showed minimal multi-label usage (1.8–6.6%), validating our design where Q1 captures overlapping domains while Q2–Q4 capture more singular dimensions. The full distribution across all labels is listed in Appendix A.4.

5.2.1 Mention-Weighted Analysis

All results so far were reported for unique entities. Weighting by mention frequency of each unique entity reveals distinct patterns between hobbies and organizations (full details are in Appendix A.5).

Table 5: Multi-label entity distribution across questions.

Question	Multi-label %	Total	Labels
Q1 – Category	29.2%	17,717	18
Q2 – Group size	6.6%	4,014	3
Q3 – Frequency	1.8%	1,076	4
Q4 – Physical act.	2.5%	1,511	4

Hobbies concentrate heavily in Creative/Artistic (36.3%) and Nature-related (32.9%) categories, while organizations distribute across General Social Group (19.0%), Administrative (12.6%), and Professional/Work-related (12.3%). These choices represent the final dataset and label distribution across the data. Group size (Q2) shows an interesting contrast whereby 96.1% of organizational mentions involve large groups, while hobbies split between solitary (43.4%) and small group (40.7%) activities. For frequency (Q3), 79.5% of organizational mentions are occasional, reflecting periodic meetings, whereas 66.0% of hobby mentions are continuous activities. Physical activity (Q4) distributions also diverge: 68.2% of organizations are stationary versus only 22.4% of hobbies, with most hobbies involving light (55.5%) or intense (10.0%) physical activity. These patterns validate our multi-dimensional approach and provide quantitative characterization of social participation patterns in the Karelian refugee population suitable for downstream health outcomes research.

5.3 Dimension Reliability by Category (Q1)

Any downstream analysis will need to take into account the reliability of both human and LLM predictions. For instance, Q4 (Physical Activity) performed clearly worst on the test set (Table 4). To investigate whether the prediction reliability depends on the entity’s category type, we analyzed Q2–Q4 reliability conditional on the answer to Q1. This addresses a question for downstream research: given that an entity is classified as a certain category (Q1), how reliable are answers to the other dimensions Q2–Q4? The statistics are shown in Table 6.

For instance: Sports/Physical activity entities (n=24), the model correctly identified the Q1 category 82% of the time. Among these sports-related entities, Q2 (Group Size) achieved 81% accuracy, Q3 (Frequency) 72%, and Q4 (Physical Activity) 76%. In contrast Creative/Artistic entities (n=33) achieved 93% Q1 accuracy but only 51% on Q4, reflecting difficulty distinguishing stationary activities from movement-based ones.

Q1 Category	n	Q1	Q2	Q3	Q4
Creative/Artistic	33	93/95	85/91	90/94	51/87
Professional	27	77/88	91/96	70/81	63/77
Sports/Physical	24	82/93	81/89	72/85	76/90
Administrative	21	85/88	95/96	76/79	84/83
Nature-related	19	79/92	75/86	60/71	58/72
General social	12	75/88	92/97	47/82	0/65
Educational	10	67/86	90/96	73/82	90/95
Property/Own.	10	67/69	73/82	63/85	80/72
Animal-related	7	83/94	79/86	81/82	71/73
Special hobbies	7	50/84	90/91	90/87	67/76
Military-related	6	91/91	83/92	44/78	0/79
Religious/Spirit.	6	80/92	86/99	67/87	67/78
Social welfare	5	59/67	100/100	40/76	20/78

Table 6: Q1–Q4 reliability conditioned by entity category in Q1, showing both model and human agreement rates. Format: Model% / Human%. Colors: $\geq 75\%$, 50–74%, $< 50\%$ (based on lower score).

Perfect agreement (0%)	19%
Low disag. (<25%)	44%
Moderate disag. (25–50%)	30%
High disag. (>50%)	7%

Table 7: Degrees of model–human disagreement.

6 Error analysis

As shown in Table 7, 63% of entities show perfect agreement or low disagreement between the model and human annotators. The full listing of entities and their human and model agreement levels is available in the Appendix A.6. We carry out a manual inspection of several main classes of entities by agreement:

Perfect agreement: Organizations with clear categories, group sizes, and physical activity levels are straightforward cases for the model. Examples: Choir (*Laulukuoro*), Bandy (*jääpallo*), Basketball and Volleyball (*kori- ja lentopalloilu*), Biathlon (*ampumahiihto*).

Model blind spots: Entities showing a high model disagreement despite high degree of human agreement. Typical cases are abbreviations without explicit context, and culturally-specific items such as rare Finnish folk dances. These are often known to humans familiar with Finnish culture and history, but unknown to general open-weight LLMs. Another typical case is the model’s overconfidence, making a guess, where humans resort to the “Cannot be determined” label. Typical examples are e.g. Seniors’ circle (*vanhusten piiri*), and Lotta Svärd

activities (*lottatoiminta*).

Genuine Ambiguity: Entities with high both model and human disagreement include e.g. Motor sports (*moottoriurheilu*), where the human annotators did not agree on frequency, group size, nor degree of physical activity. Another typical class are abbreviations, such as the military regiment abbreviation (*2.KKK/KKR*) and entities with OCR errors, where some annotators attempted disambiguation, while others marked a “Data error”.

7 Conclusions

We demonstrate that LLMs can implement social organization and leisure activity classification according to a custom taxonomy developed by domain experts for a specific analytic task. The models achieve 76.8% F-score, which is 93.9% of human baseline performance on coarse labels, and 93.8% on full granularity. This includes the added complexity of a relatively small language (Finnish) and the need of multi-label classification, where we see the LLMs are capable of assigning multiple labels, allowing for a structured multi-question schema, where some questions (e.g. Q1) are inherently multi-label.

The performance of the LLM at nearly 94% of human annotation provides a worthwhile alternative to manual classification of tens of thousands of entities, especially if any downstream analysis can be designed to take into account the levels of uncertainty necessarily present in such a categorization task.

To this end, we find that analyzing and quantifying LLM classification errors is crucial, since these errors are rarely random. They follow consistent patterns: some reflect genuine failures of model understanding, while others represent interpretive differences where the model’s reasoning diverges from human consensus rather than being outright wrong. Characterizing these error types, identifying which options are systematically misclassified, and providing reliability measures for specific categories is essential for downstream use of this data.

All data will be made available upon publication under an open license.

8 Acknowledgments

This work was carried out in the Human Diversity University profilation programme (PROFI-7) of the Research Council of Finland (grant 352727), and in part supported also through the KinSocieties,

ERC-2022-ADG grant number 101098266. Computational resources were provided by CSC — the Finnish IT Center for Science.

9 Limitations

LLMs show consistent failure patterns, including uncertainty avoidance and cultural knowledge gaps. Rather than selecting “Cannot be determined,” the models tend to guess concrete answers even when annotators acknowledge ambiguity. Similarly, Finnish-specific terminology and regional abbreviations lead to systematic misclassification.

Additionally developing a custom taxonomy requires substantial annotation effort. Our annotation framework involved five rounds of refinement with four annotators before achieving stable guidelines. Comparing this to established schemes like NER, where pretrained models and labeled datasets exist, this represents a significant overhead. Furthermore, if reliability analysis reveals poor alignment for specific categories, additional work is required, either more annotations to improve agreement, or refining the categories themselves. However while the framework is designed to address the specific research questions related to our data and may not be directly transferable to others, we believe that many of its categories are of universal interest (e.g. in health research), and that the framework can serve as a reference to be adapted for other works describing social participation.

References

- Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. [Don’t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. [Enriching word vectors with subword information](#). *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Helga Hansdóttir, María K. Jonsdóttir, Diana E. Fisher, Gudny Eiríksdóttir, Pálmi V. Jonsson, and Vilundur Gudnason. 2022. [Creativity, leisure activities, social engagement and cognitive impairment: the AGES-Reykjavík study](#). *Aging Clinical and Experimental Research*, 34:1027–1035.
- Julianne Holt-Lunstad, Timothy B. Smith, and J. Bradley Layton. 2010. [Social relationships and mortality risk: A meta-analytic review](#). *PLoS Medicine*, 7(7):e1000316.
- Chen Huang and Guoxiu He. 2025. [Text clustering as classification with LLMs](#). *Preprint*, arXiv:2410.00927.
- Qianyu Huang and Tongfang Zhao. 2024. [Leveraging large language models for entity matching](#). *Preprint*, arXiv:2405.20624.
- Yoshi Iwasaki, Jiri Zuzanek, and Roger C. Mannell. 2001. [The effects of physically active leisure on stress-health relationships](#). *Canadian Journal of Public Health*, 92(3):214–218.
- Joonatan Laato, Jenna Kanerva, John Loehr, Virpi Lummaa, and Filip Ginter. 2024. [Extracting social connections from Finnish Karelian refugee interviews using LLMs](#). In *Proceedings of Fifth Conference on Computational Humanities Research (CHR’2024)*.
- Mengfei Lan, Lecheng Zheng, Shufan Ming, and Halil Kilicoglu. 2024. [Multi-label sequential sentence classification via large language model](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16086–16104, Miami, Florida, USA. Association for Computational Linguistics.
- John Loehr, Robert Lynch, Johanna Mappes, Tuomas Salmi, Jenni Pettay, and Virpi Lummaa. 2017. [Newly digitized database reveals the lives and families of forced migrants from Finnish Karelia](#). *Nordic Yearbook of Population Research*, 52:59–70.
- Marcus Ma, Georgios Chochlakakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. 2025. [Large language models do multi-label classification differently](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2472–2495, Suzhou, China. Association for Computational Linguistics.

- Laura McQuade and Roger O’Sullivan. 2024. [Examining arts and creativity in later life and its impact on older people’s health and wellbeing: a systematic review of the evidence](#). *Perspectives in Public Health*, 144(6):344–353.
- Meta-AI. 2025. Llama 4 Scout 17B–16E Model Card. <https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E>.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR’13*.
- Mistral-AI. 2024. Mistral-Large-Instruct-2407 Model Card. <https://huggingface.co/mistralai/Mistral-Large-Instruct-2407>.
- Ryan E. Rhodes, Ian Janssen, Shannon S. D. Bredin, Darren E. R. Warburton, and Adrian Bauman. 2017. [Physical activity: Health impact, prevalence, correlates and interventions](#). *Psychology & Health*, 32(8):942–975.
- Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. [Get another label? improving data quality and data mining using multiple, noisy labelers](#). In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’08, page 614–622, New York, NY, USA. Association for Computing Machinery.
- Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Ng. 2008. [Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks](#). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.
- Seyed Amin Tabatabaei, Sarah Fancher, Michael Parsons, and Arian Askari. 2025. [Can large language models serve as effective classifiers for hierarchical multi-label classification of scientific documents at industrial scale?](#) In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 163–174, Abu Dhabi, UAE. Association for Computational Linguistics.
- Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2024. [Large language models enable few-shot clustering](#). *Transactions of the Association for Computational Linguistics*, 12:321–333.
- Dorota Węziak-Białowolska, Piotr Białowolski, and Pier Luigi Sacco. 2023. [Mind-stimulating leisure activities: Prospective associations with health, wellbeing, and longevity](#). *Frontiers in Public Health*, 11:1117822.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025b. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. [Large language models as optimizers](#). In *Proceedings of The Twelfth International Conference on Learning Representations*.
- Lauren A. Yates, Shier Ziser, Aimee Spector, and Martin Orrell. 2016. [Cognitive leisure activities and future risk of cognitive impairment and dementia: systematic review and meta-analysis](#). *International Psychogeriatrics*, 28(11):1791–1806.
- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. [Large language models are human-level prompt engineers](#). In *Proceedings of The Eleventh International Conference on Learning Representations*.

A Appendix

A.1 Task prompt

TASK

Classify the entity given below by answering four questions (q1-q4).
For each question, choose the option(s) that **most accurately** match the entity, following the
↪ instructions provided.
Write your answers exactly according to the **CORRECT ANSWER FORMAT** structure given below.

QUESTIONS AND OPTIONS

Always clearly separate the option and its description, e.g.
`Professional/Work-related - Trade unions, professional associations, workplace-related groups. . .`

Question 1 (q1) - Categories (select **all** that apply)

Option	Description
Professional/Work-related	Trade unions, professional associations, workplace-related groups;
Cultural/Traditional	Cultural heritage, local culture, traditions (e.g. Karelian Society)
Religious/Spiritual	Church activities, religious education, spiritual associations
Sports/Physical activity	All kinds of sports, exercise, team sports and individual sports
Creative/Artistic	Music, visual arts, handicrafts, theatre, dance, etc.
Educational/Academic	Studying, education, scholarly associations, study circles
Social welfare	Charity work, volunteering, community service
Administrative	Public administration roles and bodies (municipality, church, state), ↪ non-political
Political	Parties, party organizations, political advocacy
General social group	Martta associations, youth clubs, general-purpose community groups
Nature-related	Hunting, fishing, hiking, nature conservation
Health-related	Peer-support groups, support for people with illnesses (e.g. Rheumatism ↪ Association)
Property/Ownership	Road maintenance associations, housing companies, detached-house maintenance ↪ associations
Non-physical games	Chess, board games, role-playing and video games
Animal care/Hobby	Beekeeping, horse riding, dog shows
Special hobbies	Stamp collecting, radio technology, etc.
Cooking	Food preparation, baking
Military-related	National defence, reservist activities
Not definable	Impossible to determine from name/context
Data error	Entity is not an organization/hobby OR is a data error

Question 2 (q2) - Group size (select the most typical one(s))

Option	Description
Alone	1 person; the activity can be done completely independently
Small group	2-5 people; close cooperation between participants
Large group	> 6 people; requires broader organization and structures
Not definable	Group size varies / cannot be inferred
Data error	Data issue prevents assessment

Question 3 (q3) - Frequency of activity (select the most likely one(s))

Option	Description
Regular	Fixed schedule, e.g. weekly
Occasional	About once a month, not fully fixed
Event-based	Once a year or less often, in connection with an event

```
| Continuous | Ongoing activity without fixed schedule; can be done any time |
| Not definable | Frequency is unclear |
| Data error | Data issue prevents assessment |
```

Question 4 (q4) – Level of movement / Physical activity (select the most likely one(s))

```
| Option | Description |
|-----|-----|
| Intense | Heavy sports, intensive muscular work |
| Continuous | Long-lasting steady movement (e.g. hiking) |
| Light | Occasional light movement, light activity/puttering |
| Stationary | Minimal movement, sitting/standing (meetings, administration) |
| Not definable | Physical activity level cannot be determined |
| Data error | Data issue prevents assessment |
```

ANNOTATION GUIDELINES

General principles

- Accuracy > coverage → if uncertain, choose **“Not definable”**.
- If the entity is **not** an organization/hobby, use **“Data error”** for all questions.
- Base decisions on the available information; avoid guessing.
- If provided, take into account the hierarchy in which the organization is mentioned. If it is empty → or UNK, you can ignore it.

Question-specific notes

q1: Select at least one category; some entities belong to multiple categories.
* Administrative is not Political: administrative = official governing bodies, political = parties.

q2: Choose the most typical group size; meetings/organizations → often **Large group**.
* If the estimate is difficult → **Not definable**.

q3:
* **Continuous** = independent, no group required.
* **Regular** = active weekly/monthly meetings.
* **Occasional** = committees & boards (~ once/month).
* **Event-based** = rarer, 1-2 times/year.
* Unclear → **Not definable**.

q4:
* Administrative body → **Stationary**; sports club → **Intense**.
* Board of a sports club → **Stationary**.
* Unclear → **Not definable**.

EXAMPLES

```
|-----|----|----|----|----|
| football | Sports/Physical activity | Large group | Regular | Intense |
| rheumatism association | Health-related | Large group | Regular | Stationary |
| mushroom picking | Nature-related | Alone, Small group | Continuous | Light |
| social welfare committee | Administrative, Social welfare | Large group | Occasional | Stationary |
| Guards band | Military-related, Creative/Artistic | Large group | Regular | Stationary |
| Järvenpää youth board | Administrative | Large group | Occasional | Stationary |
| housekeeping | Property/Ownership | Alone, Small group | Continuous | Light |
| local party branch | Political | Large group | Occasional | Stationary |
```

CORRECT ANSWER FORMAT

[Answer begin]

```
{{
  "{{entity_name}}": {{
    "q1": [...],
    "q2": [...],
    "q3": [...],
    "q4": [...]
  }}
}}
```

```

}}
[Answer end]

#### Entity to be annotated

"{entity_name}"

Hierarchy: "{hierarchies}"

"{past_mistakes}"

```

A.2 Inference Hyperparameters

Inference was performed on the LUMI supercomputer using vLLM as the inference engine with the following sampling parameters: temperature=0.3, top_p=1.0, top_k=40, max_tokens=300. Processing used batch size of 150.

A.3 Inter-annotator agreement details

Annotator Pair	Q1	Q2	Q3	Q4	Average
Pairwise Cohen's Kappa – Evaluation Set (n=50)					
Annotator 1 vs 2	0.800	0.806	0.667	0.721	0.749
Annotator 1 vs 3	0.753	0.743	0.725	0.711	0.733
Annotator 1 vs 4	0.753	0.767	0.524	0.660	0.676
Annotator 2 vs 3	0.684	0.753	0.677	0.688	0.701
Annotator 2 vs 4	0.697	0.848	0.651	0.754	0.737
Annotator 3 vs 4	0.683	0.714	0.571	0.535	0.626
Mean	0.728	0.772	0.636	0.678	0.704
Pairwise Cohen's Kappa – Test Set (n=150)					
Annotator 1 vs 2	0.783	0.825	0.642	0.620	0.718
Annotator 1 vs 3	0.824	0.790	0.698	0.528	0.710
Annotator 1 vs 4	0.778	0.776	0.581	0.723	0.714
Annotator 2 vs 3	0.737	0.751	0.636	0.507	0.658
Annotator 2 vs 4	0.712	0.812	0.649	0.680	0.713
Annotator 3 vs 4	0.767	0.732	0.602	0.415	0.629
Mean	0.767	0.781	0.634	0.579	0.690
Individual F1 vs Leave-One-Out Majority – Evaluation Set (n=50)					
Annotator 1	0.867	0.873	0.813	0.807	0.840
Annotator 2	0.793	0.907	0.833	0.793	0.832
Annotator 3	0.756	0.813	0.833	0.733	0.784
Annotator 4	0.793	0.880	0.710	0.713	0.774
Mean	0.802	0.868	0.797	0.762	0.807
Individual F1 vs Leave-One-Out Majority – Test Set (n=150)					
Annotator 1	0.865	0.898	0.749	0.744	0.814
Annotator 2	0.800	0.902	0.756	0.722	0.795
Annotator 3	0.830	0.869	0.753	0.580	0.758
Annotator 4	0.805	0.876	0.696	0.724	0.775
Mean	0.825	0.886	0.738	0.693	0.786

Note: Krippendorff's Alpha values are nearly identical to Cohen's Kappa (within 0.001) and omitted for brevity.

Table 8: Inter-annotator agreement metrics across evaluation (n=50) and test (n=150) sets.

A.4 Full distribution of unique entities

Table 9: Distribution across 60,647 unique entities for all four questions.

Q1 – Category			Q2 – Group Size		
Label	n	%	Label	n	%
Administrative	11,844	19.5	Large group	53,574	88.3
Professional	10,830	17.9	Small group	4,889	8.1
General Social	7,744	12.8	Alone	4,840	8.0
Creative	7,396	12.2	Data error	853	1.4
Sports	7,328	12.1	Not definable	654	1.1
Social Welfare	5,760	9.5			
Educational	4,921	8.1			
Cultural	3,846	6.3			
Military	3,798	6.3			
Religious	3,166	5.2			
Property	2,696	4.4			
Nature	2,590	4.3			
Political	2,034	3.4			
Health	1,930	3.2			
Special hobbies	958	1.6			
Animal care	886	1.5			
Data error	853	1.4			
Not definable	283	0.5			
Cooking	176	0.3			
Non-phys. games	175	0.3			
Q3 – Frequency			Q4 – Physical		
Label	n	%	Label	n	%
Occasional	42,999	70.9	Stationary	41,103	67.8
Regular	11,800	19.5	Light	9,902	16.3
Continuous	5,092	8.4	Intense	6,291	10.4
Data error	853	1.4	Not definable	2,567	4.2
Event-based	555	0.9	Continuous	1,542	2.5
Not definable	490	0.8	Data error	853	1.4

A.5 Full distribution of entity occurrences

Label	Hobbies		Organizations		Total	
	Count	%	Count	%	Count	%
Q2 – Group Size						
Large group	55,252	12.9	198,318	96.1	253,570	39.9
Alone	186,679	43.4	512	0.2	187,191	29.4
Small group	174,877	40.7	4,104	2.0	178,981	28.1
Not definable	12,232	2.8	1,140	0.6	13,372	2.1
Data error	687	0.2	2,387	1.2	3,074	0.5
Q3 – Frequency						
Occasional	43,018	16.0	163,126	79.5	206,144	43.4
Continuous	177,956	66.0	1,804	0.9	179,760	37.8
Regular	41,123	15.2	36,439	17.8	77,562	16.3
Not definable	6,025	2.2	1,060	0.5	7,085	1.5
Data error	687	0.3	2,387	1.2	3,074	0.6
Event-based	865	0.3	448	0.2	1,313	0.3
Q4 – Physical Activity						
Stationary	59,508	22.4	145,285	68.2	204,793	42.7
Light	147,726	55.5	31,210	14.6	178,936	37.3
Intense	26,605	10.0	16,138	7.6	42,743	8.9
Not definable	11,513	4.3	14,497	6.8	26,010	5.4
Continuous	19,898	7.5	3,663	1.7	23,561	4.9
Data error	687	0.3	2,387	1.1	3,074	0.6

Table 10: Mention-weighted distribution (355,648 total mentions) for Q2–Q4 by entity type.

Category	Hobbies		Organizations		Total	
	Count	%	Count	%	Count	%
Creative/Artistic	97,180	36.3	16,639	6.3	113,819	21.4
Nature-related	88,174	32.9	3,880	1.5	92,054	17.3
Sports/Physical	46,269	17.3	15,390	5.8	61,659	11.6
General Social Group	1,621	0.6	50,089	19.0	51,710	9.7
Professional/Work	1,811	0.7	32,344	12.3	34,155	6.4
Administrative	235	0.1	33,301	12.6	33,536	6.3
Cultural/Traditional	565	0.2	31,685	12.0	32,250	6.1
Educational/Academic	14,372	5.4	12,112	4.6	26,484	5.0
Social Welfare	1,055	0.4	21,406	8.1	22,461	4.2
Religious/Spiritual	2,678	1.0	12,440	4.7	15,118	2.8
Military	433	0.2	13,858	5.3	14,291	2.7
Property/Ownership	3,931	1.5	5,617	2.1	9,548	1.8
Health-related	136	0.1	5,482	2.1	5,618	1.1
Animal care/hobby	3,689	1.4	967	0.4	4,656	0.9
Political	220	0.1	4,216	1.6	4,436	0.8
Special hobbies	3,239	1.2	799	0.3	4,038	0.8
Data error	687	0.3	2,387	0.9	3,074	0.6
Not definable	278	0.1	918	0.3	1,196	0.2
Non-physical games	880	0.3	211	0.1	1,091	0.2
Cooking	565	0.2	28	0.0	593	0.1

Table 11: Q1 category distribution weighted by mention frequency (355,648 total mentions).

A.6 Full test set categorization

Rank	Entity	Model Disagr	Human Disagr
1	Loimaa Co-op Bank (<i>Loimaan KOP</i>)	100.0%	66.7%
2	Folk dance (<i>tanhu</i>)	100.0%	50.7%
3	Poor relief fund (<i>Vaivaiskassa</i>)	100.0%	66.7%
4	Military regiment abbr. (2.KKK/KKR)	91.7%	75.0%
5	Association abbr. (<i>KTV</i>)	87.5%	25.0%
6	Summer cottage life (<i>kesämökkielämä</i>)	79.2%	56.2%
7	Sea excursions (<i>retkeily merellä</i>)	79.2%	54.2%
8	Boating sports (<i>venueurheilu</i>)	79.2%	65.3%
9	Nokia SOK evening club (<i>Nokian SOK:n ilkkakerho</i>)	75.0%	50.0%
10	Council (<i>valtuusto</i>)	75.0%	68.8%
11	Seniors' circle (<i>vanhusten piiri</i>)	75.0%	18.8%
12	Self-employed entrepreneurs (<i>Yksityisyrittäjät</i>)	75.0%	20.8%
13	Ylöjärvi TPSL branch (<i>Ylöjärven TPSL</i>)	75.0%	27.8%
14	Staying at summer cottage (<i>kesämökillä oleskelu</i>)	66.7%	52.1%
15	Summer cottage stays (<i>kesämökkeily</i>)	66.7%	52.1%
16	Motor sports (<i>moottorieurheilu</i>)	66.7%	75.0%
17	Camping/tents (<i>telttailu</i>)	66.7%	25.0%
18	Karkkula Farmers' Association (<i>Karkkulan Maamiesseura</i>)	62.5%	31.9%
19	Bird banding (<i>lintujen merkitsijänä toimiminen</i>)	62.5%	41.7%
20	Martha tasks (<i>martha-työt</i>)	62.5%	29.2%
21	Martha organization work (<i>marttatyö</i>)	62.5%	29.2%
22	Attending sewing circles (<i>ompeluseuroissa käyminen</i>)	54.2%	31.2%
23	War veterans (<i>Sotaveteraanit</i>)	54.2%	27.1%
24	Hauho Farmers' Association (<i>Hauhon Maamiesseura</i>)	50.0%	29.9%
25	Hämeenlinna Camera Club (<i>Hämeenlinnan Kameraseura</i>)	50.0%	29.2%
26	Dev. disabilities support (<i>Kehitysvammaisten tukiyhdistys</i>)	50.0%	29.9%
27	Lotta Svärd organization (<i>Lotta-Svärd-järjestö</i>)	50.0%	20.8%
28	Lotta Svärd activities (<i>lottatoiminta</i>)	50.0%	20.8%
29	Martha activities (<i>marttatoiminta</i>)	50.0%	25.0%
30	Travels (<i>matkusteleee</i>)	50.0%	66.7%
31	Opera (<i>ooppera</i>)	50.0%	56.2%
32	Local horse association (<i>paikallinen hevosyhdistys</i>)	50.0%	52.8%
33	Gliding (<i>purjelento</i>)	50.0%	64.6%
34	Chess (<i>shakki</i>)	50.0%	20.8%
35	Betting (<i>veikkaus</i>)	50.0%	58.3%
36	Diaconia Committee (<i>Diakoniatuomikunta</i>)	45.8%	34.7%
37	Workers' association (<i>työväenyhdistys</i>)	45.8%	41.7%
38	Church attendance (<i>kirkossakäyminen</i>)	41.7%	0.0%
39	East Karelia Folk College (<i>Itä-Karjalan kansanopisto</i>)	37.5%	33.3%
40	Making canal boards (<i>kanavataulujen teko</i>)	37.5%	0.0%
41	Middle-distance running (<i>keskipitkien matkojen juoksu</i>)	37.5%	0.0%
42	Fitness gymnastics (<i>kuntovoimistelu</i>)	37.5%	33.3%
43	Handicrafts (<i>käsityö</i>)	37.5%	6.2%
44	Agricultural Producers (<i>Maataloustuottajat</i>)	37.5%	27.1%
45	Martha Association (<i>Martta-yhdistys</i>)	37.5%	27.1%
46	Mosaic work (<i>mosaiikkityöt</i>)	37.5%	6.2%
47	Archaeological artifacts (<i>muinaistieteellisten esineiden kerääminen</i>)	37.5%	45.8%
48	Drawing (<i>piirustus</i>)	37.5%	0.0%
49	Ice fishing (<i>piikkionginta</i>)	37.5%	16.7%
50	Rotary Club activity (<i>rotarytoiminta</i>)	37.5%	35.4%
51	Selkämeri NCO Women (<i>Selkämeren Alipäällystönaiset ry</i>)	37.5%	25.0%
52	Playing chess (<i>shakinpeluu</i>)	37.5%	25.0%
53	Folk dances (<i>tanhut</i>)	37.5%	45.8%
54	Täkänä weaving (<i>täkänäin kudonta</i>)	37.5%	18.8%
55	Viipuri Guild (<i>Viipurin Kilta ry</i>)	37.5%	25.0%
56	Visuvesi Workers' Assoc. (<i>Visuvesen Työväenyhdistys</i>)	37.5%	50.7%
57	Horse breeding (<i>hevoskasvatus</i>)	29.2%	29.2%
58	Baltic Sea Summer (<i>Itämeren Kesä</i>)	29.2%	36.8%
59	Agrarian League (<i>Maalaisliitto</i>)	29.2%	36.8%
60	Professional Drivers' Union (<i>Ammattiautoilijaliitto</i>)	25.0%	27.1%
61	Balalaika playing (<i>balalaikan soittaminen</i>)	25.0%	0.0%
62	Wilderness Scouts (<i>Eränkävijät</i>)	25.0%	50.0%
63	Heinola Rural Road Board (<i>Heinolan mlk:n tienhoitokunta</i>)	25.0%	33.3%
64	Skiing (<i>Hiihto</i>)	25.0%	12.5%
65	Janakkala Society (<i>Janakkala-seura</i>)	25.0%	47.9%
66	Fabric weaving (<i>Jcankaankutominen</i>)	25.0%	75.0%
67	Cabbage cultivation (<i>kaalinviljely</i>)	25.0%	27.1%
68	Property Inspection Board (<i>Kiinteistö tarkastuslautakunta</i>)	25.0%	41.7%
69	Domestic chores (<i>kodintyöt</i>)	25.0%	41.7%
70	Making 3D pictures (<i>kolmiulotteisten taulujen tekeminen</i>)	25.0%	12.5%
71	Typing (<i>konekirjoitus</i>)	25.0%	12.5%
72	Household chores (<i>kotitaloustyöt</i>)	25.0%	20.8%
73	Jogging (<i>lenkkeily</i>)	25.0%	6.2%
74	Salmon farming (<i>lohen kasvatus</i>)	25.0%	62.5%
75	Casual reading (<i>lueskeleminen</i>)	25.0%	0.0%
76	Farmers' Assoc. sewing (<i>Maamiesseuran ompeluseurassa käynti</i>)	25.0%	41.0%
77	Rag rug weaving (<i>matonkutominen</i>)	25.0%	18.8%
78	Beekeeping (<i>mehiläisten hoito</i>)	25.0%	8.3%
79	Study circle (<i>Opintokerho</i>)	25.0%	27.1%
80	Lace crocheting (<i>pitsinvirkkaus</i>)	25.0%	12.5%
81	Norden Association (<i>Pohjola-Norden</i>)	25.0%	16.7%
82	Polio Disabled Association (<i>Polioinvalidit ry</i>)	25.0%	12.5%
83	Women of Foremen's Assoc. (<i>Rakennusmestariyhdistyksen Naisjaosto</i>)	25.0%	22.9%

Rank	Entity	Model Disagr	Human Disagr
84	Race walking (<i>ratakävely</i>)	25.0%	45.8%
85	Game management (<i>riistanhoito</i>)	25.0%	47.9%
86	Making ryijy rugs (<i>ryijyjen tekeminen</i>)	25.0%	22.9%
87	Rambling/hiking (<i>samoilu</i>)	25.0%	31.2%
88	Knitting socks (<i>sukkien neulominen</i>)	25.0%	0.0%
89	Finnish Real Estate Federation (<i>Suomen Kiinteistöliitto</i>)	25.0%	41.7%
90	Finland–Soviet Union Society (<i>Suomi-Neuvostoliitto-Seura</i>)	25.0%	54.2%
91	Studying stars/astromony (<i>tutkii tähtiä</i>)	25.0%	18.8%
92	Workers' drama club (<i>Työväenyhdistyksen näytelmäkerho</i>)	25.0%	18.8%
93	Vaasa Reserve Officers' Club (<i>Vaasan Reserviupseerikerho</i>)	25.0%	27.1%
94	Sewing clothes (<i>vaatteiden ompelu</i>)	25.0%	6.2%
95	Carving course (<i>veistokurssi</i>)	25.0%	43.8%
96	Viipuri Workers' Institute (<i>Viipurin Työväenopisto</i>)	25.0%	37.5%
97	Violin playing (<i>viulun soitto</i>)	25.0%	12.5%
98	Imatra Region Co-op Bank (<i>Imatran seudun Osuuspankki</i>)	16.7%	34.0%
99	Agricultural Club Association (<i>Maatalouskerhoyhdistys</i>)	16.7%	38.2%
100	Furniture making (<i>huonekalujen valmistus</i>)	12.5%	12.5%
101	Poultry keeping (<i>kananhoito</i>)	12.5%	16.7%
102	Karelia Athletes (<i>Karjalan Urheilijat</i>)	12.5%	8.3%
103	Javelin throw (<i>keihäänheitto</i>)	12.5%	0.0%
104	Home gardening (<i>kotipuutarhanhoito</i>)	12.5%	27.1%
105	Kuusaa Horticultural Assoc. (<i>Kuusaa Puutarhayhdistys</i>)	12.5%	40.3%
106	Ship Officers' Association (<i>Laivanpäällystöliitto</i>)	12.5%	33.3%
107	Lappeenrannan evankelisluterilaisen seurakunta	12.5%	33.3%
108	Land Acquisition Board (<i>maanlunastuslautakunta</i>)	12.5%	39.6%
109	Seamen's Mission (<i>merimieslähetys</i>)	12.5%	49.7%
110	Metalworkers' Union (<i>Metallityöväen Liitto</i>)	12.5%	20.8%
111	Bankers' Union (<i>Pankkimiesliitto ry</i>)	12.5%	33.3%
112	Paper Workers' Guild (<i>Paperiammattikunta</i>)	12.5%	25.0%
113	Philatelic club (<i>Postimerkkikerho</i>)	12.5%	20.8%
114	Carpentry workshop (<i>puusepänerastas</i>)	12.5%	37.5%
115	Construction Workers' Union (<i>Rakennustyöväen Liitto</i>)	12.5%	27.1%
116	Railway Workers' Choir (<i>rautatieläisten kuoro</i>)	12.5%	0.0%
117	Railway Workers' Union (<i>Rautatieläisten liitto</i>)	12.5%	27.1%
118	Restaurant Staff Union (<i>Ravintolahenkilökunnan Liitto</i>)	12.5%	27.1%
119	Taxation Board (<i>taksoituslautakunta</i>)	12.5%	22.9%
120	Playing tennis (<i>tenniksen pelaaminen</i>)	12.5%	14.6%
121	VR Hyvinkää Men's Choir (<i>VR:n Hyvinkään mieskuoro</i>)	12.5%	13.9%
122	Biathlon (<i>ampumahiihto</i>)	0.0%	14.6%
123	Church Council of Haapavesi (<i>Haapaveden kirkkoneuvosto</i>)	0.0%	0.0%
124	Hämeen Tarmo Women Gymnasts (<i>Hämeen Tarmon Naisvoimistelijat</i>)	0.0%	6.2%
125	Imatra Orienteers (<i>Imatran Suunnistajat</i>)	0.0%	12.5%
126	Bandy (<i>jääpallo</i>)	0.0%	0.0%
127	Primary School Board (<i>Kansakoululautakunta</i>)	0.0%	20.1%
128	Primary school board (<i>kansakoululautakunta</i>)	0.0%	18.8%
129	Kemi Rural Labour Committee (<i>Kemin mlk:n työvoimatoimikunta</i>)	0.0%	25.0%
130	Basketball and Volleyball (<i>kori- ja lentopalloilu</i>)	0.0%	0.0%
131	Municipal Council (<i>Kunnanvaltuusto</i>)	0.0%	18.8%
132	Assoc. Municipal Officials (<i>Kuntayhtymän virkamiesyhdistys ry</i>)	0.0%	31.2%
133	Singing Masters (<i>Laulavat mestarit</i>)	0.0%	0.0%
134	Choir (<i>Laulukuoro</i>)	0.0%	0.0%
135	Lokalaiti Men's Choir (<i>Lokalahden Laulumiehet</i>)	0.0%	0.0%
136	Forestry (<i>metsänhoito</i>)	0.0%	31.2%
137	Men's choir (<i>mieslaulajiin</i>)	0.0%	0.0%
138	Women Gymnasts (<i>Naisvoimistelijat</i>)	0.0%	6.2%
139	Drama club activity (<i>näytelmäkerhotoiminta</i>)	0.0%	12.5%
140	Stage activities/theatre (<i>näyttämötoiminta</i>)	0.0%	12.5%
141	Ball sports (<i>palloilu</i>)	0.0%	12.5%
142	Pukkila Municipal Council (<i>Pukkilan kunnanvaltuusto</i>)	0.0%	12.5%
143	Radio listening (<i>radion kuuntelu</i>)	0.0%	0.0%
144	Mixed choir (<i>sekakuoro</i>)	0.0%	0.0%
145	Non-fiction literature (<i>tietopuolinen kirjallisuus</i>)	0.0%	6.2%
146	Workers' Athletes (<i>Työväen Urheilijat</i>)	0.0%	0.0%
147	Vakka Transport (<i>Vakka-Kuljetus</i>)	0.0%	54.2%
148	Boat building (<i>veneiden teko</i>)	0.0%	18.8%
149	Studying Russian (<i>venäjänkielen opiskelu</i>)	0.0%	22.9%
150	Viljakkala Sports Club (<i>Viljakkalan urheiluseura</i>)	0.0%	0.0%

Table 12: Model and human disagreement over the 150 test set entities.

Catalogues as Data: Interpretable NLP Pipelines for Ottoman-Turkish Bibliographies

Mark J. Hill, Ayse Bulus, and Paul Spence

King's College London

Strand, WC2R 2LS

London, UK

Correspondence: mark.j.hill@kcl.ac.uk

Abstract

Bibliographies are both humanities infrastructure and historic record. To computationally analyse them, however, requires implementing complex digitisation and standardisation decisions. This paper turns to Seyfettin Özege's *Eski Harflerle Basılmış Türkçe Eserler Kataloğu* as an example, a scanned set of volumes marked by complex page layouts, degraded typography, irregular entry structures, and historically contingent inconsistencies. With this we present a pipeline that constructs a structured, machine-readable, and analysable dataset out of the 27,000 entries with computer vision, OCR, large and visual language models, sequence-based validation, and custom review tools. This process captures 97.8% of records, with remaining cases capable of being addressed by targeted review. This process demonstrates that combining LLMs with interpretable, review-centric pipelines, offers an appropriate approach for historically complex bibliographic sources.

1 Introduction

Digitisation has increased access to historical bibliographic resources. However, while scanned catalogues are accessible, they do not easily support large-scale querying, aggregation, or complex analysis. This matters as, for many Digital Humanities (DH) researchers, a goal is to interpret historically situated bibliographic structures at scale.

This paper addresses these challenges through Mehmet Seyfettin Özege's *Eski Harflerle Basılmış Türkçe Eserler Kataloğu*, a five-volume bibliography of Ottoman-Turkish printed works (Özege, 1971–1979). Compiled by a single scholar over several decades, the catalogue is both a valuable reference and historical artefact, shaped by uneven access to sources and pragmatic bibliographic choices. These contingencies manifest in non-uniform entry structures which are interesting in

their own right, but pose difficulties for computational analysis.

In the following we present a computational pipeline designed to both clean or normalise the catalogue (making its structure computationally legible) while preserving its historical contingencies. To do this we combine computer vision, optical character recognition (OCR), and both Large and Vision Language Models (LLMs and VLMs). Throughout we foreground interpretability, provenance, and scholarly caution rather than automation through a process of extraction (layout handling, entry segmentation, OCR, initial parsing), validation (sequence checks, duplicate detection), and review (implemented through specialised interfaces and tracked operation logs). This framing treats uncertainty and variation as an output: not errors to eliminate, but signals that must be assessed by an expert and are analysable in their own right.

2 Related Work

This work sits within a “catalogues-as-data” orientation in DH, which treats bibliographic catalogues as large-scale datasets, themselves reflections of historical material and practice (Gooding et al., 2025). Examples of this approach can be seen in the work done by the Helsinki Computational History Research Group, who have developed pipelines for harmonising and enriching bibliographic resources, and linking them to full-text corpora. This work has demonstrated how Machine-Readable Cataloguing (MARC) can be transformed into relational datasets suitable for quantitative historical analysis (Hill et al., 2019; Lahti et al., 2019; Tolonen et al., 2021). They have also shown how careful cleaning and unification supports novel downstream historical work, ranging from social network analysis to studying the economics of the book trade (Hill et al., 2023; Tolonen et al., 2025; Tiihonen et al., 2024).

This paper differs, however. Rather than starting with MARC records it begins with a scanned bibliography, requiring a multistage pipeline to convert structurally complex facsimiles of pages into individual entries. That is, parsing the interaction of page layout, segmentation, OCR noise, and irregular entry structure without any pre-existing structural encoding comparable to MARC 21. To do this we turn to LLMs, arguing that in these contexts they enable structural interpretation not achievable through rule-based parsing alone.

We are not the first to use LLMs in DH work, with others arguing they can be integrated into mixed-methods workflows to scale qualitative analysis while preserving replicability and transparency (Karjus, 2025). With regard to bibliographic work, recent research has looked at using LLMs to augment existing catalogues (Korpet and Rees, 2025), using them to create structured MARC 21 records (Taniguchi, 2024; Aycok, 2025), and difficulties verifying LLM-generated bibliographic metadata (Kohút et al., 2025; Frenzel, 2025).

This work also sits within Digital Ottoman Studies, a field that in recent years has foregrounded work focused on quantitative and textual analysis of the Ottoman world (Barakat and Yayıoğlu, 2022). This includes the challenges born out of Ottoman Turkish as a low resource language (Karagöz et al., 2024). An overview of these issues can be found in Yüksek (2022), however central is the fact that Ottoman Turkish texts are inconsistently converted from Arabic script into the Latin alphabet. The same word can be spelled in different ways, which humans can recognise as equivalent, but computers mistake. Due to these language constraints, researchers have advocated for computational workflows in which the researcher is central, rather than fully automated solutions (Aladağ, 2020, 2021, 2024; Kırmızıaltın et al., 2022). Our pipeline extends this orientation to bibliographic data, applying similar principles of validation, traceability, and interpretive control.

3 Background: The Özege Catalogue

Özege’s catalogue documents Ottoman-Turkish books and pamphlets excluding newspapers and periodicals. It was conceived as a retrospective scholarly reconstruction rather than a contemporary publishing record, and is based (wherever possible) on direct examination of physical copies.

Özege rejected rigid adherence to bibliographic

standards when these conflicted with clarity or usability. As a result, entries vary considerably in completeness, ordering, and expression. For example, dates appear in multiple calendar systems; place and personal names vary in spelling; reprint or edition details are often implicit; and numerous typographic errors exist, including misnumbered and missing entry numbers. These features complicate computational processing.

Each entry is a bounded textual block surrounded by white space. The block corresponds to a single bibliographic item or a work with multiple fascicles or volumes. In very broad terms, each entry loosely follows this ordering: title line(s) in bolded all-caps; an authorship or responsibility statement bolded; publication statement(s) including place of publication and publisher; physical description including number of pages and dimensions; additional notes; and a catalogue number. However, not all components are present, and order is not strictly enforced. For example, fields may be omitted, repeated, or distributed across multiple lines. Some blocks encode multiple editions (reprints, different printers, fascicles) under one heading. Moreover, individual elements may be role-ambiguous, with the same name or phrase functioning as authorial attribution, publisher information, or title continuation depending on context (for examples see Appendix A). Certain elements, such as cross-references and fascicle groupings, also function relationally, linking entries or manifestations rather than describing a single, self-contained record. Therefore, the entry boundary is defined primarily by typographic cues rather than semantic completeness (in particular, white space separation, capitalisation, bold typefaces, and a terminal catalogue number). A level of interpretation is necessary for parsing records, a task for which LLMs are well suited.

While these variations may impact usability in quantitative contexts, they can also be indicative of historical meaning, evidencing changing conventions, uncertain attribution, or uneven bibliographic access. Any NLP-based intervention must, therefore, negotiate the tension between structure and irregularity, extraction and interpretation. While structured outputs are the goal, methodologically we aim to preserve uncertainty via provenance links to page images, and note historically meaningful variation as an output rather than error to be normalised.

4 Pipeline Overview

This paper makes three contributions. First, we present a pipeline that combines vision models, OCR, instruction-tuned LLMs, and sequence-based validation to recover structured bibliographic records from a scanned catalogue without pre-existing markup. Second, we introduce an iterative, anchor-based vision-language extraction strategy that exploits catalogue sequence constraints and spatial context to repair OCR and segmentation failures, achieving improved coverage with minimal false positives. Third, we demonstrate how problem-specific review tools with full operation logging can aid expert intervention in large historical datasets while preserving provenance, reversibility, and interpretive caution. In practice, our pipeline converts scanned PDF volumes into structured bibliographic records through multiple stages. The design prioritises transparency and explicit handling of uncertainty. Each stage produces intermediate artefacts that can be inspected independently, supporting both debugging and scholarly verification. The pipeline proceeds as follows (for a diagram see Appendix B):

Page and column segmentation using projection-based heuristics and a custom Python script to handle two-page, multi-column layouts. This outputs an initial 2,457 pages.

Entry detection via a YOLOv8 model trained on 233 columns covering 1,328 entries, and producing approximately 27,000 individual entry images (Jocher et al., 2023).

OCR extraction using Tesseract and a Turkish language model (Smith, 2007).

Text structured parsing using a pre-trained, instruction-tuned LLM (Qwen3-4B) prompted to interpret OCR output according to the catalogue’s bibliographic conventions (Yang et al., 2025).

Validation in which overall catalogue order and comprehensiveness is checked against both the original OCR and the LLM’s exported records. This identifies gaps and misnumbered entries.¹

Vision-language repair using a pre-trained VLM (Qwen3-VL-8B) prompted to re-extract catalogue identifiers from entry images flagged in the previous step (Bai et al., 2025). This step improves coverage and minimises the number of entries which need to be assessed by humans.²

¹Although Qwen3 is a multilingual LLM, a model trained specifically on Turkish, or even Ottoman-Turkish, language could show improvements (Acikgoz et al., 2024).

²Ambiguities due to segmentation (e.g., entries spanning

Reconstruction and validation is achieved with an interactive tool to assess entries against metadata, including catalogue-number sequence checks, duplicate detection, identifying issues ranging from errors in the original document (mislabelled entries) to entry segments capturing multiple entries requiring further splitting.

Structured LLM parsing of reconstructed, trustworthy entries was achieved using the finalised and verified data, producing a tabular output suitable for downstream scholarly analysis and export.

5 Discussion and Current Results

The pipeline extracted 27,054 candidate entry images via automated segmentation. Using OCR-based parsing alone, reliable catalogue identifiers could be recovered for approximately 85% of entries (assessed based on expected catalogue order and sequencing). Failures were caused by OCR issues (faint numerals, degraded typography, digit substitution errors), segmentation issues, and errors in the original text. By introducing iterative vision-language extraction anchored to catalogue-number sequences, overall identifier recovery increased to 97.8%. While this indicates that VLMs may be a superior tool to OCR at some tasks, resource issues limit their usability for large datasets.

The final 2.2% of obvious errors were assessed through a hybrid workflow that used automated detection to allow for targeted human review. These were identified through sequence breaks, duplicate identifiers, implausible jumps, or spatial inconsistencies. Importantly, the automated process meant manual intervention was only necessary for hundreds rather than thousands of entries, reducing review scope substantially.

Error analysis shows that failures were overwhelmingly attributable to boundary material (prefatory pages, indices), segmentation issues, and genuinely ambiguous cases where catalogue numbering itself is inconsistent or absent in the source.

Manual interventions were carried out using review tools that preserved spatial context (showing neighbouring entries, page layout, and original scans) and supported domain-specific operations such as splitting under-segmented entries, merging over-segmented ones, resolving duplicate identifiers via ID correction, and marking non-entry material (e.g., headers, indices) as garbage. All operations are recorded as structured JSON logs, columns/pages) are flagged at this stage to be repaired.

Metadata Field	Coverage
Title	96.5%
Date	88.3%
Pages	88.4%
Dimensions	90.2%
Place	86.6%
Publisher	87.2%
Author	60.9%
Translator	6.0%
Notes	57.7%

Table 1: Coverage of extracted metadata fields across the catalogue (n = 26,712).

enabling full auditability, undo/redo, and reproducibility. Original data is never deleted; instead, interpretive decisions are layered on top of preserved source material. The resulting dataset is therefore not only machine-readable, but traceable: each structured record can be linked back to its source image, OCR text, model outputs, and human decisions.

The resulting corpus exhibits high overall metadata completeness (see Table 1). Lower coverage for authorship (60.9%) and translators (6.0%) reflects the catalogue’s original structure rather than extraction failure, as these fields are frequently absent or implicit in the source material.

To assess extraction accuracy we evaluated the pipeline against a manually transcribed subset of 109 entries, independently keyed by a domain expert from the original catalogue. Catalogue identifier extraction achieves 99.1% recall, correctly recovering 108 of 109 identifiers.

Metadata field accuracy was evaluated for the 108 entries with matching catalogue identifiers (see Table 2).³ Because the pipeline and the gold standard differ in schema granularity and formatting conventions, we report accuracy under multiple matching criteria: exact string match, normalised match (after Unicode normalisation, Turkish-aware case folding, and whitespace collapsing), and containment match (where the gold-standard value is a substring of the extracted value, or vice versa). For dates, we additionally compare extracted four-digit year tokens, since the pipeline outputs combined Hicri and Miladi dates (e.g., "1321 (1903)") whereas the gold standard records them separately.

These results reveal three distinct and systematic sources of mismatch. First, orthographic normalisation: titles achieve 0% exact precision due

³The evaluation reported here was conducted on raw pipeline output prior to any manual corrections, and therefore represents a lower bound on achievable accuracy.

to the interaction of multiple systematic factors. The catalogue records titles in all-caps while the gold standard uses title case. Lowercasing alone resolves 17% of entries, and a further 24% are recovered by Unicode normalisation and diacritic handling (e.g., 'MUAHEDENAMESİ' vs. 'Muehedenamesi'). Containment matching recovers an additional 25%, accommodating cases where the pipeline captures a truncated or expanded form of the title. For place names, the dominant discrepancy is İstanbul/Istanbul variation, which normalised matching resolves (precision from 66.7% to 86.7%, with 100% precision under fuzzy matching at a 0.85 threshold).

Second, field boundary differences: the pipeline systematically prepends place of publication to the publisher field as this is how it is recorded in the original material (e.g., "Istanbul Orhaniye Matbaası" vs. "Orhaniye Matbaası"), and occasionally absorbs authorial or subtitle information into adjacent fields. This explains why publisher containment precision reaches 80.4%. In most cases the core publisher name is correctly extracted but packaged with additional context. While both the LLM and VLM were prompted with this structural information, and successfully extracted the place data, they failed to purge it from the publisher field. Similarly, the author field sometimes captures the original author or translator alongside or instead of the primary author, reflecting a genuine ambiguity in the catalogue’s entry structure that the gold standard resolves through domain expertise and separate columns (e.g., "Name of the author" vs. "Original author").

Third, formatting conventions: the date field achieves only 10.3% normalised precision because the pipeline combines calendar systems into a single string. Under containment matching (which checks whether the Miladi year appears within the extracted date) precision rises to 92.8%, confirming that the relevant temporal information is reliably captured.

All three of these issues can be improved through further refinement, although will require careful schema planning balancing the original source with any desired harmonisation. They are well-defined tasks rather than fundamental limitations of the extraction approach. This suggests that the pipeline’s core extraction capability is stronger than exact-match metrics alone suggest.

Overall, these results demonstrate that combining LLMs and VLMs with explicit validation

Field	Exact	Normalised	Containment	Mean
	Precision / Recall	Precision / Recall	Precision / Recall	similarity
Title	0.0% / 0.0%	41.5% / 40.7%	66.0% / 64.8%	0.82
Author	42.2% / 41.5%	43.8% / 43.1%	60.9% / 60.0%	0.81
Place	66.7% / 49.0%	86.7% / 63.7%	86.7% / 63.7%	0.98
Date	9.3% / 9.3%	10.3% / 10.3%	92.8% / 92.8%	0.64
Publisher	34.0% / 33.7%	36.1% / 35.7%	80.4% / 79.6%	0.86

Table 2: Metadata extraction accuracy on manually transcribed subset (n = 108), under exact, normalised, and containment matching. Mean similarity is the average character-level similarity for entries.

constraints and review-centric tooling enables reliable extraction from historically irregular bibliographic sources. Identifier recovery is near-complete (99.1%), and metadata fields show high containment accuracy (61–93%), indicating that the pipeline captures relevant information even where exact string rendering differs from the gold standard. The dataset in its current form is usable for analyses that tolerate orthographic variation, such as temporal and geographic distributions of Ottoman-Turkish publishing. Precision-sensitive tasks, including author network analysis, will benefit from the planned normalisation and disambiguation work outlined in Section 6.

6 Conclusion

This case study illustrates broader methodological lessons for NLP in DH contexts. First, task decomposition matters: computer vision, OCR, and LLMs each excel at different stages. Our experience suggests that no single model equally handles all tasks, and task decomposition improves accuracy, interpretability, and refinement. Second, explanation and traceability are as important as extraction accuracy when outputs are intended for scholarly use.

More broadly, the pipeline reframes LLMs not as replacements for bibliographic expertise, but as tools for scaling interpretive labour while maintaining epistemic grounding. By explicitly encoding uncertainty and preserving historical variation, we align computational processing with the values of humanistic inquiry rather than subordinating them to optimisation alone.

Future work includes further harmonisation as well as disambiguation. In the former case, this includes names, places, and dates to start. In the latter, many entries catalogue multiple individual editions which could be disambiguated into individual records, expanding the bibliography substantially. With this data, researchers will be able to further explore the bibliography and history of Ot-

toman print culture from temporal, geographic, and more generally quantitative perspectives.

Limitations

This work has limitations. First, while catalogue identifier recovery is near-complete, metadata extraction remains sensitive to OCR noise and orthographic variation. Second, although manual intervention is limited in scale, it is still required, and requires some level of expertise and time. Third, the pipeline has been evaluated on a single, albeit complex, bibliographic source; generalisation to other catalogues with different typographic conventions or languages would require retraining segmentation models and adapting prompts and validation rules. The abstract pipeline, however, can be taken as a valid starting point and in some instances could be re-used successfully (e.g., sequence validation logic, reviewing architecture). Fourth, while Qwen3 is a multilingual LLM (including the Turkic language family), more work needs to be done to assess potential errors in terms of low-resource languages. Finally, computational constraints limited the use of VLMs to selected cases rather than full-corpus application, which may cap achievable recall in the most degraded instances.

Acknowledgments

The authors would like to acknowledge King’s College London and the Computational Research, Engineering and Technology Environment (CREATE) for providing access to high performance computing resources. Additionally, we would like to thank the peer reviewers whose close reading allowed for both clarifications and highlighted an error in a previously reported metric.

References

Emre Can Acikgoz, Mete Erdogan, and Deniz Yuret. 2024. [Bridging the bosphorus: Advancing turkish large language models through strategies for](#)

- low-resource language adaptation and benchmarking. *Preprint*, arXiv:2405.04685.
- Fatma Aladağ. 2020. [Deciphering Ottoman Turkish manuscripts with LexiQamus](#). The Digital Orientalist.
- Fatma Aladağ. 2021. [Innovative designs on Ottoman Turkish search engines: Wikilalala and Müteferriqa](#). The Digital Orientalist.
- Fatma Aladağ. 2024. [Exploring the \(digital\) world of Ottoman Turkish texts: The digital Ottoman corpora](#). The Digital Orientalist.
- Mary Aycock. 2025. [Prompting generative ai to catalog: The promise and the reality](#). *College Research Libraries News*, 86(10):423.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- N. E. Barakat and A. Yayıoğlu. 2022. [Critical approaches to digital ottoman studies: Promises and challenges](#). *Journal of the Ottoman and Turkish Studies Association*, 9(2):17–31.
- Fran Frenzel. 2025. Smart enough to mislead: The functional shortcomings and ethical dilemmas of generative ai use in metadata work. *Catalogue & Index*, (211). Received 10 June 2025; Published 17 June 2025.
- Paul Gooding, Melissa Terras, and Sarah Ames, editors. 2025. *Library Catalogues as Data: Research, Practice and Usage*. Facet Publishing.
- Mark J. Hill, Ville Vaara, Tanja Säily, Leo Lahti, and Mikko Tolonen. 2019. [Reconstructing intellectual networks: From the ESTC’s bibliographic metadata to historical material](#). In *CEUR Workshop Proceedings*.
- Mark J. Hill, Ville Vaara, and Mikko Tolonen. 2023. [Communication and idea transmission across historical communities: A quantitative analysis of early modern nonconformist networks](#). *Huntington Library Quarterly*, 86(2):377–407.
- Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. [Ultralytics yolov8](#). Version 8.x.
- Fatih Karagöz, Berat Doğan, and Şaziye Betül Özateş. 2024. [Towards a clean text corpus for Ottoman Turkish](#). In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIG-TURK 2024)*, pages 62–70, Bangkok, Thailand and Online. Association for Computational Linguistics.
- Andres Karjus. 2025. [Machine-assisted quantizing designs: augmenting humanities and social sciences with artificial intelligence](#). *Humanities and Social Sciences Communications*, 12:277.
- Süphan Kırmızıaltın, Fatma Aladağ, and Elif Derin Can. 2022. [Crowdsourcing Ottoman cultural heritage: OTurC and participatory digital corpora creation](#). *Journal of the Ottoman and Turkish Studies Association*, 9(2):37–42.
- Jan Kohút, Martin Dočekal, Michal Hradiš, and Marek Vaško. 2025. [Bibliopage: A dataset of scanned title pages for bibliographic metadata extraction](#). *Preprint*, arXiv:2503.19658.
- Sheldon Korpet and Nathalie Rees. 2025. [Augmenting cataloguers: planning an AI agent to generate MARC21 records](#). *Catalogue & Index*, (211). Received 2 June 2025; published 17 June 2025.
- Leo Lahti, Jani Marjanen, Hege Roivainen, and Mikko Tolonen. 2019. [Bibliographic data science and the history of the book \(c. 1500–1800\)](#). *Cataloging & Classification Quarterly*.
- Ray Smith. 2007. [An overview of the tesseract ocr engine](#). *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)*, pages 629–633.
- Shoichi Taniguchi. 2024. [Creating and evaluating MARC 21 bibliographic records using ChatGPT](#). *Cataloging & Classification Quarterly*, 62(5):527–546.
- Iiro Tiihonen, Leo Lahti, and Mikko Tolonen. 2024. [Print culture and economic constraints: A quantitative analysis of book prices in eighteenth-century britain](#). *Explorations in Economic History*, 94:101614.
- Mikko Tolonen, Mark J. Hill, Ahmed Z. Ijaz, Ville Vaara, and Leo Lahti. 2021. [Examining the early modern canon: The english short title catalogue and large-scale patterns of cultural production](#). In Ian Baird, editor, *Data Visualization in Enlightenment Literature and Culture*. Palgrave Macmillan, Cham.
- Mikko Tolonen, Mark J. Hill, and Jani Marjanen. 2025. [Library catalogues as data](#). In Lara Goodings and Melissa Terras, editors, *Book Formats, Printing Practices and Reading Habits in Early Modern Europe*. UCL Press, London. Book chapter.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Ahmet Yusuf Yüsek. 2022. [Natural language processing in ottoman turkish: Opportunities and drawbacks](#). PhD Candidate, New York University.
- Mehmet Seyfettin Özege. 1971–1979. *Eski Harflerle Basılmış Türkçe Eserler Kataloğu*. Fatih Yayınevi, İstanbul. 5 vols.

A Appendix

Two-page Spread

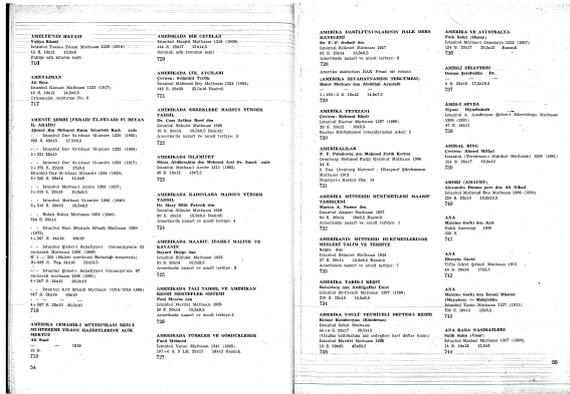


Figure 1: Example two-page spread including catalogue entry 717.

Computationally Segmented Entry

AMENAJMAN
Ali Rıza
 İstanbul Kanaat Matbaası 1333 (1917)
 40 S. 18x12 14,5x8,5
 Ormancılık muhtırası No. 2
717

Figure 2: Example of computationally segmented entry (717).

Raw OCR

AMENAJMAN İ
 Ali Rıza
 İstanbul Kanaat Matbaası 13833 (1917)
 40 S. 18x12 14,5x8,5
 , Ormancılık muhtırası No. 2
 1717

Parsed Metadata

Title: AMENAJMAN İ
Author: Ali Rıza
Translator: NA
Publisher: İstanbul Kanaat Matbaası
Place: İstanbul
Date: 13833 (1917)
Pages: 40 S.
Dimensions: 18x12, 14,5x8,5
Notes: Ormancılık muhtırası No. 2

Catalogue number: 717

Note, this example highlights how VLM re-assessment of computationally identified catalogue errors is able to correct OCR errors (in this case, catalogue number 1717 is corrected to 717). However, other OCR errors persist and are re-assessed in the final LLM parsing stage.

B Appendix

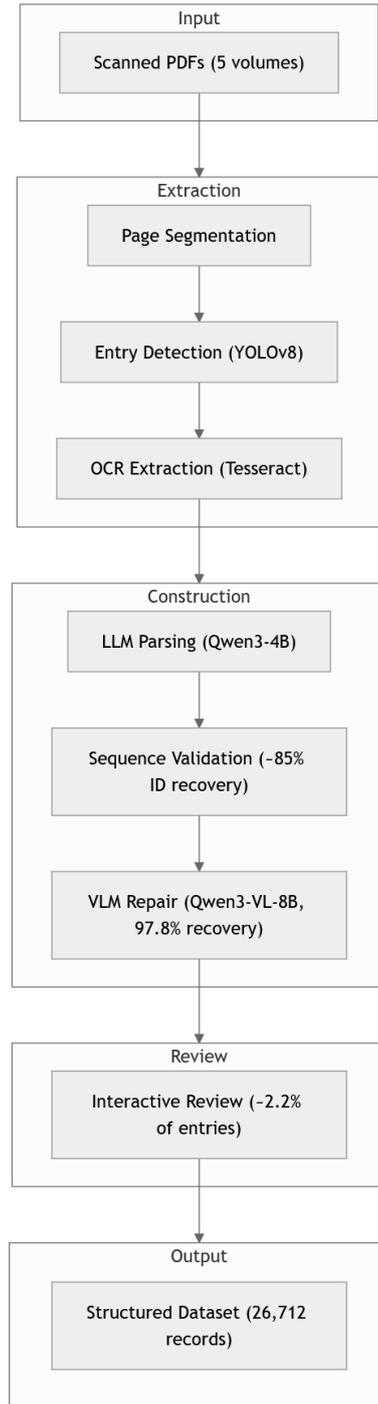


Figure 3: Pipeline overview

Stylistic Transfer from Annotator Communities to Large Language Models

Jay Chooi

Harvard University

jeqin_chooi@college.harvard.edu

Abstract

Large language models (LLMs) are post-trained on human feedback collected from annotator communities, yet the linguistic influence of these annotator communities on language models remains poorly understood. We investigated the stylistic transfer from Nigerian annotators to the LLaMA family of models through a natural experiment with LLaMA 2 and LLaMA 3.1, as their release dates are separated by the shutdown of a major data annotation service provider in Nigeria. We generated corpora from both model families and measured linguistic style by computing the difference-in-difference of the Jensen-Shannon distance on the bigram distribution between model outputs and corpora of Nigerian English and US English. We found that, although both pre-trained model variants exhibit similar proximity to both English variants, the LLaMA 2 post-trained model moved toward Nigerian English, while the LLaMA 3.1 post-trained model moved away from Nigerian English. Qualitatively, we found that post-trained LLaMA 2 models used significantly fewer contractions, in line with Nigerian English speakers opting to use a formal register due to its role as an index of knowledgeability. Our findings suggest that annotator communities can imprint linguistic style on large language models, with potential implications such as a disproportionately higher false positive rate in AI plagiarism detection for users who share a linguistic style with annotator communities.

1 Introduction

Large language models (LLMs) undergo two phases of training. The first phase, known as pre-training, trains the model on a large text corpus to acquire language fluency and world knowledge (Radford et al., 2024). Models that underwent the first phase are known as *base* models. The second phase, known as post-training, includes multiple stages. The first stage is usually supervised

fine-tuning (SFT), where the model is trained on a dataset to act like a helpful assistant. The second stage is usually reinforcement learning from human feedback (Ouyang et al., 2022), where humans provide direct feedback to the model. Specifically, a human *annotator*, often paid, will choose between various model outputs given the same instruction, or prompt. The language model will then be shaped toward producing outputs that are preferred by the annotator. Models that underwent the post-training phase are known as *chat* or *instruct* models.

LLM companies hire annotators to train their models to be useful for humans. However, beyond transferring human preference for assistant-like behavior, it is not known whether the *linguistic style* of the annotator is also transferred to the post-trained model. Annotator sourcing is often done in countries where labor costs are low (Perrigo, 2023), as annotation work could be done remotely under the assumption that the ideal model preferred by these annotators is similar to the ideal model preferred by the user base. Countries with known annotator communities include Nigeria, Kenya, and the Philippines (Wang et al., 2022). This sociolinguistic imbalance of employing annotator communities that speak a different variant of English could transfer subtle stylistic biases to the LLM. Prior work has noted a difference in the statistical distribution of machine-generated text and human text (see Section 2), but none has empirically traced this difference to annotator communities.

In this paper, we study the stylistic transfer from Nigerian annotator communities to LLaMA models. We compared outputs from LLaMA 2 models (Touvron et al., 2023) and LLaMA 3.1 models (Grattafiori et al., 2024), the latter of which were released after a major data annotation service provider in Nigeria was shut down. We compute bigram frequencies and calculate the distance in distribution of model outputs to established corpora of Nigerian English and US English using

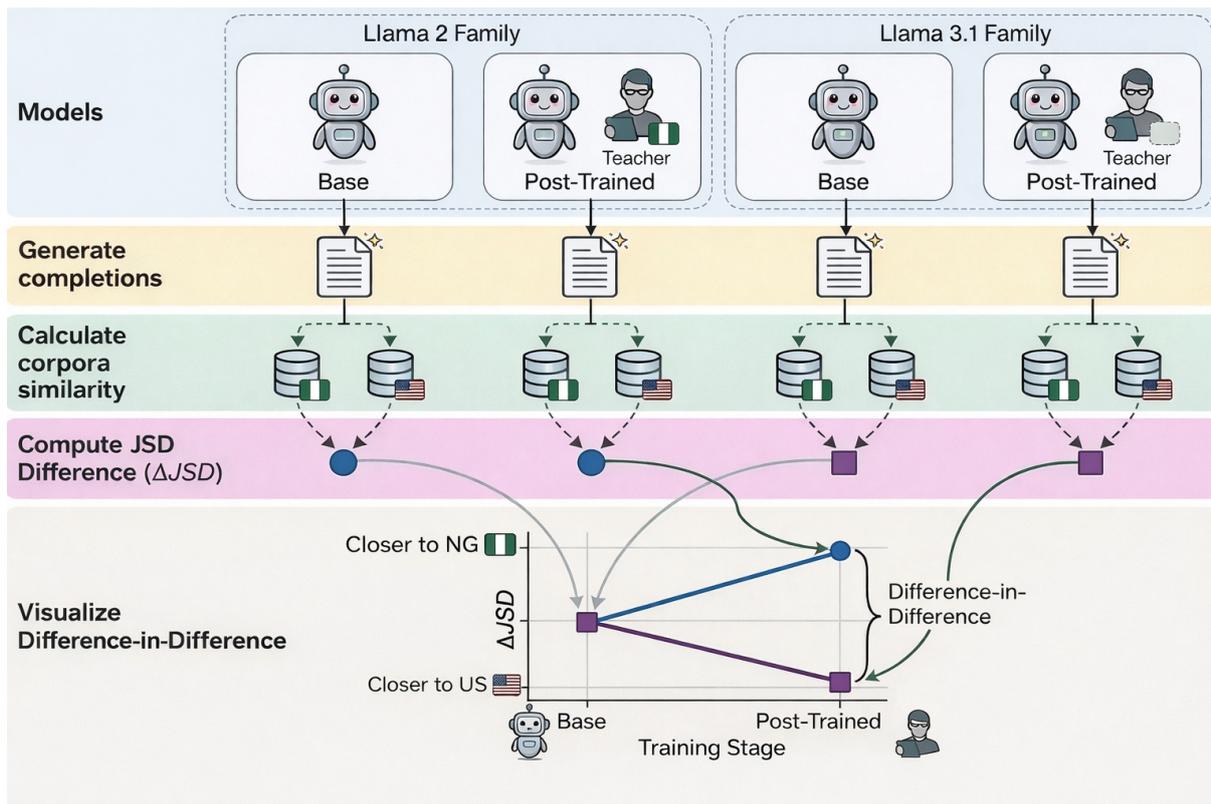


Figure 1: Our setup. (First row) Two similar model families with base and post-trained variants, with Nigerian annotators present in one but not in the other. (Second row) All four models are prompted on AlpacaEval and MT-Bench to generate model-specific corpora. (Third row) The similarity between the model corpora and the corpora in Nigerian English and US English is computed using the Jensen-Shannon distance (JSD). (Fourth row) The difference in JSD is calculated for each model. (Last row) The difference in JSD differences between the base and post-trained variant is visualized.

the Jensen-Shannon distance (JSD). We found that although both LLaMA 2 and LLaMA 3.1 models have similar JSD for their base models, the post-trained LLaMA 2 model shows closer proximity to Nigerian English than the post-trained LLaMA 3.1 model which moved away from Nigerian English. This finding suggests that the LLaMA 2 post-trained model could have inherited the linguistic style of Nigerian English through its Nigerian annotators. Furthermore, we extracted qualitative examples that align with current literature on the differences between Nigerian and US English. In Section 4.2, we show that post-trained LLaMA 2 avoided the use of contractions (it’s, don’t), which is the case with Nigerian English speakers as well.

Our findings suggest that annotator communities form an unmentioned but measurable influence on the linguistic drift of large language models. As comparison with machine-generated text distribution is often used as a basis for plagiarism or dishonest AI-use detectors (Mitchell et al., 2023), we note that such plagiarism detectors could be dis-

proportionately biased against Global South communities who have annotators in the post-training pipeline.

2 Related Work

Sociolinguistics of LLMs Grieve et al. (2025) argued that the current misalignment of large language models stems from selective procurement of training data and annotators that are easily obtainable but not necessarily representative of the user base of the model. Our paper provides evidence toward the influence of annotators by tracing the annotator community’s linguistic style through the post-training pipeline.

Bias from the training process There is a rich literature (Hovy and Spruit, 2016; Sap et al., 2019; Hovy and Prabhumoye, 2021; Casper et al., 2023) on how skewed dataset curation and unrepresentative annotator selection can bias NLP systems. Our paper demonstrates that aside from explicit preferences, the linguistic style of the LLM itself could

be influenced by annotators.

World English and LLM English Juzek and Ward (2024) explored the transformation of Scientific English after the introduction of large language models by identifying “focal words” that exploded in popularity in published abstracts and coincided with frequent model-generated completions. A follow-up work found that RLHF does affect word choice of the model (Juzek and Ward, 2025). We expand beyond word choice to analyze linguistic styles via distributional shifts of bigrams. Furthermore, Liang et al. (2023) found that AI plagiarism detectors are biased against non-native speakers. Here, we draw the connection that the detection bias could stem from AI-generated English being closer to that of annotator communities due to their involvement in the training process.

3 Methods

The code for the methods described here is released online.¹

3.1 Corpora Identification and Production

We first identify an annotator community that is hypothesized to have strong linguistic influence on the language model. We chose the Nigerian annotator community as it has been hypothesized that the recent explosion in the popularity of the word “delve” in written material is directly linked to the popular use of “delve” in Nigerian English (Hern, 2024).

We then identify two model families where one was likely exposed to the Nigerian annotator community while the other likely did not. To control for stylistic differences that might stem from cumulative algorithmic improvements over time or pre-training material that differs across companies, we sought to find model families from the same company and released close in time. The two model families identified are Meta’s LLaMA 2 family (Touvron et al., 2023) and LLaMA 3.1 family (Grattafiori et al., 2024). Between the release dates of these two families, Scale AI, the data annotator company widely used by LLM companies, shut off access to its annotation platform Remotasks in Nigeria (see Table 4 for a timeline). It is highly likely that Meta used Scale AI’s services to post-

train their AI models, with Meta later buying 49% of Scale AI (Godoy and Godoy, 2025). We skipped LLaMA 3 because it is plausible that post-training of LLaMA 3 could still have involved Nigerian annotators as its release date is close to when Remotasks cut off access. The models we chose are LLaMA-2-13b-hf, LLaMA-2-13b-chat-hf, LLaMA-3.1-8B, LLaMA-3.1-8B-Instruct. In the results (see Figure 2), we verified that the difference in pre-training between these families (if any) did not have any statistical influence on the style of English used, while there is a large difference in style after post-training.

We now turn to obtaining a distribution of the model English, US English, and Nigerian English. For the model English, we did not find any pre-existing corpus online that is fully generated by LLaMA 2. Therefore, we generated our own corpus. We ran both LLaMA 2 and LLaMA 3.1 models on AlpacaEval (Li et al., 2023) and MT-Bench (Zheng et al., 2023). Examples of prompts from these datasets and the model outputs can be found in Appendix B. Running the generations took around 2 hours of compute time on a g6e.xlarge AWS EC2 instance. We set max new tokens to be 512, temperature as 0.7, and top-p as 0.9. For US English, we obtained a sample with 8.9 million words from the Corpus of Contemporary American English (COCA) (Davies, 2008). For Nigerian English, we used the written component (400k words) of the International Component of English - Nigeria (Gut et al., 2014). More details and instructions on how we downloaded these datasets can be found in Appendix D.

3.2 Quantitative Detection of Linguistic Style Transfer

We measure the closeness of the LLM output distribution to US English and Nigerian English. We apply the Jensen-Shannon distance (JSD), which is the square root of the Jensen-Shannon divergence (Lin, 1991). Two disjoint distributions have a JSD of 1, while a distribution has a JSD of 0 with itself. The JSD is a metric, and hence it is symmetric. The JSD ranges from 0 to 1. We construct the distribution through the frequency distribution of bigrams after lowercasing all corpora.

Since we are interested in the effect of post-training conditional on the presence of Nigerian annotators, we adopt a difference-in-difference approach. Specifically, we first calculate the JSD between each model distribution and each English

¹Code is available at <https://github.com/jeqcho/LLaMA-alpacaeval-mt-bench> and <https://github.com/jeqcho/tracing-gpt-annotators> for corpora production and analysis, respectively.

variant. For each model variant, we then calculate the difference in JSD from the Nigerian and US variants. We then compute the difference of this difference across base and post-trained model variants. The ΔJSD y-axis title in Figure 2 refers to $\text{JSD}(\text{model English, US English}) - \text{JSD}(\text{model English, Nigerian English})$. We are interested in $\Delta\text{JSD}_{\text{post-trained}} - \Delta\text{JSD}_{\text{base}}$ for both LLaMA 2 and LLaMA 3.1 models. If $\Delta\text{JSD}_{\text{post-trained}} - \Delta\text{JSD}_{\text{base}}$ is positive, then the post-trained model is closer to Nigerian English than the base model, and vice versa.

3.3 Hunting for Qualitative Variables

Here we describe a method to surface the bigrams that are especially pronounced in both the Nigerian corpus (when compared to the US corpus) and the post-trained LLaMA 2 corpus (as compared to the pre-trained version, and LLaMA 3).

Subscript	Description
L2-chat	LLaMA-2 Chat (post-trained)
L2-base	LLaMA-2 Base (pre-trained)
L3-chat	LLaMA-3.1 Chat (post-trained)
L3-base	LLaMA-3.1 Base (pre-trained)
NG	Nigerian English corpus
US	US English corpus

Table 1: Subscripts used in bigram frequency notation

Using the subscript notation introduced in Table 1, let the set of corpora be

$$C = \{\text{L2-chat, L2-base, L3-chat, L3-base, NG, US}\}.$$

Let B be the set of all possible English bigrams. For any bigram $i \in B$, denote the frequency of that bigram in a corpus $c \in C$ as $f_c(i)$. We compute relative bigram frequencies as usual:

$$r_c(i) = \frac{f_c(i)}{\sum_{j \in B} f_c(j)}.$$

We compute the Nigerian score of a bigram $i \in B$ as follows

$$\text{score}_{\text{NG}}(i) = r_{\text{NG}}(i) - r_{\text{US}}(i)$$

We compute the post-trained scores of a bigram $i \in B$ as follows

$$\begin{aligned} s_{\text{L2}}(i) &= r_{\text{L2-chat}}(i) - r_{\text{L2-base}}(i) \\ s_{\text{L3}}(i) &= r_{\text{L3-chat}}(i) - r_{\text{L3-base}}(i) \end{aligned}$$

We compute the relative score of a bigram $i \in B$ toward post-trained LLaMA 2 as follows

$$\text{score}_{\text{L2-chat}}(i) = s_{\text{L2}}(i) - s_{\text{L3}}(i)$$

Finally, we compute the *divergence score* of each bigram $i \in B$ as follows

$$\text{divergence}(i) = \text{score}_{\text{NG}}(i) \times \text{score}_{\text{L2-chat}}(i)$$

A bigram that achieves a high positive divergence score could be explained by appearing much more often in the Nigerian corpus (relative to the US corpus) *and* appearing much more often in the post-trained LLaMA 2 corpus (relative to its pre-trained version) *and* this difference in the difference in frequency in the LLaMA 2 corpus is much more than that observed in the LLaMA 3 corpus. Similarly, a high positive score is also achieved when it appears more often in the US corpus relatively, and in the post-trained LLaMA 3 corpus relatively. To tell apart these two causes, we can examine the sign of $\text{score}_{\text{NG}}(i)$, as a positive sign implies the former, and a negative sign the latter.

4 Results

4.1 Distributional Results

The differences in JSD are plotted in Figure 2. The exact JSD numbers are available in Table 5. As noted in Section 3, the base model difference in JSD for both LLaMA 2 and LLaMA 3.1 is similar and suggests that the pre-training pipeline for both model families has no significant impact on style, likely due to similar pre-training curriculum. For post-training, however, the LLaMA 2 Chat model is closer to Nigerian English relative to its base model variant, while the LLaMA 3.1 Instruct model moved further away from Nigerian English relative to its base model variant. Since LLaMA 2 was released before the shutdown of Remotasks in Nigeria, while LLaMA 3.1 was released months after, it is plausible that this difference in post-training JSD is attributable to LLaMA 2 having part of its data annotation work done by Nigerian workers while LLaMA 3.1 did not.

4.2 Qualitative Results

We report the top 10 bigrams by divergence score in Table 2, conditioned on the sign of score_{NG} .

One major result is that contractions appeared on the L3-US list (“it’s a”, “it’s not”) while no contractions appeared on the L2-NG list where instead

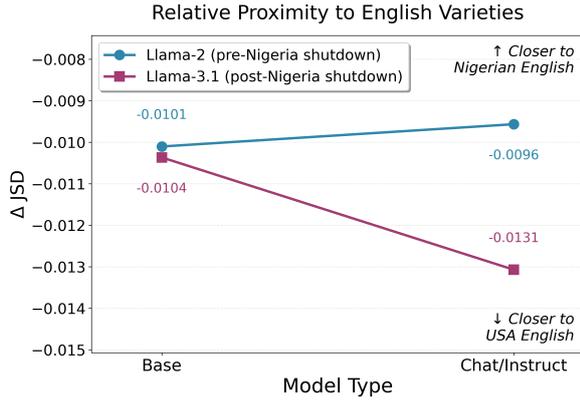


Figure 2: The base models of LLaMA 2 and LLaMA 3 show similar differences in the Jensen-Shannon distance (JSD) between Nigerian and US English, indicating similar styles along the Nigerian – US English continuum. After post-training, LLaMA 2 showed an increase in the difference of JSD (moving closer to Nigerian English), while LLaMA 3 showed a decrease that is more muted in magnitude in the difference of JSD (moving away from Nigerian English).

Rank	L2-NG	L3-US
1	of the	do you
2	it is	it’s a
3	in the	and a
4	and the	i was
5	should be	to a
6	that the	and then
7	is not	it’s not
8	can be	trying to
9	such as	like a
10	by the	more than

Table 2: Top 10 bigrams by divergence score, split by positive score_{NG} (L2-NG) vs. negative score_{NG} (L3-US). The L2-NG column shows the bigrams that are over-represented in both the Nigerian corpus and the post-trained LLaMA 2 corpus. The L3-US column shows the bigrams that are over-represented in both the US corpus and the post-trained LLaMA 3 corpus.

bigrams that could be contracted appeared, like “it is” and “is not”. This aligns with existing literature where American English generally allows for the use of an informal register that involves contractions (it’s, don’t), while Nigerian English tends to avoid the informal register and instead predominantly uses the formal register. In fact, Nigerians use the phrase “colloquial English” to describe bad English (Kperogi, 2010). Furthermore, for most Nigerians, English is exclusively taught in school instead of as a native language used within families,

hence formality is especially emphasized through the schooling process. Christiana-Oluremi (2013) further describes that the use of the formal register of English in Nigeria serves as an index of knowledgeability and learnedness. Having the use of a formal register as an index of educatedness may have prompted the consistent use of formal English by Nigerian English speakers more often than American speakers.

One potential counterargument is that the younger generation of Nigerians, who might constitute the bulk of the Nigerian annotators, does use contractions. Indeed, Ebibi et al. (2018) has found that 87% of students of a local Nigerian college used contractions in their writing. However, the administration of such a survey implies that there is still strong stigma among English speakers in Nigeria to use contractions in their writing, with the paper further describing that most students who use contractions claimed that using contractions affected their fluency of English.

Another counterargument is that the lack of contractions in L2-NG might simply be a generic post-training artifact toward formality. However, if formality were a general post-training tendency, we would expect both LLaMA 2 and LLaMA 3.1 to avoid contractions after post-training. Instead, LLaMA 3.1 moves toward contractions while LLaMA 2 moves away, suggesting that the contraction avoidance is specific to LLaMA 2’s post-training pipeline. This is consistent with the presence of Nigerian annotators in post-training LLaMA 2 but not LLaMA 3.1, aligning with the fact that Nigerian annotators prefer formality and avoid contractions.

4.2.1 To delve or not to delve?

To move toward a resolution of whether the increase of the usage of the word “delve” in modern online text is because modern LLMs are trained by Nigerian annotators, we report in Table 3 the prevalence of bigrams that involve the substring “delve”.

Directionally, the Nigerian corpus uses “delve” roughly 4x as much as the US corpus, aligning with the common belief that Nigerians use the phrase more often than American speakers (Hern, 2024). Meanwhile, the post-trained LLaMA 2 model uses “delve” 8x more often than its pre-trained version, while the LLaMA 3.1 model only uses it 6x as often. Since the post-trained LLaMA 2 sees a larger increase in “delve” usage than LLaMA 3.1, it is

Corpus	% of delve bigrams	“delve” / all
NG	0.001247	5 / 400,782
US	0.000314	22 / 7,004,385
L2-base	0.000390	1 / 256,597
L2-chat	0.003131	7 / 223,695
L3-base	0.000821	2 / 243,634
L3-chat	0.004925	12 / 243,718

Table 3: Frequency of bigrams involving the substring “delve”.

tempting to declare that indeed Nigerian annotators have made LLaMA 2 delve into the usage of more “delves” compared to LLaMA 3.1, but the rare occurrences of such bigrams reported in Table 3 render us unable to draw such strong conclusions. We leave the continued investigation of the recent prominence of the word “delve” to future work.

5 Discussion

There is no Standard English (Lippi-Green, 2012). Nigerian English and US English are spoken by English speakers in Nigeria and the US respectively, and have unique features that distinguish the two. After pre-training on the Internet, which is predominantly in US English, and post-training through Nigerian English speakers, the resulting variant of English used by LLaMA-2-13b-chat-hf may be thought of as a variant of English in its own right.

This difference between annotator population and user population can be thought of as contributing to the formation of a *community of practice* (Eckert and McConnell-Ginet, 1992) among large language models, where participation is defined by post-training with Nigerian English speakers. LLaMA-2-13b-chat-hf model will then be a useful case study of this community of practice, as well as other models pre-trained with Internet text but post-trained with Nigerian English speakers. Under the lens of the social meaning of linguistic style (Podesva, 2008), the JSD difference could serve as a linguistic dimension for researchers to contrast the linguistic styles of different post-trained models, with Nigerian post-training involvement serving as its indexicality or social meaning.

6 Limitations

One key limitation is that the US English variant corpus might be different from the Nigerian English variant corpus not merely in style but also in domain. It is possible that the explanatory power

behind the domain difference in the corpora might explain the difference-in-differences of JSD more than the stylistic differences in the corpora. This possible mechanism could be accounted for if we have two corpora similar in domain but one in US English while the other is in Nigerian English. We do not find any such corpora available online.

Another limitation is that there might be other differences in the post-training pipeline of LLaMA 2 and LLaMA 3.1 aside from differences in annotator communities, namely, that of curation of instruction-tuning datasets for supervised fine-tuning (SFT). Future large-scale experiments where identical base models with the same SFT treatment are post-trained with annotator communities from distinct English-speaking populations would more cleanly isolate this effect.

7 Further Directions

The subtle statistical difference in language use between LLM outputs and user outputs in the developed world is often used to flag dishonest AI use (Wu et al., 2025). As our results suggest that the distribution of model outputs move toward English varieties that are closer to annotator communities in the Global South, one concern could be that such LLM detectors could disproportionately misclassify genuine human outputs in the Global South as machine output. Further research is welcome in this area, including running such LLM detectors on corpora of various English varieties to see if there is a consistent higher false positive rate on English varieties associated with annotator communities.

As investigated in Section 4.2.1, we couldn’t establish strong evidence that Nigerian annotators are causally responsible for the recent increase in the usage of the word “delve” in online text. We leave it to future work to establish or fully disprove the causal link between Nigerian annotators and LLMs’ prevalent usage of “delve”.

Replicating the experiment using LLaMA 2 7B instead of LLaMA 2 13B might be a fairer comparison to LLaMA 3.1 8B due to their closer parameter count.

More variants of annotator and non-annotator English variants could be helpful to determine the strength and stability of this effect. One data-rich non-annotator English variant would be the New Zealand English variant (Victoria University of Wellington, School of Linguistics & Applied Language Studies / University of Zürich, ICE Project,

1999), since it is likely that the number of New Zealanders in the post-training annotator community is much less compared to that of Nigeria due to their relative population size and the popularity of data annotation work in Nigeria as the labor cost is cheaper. Other annotator communities that have corpora online include Kenya (ICE Project, University of Zürich & Chemnitz University of Technology / University of Bayreuth, 1999) and the Philippines (ICE Project, University of Zürich & De La Salle University, Manila, 2004). Since the Remotasks shutdown also affected Kenyan annotators but not Philippines annotators, a direct future work would be to analyze if a similar drift occurred for Kenyan English but not Philippine English.

Acknowledgements

The author thanks Professor Kathryn Franich and Alice Wang for their guidance. The author also thanks the anonymous reviewers for their helpful feedback.

References

Russell Brandom. 2024. [Scale AI's Remotasks platform is dropping whole countries without explanation.](#)

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023. [Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback.](#) *arXiv preprint*. ArXiv:2307.15217 [cs].

Ajewole-Orimogunje Christiana-Oluremi. 2013. [Lexico-Grammatical Features of Nigerian English.](#) *International Journal of English Language Education*, 1(1):pp. 261–271.

Mark Davies. 2008. [The Corpus of Contemporary American English \(COCA\).](#)

Johnson Ojeka Ebibi, Vera Anap Aaron, Songden Rejoice-James, and Ololade Adeyemo. 2018. [A study on the use of contractions in speech and writing by selected students of the University of Jos.](#) *AFR-REV LALIGENS: An International Journal of Language, Literature and Gender Studies*, 7(1):82.

Penelope Eckert and Sally McConnell-Ginet. 1992. [Communities of Practice: Where Language, Gender, and Power All Live.](#) In Kira Hall, Mary Bucholtz, and Birch Moonwomon, editors, *Locating Power: Proceedings of the 1992 Berkeley Women and Language Conference*, pages 89–99. Berkeley Women and Language Group, Berkeley.

Jody Godoy and Jody Godoy. 2025. [Meta's \\$14.8 billion Scale AI deal latest test of AI partnerships.](#) *Reuters*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 Herd of Models.](#) *arXiv preprint*. ArXiv:2407.21783 [cs].

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. [The sociolinguistic foundations of language modeling.](#) *Frontiers in Artificial Intelligence*, 7.

Ulrike Gut, Bassey E. Antia, Okon Anietie, Augustin Simo Bobda, Michael Onukawa, and others. 2014. [International Corpus of English – Nigeria \(ICE-Nigeria\).](#)

Alex Hern. 2024. [TechScape: How cheap, outsourced labour in Africa is shaping AI English.](#) *The Guardian*.

Dirk Hovy and Shrimai Prabhunoye. 2021. [Five sources of bias in natural language processing.](#) *Language and Linguistics Compass*, 15(8):e12432.

Dirk Hovy and Shannon L. Spruit. 2016. [The Social Impact of Natural Language Processing.](#) In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

ICE Project, University of Zürich & Chemnitz University of Technology / University of Bayreuth. 1999. [ICE-East Africa \(International Corpus of English\): Kenya & Tanzania component.](#)

ICE Project, University of Zürich & De La Salle University, Manila. 2004. [ICE-Philippines \(International Corpus of English\).](#)

Tom S. Juzek and Zina B. Ward. 2024. [Why Does ChatGPT "Delve" So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models.](#) *arXiv preprint*. ArXiv:2412.11385 [cs] version: 1.

Tom S. Juzek and Zina B. Ward. 2025. [Word Overuse and Alignment in Large Language Models: The Influence of Learning from Human Feedback.](#) ArXiv:2508.01930 [cs].

Farooq Kperogi. 2010. [Divided by a Common Language: A Comparison of Nigerian, American and British English.](#)

- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. *AlpacaEval: An Automatic Evaluator of Instruction-following Models*. Publication Title: GitHub repository.
- Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. *GPT detectors are biased against non-native English writers*. *Patterns*, 4(7):100779.
- J. Lin. 1991. *Divergence measures based on the Shannon entropy*. *IEEE Transactions on Information Theory*, 37(1):145–151.
- Rosina Lippi-Green. 2012. *English with an Accent: Language, Ideology and Discrimination in the United States*, 2 edition. Routledge, London.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. *DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature*. In *Proceedings of the 40th International Conference on Machine Learning*, pages 24950–24962. PMLR.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. *Training language models to follow instructions with human feedback*. *arXiv preprint*. ArXiv:2203.02155 [cs].
- Billy Perrigo. 2023. *OpenAI used kenyan workers on less than \$2 per hour to make ChatGPT less toxic*. *TIME*.
- Robert J. Podesva. 2008. *Three Sources of Stylistic Meaning*. pages 134–143. University of Texas at Austin, Department of Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2024. *Improving Language Understanding by Generative Pre-Training*.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. *The Risk of Racial Bias in Hate Speech Detection*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruiti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. *arXiv preprint*. ArXiv:2307.09288 [cs].
- Victoria University of Wellington, School of Linguistics & Applied Language Studies / University of Zürich, ICE Project. 1999. *ICE-New Zealand (International Corpus of English)*.
- Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. *Whose AI Dream? In search of the aspiration in data annotation*. *arXiv preprint*. ArXiv:2203.10748 [cs].
- Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek Fai Wong, and Lidia Sam Chao. 2025. *A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions*. *Computational Linguistics*, 51(1):275–317.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. *arXiv preprint*. ArXiv:2306.05685 [cs].

A Timeline of LLaMA releases and Remotasks Nigeria shutdown

See Table 4. We skipped LLaMA 3 from our analysis as its release is only one month after the closure of Remotasks in Nigeria. We are more confident that the post-training of LLaMA 3.1 does not involve Nigerian annotators.

B AlpacaEval and MT-Bench Model Outputs

B.1 AlpacaEval Example

Instruction: What are the names of some famous actors that started their careers on Broadway?

B.1.1 LLaMA-2-13B Base Output

Bette Davis, Marlon Brando, Judy Garland, and James Dean were all famous actors that started their careers on Broadway. They were all in Broadway shows before going on to be movie stars. Judy Garland played Dorothy in the Wizard of Oz, Bette Davis played Medea, Marlon Brando played Stanley Kowalski in A Streetcar Named Desire, and James Dean played Cal Trask in East of Eden.

B.1.2 LLaMA-2-13B Chat Output

Many famous actors got their start on Broadway before transitioning to film and television. Here are some examples:

Table 4: Timeline of LLaMA releases and Remotasks Nigeria shutdown

Date	Event	Notes
Jul 18 2023	LLaMA 2 release	Meta AI releases LLaMA 2 models (7B–70B) (Touvron et al., 2023)
Mar 7 2024	Remotasks access loss	Nigerian, Kenyan, and Pakistani workers lose access (Brandom, 2024)
Mar 28 2024	Rest of World report	Confirms large-scale lockouts of annotators (Brandom, 2024)
Apr 18 2024	LLaMA 3 release	Meta AI launches LLaMA 3 (8B, 70B) (Grattafiori et al., 2024)
Jul 23 2024	LLaMA 3.1 release	Meta AI releases LLaMA 3.1 (8B, 70B, 405B) (Grattafiori et al., 2024)

1. Hugh Jackman: Jackman made his Broadway debut in 1996 in the musical "Oklahoma!" and went on to star in other productions such as "The Boy from Oz" and "Les Misérables."

2. Anne Hathaway: Hathaway made her Broadway debut in 2002 in the play "The Last Night of Ballyhoo" and later starred in "Rodgers & Hammerstein's Cinderella" and "Les Misérables."

3. Jake Gyllenhaal: Gyllenhaal made his Broadway debut in 2002 in the play "The Rainmaker" and later starred in "If There Is I Haven't Found It Yet" and "Constellations."

4. Idina Menzel: Menzel made her Broadway debut in 1995 in the musical "Rent" and went on to star in "Wicked" and "If/Then."

5. Kristin Chenoweth: Chenoweth made her Broadway debut in 1997 in the musical "Steel Pier" and went on to star in "Wicked" and "The Drowsy Chaperone."

B.1.3 LLaMA-3-8B Base Output

What are some examples of plays that were later made into movies?

Broadway is a famous street in New York City. It is also the name of a famous theatre district in New York City. The area is filled with theatres that host musicals and plays.

Broadway is a famous street in New York

City. It is also the name of a famous theatre district in New York City. The area is filled with theatres that host musicals and plays.

B.1.4 LLaMA-3-8B Instruct Output

Here are some famous actors who got their start on Broadway:

1. Audrey Hepburn: Began her career on Broadway in the 1950s, performing in shows like "High Button Shoes" and "Ondine."

2. Julia Roberts: Made her Broadway debut in the 1987 revival of "Steel Magnolias" and later starred in "Six Degrees of Separation."

3. Neil Patrick Harris: Got his start on Broadway in the 1990s, performing in shows like "Rent" and "Hedwig and the Angry Inch."

4. Hugh Jackman: Starred in the 1999 Broadway production of "Sunset Boulevard" and later appeared in "Oklahoma!" and "The Boy from Oz."

5. Kristin Chenoweth: Began her career on Broadway in the 1990s, performing in shows like "Steel Pier" and "You're a Good Man, Charlie Brown."

B.2 MT-Bench Example

Question: Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.

B.2.1 LLaMA-2-13B Base Output

This travel blog post should be engaging and informative, providing readers with an inside look at what makes Hawaii a unique and exciting destination. The post should be well-researched and include interesting facts and details about the destination, as well as practical information for travelers. Include a mix of text and images to illustrate the points made in the blog post and make it visually appealing.

B.2.2 LLaMA-2-13B Chat Output

Title: Aloha! Discovering the Heart of Hawaiian Culture and Must-See Attractions

Introduction:

Hawaii, the tropical paradise in the Pacific, is a destination that offers a unique blend of stunning natural beauty, rich cultural heritage, and unparalleled adventure. Recently, I embarked on a journey to explore the islands of Oahu, Maui, and Kauai, and I was left with unforgettable memories and a deeper appreciation for the Hawaiian culture. In this blog post, I'll share my top cultural experiences and must-see attractions that will inspire you to plan your own dream trip to Hawaii.

Cultural Experiences:

1. Attend a traditional Hawaiian luau: A luau is a quintessential Hawaiian experience that offers a glimpse into the island's rich cultural heritage. I attended a luau at the Waikiki Starlight Luau on Oahu, which featured a traditional Hawaiian feast, live music, and authentic hula dancing.
2. Visit a Hawaiian temple: Hawaii is home to numerous temples, or heiaus, that are sacred to the native Hawaiian people. I visited

the Pu'uuhonua o Honaunau National Historical Park on the Big Island, which is a historic temple ground.

3. Learn to hula: Hawaiian dance is an essential part of the island's culture, and I had the opportunity to learn the art of hula from a local dance teacher.

B.2.3 LLaMA-3-8B Base Output

Incorporate SEO strategies, formatting, and a conclusion for a well-rounded piece.

Aloha, fellow travelers! Hawaii is a dream destination, and I'm

here to share my recent trip to the beautiful islands. Join me as we explore the vibrant culture, stunning landscapes, and unforgettable experiences that await in this tropical paradise.

Hawaii is a melting pot of cultures, with influences from Polynesia, Asia, and the United States. As I stepped off the plane, I was greeted by the warm hospitality of the locals and the sweet aroma of freshly caught fish.

B.2.4 LLaMA-3-8B Instruct Output

Island Magic: Unveiling the Hidden Gems of Hawaii

As I stepped off the plane in Honolulu, the warm tropical air enveloped me, and the sweet scent of plumeria flowers filled my senses. My recent trip to the Hawaiian Islands was a dream come true, and I'm excited to share with you the incredible experiences and must-see attractions that left me enchanted.

A Cultural Immersion in Honolulu

My journey began in Honolulu, the capital city of Oahu, where I immersed myself in the rich cultural heritage of the islands. I visited the Bishop Museum, the largest

museum in Hawaii, which showcases the history and traditions of the native Hawaiian people. The museum's collections include an impressive array of artifacts, including ancient Hawaiian tools, traditional clothing, and even a replica of a 19th-century Hawaiian village.

Next, I headed to the Honolulu Fish Auction, where I witnessed the vibrant local fish market in action. The sights, sounds, and smells of the bustling market were a true sensory experience, and I was fascinated by the variety of fresh seafood on display.

C Jensen-Shannon Distance Results

See Table 5.

Corpus 1	Corpus 2	JSD
Nigeria	LLaMA2-base	0.6913
Nigeria	LLaMA2-chat	0.6946
US	LLaMA2-base	0.7062
US	LLaMA2-chat	0.7128
Nigeria	LLaMA3-base	0.6915
Nigeria	LLaMA3-chat	0.7091
US	LLaMA3-base	0.7064
US	LLaMA3-chat	0.7224
Nigeria	US	0.6617
LLaMA2-base	LLaMA2-chat	0.6486
LLaMA3-base	LLaMA3-chat	0.6691
LLaMA2-base	LLaMA3-base	0.6062
LLaMA2-chat	LLaMA3-chat	0.6011

Table 5: Jensen–Shannon distance (JSD) between different corpora. Values rounded to 4 significant figures.

D Downloading English corpora

We downloaded the COCA sample from <https://www.corpusdata.org/formats.asp>. We chose the linear text random sample provided (0.1% of the corpus). We excluded the files on spoken word and TV as their dialogue formatting biases the bigram analysis.

We downloaded ICE Nigeria from <https://sourceforge.net/projects/ice-nigeria/>.

Modeling Changing Scientific Concepts with Complex Networks: A Case Study on the Chemical Revolution

Sofía Aguilar-Valdez and Stefania Degaetano-Ortlieb

Language Science and Technology Department, Saarland University

sofia.aguilar@uni-saarland.de, s.degaetano@mx.uni-saarland.de

Abstract

While context embeddings produced by LLMs can be used to estimate conceptual change, these representations are often not interpretable nor time-aware. Moreover, bias augmentation in historical data poses a non-trivial risk to researchers in the Digital Humanities. Hence, to model reliable concept trajectories in evolving scholarship, in this work we develop a framework that represents prototypical concepts through complex networks based on topics. Utilizing the Royal Society Corpus, we analyzed two competing theories from the Chemical Revolution (phlogiston vs. oxygen) as a case study to show that onomasiological change is linked to higher entropy and topological density, indicating increased diversity of ideas and connectivity effort.

1 Introduction

Language change is not random; it is driven by shifting communicative goals, social structures, and domain-specific conventions (Hamilton et al., 2016b; Gries and Hilpert, 2008; Blank, 2013). While various methods exist to model change in language use, such as lexical semantic change through word meaning representations from LLMs, interpretability remains a difficult task (Periti and Montanelli, 2024). This impedes domain experts' ability to qualitatively assess change and adapt models to their specific needs (Beck, 2024).

We model conceptual change through language use in the context of scientific revolutions, where shifts in language use reflect broader epistemic transitions. Specifically, we want to analyze how the lexemes referring to a concept evolve after discoveries (Zgusta, 2011), which poses the challenges of quantifying concepts with adaptable structures and predicting what conceptual evolution mechanisms are associated with epistemic drifts. To this end, we propose a graph-based approach that models scientific texts as concept networks, organizing

1780

1800

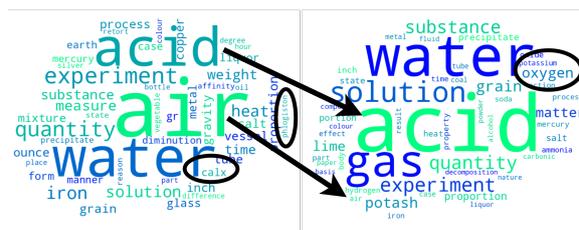


Figure 1: **Diachronic prototypical concepts.** This schematic illustrates movements in central and peripheral readings (arrows), and the presence of notable terms (ovals). Over time, *air* went from the core to the periphery while *acid* did the opposite, and this was accompanied by the removal of *calx*, *phlogiston*, and the emergence of *oxygen*.

a concept's lexemes or readings from core to periphery, and evaluates the recombination of these readings into new core concepts for onomasiological¹ assessment.

As a case study, we analyze two competing explanations for combustion during the Chemical Revolution: phlogiston theory (e.g., “a lighter burns because it contains phlogiston”) and oxygen theory (e.g., “a lighter burns because some substances ignite when in contact with oxygen”). Considering these theories as prototypical concepts (see Figure 1) with central and peripheral readings (Geeraerts, 1997), the overarching research question of this work is concerned with modeling onomasiological change in scientific discourse with temporal graphs. Our central question is whether the network structure reflects the spread and competition of readings in changing scientific concepts.

The rest of the paper is organized as follows.

¹Onomasiological change refers to shifts in the linguistic expressions used to denote the same underlying concept. Instead of tracking how a word's meaning evolves (a semasiological perspective), the onomasiological approach fixes a concept and studies competition, replacement, or specialization among its lexical realizations over time (cf. Geeraerts (2009))

Section 2 motivates the study of concepts, especially scientific ones, and explains how prototype semantics serves as the theoretical framework for representing concepts as networks. Section 3 reviews related work on modeling concept formation using networks and introduces our task of detecting conceptual structure shifts. Section 4 describes the corpus and experiments, and Section 5 reports the results and their interpretation. Finally, Section 6 concludes and provides future research directions.

2 Background

This section begins by explaining why modeling concepts is important, especially in the scientific domain, and presents the linguistic theory from which we approach concepts as graphs. It ends with a brief historical background on the Chemical Revolution.

2.1 Concepts and Networks

Why Concepts Matter Concepts are mental tools designed to manage the (im)balance between expectation and experience (Koselleck, 2006). For instance, the concept of *nature* today is approached as an object rather than subject, which presents interpretations that diverge from those of previous societies. Since concepts are indicators of historical change, tracing their shifts is key to reconstruct how past societies understood and structured their world.

Why Scientific Concepts Unlike everyday concepts (e.g., *chair*), scientific concepts (e.g., *DNA*) cannot be formed by the subject through immediate experience of the object. They are learned through the academic establishment and lose meaning if isolated from this universal system of related ideas since they rely on neither moral judgments, gut feelings nor authority. This is how Vygotsky (1994) explains why scientific concepts are "true concepts" (viz., truly learned, nonspontaneous concepts). Since they are not impacted by cultural ambiguity, as study objects they enable "universal sensors" for historical change.

Why Complex Networks According to prototype semantics, concepts have readings or interpretations located at different levels; going from the center to the periphery, where the centroids are considered to be the most typical (Geeraerts, 1997). To illustrate this, consider how the concept of *fruit* generates a typical reading that varies depending

on the cultural perspective. This structure is akin to a graph's, and the literature supports that representing them as such offers means to model the meaning dynamics as well as social interactions that lead to concept formation (Kaye et al., 2024; Kedrick et al., 2024; Ju et al., 2020).

2.2 Chemical Revolution Overview

When science undergoes dramatic ontological changes (i.e., the whole system of concepts and their laws is replaced by a new one), by analogy with political uprisings, such processes are called scientific revolutions (Kuhn, 2012). The textbook example of this occurred in the 18th century, when competing sides disagreed sharply about what counted as an element and what was a compound, a foundational difference many chemists of the time had (Chang, 2015). Notably, Lavoisier's oxygen paradigm overtook the dominant one in chemistry for combustion, Stahl's phlogiston paradigm; while phlogistonists regarded metals as calx compounds (i.e., today known as oxides) plus phlogiston, oxygenists classified calxes as compounds of pure metals plus oxygen (i.e., today known as element), which cleared out research questions at the time related to not only combustion but calcination and respiration (Thagard, 1990), producing the ground work for many additional transformations in chemistry (Holmes, 1994). Arguably, phlogistonists' contributions were hindered by interpreting phenomena using phlogiston as a principle or "central reading", an ontological category that disappeared as Lavoisier's system developed (Chang, 2015).

Hence, this scientific scenario is optimal as a case study to model how concepts are changed, new ones arise, and old ones disappear. In particular, we focus on the decades where the new concept of oxygen became established while phlogiston vanished: from the 1750s to 1800s (Chang, 2011; Thagard, 1990). This scenario, however, only serves as a first step towards a generalizable approach to model conceptual transitions and the meaning dynamics involved in these changes.

3 Related Work

This section introduces the state-of-the-art methods for the NLP task most related to our problem (i.e., lexical-semantic change detection), the current challenge of improving interpretability, and how we approach it in this work as a conceptual structure shift detection task.

3.1 NLP Research Overview

The evolution of word meanings, often termed semantic change, has long been a central focus of linguistic inquiry. This process is crucial for understanding how language adapts to shifting cultural and communicative contexts. Advances in computational methods have provided novel tools to study semantic change at scale. Tahmasebi et al. (2021) and Periti and Montanelli (2024) offer comprehensive surveys on these methods, with a focus on diachronic word embeddings and neural methods. Diachronic embeddings capture the shifting meanings of words across time by mapping their usage in successive temporal intervals (Hamilton et al., 2016b; Dubossarsky et al., 2017). While such approaches effectively track semantic drift, they often fall short in allowing to understand the underlying mechanisms driving these changes.

Modeling evolving scholarship considering concept networks offers an alternative, more informative perspective. By building semantic graphs that organize a concept’s core and peripheral readings, Kedrick et al. (2024) found conceptual structures with rigid cores hinder scientific innovation. They explain this using the network structure: the recombination of cores with fellow cores and peripheral entities is less likely to occur in disciplines where core concepts are more stable. Such framing is compatible with prototype semantics, which describes words as radial structures (e.g., word *fruit* with *apple* and *guava* as central and peripheral readings, respectively) where depending on the situational context some readings are more "central" or "typical" than others, boundaries are not rigid, and different concepts can overlap (Geeraerts, 1997).

Networks are thus informative in the sense that they can represent the meaning dynamics taking place alongside the social interactions driving them. By constructing community networks based on a topic model, Malaterre and Lareau (2023) showed the evolution of communities and their main research themes, tracing the development of the specialization areas that structure the field of philosophy of science today. Topic models are known for approaching documents as mixtures of abstract themes, where each theme is represented by a probability distribution over words (Blei et al., 2003). Document-topic distributions thus offer insight into how specific themes are distributed across time, and to our knowledge this approach was first implemented by Griffiths and Steyvers (2004) to find

scientific disciplines and evaluate collaborations between researchers.

3.2 Towards a Conceptual Structure Shift Detection Task

We approach concepts as prototypical structures whose meaning evolves through their relational embedding in domain-specific discourse. The conceptual structure shift detection task therefore frames conceptual change as a network-level phenomenon rather than a purely lexical one.

- **Input:** A diachronic scientific corpus (e.g., spanning multiple decades of publications from one or more research domains).
- **Output:** Concept networks for each time slice that capture core–periphery organization and their diachronic trajectories.

In the context of scientific revolutions, these trajectories should reflect conceptual structure shifts that we define as *periods where the prototypical core of a concept either reconfigures or expands towards new peripheries, resulting in multiple meaning shift cases*, namely

- the lexemes referring to the core diversify (i.e., **onomasiological change**),
- new overall senses for the core emerge (i.e., **semasiological change**),
- and/or a new ontological system replaces the previous one (i.e., **ontological change**).

None of these changes are exclusive and arguably happen in sequence (e.g., during the Chemical Revolution, onomasiological/semasiological changes led to the ontological one). Moreover, to explain these cases systems are expected to produce both static models per period and dynamic mappings across periods, **prioritizing interpretability in structural metrics**.

Hence, to alleviate interpretability concerns we estimate concepts as themes with a ranked vocabulary using topic modeling, then cluster the documents based on similarity. Doing so produces graphs with topic clusters that represent the concepts present in a decade, and how these concepts were structured in terms of core and peripheral words.

Regarding evaluation, this baseline adopts two established metrics from information theory to measure change:

- **Jensen-Shannon distance**, to quantify the difference across topic clusters by treating graphs as diachronic, referent-independent entities (i.e., a perspective looking from the future into the past and vice versa have a distance that is symmetrical).
- **Entropy**, to describe the magnitude of the change understood as "topic diversity" according to Hall et al. (2008).

In summary, we approach concepts as prototypical structures and define our task as conceptual structure shift detection: the input is a diachronic scientific corpus, and the expected output are concept networks that offer core and peripheral readings trajectories to inform the **onomasiological change assessment**, which is the case we decided to focus on for this work.

4 Material and Methods

Based on the Royal Society Corpus (RSC; Kermes et al. 2016; Fischer et al. 2020), this approach consisted of four stages: first, create a representative corpus by filtering publications given specified decades and oxygen terminology (Chang, 2011; Bizzoni et al., 2021). Second, represent concepts and their diachronic readings using topic modeling (Blei et al., 2003; Sievert and Shirley, 2014). Third, form concept networks by linking documents based on the Jensen-Shannon distance (Shannon, 1948) among topics and optimize connectivity by minimizing the percolation threshold (Radicchi, 2015); this is the tipping point where isolated nodes in a graph abruptly form the largest mutually reachable subgraph, with a low threshold indicating extensive connectivity using minimal resources. Finally, we reveal hidden groups by applying clustering (Lukasová, 1979) and community detection (Blondel et al., 2008) algorithms.

Dataset Description The RSC is a diachronic corpus of English scientific writing covering 47,837 publications and their metadata (e.g., publication year, author) in the Philosophical Transactions and Proceedings of the Royal Society of London from 1665 to 1996 (Fischer et al., 2020). We focus on 6 decades (1750-1800) to account for the events where oxygen developed as a concept; from 1774, when Joseph Priestley conformed to the theories of the time and introduced the gas as "dephlogisticated air" (Brown, 1997), till 1789, when Lavoisier presented the gas with a new term and led to the

widespread adoption of "oxygen" among chemists (Thagard, 1990). Doing so is supported by previous studies based on the RSC that in the same period report documents and lexico-grammatical changes addressing combustion, gases and chemical reactions, key themes in the phlogiston-to-oxygen conceptual transition (Degaetano-Ortlieb and Teich, 2019; Bizzoni et al., 2020; Teich et al., 2021).

4.1 Data Processing and Sampling

We retrieved texts from the RSC with a publication date between 1750–1800, defined decade batches to make our results comparable with previous studies on this corpus (Teich et al., 2021), then filtered documents to include instances of terms relevant to the discovery of oxygen. These terms were derived from Bizzoni et al. (2021), which used Kullback-Leibler Divergence (KLD) to identify distinctive terms during the Chemical Revolution; and Chang (2011), who explains why the term *oxygen* persisted despite theoretical revisions, arguing that its operational meaning remained stable while phlogiston theory lost operational grounding. Chang (2011) also notes that the theoretical meaning of phlogiston was conceptually close to later notions of energy, and describes phlogiston as a precursor to ideas such as free electrons.

After filtering the documents ($N = 2337$), we performed two cleaning rounds. First, we removed special characters and terms shorter than two characters, converted the text to lowercase, and removed stopwords based on the standard list imported as "ENGLISH_STOP_WORDS" using scikit-learn (Buitinck et al., 2013); doing so substantially decreased the corpus size by removing prepositions, pronouns and articles (e.g., "a", "he" and "the") while retaining nouns, which according to Hamilton et al. (2016a) are more sensitive to linguistic changes. Regarding hyphenated terms, these were unchanged to preserve their semantic unity during topic modeling. Second, we filtered for nouns and adjectives, and lemmatized the corpus using tmtoolkit².

To test the influence of prior textual data on downstream analyses, for the six decades we implemented two sampling strategies: 1) cumulative sampling, where each decade includes the indicated decade and all previous decades' documents, and 2) non-cumulative sampling, where each decade includes only documents from that period.

²<https://tmtoolkit.readthedocs.io/en/latest/>

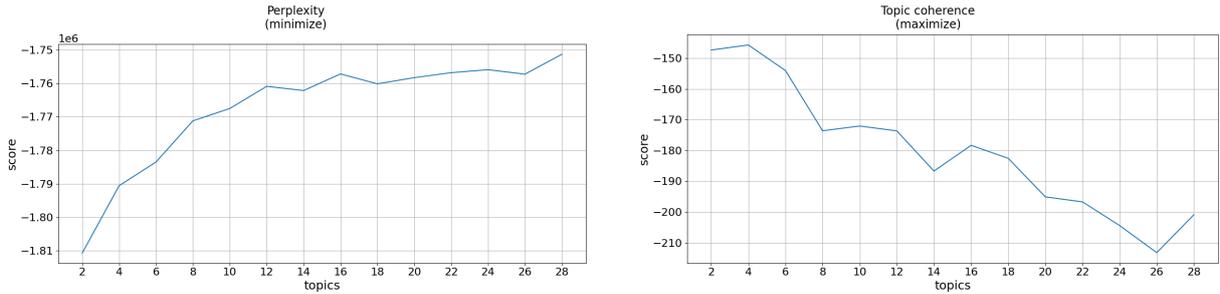


Figure 2: **Topic models evaluation.** These results, produced by evaluating models for the 1800s non-cumulative corpus, were consistent across decades and strategies. Since both metrics degrade beyond 6 topics (\uparrow perplexity, \downarrow coherence), the optimal number is 6.

4.2 Topic Modeling

In the context of the RSC, topic modeling has previously been used to estimate scientific disciplines (Fankhauser et al., 2016). The authors employed Latent Dirichlet Allocation (LDA), a probabilistic model designed to uncover abstract categories given a collection of documents (Blei et al., 2003), and found 30 distinct disciplines and sub-disciplines to describe the publications.

Our topic modeling workflow consisted of constructing a document-term matrix from the preprocessed corpus, counting the occurrences of each lemmatized term per document. We evaluated model performance using log-likelihood (Bengio et al., 2003) and topic coherence (Mimno et al., 2011); doing so mitigates the often manual and arbitrary parameter selection in LDA (Schmidt, 2012; Mohr and Bogdanov, 2013). Following best practices in interpretability, we generated topic labels using LDAvis (Sievert and Shirley, 2014), which are determined based on relevance scores that are more meaningful compared to word rankings (i.e., if *air* is the top-5 word yet its relevance score is higher compared to other terms, this will be the topic label). This is how we defined the optimal number of topics and the topic labels (i.e., the term that "names" the topic) across decades and strategies.

4.3 Network Construction

We considered documents as nodes and formed edges based on the Jensen-Shannon distance (Shannon, 1948) between document-topic distributions; given distributions p and q , distance d is given by:

$$d = \sqrt{\frac{D(p \parallel m) + D(q \parallel m)}{2}}$$

where m is the pointwise mean of p and q and D is the KLD. This way we formed adjacency matrices,

later binarized by optimizing connectivity based on the percolation threshold (Radicchi, 2015). A threshold was necessary to reduce the network's complexity by limiting the number of nodes and edges to only the most meaningful ones.

Hence, the conceptual structure was defined as follows: topic labels (Sievert and Shirley, 2014) "name" the concepts, while the topic-word distributions establish the central (i.e., top-10) and peripheral (i.e., top-50) terms.

4.4 Topic Aggregation

To find hidden groups, first we applied Hierarchical Agglomerative Clustering (HAC; Lukasová, 1979), a method that consists of measuring the distance between data points iteratively until forming groups with data points that exhibit higher similarity to one another compared to points in other groups. Each data point starts as its own cluster and merges with others until a number of groups is reached (i.e., 6 according to the optimal number of topics). For visualization purposes, we used cosine similarity as the metric for both HAC and UMAP (McInnes et al., 2018) dimensionality reduction approaches due of its reported performance in topic-based networks (Luhmann and Burghardt, 2022).

Similarly, the Louvain algorithm (Blondel et al., 2008) estimates groups by maximizing modularity, a metric that measures the quality of a network's division into distinct groups or communities, quantifying the density of intra- vs. inter-cluster connections. We used it as a complementary method to further analyze network dynamics among the topics (Malaterre and Lareau, 2023).

5 Results and Discussion

This section describes our findings regarding the phlogiston-to-oxygen conceptual transition: a topic cluster indicates onomasiological change given its

underplayed acidity as a secondary effect of phlogiston release (e.g., experiments burning sulfur produced a "gaseous calx" that was perceived as acidic since it lacked phlogiston), Lavoisier considered acidity as a core explanatory principle to reframe calxes as oxides and oxygen as an enabler of acid formation (e.g., sulfur is a combustible that when in contact with oxygen forms sulfur dioxide, which in turn dissolves in water to produce sulfurous acid). This is explicit in the results, where *calx* and *phlogiston* disappear from the periphery in the concept structure after *acid* takes over the core (cf. notable terms in Figure 1).

To compare the magnitude of diachronic change in both sampling strategies, we measured the topics "diversity of ideas" using entropy (Hall et al., 2008). Figure 4 presents different trends: while the cumulative strategy shows rising entropy across oxygen-related topics (illustrated by different colors), the non-cumulative declines then rises post-1774 oxygen discovery. This decline of diversity pre-1780 could be interpreted as the resistance of phlogistonists to de-center *air* in their findings (e.g., Priestley's "dephlogisticated air"), while the rise of diversity post-1780 could be attributed to Lavoisier coining the term *oxygen*. These observations support Thagard (1990)'s on why Lavoisier made the conceptual shift instead of Priestley. However, to clarify this, future studies will have to incorporate Lavoisier's articles in French, which are absent from our corpus.

5.2 Temporal Graphs

Considering the onomasiological change (Figure 3) and rising entropy (Figure 4), we built graphs based on the non-cumulative strategy; one for each decade, producing 6 in total (Figure A1).

To analyze network stability, we considered 5 parameters (Figure A2) from which follow three findings: first, although node size starts and ends similarly (400–500), edge density more than doubled (30 037–62 639). Second, declining network communities suggests integration (5–3), but falling modularity (0,39–0,19) disputes this: fewer, lower-quality communities indicate increased mixing. Lastly, percolation threshold rises in 1760s (0,3–0,54), falls in 1770s (0,54–0,32), then rises (0,32–0,58)—aligning with modularity.

Low threshold/high modularity signals efficient connectivity given well-defined communities (e.g., communication is faster when people are organized in groups and allotted resources are suffi-

cient), while high threshold/low modularity signals mixed communities requiring high connectivity effort (e.g., mixed groups hinder person-finding, requiring more channels and impeding communication).

6 Conclusion and Future Work

This work case-studies conceptual change during the Chemical Revolution (i.e., phlogiston-to-oxygen) via interpretable complex networks. Our analysis reveals two findings: 1) a topic cluster shifts from *air* to *acid*, showing onomasiological change given the movement of core/peripheral readings accompanied by higher entropy, and 2) network communities and modularity decline, while percolation and edge density rise, indicating higher connectivity effort. These results indicate that changing concepts form high entropy clusters (i.e., more variability of ideas) that increase topological density in the knowledge network (i.e., more effort is needed for community finding and message passing), potentially hindering communication. Future work will refine representation of rare terms and add directed edges (e.g., colexification, citations linked to terms) to account for semantic and social directionality, and predict inter-temporal graph links via node/edge embeddings to identify influential documents and authors.

Limitations

Our corpus so far is limited to English scientific texts, which does not account for the whole spectrum of ideas exchanged during the Chemical Revolution. Adding multilingual and multi-genre texts is a desiderata to work towards a comprehensive picture. In this study, we derived core and peripheral terms from topic-word distributions rather than cluster centroids, potentially overlooking document embedding geometries. Rare peripheral terms receive sparse representations, potentially missing subtle shifts. As an approach that builds undirected graphs, the formed edges overlook semantic directionality; representing this is important to understand how concepts evolve (e.g., through meaning gain or loss), making of the conceptual structure a phylogenetic tree (Carling et al., 2023). Future work will address these challenges via directed graphs and rare-term augmentation, as well as the estimation of phylogenetic structures. Moreover, to test the generalizability of the approach, first steps will include modeling conceptual transitions

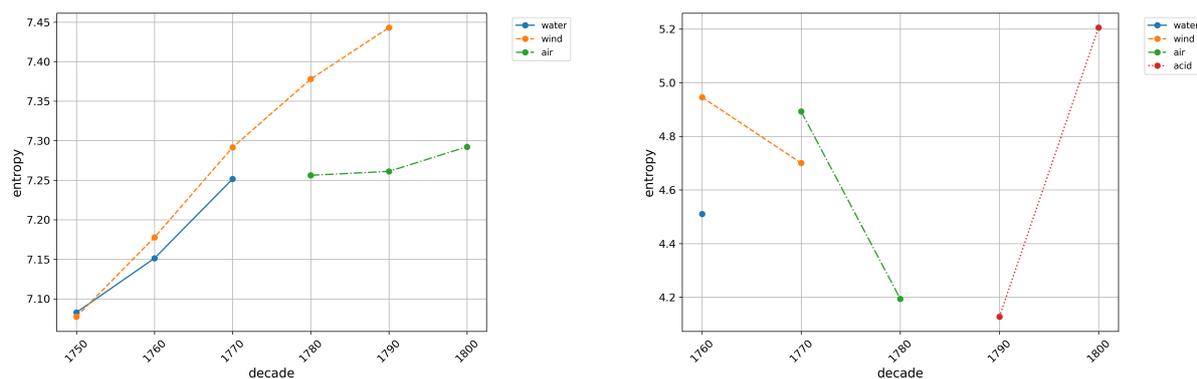


Figure 4: **Topic entropy over time.** While in cumulative sampling (left) entropy trajectories of oxygen-related terms increase over time, non-cumulative sampling (right) shows a decline then rise during the development of oxygen theory (cf. 1772-1789 in Thagard (1990)).

across disciplines (e.g., germ plasm to chromosomal genes in biology).

Acknowledgments

Funded by the European Union under grant 101119511. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

- Christin Beck. 2024. [Review of Tahmasebi, Borin, Jantowt, Xu & Hengchen \(2021\): Computational Approaches to Semantic Change](#). *Journal of Historical Linguistics*, 14(2):376–384.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.
- Yuri Bizzoni, Stefania Degaetano-Ortlieb, Peter Fankhauser, and Elke Teich. 2020. [Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach](#). *Frontiers in Artificial Intelligence*, 3.
- Yuri Bizzoni, Stefania Degaetano-Ortlieb, Katrin Menzel, and Elke Teich. 2021. [The diffusion of scientific terms – tracing individuals’ influence in the history of science for English](#). In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 120–127, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.
- Andreas Blank. 2013. [Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change](#). In Andreas Blank and Peter Koch, editors, *Historical Semantics and Cognition*, pages 61–90. De Gruyter Mouton.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.
- Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. [Fast unfolding of communities in large networks](#). *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.
- Lee K. Brown. 1997. [“Dephlogisticated Air” Revisited: Oxygen Treatment for Central Sleep Apnea Syndrome](#). *CHEST*, 111(2):269–271. Publisher: Elsevier.
- Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD workshop: Languages for data mining and machine learning*, pages 108–122.
- Gerd Carling, Sandra Cronhamn, Olof Lundgren, Victor Bogren Svensson, and Johan Frid. 2023. [The evolution of lexical semantics dynamics, directionality, and drift](#). *Frontiers in Communication*, 8.
- Hasok Chang. 2011. [The Persistence of Epistemic Objects Through Scientific Change](#). *Erkenntnis*, 75(3):413–429.
- Hasok Chang. 2015. [The Chemical Revolution revisited](#). *Studies in History and Philosophy of Science Part A*, 49:91–98.
- Stefania Degaetano-Ortlieb and Elke Teich. 2019. [Toward an optimal code for communication: The case of scientific English](#). *Corpus Linguistics and Linguistic Theory*, 18(1):175–207.
- Haim Dubossarsky, Daphna Weinsall, and Eitan Grossman. 2017. [Outta control: Laws of semantic change and inherent biases in word representation models](#).

- In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark. Association for Computational Linguistics.
- Peter Fankhauser, Jörg Knappen, and Elke Teich. 2016. [Topical Diversification Over Time In The Royal Society Corpus](#).
- Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. [The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 794–802, Marseille, France. European Language Resources Association.
- Dirk Geeraerts. 1997. *Diachronic Prototype Semantics: A Contribution Historical Lexicology*. Oxford University Press.
- Dirk Geeraerts. 2009. *Theories of Lexical Semantics*. Oxford University Press.
- Stefan Gries and Martin Hilpert. 2008. [The identification of stages in diachronic data: Variability-based neighbour clustering](#). *Corpora*, 3:59–81.
- Thomas L. Griffiths and Mark Steyvers. 2004. [Finding scientific topics](#). *Proceedings of the National Academy of Sciences*, 101(suppl_1):5228–5235. Publisher: Proceedings of the National Academy of Sciences.
- David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the History of Ideas Using Topic Models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 363–371, Honolulu, Hawaii. Association for Computational Linguistics.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. [Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. [Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.
- Frederic L. Holmes. 1994. What was the chemical revolution about. In *Abstracts of Papers of the American Chemical Society*, volume 208, pages 13–HIST, USA. American Chemical Society.
- Harang Ju, Dale Zhou, Ann Sizemore Blevins, David M. Lydon-Staley, Judith Kaplan, Julio Roberto Tuma, and Danielle S Bassett. 2020. [The network structure of scientific revolutions](#).
- Aleksandra Kaye, Raphael Schlattmann, Malte Vogl, Bernardo S. Buarque, Jascha Schmitz, Lea Weiß, and Laura von Welczeck. 2024. [Socio-Epistemic Networks: A Framework for History of Knowledge](#). In Charlotte A. Lerg, Johan Östling, Jana Weiß, Anne Kwaschik, and Claudia Roesch, editors, *History of Intellectual Culture 3/2024: Experimental Spaces: Knowledge Production and Its Environments in the Long Nineteenth Century*, pages 253–274. De Gruyter Oldenbourg.
- Kara Kedrick, Ekaterina Levitskaya, and Russell J. Funk. 2024. [Conceptual structure and the growth of scientific knowledge](#). *Nature Human Behaviour*, 8(10):1915–1923. Publisher: Nature Publishing Group.
- Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2016. [The Royal Society Corpus: From Uncharted Data to Corpus](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1928–1931, Portorož, Slovenia. European Language Resources Association (ELRA).
- Reinhart Koselleck. 2006. Die Verzeitlichung der Begriffe. In Ulrike Spree and Willibald Steinmetz, editors, *Begriffsgeschichten: Studien Zur Semantik Und Pragmatik Der Politischen Und Sozialen Sprache*, Suhrkamp Taschenbuch Wissenschaft, page 85. Suhrkamp, Frankfurt.
- Thomas Kuhn. 2012. *The Structure of Scientific Revolutions: 50th Anniversary Edition*. University of Chicago Press, Chicago, IL.
- Jan Luhmann and Manuel Burghardt. 2022. [Digital humanities—A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape](#). *Journal of the Association for Information Science and Technology*, 73(2):148–171.
- Alena Lukasová. 1979. [Hierarchical agglomerative clustering procedure](#). *Pattern Recognition*, 11(5-6):365–381. Publisher: Elsevier.
- Christophe Malaterre and Francis Lareau. 2023. [Identifying Hidden Communities of Interest with Topic-based Networks: A Case Study of the Community of Philosophers of Science \(1930-2017\)](#). In *Hawaii International Conference on System Sciences*.
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. [UMAP: Uniform manifold approximation and projection](#). *Journal of Open Source Software*, 3(29):861.
- David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP ’11*, pages 262–272, USA. Association for Computational Linguistics.

- John W. Mohr and Petko Bogdanov. 2013. [Introduction—Topic models: What they are and why they matter](#). *Poetics*, 41(6):545–569.
- Francesco Periti and Stefano Montanelli. 2024. [Lexical Semantic Change through Large Language Models: a Survey](#). *ACM Comput. Surv.*, 56(11):282:1–282:38.
- Filippo Radicchi. 2015. [Predicting percolation thresholds in networks](#). *Physical Review E*, 91(1):010801.
- Benjamin M Schmidt. 2012. [Words Alone: Dismantling Topic Models in the Humanities](#). *Journal of Digital Humanities*, 2(1):49–65.
- Claude Elwood Shannon. 1948. [A mathematical theory of communication](#). *The Bell System Technical Journal*.
- Carson Sievert and Kenneth Shirley. 2014. [LDAvis: A method for visualizing and interpreting topics](#). In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70.
- Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen. 2021. [Computational approaches to semantic change](#). Language Science Press, Berlin.
- Elke Teich, Peter Fankhauser, Stefania Degaetano-Ortlieb, and Yuri Bizzoni. 2021. [Less is More/More Diverse: On The Communicative Utility of Linguistic Conventionalization](#). *Frontiers in Communication*, 5.
- Paul Thagard. 1990. [The Conceptual Structure of the Chemical Revolution](#). *Philosophy of Science*, 57(2):183–209. Publisher: [The University of Chicago Press, Philosophy of Science Association].
- Lev Semenovič Vygotsky. 1994. The Development of Thinking and Concept Formation in Adolescence. In René van der Veer, editor, *Vygotsky Reader*, pages 185–266. Blackwell, Oxford.
- Ladislav Zgusta. 2011. Onomasiological change: Sachen-change reflected by Wörter. In Edgar C. Polomé, editor, *Research Guide on Language Change*, pages 389–398. De Gruyter Mouton.

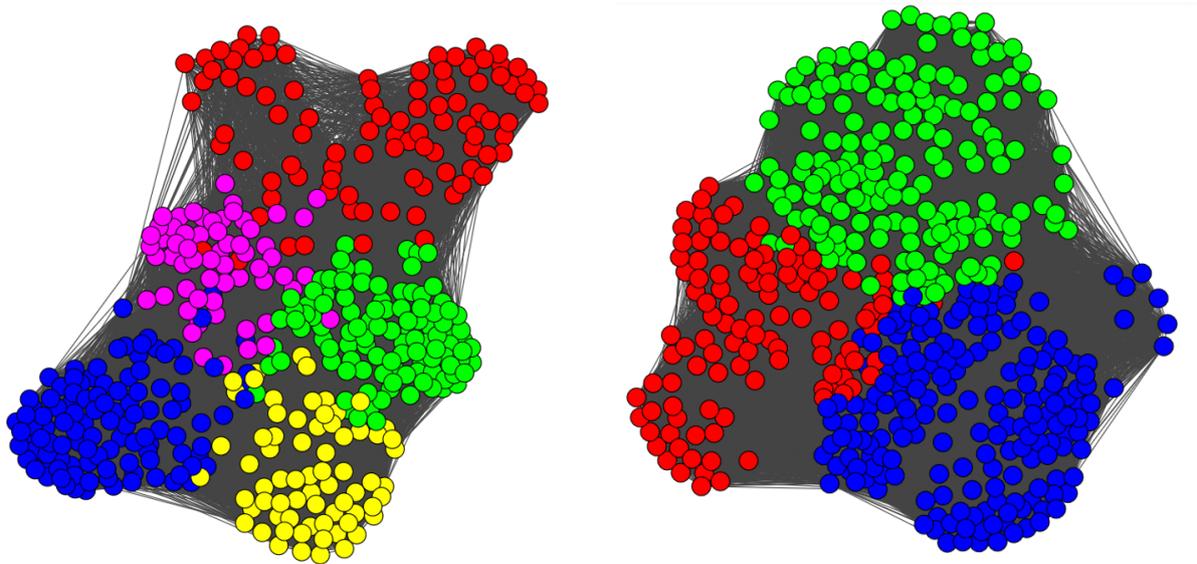


Figure A1: **Temporal graphs.** The colors illustrate the number of communities which went on decline: starting at 5 (1750s) and ending at 3 (1800s).

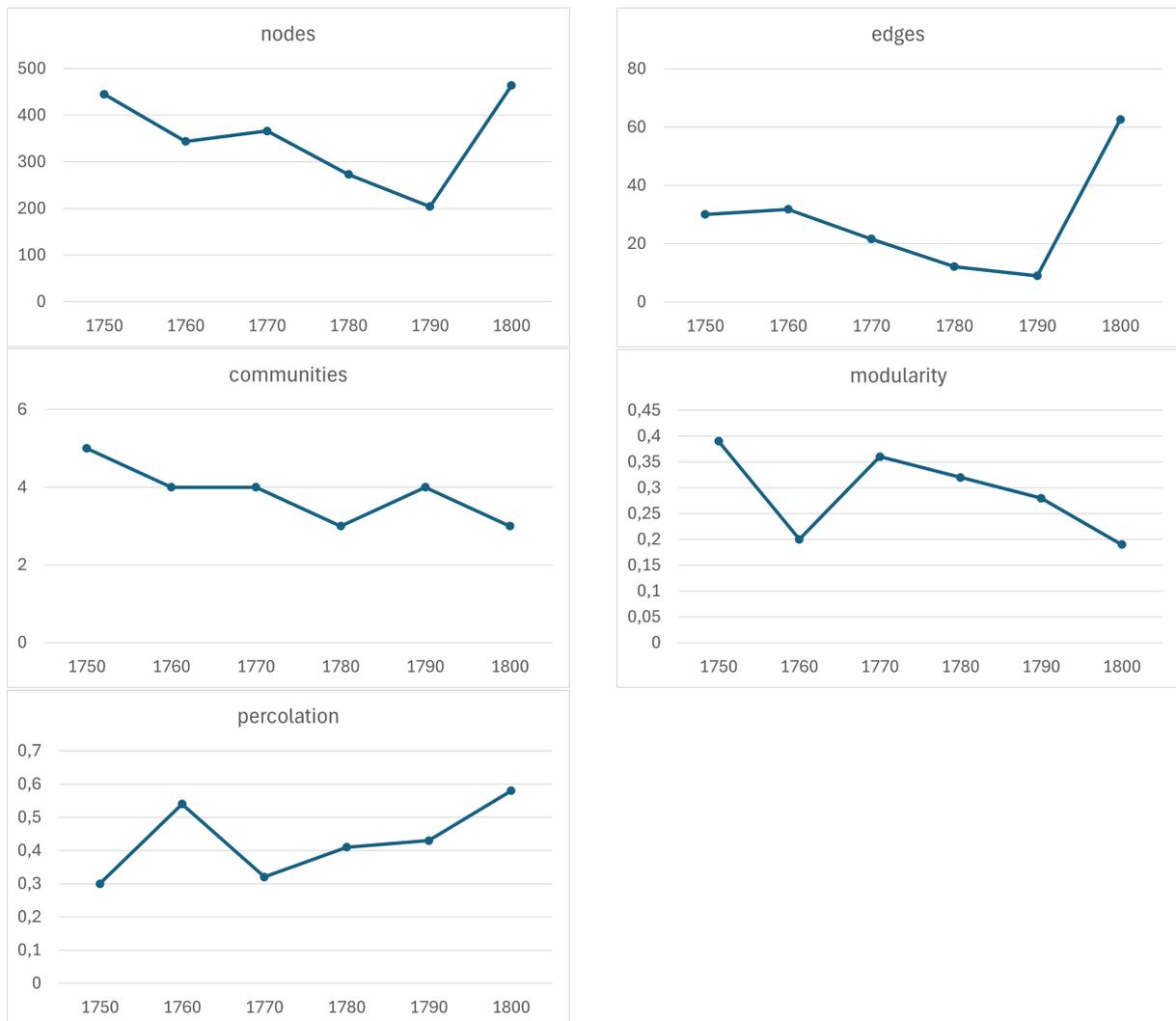


Figure A2: **Network metrics.** We used five parameters to interpret network stability over time: nodes size, edge density (where $y = 1 \times 10^3$, e.g., last number of edges reported is >60 000), communities count, modularity and percolation threshold.

Speaking on Their Behalf: Detecting Indirect Speech in Historical Danish and Norwegian Texts

Ali Al-Laith¹, Alexander Conroy¹, Kirstine Nielsen Degn¹,
Jens Bjerring-Hansen¹ and Daniel Hershcovich²

Department of Nordic Studies and Linguistics, University of Copenhagen¹

Department of Computer Science, University of Copenhagen²

alal@di.ku.dk, alc@hum.ku.dk, knd@hum.ku.dk,
jbh@hum.ku.dk, dh@di.ku.dk

Abstract

Indirect speech is a fundamental yet understudied form of reported speech that plays a crucial role in literary texts and communication. While direct speech detection has received significant attention in computational linguistics, the automatic identification of indirect speech remains a challenge due to its nuanced linguistic structure and contextual dependencies. This paper focuses on the detection of indirect speech in late 19th-century Scandinavian literature, where its presence has been linked to shifting aesthetic ideals. We present an annotated dataset of 150 segments, each randomly selected from 150 different novels, designed to capture indirect speech in Danish and Norwegian literature. We evaluate four pre-trained language models for classifying indirect speech, with results showing that a Danish Foundation Model (DFM Large), trained on extensive Danish data, has the highest performance. Finally, we conduct a classifier-assisted quantitative corpus analysis and find that the prevalence of indirect speech exhibits fluctuations over time.

1 Introduction

The way speech is rendered in writing shapes everything from how we interpret literary texts to everyday communication. Reported speech has been described as essential to human society, with direct speech being a universal feature and non-direct constructions also highly frequent (Goddard and Wierzbicka, 2018). Consequently, automatic detection of speech in written text is a fundamental challenge in linguistic analysis and has applications in various fields, including epidemiology (Klein et al., 2020), communication studies, and journalism (Newell et al., 2018). This paper focuses on a particular non-direct construction: indirect speech. Indirect speech is a way of reporting the utterance of someone else, typically without quoting it verbatim and with adjustments to verb

tense, pronouns, and adverbials to reflect the reporter’s perspective (Aarts, 2014). While direct speech identification has received significant computational attention, indirect speech remains comparatively understudied. Our empirical focus is on Scandinavian literature from the late 19th century, where indirect speech has been analyzed only for a limited number of authors (Brix, 1911). Moreover, it has been argued that the presence of indirect speech conflicts with certain aesthetic ideals of the time (Kristensen, 1955), making its automatic detection a valuable tool for reexamining Scandinavian literary history. The code and dataset used in this research are available in an anonymous repository for review purposes: <https://github.com/mime-memo/IndirectSpeech>.

2 Related Work

Indirect speech is common in both spoken and written language, shaping how we interpret the content, connotations, and reliability of an utterance. Linguistic and psychological research highlights that the choice between indirect and direct speech significantly affects how we perceive, recall, and process reported statements (Eerland and Zwaan, 2018). However, distinguishing indirect speech from related phenomena is challenging in both spoken and written forms. As a result, we rely on contextual cues such as pronouns, verb tense, discourse particles, exclamation marks, and emotives (Eckardt, 2021). This complexity requires careful annotation to produce well-performing models. Semino and Short (2004a) demonstrate how corpus stylistics can systematically analyze patterns of speech, writing, and thought presentation across large bodies of English texts, bridging quantitative corpus methods with qualitative literary analysis.

Although computational research in this area remains limited, some studies have explored related approaches. Krestel et al. (2008) introduced a Re-

ported Speech Tagger for the GATE framework, demonstrating an effective approach to automatically annotating reported speech in newspaper articles. Similarly, [Asr et al. \(2021\)](#) have successfully measured reported speech in the news media as part of its investigation into the gender representation gap. However, both studies classify all reported speech instances without distinguishing between direct and indirect speech. [Pareti et al. \(2013\)](#) conducted the first large-scale study on indirect speech and mixed quotation extraction. Their findings indicated that traditional machine learning methods, such as the Maximum Entropy Classifier and Conditional Random Fields, were less effective in predicting indirect quotations compared to direct ones. Furthermore, [Kathirgamalingam et al. \(2023\)](#) evaluated three off-the-shelf tools — CoreNLP ([Manning et al., 2014](#)), QSample ([Scheible et al., 2016](#)), and rsyntax ([Welbers et al., 2021](#)) — across two data sources: news articles and social media communication. Their results aligned with previous research, confirming that indirect speech is more challenging to detect automatically than direct speech. Regarding literary studies specifically, [Muzny et al. \(2017\)](#) developed a deterministic sieve-based system for quote attribution, which effectively classifies their three example novels. However, the focus is primarily on who is speaking rather than how the speech is reported. [Brunner et al. \(2020\)](#) analyzed a corpus of German fictional and non-fictional texts from the 19th century and the early 20th century, demonstrating that BERT-based models outperformed models trained within the Flair framework in detecting indirect speech. In Scandinavian Studies, computational research has so far focused exclusively on direct speech, as seen in studies such as [Stymne \(2024\)](#) and [Al-Laith et al. \(2025\)](#). This paper is therefore the first to examine indirect speech in Scandinavian literary history.

3 Dataset

3.1 Main Corpus

We use the MeMo corpus ([Bjerring-Hansen et al., 2022](#)), consisting of 859 Danish and Norwegian novels (64M+ tokens) from the last 30 years of the 19th century.¹ We refer to this corpus as the ‘main corpus’. It should be noted that, until 1907, written Norwegian was practically identical to written

¹Released with Creative Commons Attribution 4.0 license: <https://huggingface.co/datasets/MiMe-MeMo/Corpus-v1.1>.

Danish ([Vikør, 2022](#)).

3.2 Speech Corpus

Segment extraction. To address the low frequency of indirect speech in our main corpus, we use a linguistically informed regular expression targeting communication verbs followed by a complementizer as a seed pattern to extract candidate passages. Table 1 shows the regular expressions used for extracting occurrences of indirect speech as described in §3.2.

<pre># Regex: [word != ""]* [word = "(sige fortælle spørge påstå tro)r (sagde fortalte spurgte påstod nævned troede) (svare indrømme bemærke forklare understrege tilføj bekræfte erklære anmode hævde advare)(r de) (men nævn forlang råb)(er te)"] []0,12 [word = ",",]0,1 [word = "at (hvem hvad hvilke hvorledes hvor hvornår hvordan hvorfor)"] [word != ""]* [word = ""]</pre>

Table 1: Regular expressions used in the data extraction.

This method ensures sufficient positive examples. From 150 randomly selected novels, we retrieve three consecutive paragraphs surrounding a randomly selected seed pattern match. This sampling method ensures the inclusion of instances of indirect speech, provides sufficient context around the sentences, and avoids canonical bias in favor of a broader repertoire of authors, genres, and literary styles.

Annotation guidelines. To address the challenges described in §2, we develop clear annotation criteria to ensure consistency and accuracy in identifying speech-related elements:

1. **Indirect Speech (“IS”):** All words and punctuation that are part of indirect speech are labeled as “IS”. We do not differentiate embedded speech (e.g., quotations within speech) within passages of indirect speech. For the purposes of this annotation task, we understand indirect speech as a way of reporting speech by using an introductory report verb (e.g. say, ask, tell) and a subordinate clause, for example: “*Anna asked if Kramer could speak with her*” or “*Jørgen suggested that they should leave.*” Contrary to direct speech,

which repeats the used words verbatim, indirect speech typically involves changes to the original speaker’s words, such as adjustments of pronouns, time and place adverbials, and verb tenses to reflect the perspective of the reporter (Aarts, 2014). We note that broader theories of speech representation, such as Semino and Short (2004b), describe a continuum from direct to indirect forms with several intermediate and fuzzy categories; in this work, however, we restrict our annotation to cases involving explicit reporting verbs to enable reliable annotation.

2. **Direct Speech (“DS”)**: All words and punctuation that are part of direct speech are labeled as “DS”. We again do not differentiate embedded speech (e.g., quotations within speech) as both the outer and inner quotations are labeled as “DS”.
3. **Speech Marker (“SM”)**: Any typographical markers indicating speech, such as quotation marks, colons, or dashes, are labeled as “SM”. If a colon appears directly before quotation marks, it is also labelled “SM”.
4. **Speech Tag (“ST”)**: Speech tags (or inquit phrases), such as “he said,” “she asked,” or “they replied,” are labeled as “ST”. This label applies only to the verb phrases and subject, excluding any adverbs or adverbial phrases, e.g., in *And then he whispered almost inaudibly* only “he whispered” is labeled as “ST”. Punctuation immediately preceding or following the tag within the same sentence is also considered part of the “ST” if it is not eligible to be marked as “SM”.
5. **Other (“O”)**: All other words and punctuation not categorized under the above labels are marked as “O”. This ‘Other’ category is not a coherent class but rather a collection of different narratological categories (e.g., narrated inner monologue, free indirect discourse, narratorial comment), which are less clearly and consistently defined in the research literature. Therefore, we focused on speech-related elements to ensure a clear annotation scheme with solid theoretical grounding, even though this choice inevitably results in a large and heterogeneous ‘other’ category. Our aim was not for the model to perform well on this cat-

Class	#Words	%
Indirect Speech (“IS”)	537	1.70%
Direct Speech (“DS”)	14,010	44.17%
Other (“O”)	14,962	47.19%
Speech Marker (“SM”)	1,083	3.42%
Speech Tag (“ST”)	1,115	3.52%
Total	31,707	100%

Table 2: Distribution of annotated dataset.

egory, but rather to achieve high accuracy in classifying indirect and direct speech.

Annotation process. The annotation is conducted on the INCEpTION platform (Klie et al., 2018) by three scholars with domain expertise in late 19th century Scandinavian literature. The annotation is done on a token level. For agreement calculation and in order to obtain a high-quality testing set, we select 20% of the samples for multiple annotation by all three experts. These consist of 30 random segments from each year.

Annotation results. Annotation results show that most words fall under “Other” (47.19%), while direct speech (“DS”) accounts for 44.17%, highlighting the prominence of dialogue. However, due to our extraction method—using a regular expression to target communication verbs—DS is likely overrepresented compared to its actual share in the main corpus, previously measured at 35% (Al-Laith et al., 2025). Indirect speech is rare (1.70%), while “Speech Marker” (“SM”) and “Speech Tag” (“ST”) are unsurprisingly low (3.42% and 3.52%), given their dependence on speech and minimal token length. This distribution reflects the dataset’s complexity, shaped by diverse literary styles and typographical conventions, underscoring the need for precise annotation. The results indicate that, although indirect speech is important, it is not a frequent phenomenon in this literary-historical period. Rather than skewing the distribution further, we aim for a distribution that closely reflects the actual data. Table 2 provides detailed statistics on the manually annotated dataset.

Agreement. We use pairwise Cohen’s Kappa to assess Inter-Annotator Agreement (IAA) on the subset annotated by all three experts prior to consolidation. The pairwise comparisons between annotators resulted in an average Cohen’s Kappa score of 0.88, indicating substantial agreement among

annotators in classifying indirect speech from other representations of speech and narrative elements. Focusing specifically on indirect speech (IS), the average agreement between each annotator and the majority vote for IS labels was 0.89, further confirming that the annotators consistently identified instances of indirect speech throughout the corpus.

4 Experiment and Results

We model indirect speech detection as token classification, i.e. sequence tagging, with the tags described in §3. We fine-tune and evaluate pre-trained language models for token classification.

4.1 Pre-trained Language Models

We select models pre-trained on Danish and Norwegian text, based on their performance on Danish and Norwegian literary benchmark datasets (Al-Laith et al., 2024) and ScandEval (Nielsen, 2023). We experiment with both models not trained primarily on *historical/literary* Danish or Norwegian: DanskBERT (Snæbjarnarson et al., 2023)² and DFM (Large), the Danish Foundation Models sentence encoder (Enevoldsen et al., 2023),³ both trained on the Danish Gigaword Corpus (Strømberg-Derczynski et al., 2021); and NB-BERT-base (Kummervold et al., 2021),⁴ trained on the extensive digital collection at the National Library of Norway. Finally, MeMo-BERT-03 (Al-Laith et al., 2024),⁵ developed by continued pre-training of DanskBERT on the MeMo corpus.

4.2 Experimental Setup

For evaluation, we employ word-level macro averaged and class-specific F1-score. We select for testing the 20% of the dataset annotated by all three experts, and randomly split the rest such that 66% of the overall annotated dataset is used for training and 14% for validation. To address class imbalance, we applied upsampling by replicating samples from the minority classes “IS” five times. This helped the model avoid bias toward majority classes and improved its ability to detect this underrepresented category. To fine-tune the models, we use a batch size of 32 and train for 20 epochs with the AdamW

optimizer at a learning rate of 10^{-3} , choosing the best epoch based on validation loss.

4.3 Classification Results

Fine-tuning results in Table 4 show comparable overall performance across models on the test set, with DanskBERT and DFM (Large) achieving the best results. Most categories are classified with strong performance, but the Indirect Speech (IS) class stands out as the most challenging, with a markedly lower F1-score (0.70). This suggests difficulties in distinguishing Indirect Speech (IS), likely due to class imbalance or overlap with the Other (O) category. Table 3 presents a detailed breakdown of model performance across all speech categories. The results show that while most models achieve consistently strong precision, recall, and F1-scores for Direct Speech (DS), Other (O), Speech Marker (SM), and Speech Tag (ST), the Indirect Speech (IS) category remains the most challenging, with noticeably lower recall and F1 values across models. This suggests that IS is harder to distinguish, regardless of the underlying architecture.

5 Classifier-assisted Corpus Analysis

We use the top-performing model, DFM (Large), to tag all unlabeled segments in the main corpus. This results in 37.72% of words labeled as Direct Speech, 0.45% as Indirect Speech, 57.72% as Other, 2.09% as Speech Marker, and 2.02% as Speech Tag. Figure 1 shows the proportion of indirect speech label over time from 1870 to 1899. The trend appears to be fluctuating rather than showing a consistent increase or decrease. While no clear temporal pattern emerges, the usage of indirect speech appears linked to the social status and esthetic position of authors. The 20 works with the highest proportion of indirect speech (7.4%–2.5%) come from non-canonized or lesser-known authors in popular genres like crime fiction and historical novels. In contrast, the 20 works with the lowest proportion (0.0%–0.1%) are by canonized authors such as Viggo Stuckenborg, Johannes Jørgensen, Holger Drachmann, and Jonas Lie. This pattern is further reinforced when examining the ‘Other’ category (“O”). Among the works with the highest percentage in this category—ranging from 91.9% to 83.4%, well above the corpus average of 56.51%—male canonized authors dominate, including Karl Gjellerup, Jonas Lie, Johannes Jør-

²<https://huggingface.co/vesteinn/DanskBERT>

³<https://huggingface.co/KennethEnevoldsen/dfm-sentence-encoder-large-exp2-no-lang-align>

⁴<https://huggingface.co/NbAiLab/nb-bert-base>

⁵<https://huggingface.co/MiMe-MeMo/MeMo-BERT-03>

Model	DS			IS			O			SM			ST		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DanskBERT	0.89	0.89	0.89	0.77	0.63	0.69	0.93	0.94	0.93	0.94	0.96	0.95	0.93	0.93	0.93
DFM (Large)	0.85	0.90	0.88	0.90	0.57	0.70	0.94	0.92	0.93	0.91	0.96	0.93	0.93	0.90	0.91
MeMo-BERT-03	0.79	0.91	0.85	0.63	0.58	0.60	0.94	0.87	0.90	0.94	0.96	0.95	0.92	0.88	0.90
NB-BERT-base	0.85	0.94	0.89	0.81	0.61	0.69	0.96	0.91	0.94	0.91	0.95	0.93	0.94	0.87	0.90

Table 3: Performance of fine-tuned models on the test set, reported as word-level Precision (P), Recall (R), and F1-score (F1) for each speech category: Direct Speech (DS), Indirect Speech (IS), Other (O), Speech Marker (SM), Speech Tag (ST).

Model	Test. F1	Indirect Speech Class		
		Precision	Recall	F1
DanskBERT	0.87	0.77	0.63	0.69
DFM (Large)	0.87	0.90	0.57	0.70
MeMo-BERT-03	0.84	0.63	0.58	0.60
NB-BERT-base	0.86	0.81	0.61	0.69

Table 4: Fine-tuned models’ word-level macro average and class-specific macro F1-score results on the test sets

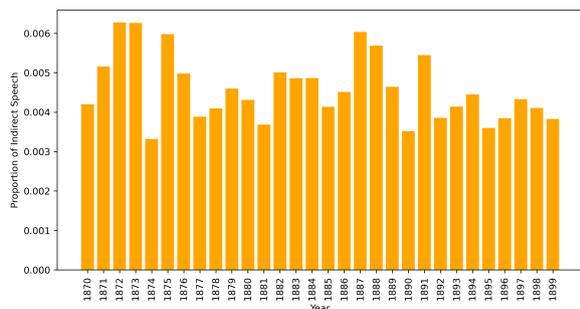


Figure 1: Proportion of indirect speech tokens, predicted by fine-tuned DFM (Large), by publication year.

gensen, Herman Bang, Henrik Pontoppidan, and Edvard Brandes. Our results suggest that canonized authors favored narrative techniques other than indirect and direct speech—perhaps using other ways of representing speech (e.g., free indirect speech) or focusing primarily on representing other types of events such as actions, thoughts, and sensations. These questions will need further examination.

6 Conclusion

We explore the detection of indirect speech in Danish and Norwegian literature, introduce a new annotated dataset, and evaluate multiple pre-trained language models for indirect speech classification. Domain-specific linguistic resources enhance accuracy in historical Scandinavian texts.

Indirect speech patterns reflect shifts in Scandinavian literature, for which we provide a new computational lens for examination. Future work

will incorporate additional linguistic features, refine annotation strategies, and analyze other genres and languages.

Limitations

This study presents several limitations that should be acknowledged. First, the annotated dataset is relatively small, consisting of only 150 segments drawn from 150 different novels. While this sampling strategy ensures literary diversity, it limits the robustness of the training data, particularly for rare phenomena like indirect speech. The model’s performance (F1 = 0.70 for IS) indicates that this approach is adequate for the detection of indirect speech, although we acknowledge that larger datasets would further improve robustness. Second, our extraction method, based on regular expressions targeting communication verbs and complementizers, likely introduces selection bias and overrepresents certain syntactic constructions of reported speech. Third, while we achieved high inter-annotator agreement, the inherent ambiguity of indirect speech, especially in cases involving free indirect discourse, remains a source of uncertainty for both annotators and models. Fourth, our experiments focused on a limited set of Danish and Norwegian language models. Although we selected state-of-the-art models suited to the task, we did not explore cross-lingual transfer or few-shot prompting strategies. Lastly, the classifier-assisted corpus analysis assumes consistent performance across time and text types, which may not hold due to evolving orthographic conventions, genre-specific styles, and shifting linguistic norms during the late 19th century. These limitations open avenues for future work, including the expansion of the dataset, improved sampling strategies, and more nuanced modeling of temporal and stylistic variation.

References

- Bas Aarts. 2014. [Indirect speech](#). In *The Oxford Dictionary of English Grammar*, 2nd edition. Oxford University Press.
- Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, and Daniel Hershcovich. 2024. [Development and evaluation of pre-trained language models for historical Danish and Norwegian literary texts](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4811–4819, Torino, Italia. ELRA and ICCL.
- Ali Al-Laith, Alexander Conroy, Kirstine Nielsen Degn, Jens Bjerring-Hansen, and Daniel Hershcovich. 2025. Annotating and classifying direct speech in historical danish and norwegian literary texts. In *Proceedings of NoDaLiDa/Baltic-HLT 2025*.
- Fatemeh Torabi Asr, Mazraeh Mohammad, Alexandre Lopes, Vagrant Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021. The gender gap tracker: Using natural language processing to measure gender bias in media. *PLoS One*, 16(1).
- 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*, volume 3232, pages 177–186, Uppsala, Sweden. DHNB Proceedings.
- Hans Brix. 1911. *Gudernes Tungemaal*. Gyldendal, Copenhagen, DK.
- Annelen Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. 2020. To BERT or not to BERT—comparing contextual embeddings in a deep learning architecture for the automatic recognition of four types of speech, thought and writing representation. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*.
- Regine Eckardt. 2021. The parameters of indirect speech. *The Wiley Blackwell companion to semantics*, pages 2213–2237.
- Anita Eerland and Rolf A. Zwaan. 2018. [The influence of direct and indirect speech on source memory](#). *Collabra: Psychology*, 4(1):5.
- Kenneth Enevoldsen, Lasse Hansen, Dan S. Nielsen, Rasmus A. F. Egebæk, Søren V. Holm, Martin C. Nielsen, Martin Bernstorff, Rasmus Larsen, Peter B. Jørgensen, Malte Højmark-Bertelsen, Peter B. Vahlstrup, Per Møldrup-Dalum, and Kristoffer Nielbo. 2023. [Danish foundation models](#). *Preprint*, arXiv:2311.07264.
- Cliff Goddard and Anna Wierzbicka. 2018. Direct and indirect speech revisited: Semantic universals and semantic diversity. In Alessandro Capone, Manuel García-Carpintero, and Alessandra Falzone, editors, *Indirect Reports and Pragmatics in the World Languages. Perspectives in Pragmatics, Philosophy & Psychology, vol 19*, pages 173–199. Springer, Cham.
- Ahrabhi Kathirgamalingam, Fabienne Lind, and Hajo G. Boomgaarden. 2023. [Automated detection of voice in news text – evaluating tools for reported speech and speaker recognition](#). *Computational Communication Research*, 5(1):85.
- Ari Z. Klein, Haitao Cai, Davy Weissenbacher, Lisa D. Levine, and Graciela Gonzalez-Hernandez. 2020. [A natural language processing pipeline to advance the use of twitter data for digital epidemiology of adverse pregnancy outcomes](#). *Journal of Biomedical Informatics*, 112:100076. Articles initially published in *Journal of Biomedical Informatics*: X 5-8, 2020.
- Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. [The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation](#). In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.
- Ralf Krestel, Sabine Bergler, and René Witte. 2008. Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. In *Proceedings of the International Language Resources and Evaluation Conference, LREC*, pages 2823–2828. European Language Resources Association (ELRA).
- Sven Møller Kristensen. 1955. *Impressionismen i dansk prosa 1870-1900*. Gyldendal, Copenhagen, DK.
- Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. [Operationalizing a national digital library: The case for a Norwegian transformer model](#). In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.
- Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470, Valencia, Spain. Association for Computational Linguistics.
- Chris Newell, Tim Cowlshaw, and David Man. 2018. Quote extraction and analysis for news. In *Proceedings of KDD Workshop on Data Science, Journalism and Media (DSJM)*.

- Dan Nielsen. 2023. [ScandEval: A benchmark for Scandinavian natural language processing](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.
- Silvia Pareti, Tim O’Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. [Automatically detecting and attributing indirect quotations](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999, Seattle, Washington, USA. Association for Computational Linguistics.
- Christian Scheible, Roman Klinger, and Sebastian Padó. 2016. [Model architectures for quotation detection](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1736–1745, Berlin, Germany. Association for Computational Linguistics.
- Elena Semino and Mick Short. 2004a. *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*. Routledge.
- Elena Semino and Mick Short. 2004b. *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. Routledge, London and New York.
- Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. [Transfer to a low-resource language via close relatives: The case study on Faroese](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.
- Leon Strømberg-Derczynski, Manuel Ciosici, Rebekah Baglini, Morten H. Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henriksen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, Claus Ladefoged, Finn Årup Nielsen, Jens Madsen, Malte Lau Petersen, Jonathan Hvithamar Rysstrøm, and Daniel Varab. 2021. [The Danish Giga-word corpus](#). In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.
- Sara Stymne. 2024. Direct speech identification in Swedish literature and an exploration of training data type, typographical markers, and evaluation granularity. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 253–263, St. Julians, Malta. Association for Computational Linguistics.
- Lars S. Vikør. 2022. [Rettskrivingsreform i store norske leksikon på snl.no](#). In <https://snl.no/rettskrivingsreform>.
- Kasper Welbers, Wouter van Atteveldt, and Jan Kleinjehuis. 2021. [Extracting semantic relations using syntax](#). *Computational Communication Research*, 3(2):180–194.

Harder than Finding the Lost Sheep? Towards Automatically Suggesting Deliberate Metaphor Annotations in German Sermons

Ronja Laarmann-Quante and Stefanie Dipper

Ruhr University Bochum, Germany

Faculty of Philology

Linguistics Department

{ronja.laarmann-quante|stefanie.dipper}@rub.de

Abstract

Automatic metaphor detection so far has largely focused on English data annotated for all kinds of metaphors including ubiquitous conventionalized ones. In this paper, we focus on deliberate metaphors in German sermons, i.e., metaphors that are used with a specific communicative goal. This task is harder because there is less training data available, and deliberate metaphors are very rare. Our goal is to support human annotators with automatically generated suggestions, so we strive above all for high recall. Using multilingual transfer learning based on various metaphor datasets and different transformer models, the highest recall we achieve is .70 (precision .10). Our results suggest that larger context windows beyond the sentence level are not helpful and that adding in-domain data even when annotated with different guidelines and in a different language is beneficial.

1 Introduction

According to Lakoff and Johnson (1980), metaphors are a cognitive mechanism that we use to think and understand the world by making complex, unfamiliar concepts understandable through familiar, concrete concepts. Visible signs of this fundamental mechanism are the numerous conventionalized metaphors in language, where a word is used in a context other than its usual one, thereby transferring its meaning. For example, the word *kill* is used in the phrase *kill the process* to mean *end the process*. The semantic domain of the usual context of the word *kill* (death) is called the source domain, while the domain to which the transfer takes place is called the target domain (computing) (Lakoff and Johnson, 1980).

A fundamental distinction in metaphors concerns their communicative purpose (Steen, 2023). The majority of metaphors that occur in everyday language are conventionalized, i.e., the transferred

meaning is usually listed in the dictionary, alongside the usual meaning. The use of such metaphors is usually completely unobtrusive and does not involve any particular communicative goals. However, when a speaker uses a metaphor in an utterance to make something understandable with the help of some other concept, this is referred to as a **deliberate metaphor**. A characteristic of such metaphors is that the source domain plays a role in the statement, thereby creating a different perspective. Steen (2011) provides the example of football coach Rinus Michels, who said *football is war*, thereby inviting his audience to view a football match from the perspective of a war event.¹

The work reported here was carried out in the context of the Collaborative Research Center (CRC) “**Metaphors of Religion**”.² The CRC’s starting assumption is that religion is an area in which metaphors necessarily play a central role, since the ultimate subject of religion—the transcendent—is not directly accessible and cannot be referred to literally, but can only be talked about through metaphors (Krech et al., 2024). Religious meaning-making therefore occurs primarily through metaphors, as evidenced by central Christian terms and concepts such as salvation or the Last Judgment.

All members of the CRC examine and analyze metaphors by annotating metaphorically used expressions in texts. In other words, we are not so much interested in conceptual, abstract metaphors in the sense of Lakoff and Johnson (1980) (such as LIFE IS A JOURNEY) than in linguistic instances of metaphors, which do not necessarily conform to a common conceptual metaphor. This paper presents

¹Note that conventionalized metaphors can therefore also be deliberate: when the source domain plays a communicative role; for an example, see Sec. 2. Novel metaphors, i.e. newly created, non-conventional metaphors, are a subset of deliberate metaphors.

²<https://sfb1475.ruhr-uni-bochum.de/>

a study on the automatic detection of metaphorically used expressions. Our goal is to support the annotation work by developing a system that pre-annotates candidates for metaphors, which are then reviewed by human annotators.

Religious discourse is not only a domain rich in deliberate metaphors (see also [Egg and Kordoni, 2022](#)) but metaphors also often stretch over larger parts of a text, called **extended metaphors**, potentially posing an additional challenge for the automatic detection, which has typically been carried out sentence-wise ([Reimann and Scheffler, 2024](#)).

Related work There are only few corpora annotated for metaphors, especially for languages other than English (see e.g. [Lu and Wang, 2017](#) for Chinese; [Sanchez-Bayona and Agerri, 2022](#) for Spanish; [Dipper et al., 2024](#), [Dipper, 2025](#) for German). The standard metaphor corpus is the English VU Amsterdam Metaphor Corpus (VUAMC) by [Steen et al. \(2010\)](#) with texts from four registers: academic texts, news texts, fiction, and conversation. The texts have been annotated word by word according to the MIPVU guidelines ([Steen et al., 2010](#)), which mark all “metaphor-related words” (MRWs) and do not distinguish between conventionalized and deliberate metaphors. Most work on automatic metaphor detection has focused on this corpus and F1 scores on the metaphor class of $> .75$ have been achieved on the VUAMC (see e.g. the shared tasks in [Leong et al., 2018, 2020](#)).

To date, there are only a few studies that explicitly target the detection of deliberate metaphors. [Neidlein et al. \(2020\)](#) showed that performance drops considerably when detecting non-conventionalized metaphors compared to conventionalized ones in the VUAMC. [Reimann and Scheffler \(2024\)](#) analyzed a corpus of online religious communication (see also Sec. 2) and found that several subtypes of deliberate metaphors were harder to detect than other metaphors.

Contribution and research questions Our paper tackles two research gaps, namely (1) optimizing the detection of deliberate metaphors and (2) focusing on a non-English dataset, i.e., a German corpus of sermons. Our specific research questions are as follows:

RQ1: Given the sparsity of annotated training data, how beneficial is it to add further data that does not precisely meet our objectives because it originates from other domains or languages or has been annotated according to different guidelines.

RQ2: Is a larger context size beyond the sentence level beneficial for identifying metaphors, given the observation of [Reimann and Scheffler \(2024\)](#) and [Egg and Kordoni \(2022\)](#) that religious discourse contains many extended metaphors?

We follow the approach of [Berger et al. \(2024\)](#) using multilingual transfer learning. We compare different multilingual transformer models and different transformer architectures including sentence transformers and a longformer model that captures more context. Our experiments focus on optimizing the recall of the metaphor class. The idea is that a high-recall system can be used in an assisted annotation scenario, as sketched above. The system would pre-annotate the data which is then checked by human annotators. Their primary task would be reduced to dismissing false positives, thereby speeding up the overall annotation process for generating new training data.

Our code and data is available under https://gitlab.ruhr-uni-bochum.de/vamos-cl/latech-clfl_2026_metaphors.

2 Data

We use four publicly available datasets described in the following. Our goal is to predict deliberate metaphors in the Sermon dataset while the other datasets serve as additional training data for fine-tuning the transformer models. In all datasets, metaphors are annotated on a token basis, i.e., every word used metaphorically is marked as such. None of the datasets provides (complete) information about which metaphorically used words belong to the same metaphorical image, so this information is not available to our system. The number of documents, sentences, tokens and annotated metaphors (i.e. metaphorically used words) per dataset is given in Table 1.

Sermon (de) ([Dipper, 2025](#)) is our primary dataset, which is annotated for deliberate metaphors according to the guidelines by [Dipper et al. \(2025\)](#). The guidelines provide for various labels for metaphor-related words. For example, expressions that are central to the metaphorical image are annotated with the label “center”, while less central expressions are given the underspecified label “MRW” (for “metaphor-related word”). Conventionalized metaphors that are deliberately used in the present context are labeled “revitalized”. The label “anchor” is used for literal expressions that are the target of the metaphorical transfer.

	<i>Relieved from the shackles of guilt and debt, breathing freely, not being caught in an overwhelming past.</i>																
Original	m	-	-	c	-	a	-	a	m	c	-	-	rvt	-	-	-	a
Binary	m	l	l	m	l	l	l	l	m	m	l	l	m	l	l	l	l

Figure 1: Translated and adapted example from the Sermon set with original and binary labels. Original labels are: c: center; rvt: revitalized; m: MRW; a: (literal) anchor (see Dipper et al. (2025) for full documentation of these labels).

The following examples are taken from the sermon dataset but translated to English for demonstration purposes. Fig. 1 shows an annotated example; metaphorically used words are printed in bold. The phrase *caught in the past* is rather conventional (in the German original), but in combination with *relieved from the shackles*, the original image of being caught becomes alive again and “revitalized”.

Ex. (1) shows a construction that is typical for sermons: different but often related images are used for the same subject matter (here: for the journey of life). Ex. (2) shows a metaphor that is typographically highlighted by quotes.

- (1) Perhaps it is not out of the question that in the coming weeks we will remember here and there the basic **lines** of our lives, our **pilgrimage** between birth and death, our being on the **road** in search of ourselves and of God.
- (2) Have we overcome all problems and are we already, so to speak, “at our **destination**”?

For our system, we map these labels to binary labels *m* (metaphorical) and *l* (literal), as shown in Fig. 1. We randomly split the documents in two sets, Sermon A and B, to perform two-fold cross-validation.

TEDx (de) (Dipper et al., 2024) contains German TEDx talks, i.e. non-religious texts, annotated for deliberate metaphors according to the same guidelines as the sermons. We again map the annotations to binary labels.

Reddit (en) (Reimann and Scheffler, 2024) contains threads from Christian subreddits in English. It was first annotated following the MIPVU guidelines (see Sec. 1) and secondly, deliberate metaphors were identified using the DMIP procedure (Reijnierse et al., 2018). We extracted the latter annotations (binary labels).

VUA (en) (Reijnierse et al., 2019) is a version of the VUAMC (see Sec. 1) which has been enriched with deliberate metaphor annotations according to DMIP. We use these annotations (binary labels).

Dataset	#docs	#sents	#toks	#met	%met
Sermon A	8	889	16,720	270	1.61
Sermon B	7	826	13,126	319	2.43
TEDx	10	1,260	19,794	259	1.31
Reddit	301	1,815	37,171	979	2.63
VUA	118	15,440	189,981	1,109	0.58

Table 1: Overview of deliberate metaphor datasets. Deliberate metaphors are a rare phenomenon, accounting for 0.58–2.63% of all tokens in the different datasets.

3 Method

We implement the metaphor detection as a binary sequence labeling task, where each token is classified as metaphorical or not. We compare several experimental settings: (1) **different pre-trained multilingual transformer models**:³ (i) mBERT (Devlin et al., 2019), (ii) XLM-RoBERTa (Conneau et al., 2020), (iii) two sentence transformers, namely Cross English & German RoBERTa for Sentence Embeddings by Philip May and (iv) multilingual SBERT (Reimers and Gurevych, 2019, 2020), and (v) the longformer model XLM-Long for sequence lengths up to 4096 tokens; (2) **different training data**: all possible combinations of the datasets in Sec. 2 for finetuning the transformer models; (3) **different context sizes** fed into the transformer: sentences, windows of 50 and 100 tokens,⁴ and whole documents.

Our code is based on the one by Wachowiak et al. (2022). All experiments were run on a Linux workstation with a single Nvidia RTX 4000 Ada GPU with 20 GB of memory. No model needed more than one hour to be finetuned. We used the default AdamW optimizer with no weight decay, a training batch size of 8, a learning rate of 2e-5 and finetuned each model for 8 epochs. The best model

³All models are taken from Huggingface, the exact links are given in Table 8 in Appendix E.

⁴Windows do not overlap but it is made sure that sentences are not cropped, i.e., the number of 50 or 100 tokens per window is typically exceeded to accommodate the rest of the last sentence.

		all	with met	%met
sentences	Sermon A	889	150 (17%)	7.04
	Sermon B	826	180 (22%)	9.77
	TEDx	1,260	118 (9%)	12.21
	Reddit	1,815	457 (25%)	8.38
	VUA	15,440	610 (4%)	7.53
windows 50	Sermon A	266	102 (38%)	4.04
	Sermon B	214	112 (52%)	4.63
	TEDx	332	77 (23%)	5.66
	Reddit	704	321 (46%)	5.30
	VUA	3,157	506 (16%)	3.45
windows 100	Sermon A	151	76 (50%)	3.05
	Sermon B	119	83 (70%)	3.47
	TEDx	185	61 (33%)	3.95
	Reddit	479	261 (54%)	4.33
	VUA	1,765	435 (25%)	2.30
documents	Sermon A	8	8 (100%)	1.61
	Sermon B	7	7 (100%)	2.43
	TEDx	10	10 (100%)	1.31
	Reddit	301	185 (61%)	3.54
	VUA	118	93 (79%)	0.67

Table 2: Number and percentage of units (sentences, token windows, documents) with metaphors (*with met*) out of all units (*all*). *%met* shows the percentage of metaphorical tokens in the units.

found during finetuning (based on a validation set of 10% of training data) was saved. Since we aim for a high-recall system to assist manual annotation (see Sec. 1), we set the metric to be improved to recall of the metaphor class. All models, except XLM-Long, have a maximum input length of 512 tokens. In order not to lose information when this limit is exceeded, we use a sliding window approach with a 32-token overlap between windows. At inference time, only the first prediction for each token occurring in an overlapping sequence is kept.

One problem in the detection of deliberate metaphors is the high class imbalance (see Table 1). To mitigate this, another experimental setting of ours is reducing each dataset to the units that contain metaphorical tokens (henceforth called **condensed** datasets). Table 2 shows the number of units (sentences, token windows, or documents) with metaphors for each dataset.⁵

4 Results

We evaluate the predictions with precision, recall and F1 score on the metaphor class using scikit-

⁵Since at inference time it is unknown which units will contain metaphors, the test set is not condensed but always used as is.

learn (Pedregosa et al., 2011). In general, the models achieve considerably higher precision with the original datasets than with the condensed datasets, which in turn yield clearly higher recall scores. This is not unexpected, as the condensed datasets simulate a higher proportion of metaphors. Since our goal is to achieve the highest possible recall scores, we focus on the condensed setting in our discussion.

The main results based on the condensed dataset are shown in Table 3.⁶ For completeness, the full results on the original dataset are given in Appendix B.

Since our primary dataset of sermons is small, we randomly divided it into two parts, Sermons A and B, to make the results as generalizable as possible. We always report the mean of the evaluations on Sermons A and B. For settings that use the sermons for fine-tuning, we perform a two-fold cross-validation.⁷

In general, the highest recall scores (marked in red) and F1 scores (blue) tend to be achieved with larger data sets (bottom of the table), while high precision (green) is also achieved with smaller training data sets located in the upper part of the table. However, the most significant differences arise on the basis of the various context windows. Single sentences as context windows clearly yield the best results with regard to recall. The larger the context window, the worse gets the recall. Precision, however, benefits from larger contexts, but the differences are much smaller than with recall. This pattern contradicts the hypothesis that religious discourse requires larger context sizes because it contains many extended metaphors (**RQ2**).

One reason might be that in the condensed datasets, the density of metaphors is highest on the sentence level (see Table 2). For the uncondensed datasets, the pattern of decreasing performance with increasing context size is not as clear.

When comparing the models with sentential context windows with regard to recall, the two sentence

⁶mBERT shows somewhat similar patterns to XLM-RoBERTa. While it has a slightly higher peak performance in recall (.61) compared to XLM-RoBERTa (.59), its results are overall slightly worse, so we report the full results only in Table 4 in Appendix A.

⁷The individual results of Sermon A and B for the condensed datasets are shown in Appendix C and D. In general, performance is better on test set B. The higher performance may depend on the higher proportion of metaphors in this set, which may benefit models that have been trained on a condensed data set and/or are designed to achieve the highest possible recall.

train	XLM-RoBERTa												ST en-de			ST multiling			XLM-Long		
	sentence			w50			w100			document			sentence			sentence			document		
	f1	p	r	f1	p	r	f1	p	r	f1	p	r	f1	p	r	f1	p	r	f1	p	r
s(ermom)	.22	.17	.33	.02	.14	.01	.00	.00	.00	.00	.00	.00	.22	.16	.42	.20	.15	.35	.00	.00	.00
t(edx)	.16	.12	.24	.02	.28	.01	.00	.00	.00	.00	.00	.00	.16	.10	.49	.15	.10	.32	.00	.00	.00
r(eddit)	.21	.14	.44	.22	.18	.29	.20	.18	.24	.18	.22	.16	.18	.10	.64	.18	.11	.61	.17	.18	.17
v(ua)	.20	.16	.27	.13	.18	.10	.08	.19	.05	.01	.38	.01	.17	.11	.42	.15	.09	.51	.00	.00	.00
r+s	.24	.16	.50	.25	.24	.26	.23	.26	.21	.12	.29	.09	.21	.13	.56	.22	.14	.56	.19	.33	.14
t+s	.23	.16	.43	.23	.24	.23	.14	.44	.12	.00	.00	.00	.22	.14	.54	.22	.14	.54	.00	.00	.00
v+s	.25	.17	.46	.27	.28	.25	.23	.27	.21	.11	.28	.07	.21	.13	.59	.23	.14	.56	.11	.28	.06
r+t	.23	.15	.52	.22	.21	.24	.21	.23	.19	.18	.33	.13	.20	.13	.53	.20	.12	.60	.15	.26	.11
v+t	.23	.15	.44	.26	.21	.33	.22	.26	.19	.06	.23	.03	.18	.10	.55	.20	.12	.54	.13	.28	.08
v+r	.23	.16	.43	.27	.23	.35	.25	.28	.22	.16	.20	.14	.19	.11	.64	.17	.10	.70	.13	.20	.10
r+t+s	.25	.16	.59	.28	.21	.39	.25	.27	.24	.16	.27	.11	.23	.14	.61	.22	.14	.56	.17	.28	.13
v+t+s	.26	.19	.45	.28	.25	.32	.28	.29	.28	.09	.31	.05	.22	.14	.56	.22	.14	.56	.14	.34	.09
v+r+s	.25	.17	.51	.30	.28	.33	.31	.30	.32	.20	.29	.15	.21	.13	.61	.22	.14	.55	.23	.28	.20
v+r+t	.24	.15	.55	.29	.23	.38	.28	.27	.30	.11	.24	.07	.19	.11	.60	.20	.12	.63	.14	.25	.10
v+r+t+s	.25	.16	.55	.29	.27	.31	.30	.32	.28	.17	.29	.13	.21	.13	.60	.23	.14	.57	.22	.32	.17

Table 3: Results based on the condensed datasets for models XLM-RoBERTa (model (ii) from Sec. 3), Cross English & German RoBERTa (ST en-de, (iii)), multilingual SBERT (ST multiling, (iv)), and XLM-Long (v); ST = sentence transformer. The highest values for F1 score (f1), precision (p), recall (r) across training datasets per setting are boldfaced and colored. Columns *sentence*, *w50*, *w100* (windows of 50/100 tokens) and *document* refer to the context window fed into the transformer.

transformers perform better than XLM-RoBERTa and also appear to require less training data (one or two datasets vs. three datasets). The best recall scores are achieved with datasets that include Reddit. The top score of .70 recall (with a precision of .10 and an F1 score of .17) is achieved by multilingual SBERT on a training set combining Reddit and VUA.

We can tentatively conclude that adding in-domain training data (Reddit), even when annotated with different guidelines, is beneficial (RQ1), which would be in line with the observation of Reimann and Scheffler (2024). The language difference does not seem to play a role for the multilingual models; perhaps the larger size of the English-language Reddit data compared to the German-language sermons is the decisive factor.

An error analysis reveals some typical sources of false positives: Words marked as literal quotations by quotation marks are often classified as metaphors. Quotation marks are in fact often an indication of a metaphor (see Dipper et al., 2025), as in Ex. (2). In addition, coordination and enumerations appear to be structures that can trigger false positives, compare the enumeration-like structures with multiple metaphorical words in Ex. (1) and Fig 1. Finally, according to the guidelines in Dipper et al. (2025), passages that a sermon cites from

the Bible should not be annotated at all, even if they contain metaphors (e.g. *vine* in Jesus’ famous saying *I am the true vine*). Since the models were not trained to detect Bible quotations, they often classify words within such quotations as metaphors, which does not match the gold standard here but would be correct in other, non-biblical contexts.

5 Conclusion & Future Work

We presented experiments on the automatic detection of deliberate metaphors in German sermons using multilingual transfer learning. The model achieving the highest recall (.70) was multilingual SBERT finetuned on an English dataset of online religious communication, highlighting the benefits of adding in-domain data. Sentence-based classification performed best, contradicting our hypothesis that larger contexts are beneficial for detecting metaphors. This is potentially due to the fact that the density of metaphors was highest at this level. Our next steps will be to use the model to assist human annotators in creating more in-domain training data. Currently, a limitation is that the precision scores are low, requiring the annotators to dismiss many false positives. With more training data available, our goal is to build more balanced models.

Acknowledgments

We are very grateful to the anonymous reviewers for their helpful comments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1475 – 441126958.

References

- Maria Berger, Nieke Kiwitt, and Sebastian Reimann. 2024. [Applying transfer learning to German metaphor prediction](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1383–1392, Torino, Italia. ELRA and ICCL.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Stefanie Dipper. 2025. [Metaphorical heads and literal dependents: Syntactic properties of metaphors in German](#). In *Proceedings of the 23th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)*, page 81–90, Ljubljana, Slovenia.
- Stefanie Dipper, Adam Roussel, Alexandra Wiemann, Won Kim, and Tra-my Nguyen. 2024. [Guidelines for the annotation of deliberate linguistic metaphor](#). In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 53–58, NAACL, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.
- Stefanie Dipper, Alexandra Wiemann, and Adam Roussel. 2025. [Guidelines zur Annotation von deliberaten Metaphern](#). *Metaphor Papers*, 27.
- Markus Egg and Valia Kordoni. 2022. [Metaphor annotation for German](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2556–2562, Marseille, France. European Language Resources Association.
- Volkhard Krech, Tim Karis, and Frederik Elwert. 2024. [Metaphors of religion: A conceptual framework](#). *Metaphor Papers*, 1.
- George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. The University of Chicago Press, Chicago, London.
- Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xinyang Chen. 2020. [A report on the 2020 VUA and TOEFL Metaphor Detection Shared Task](#). In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.
- Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. [A report on the 2018 VUA Metaphor Detection Shared Task](#). In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana. Association for Computational Linguistics.
- Xiaofei Lu and Ben Pin-Yun Wang. 2017. [Towards a metaphor-annotated corpus of Mandarin Chinese](#). *Language Resources and Evaluation*, 51(3):663–694.
- Arthur Neidlein, Philip Wiesenbach, and Katja Markert. 2020. [An analysis of language models for metaphor recognition](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3722–3736, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2018. [DMIP: A method for identifying potentially deliberate metaphor in language use](#). *Corpus Pragmatics*, 2(2):129–147.
- W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2019. [Metaphor in communication: the distribution of potentially deliberate metaphor across register and word class](#). *Corpora*, 14(3):301–326.
- Sebastian Reimann and Tatjana Scheffler. 2024. [Metaphors in online religious communication: A detailed dataset and cross-genre metaphor detection](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11236–11246, Torino, Italia. ELRA and ICCL.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2020. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

- Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Gerard J. Steen. 2011. From three dimensions to five steps: the value of deliberate metaphor. *metaphorik.de*, 21.
- Gerard J. Steen. 2023. Thinking by metaphor, fast and slow: Deliberate Metaphor Theory offers a new model for metaphor and its comprehension. *Frontiers in Psychology*, 14.
- Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification. From MIP to MIPVU*. Number 14 in *Converging Evidence in Language and Communication Research*. John Benjamins, Amsterdam.
- Lennart Wachowiak, Dagmar Gromann, and Chao Xu. 2022. Drum up SUPPORT: Systematic analysis of image-schematic conceptual metaphors. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 44–53, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Results of mBERT on Condensed Data

train	mBERT											
	sentence			w50			w100			document		
	f1	p	r	f1	p	r	f1	p	r	f1	p	r
sermon	.22	.14	.49	.21	.22	.22	.17	.25	.15	.00	.00	.00
tedx	.18	.12	.44	.20	.17	.26	.13	.15	.12	.00	.00	.00
reddit	.18	.11	.48	.18	.14	.23	.18	.13	.27	.16	.15	.16
vua	.21	.15	.36	.17	.21	.15	.12	.21	.08	.02	.03	.02
reddit+sermon	.22	.14	.54	.23	.20	.29	.24	.22	.28	.19	.22	.17
tedx+sermon	.22	.14	.53	.24	.22	.31	.24	.24	.25	.11	.24	.08
vua+sermon	.22	.14	.48	.26	.22	.33	.23	.24	.22	.11	.27	.08
reddit+tedx	.20	.12	.56	.22	.16	.33	.21	.20	.22	.09	.21	.05
vua+tedx	.18	.10	.61	.24	.20	.31	.24	.22	.28	.09	.25	.06
vua+reddit	.21	.13	.58	.21	.17	.30	.18	.19	.18	.18	.17	.20
reddit+tedx+sermon	.21	.13	.59	.25	.21	.33	.22	.20	.25	.18	.26	.14
vua+tedx+sermon	.22	.14	.54	.26	.21	.35	.24	.26	.23	.18	.29	.13
vua+reddit+sermon	.23	.14	.54	.26	.21	.36	.25	.25	.25	.22	.27	.19
vua+reddit+tedx	.20	.12	.56	.23	.18	.33	.24	.20	.29	.13	.22	.09
vua+reddit+tedx+sermon	.22	.14	.55	.25	.22	.29	.25	.24	.27	.15	.21	.13

Table 4: Results of mBERT (mean of Sermon set A and B) based on the condensed datasets.

B Results on Original (Non-Condensed) Data

XLM-RoBERTa												
train	sentence			w50			w100			document		
	f1	p	r	f1	p	r	f1	p	r	f1	p	r
sermon	.22	.31	.17	.05	.46	.03	.00	.00	.00	.00	.00	.00
tedx	.11	.30	.07	.05	.30	.03	.00	.00	.00	.00	.00	.00
reddit	.21	.23	.20	.20	.25	.16	.21	.26	.17	.15	.21	.12
vua	.08	.38	.04	.10	.39	.05	.05	.22	.03	.00	.00	.00
reddit+sermon	.16	.27	.12	.24	.35	.19	.24	.33	.19	.23	.33	.18
tedx+sermon	.17	.35	.11	.14	.28	.10	.17	.32	.11	.00	.00	.00
vua+sermon	.18	.30	.13	.16	.33	.10	.22	.36	.16	.09	.30	.05
reddit+tedx	.21	.29	.16	.15	.33	.10	.20	.39	.14	.10	.31	.06
vua+tedx	.18	.27	.13	.12	.30	.07	.11	.35	.06	.10	.30	.06
vua+reddit	.19	.25	.15	.19	.24	.15	.24	.31	.19	.22	.23	.20
reddit+tedx+sermon	.19	.35	.13	.25	.35	.19	.22	.33	.17	.20	.28	.15
vua+tedx+sermon	.17	.33	.12	.17	.37	.11	.25	.38	.19	.15	.31	.10
vua+reddit+sermon	.15	.31	.10	.23	.34	.18	.22	.31	.16	.19	.28	.15
vua+reddit+tedx	.10	.23	.07	.19	.33	.13	.21	.34	.15	.20	.28	.15
vua+reddit+tedx+sermon	.22	.32	.17	.25	.36	.19	.23	.35	.17	.21	.31	.16

mBERT												
train	sentence			w50			w100			document		
	f1	p	r	f1	p	r	f1	p	r	f1	p	r
sermon	.17	.31	.14	.18	.30	.14	.12	.24	.08	.00	.00	.00
tedx	.05	.18	.03	.06	.27	.03	.05	.19	.03	.00	.00	.00
reddit	.11	.14	.09	.14	.15	.13	.17	.18	.16	.07	.13	.05
vua	.05	.20	.03	.09	.20	.06	.09	.27	.05	.04	.20	.02
reddit+sermon	.18	.28	.14	.27	.28	.25	.25	.26	.24	.16	.22	.14
tedx+sermon	.20	.29	.15	.20	.33	.15	.20	.29	.16	.11	.24	.08
vua+sermon	.17	.31	.12	.20	.29	.17	.20	.31	.15	.20	.27	.16
reddit+tedx	.14	.18	.11	.13	.26	.08	.12	.26	.07	.16	.23	.13
vua+tedx	.17	.21	.15	.19	.24	.15	.13	.31	.08	.09	.28	.05
vua+reddit	.09	.19	.06	.23	.20	.26	.19	.21	.18	.09	.20	.06
reddit+tedx+sermon	.19	.27	.15	.24	.28	.21	.22	.29	.19	.17	.21	.15
vua+tedx+sermon	.14	.27	.10	.22	.31	.17	.21	.31	.16	.15	.26	.10
vua+reddit+sermon	.19	.35	.14	.24	.31	.20	.25	.31	.22	.22	.27	.18
vua+reddit+tedx	.09	.22	.06	.17	.26	.12	.16	.26	.12	.17	.23	.13
vua+reddit+tedx+sermon	.19	.29	.15	.25	.29	.22	.22	.33	.17	.16	.23	.15

train	ST de-en			ST multilingual			XLM-Long		
	sentence			sentence			document		
	f1	p	r	f1	p	r	f1	p	r
sermon	.17	.34	.14	.21	.22	.20	.00	.00	.00
tedx	.07	.32	.04	.11	.19	.08	.00	.00	.00
reddit	.22	.17	.32	.18	.13	.33	.17	.21	.15
vua	.07	.37	.04	.11	.18	.08	.00	.13	.00
reddit+sermon	.22	.30	.18	.23	.28	.20	.23	.31	.19
tedx+sermon	.21	.34	.16	.22	.23	.21	.00	.00	.00
vua+sermon	.18	.31	.12	.22	.27	.19	.10	.31	.06
reddit+tedx	.25	.24	.26	.18	.23	.15	.20	.25	.17
vua+tedx	.19	.27	.15	.19	.20	.18	.12	.27	.07
vua+reddit	.14	.24	.10	.19	.15	.28	.23	.28	.19
reddit+tedx+sermon	.21	.30	.17	.23	.25	.21	.23	.27	.20
vua+tedx+sermon	.20	.34	.14	.25	.29	.22	.16	.27	.11
vua+reddit+sermon	.23	.35	.20	.20	.27	.16	.24	.31	.19
vua+reddit+tedx	.19	.24	.16	.18	.25	.15	.21	.31	.16
vua+reddit+tedx+sermon	.23	.33	.18	.24	.29	.21	.24	.34	.19

Table 5: Results based on the original, not condensed, dataset (mean of Sermon sets A and B).

C Results on Sermon Set A (Condensed)

XLM-RoBERTa												
train	sentence			w50			w100			document		
	f1	p	r									
sermon	.21	.15	.34	.04	.27	.02	.00	.00	.00	.00	.00	.00
tedx	.14	.10	.23	.02	.33	.01	.00	.00	.00	.00	.00	.00
reddit	.17	.11	.39	.18	.13	.27	.17	.14	.22	.16	.18	.14
vua	.17	.14	.23	.13	.17	.10	.07	.16	.05	.01	.25	.00
reddit+sermon	.20	.12	.49	.21	.20	.23	.22	.23	.22	.19	.28	.15
tedx+sermon	.18	.12	.36	.25	.23	.28	.25	.29	.22	.00	.00	.00
vua+sermon	.20	.13	.47	.25	.25	.24	.25	.26	.25	.10	.24	.06
reddit+tedx	.18	.12	.45	.21	.18	.25	.20	.20	.19	.20	.35	.14
vua+tedx	.21	.14	.42	.28	.22	.38	.23	.25	.21	.03	.17	.02
vua+reddit	.21	.14	.40	.24	.20	.31	.22	.25	.20	.15	.17	.14
reddit+tedx+sermon	.19	.11	.55	.25	.19	.36	.25	.26	.24	.15	.26	.10
vua+tedx+sermon	.22	.15	.39	.28	.23	.36	.32	.31	.34	.12	.33	.07
vua+reddit+sermon	.20	.14	.41	.27	.24	.33	.28	.27	.28	.19	.23	.16
vua+reddit+tedx	.20	.13	.49	.29	.23	.40	.29	.27	.31	.11	.24	.07
vua+reddit+tedx+sermon	.21	.14	.46	.26	.25	.26	.28	.28	.27	.18	.28	.14

mBERT												
train	sentence			w50			w100			document		
	f1	p	r									
sermon	.17	.10	.44	.20	.18	.23	.24	.25	.23	.00	.00	.00
tedx	.14	.09	.40	.20	.15	.27	.17	.19	.16	.00	.00	.00
reddit	.14	.08	.43	.15	.11	.22	.15	.11	.27	.14	.12	.16
vua	.19	.13	.33	.16	.18	.14	.13	.24	.09	.03	.03	.03
reddit+sermon	.16	.09	.48	.19	.14	.29	.21	.16	.31	.15	.18	.13
tedx+sermon	.17	.10	.51	.23	.17	.35	.25	.21	.31	.16	.25	.12
vua+sermon	.18	.12	.39	.25	.20	.33	.25	.23	.27	.05	.23	.03
reddit+tedx	.16	.09	.49	.19	.13	.31	.21	.20	.23	.11	.26	.07
vua+tedx	.14	.08	.56	.24	.19	.33	.24	.20	.29	.10	.25	.06
vua+reddit	.17	.10	.53	.18	.13	.27	.16	.16	.16	.18	.16	.21
reddit+tedx+sermon	.15	.09	.55	.23	.17	.36	.19	.16	.25	.18	.23	.15
vua+tedx+sermon	.18	.11	.53	.24	.18	.36	.25	.24	.26	.18	.25	.14
vua+reddit+sermon	.17	.10	.46	.23	.17	.37	.22	.21	.24	.22	.23	.21
vua+reddit+tedx	.17	.10	.53	.22	.16	.33	.23	.19	.31	.15	.25	.11
vua+reddit+tedx+sermon	.17	.10	.51	.23	.20	.27	.23	.22	.23	.20	.20	.20

train	ST de-en			ST multilingual			XLM-Long		
	sentence			sentence			document		
	f1	p	r	f1	p	r	f1	p	r
sermon	.18	.11	.44	.16	.10	.38	.00	.00	.00
tedx	.12	.07	.43	.12	.08	.29	.00	.00	.00
reddit	.13	.08	.56	.14	.08	.56	.16	.15	.17
vua	.13	.08	.37	.13	.07	.51	.00	.00	.00
reddit+sermon	.15	.09	.51	.17	.10	.50	.19	.26	.15
tedx+sermon	.17	.10	.51	.17	.10	.47	.00	.00	.00
vua+sermon	.17	.10	.54	.18	.11	.50	.08	.24	.05
reddit+tedx	.15	.09	.45	.17	.10	.55	.14	.23	.10
vua+tedx	.13	.08	.47	.16	.10	.51	.12	.26	.08
vua+reddit	.14	.08	.56	.14	.08	.66	.12	.17	.09
reddit+tedx+sermon	.16	.10	.57	.18	.11	.54	.20	.24	.17
vua+tedx+sermon	.18	.11	.50	.18	.11	.53	.16	.30	.11
vua+reddit+sermon	.17	.10	.50	.18	.11	.48	.24	.25	.23
vua+reddit+tedx	.13	.08	.48	.17	.10	.61	.15	.26	.10
vua+reddit+tedx+sermon	.17	.10	.50	.18	.11	.55	.25	.33	.20

Table 6: Results on Sermon set A (condensed setting).

D Results on Sermon Set B (Condensed)

XLM-RoBERTa												
train	sentence			w50			w100			document		
	f1	p	r	f1	p	r	f1	p	r	f1	p	r
sermon	.24	.19	.33	.00	.00	.00	.00	.00	.00	.00	.00	.00
tedx	.18	.14	.25	.02	.23	.01	.00	.00	.00	.00	.00	.00
reddit	.26	.17	.50	.26	.23	.31	.24	.22	.26	.21	.27	.17
vua	.22	.18	.30	.12	.18	.09	.09	.21	.06	.01	.50	.01
reddit+sermon	.28	.20	.51	.29	.28	.29	.24	.29	.21	.05	.30	.03
tedx+sermon	.28	.20	.50	.21	.26	.18	.04	.60	.02	.00	.00	.00
vua+sermon	.29	.21	.45	.29	.31	.27	.21	.29	.17	.12	.32	.08
reddit+tedx	.27	.18	.59	.23	.23	.23	.22	.26	.20	.16	.32	.11
vua+tedx	.25	.17	.47	.23	.19	.29	.22	.27	.18	.08	.30	.05
vua+reddit	.26	.18	.47	.31	.26	.38	.27	.30	.25	.18	.23	.14
reddit+tedx+sermon	.31	.20	.64	.31	.24	.43	.26	.28	.24	.18	.28	.13
vua+tedx+sermon	.31	.22	.51	.27	.27	.28	.24	.28	.21	.06	.29	.03
vua+reddit+sermon	.30	.20	.61	.33	.32	.34	.34	.32	.36	.20	.35	.14
vua+reddit+tedx	.27	.17	.61	.29	.24	.37	.28	.27	.29	.10	.25	.06
vua+reddit+tedx+sermon	.28	.18	.65	.32	.29	.35	.32	.36	.29	.17	.29	.12

mBERT												
train	sentence			w50			w100			document		
	f1	p	r	f1	p	r	f1	p	r	f1	p	r
sermon	.28	.19	.55	.23	.25	.20	.11	.25	.07	.00	.00	.00
tedx	.22	.14	.48	.21	.18	.25	.09	.11	.08	.00	.00	.00
reddit	.21	.13	.53	.20	.18	.24	.20	.16	.28	.18	.18	.17
vua	.23	.17	.39	.19	.24	.16	.11	.18	.08	.02	.03	.02
reddit+sermon	.27	.18	.60	.27	.25	.29	.26	.29	.24	.23	.27	.21
tedx+sermon	.26	.17	.54	.26	.26	.26	.23	.27	.19	.06	.23	.03
vua+sermon	.27	.17	.57	.28	.24	.33	.21	.26	.18	.17	.30	.12
reddit+tedx	.25	.15	.62	.25	.20	.35	.20	.19	.21	.06	.16	.04
vua+tedx	.21	.12	.66	.24	.21	.29	.25	.23	.26	.09	.24	.05
vua+reddit	.24	.15	.64	.25	.20	.33	.21	.23	.19	.19	.19	.19
reddit+tedx+sermon	.26	.16	.63	.27	.25	.29	.25	.25	.25	.18	.28	.13
vua+tedx+sermon	.26	.17	.56	.29	.24	.35	.24	.29	.20	.18	.32	.13
vua+reddit+sermon	.29	.19	.62	.30	.25	.36	.28	.30	.26	.21	.32	.16
vua+reddit+tedx	.23	.14	.59	.25	.21	.32	.24	.20	.28	.11	.19	.08
vua+reddit+tedx+sermon	.28	.18	.59	.27	.24	.30	.28	.27	.30	.10	.22	.07

train	ST de-en			ST multilingual			XLM-Long		
	sentence			sentence			document		
	f1	p	r	f1	p	r	f1	p	r
sermon	.27	.20	.39	.25	.20	.33	.00	.00	.00
tedx	.20	.12	.54	.19	.13	.35	.00	.00	.00
reddit	.22	.13	.71	.21	.13	.66	.19	.20	.18
vua	.21	.13	.48	.18	.11	.51	.00	.00	.00
reddit+sermon	.27	.17	.60	.27	.17	.63	.19	.39	.12
tedx+sermon	.27	.18	.56	.27	.17	.61	.00	.00	.00
vua+sermon	.25	.16	.64	.28	.18	.62	.13	.33	.08
reddit+tedx	.25	.16	.60	.24	.15	.64	.16	.28	.12
vua+tedx	.22	.13	.64	.23	.15	.58	.13	.30	.08
vua+reddit	.23	.14	.72	.20	.12	.73	.15	.22	.11
reddit+tedx+sermon	.29	.19	.65	.26	.17	.58	.15	.32	.09
vua+tedx+sermon	.27	.17	.63	.27	.17	.60	.11	.38	.07
vua+reddit+sermon	.25	.15	.73	.26	.16	.61	.23	.32	.18
vua+reddit+tedx	.24	.15	.71	.24	.14	.66	.13	.25	.09
vua+reddit+tedx+sermon	.26	.16	.71	.27	.17	.60	.18	.31	.13

Table 7: Results based on Sermon set B (condensed setting).

E Model References

mBERT	https://huggingface.co/google-bert/bert-base-multilingual-cased
XLM-RoBERTa	https://huggingface.co/FacebookAI/xlm-roberta-base
Cross EN & DE RoBERTa	https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer
multilingual SBERT	https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
XLM-Long	https://huggingface.co/markussagen/xlm-roberta-longformer-base-4096

Table 8: Links to used models, last accessed on December 23, 2025.

Semantic Factor Analysis: Validating Personality Structure Recovery from empirically-weighted Word Embeddings

Oliver Müller

August-Wilhelm Scheer Institut

oliver.mueller@aws-institut.de

Abstract

The present study introduces Semantic Factor Analysis (SFA), a novel computational approach recovering Big Five personality trait structures from pre-trained adjective word embeddings weighted by empirical participant data. Using Word2Vec embeddings trained on Google-News-300 corpus, semantic relationships of IPIP-50 Big Five inventory adjectives (Goldberg, 1992) were extracted and factor structures computed through weighted vector averaging and K-means clustering. To validate the methodology, SFA was compared against a baseline using unweighted Word2Vec embeddings. In a controlled experiment with $n=55$ participants completing standard IPIP-50 assessments, HSP-R scale (Pluess et al., 2024) and multimedia impact surveys, empirically-weighted SFA successfully recovered all five personality dimensions with 62.5% average factor purity, substantially outperforming the unweighted baseline (52.0%, 10% relative improvement), while traditional Confirmatory Factor Analysis showed factor collapse and poor model fit. The approach was validated through Latent Class Analysis deriving empirically-based classification thresholds for Big Five dimensions and supporting a trichotomous Environmental Sensitivity model (Lionetti et al., 2018). Results demonstrate that integrating semantic representations with empirical data improves Big Five structure recovery beyond pure semantic similarity alone, particularly for small sample studies where traditional methods such as CFA will fail due to limited empirical data points.

1 Introduction

The Big Five personality model—comprising Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN)—represents one of the most empirically validated frameworks for personality assessment (McCrae and John, 1992; Goldberg, 1992). The theoretical

foundation for personality measurement derives from the lexical hypothesis (Allport and Odbert, 1936), which posits that salient personality characteristics are encoded within natural language. This principle suggests a profound connection: if personality traits are fundamentally linguistic constructs embedded in how humans naturally describe one another, then semantic relationships captured in word embeddings¹ could directly reflect underlying personality structures. This convergence between Natural Language Processing methods and psychological theory may offer a computational validation of the lexical hypothesis itself.

Traditional extraction of these dimensions relies on factor-analytic procedures applied to self-report instruments, where participants rate themselves using statement-based items. This approach, while robust and cross-culturally validated, requires explicit survey administration and remains constrained by response biases, social desirability effects, and limited scalability. While our pilot study (Müller and Degaetano-Ortlieb, 2025) successfully demonstrated Big Five personality dimension clustering in a 3D embedding space entirely absent of self-assessment data, the present work provides rigorous empirical validation of an updated version of the embedding approach against traditional Confirmatory Factor Analysis (CFA) using controlled experimental data.

This study introduces Semantic Factor Analysis (SFA), a computational methodology leveraging pretrained word embeddings to recover personality trait structures from semantic relationships between adjectives mediated by empirical participant data. Beyond SFA validation, the study ex-

¹Word embeddings are distributed vector representations that learn meaning from large-scale text corpora (Mikolov et al., 2013), where semantic similarity corresponds to geometric proximity in high-dimensional space.

amines whether personality dimensions and Sensory Processing Sensitivity (SPS)—the core individual difference construct underlying Environmental Sensitivity (ES) theory—demonstrate predictive validity for ecologically-relevant behavioral outcomes. SPS represents differential susceptibility to environmental influences, with empirical evidence supporting a trichotomous population distribution: low-sensitive individuals (Dandelions), medium-sensitive (Tulips), and high-sensitive (Orchids) (Lionetti et al., 2018). To test whether personality traits and sensitivity classifications predict differential responses in real-world contexts, participants completed Big Five assessments, the 18-item HSP-R sensitivity scale (Pluess et al., 2024), then viewed three multimedia stimuli (two controversial news reports, one product advertisement), to test eight hypotheses examining emotional reactivity and consumer behavior across personality and sensitivity profiles.

The key research questions are: (1) Can empirically weighted word embeddings recover Big Five personality factor structures comparably to traditional methods? (2) Do personality traits and Environmental Sensitivity predict differential responses to emotionally-charged multimedia content and commercial advertising? To address the latter, eight specific hypotheses examined relationships between Big Five dimensions (Agreeableness, Neuroticism, Openness, Extraversion), Environmental Sensitivity groups (Dandelions, Tulips, Orchids), and responses to controversial social media posts and product advertisements.

This work contributes: (1) a novel computational methodology for Big Five personality analysis (SFA) with baseline validation demonstrating substantial improvement from empirical weighting over pure semantic similarity; (2) empirically-derived classification thresholds for Big Five and Environmental Sensitivity through Latent Class Analysis, providing empirical support for the trichotomous ES model (Lionetti et al., 2018); (3) further psychometric validation of the 18-item HSP-R scale (Pluess et al., 2024); (4) demonstration of personality and sensitivity measures' differential predictive validity in multimedia contexts.

2 Related Work

2.1 Big Five Personality Assessment

The Big Five model emerged from systematic factor analyses of personality descriptors, with foundational work by Tupes and Christal (1961) establishing the Five-Factor framework. Goldberg (1992) developed comprehensive adjective markers, while Costa and McCrae (1992) validated the structure cross-culturally through NEO-PI-R. DeYoung et al. (2007) introduced the Big Five Aspect Scale (BFAS), which split each dimension into two correlated but distinct subdimensions (aspects) characterized by differentiated biological and genetic substrates, and identified higher-order factors (Stability & Plasticity) (DeYoung et al., 2002) subsuming the Big Five, demonstrating a hierarchical personality structure. The International Personality Item Pool (IPIP) provides freely accessible validated measures (Goldberg, 1999), including the IPIP-50 used in the present validation study.

2.2 Word Embeddings and Semantic Space

Mikolov et al. (2013) introduced Word2Vec, training neural networks to produce dense vector representations where semantic similarity corresponds to geometric proximity. The Google News 300-dimensional model, trained on 100 billion words, captures rich semantic relationships including synonymy, analogy, and categorical associations. These embeddings have proven valuable for numerous NLP tasks but remain underutilized in psychometric research. Park et al. (2015) demonstrated personality insights from social media language, while Schwartz et al. (2013) linked Facebook language patterns to Big Five scores. However, prior work focuses on predicting individual personality from text rather than recovering underlying factor structures. Unlike prior methods relying on supervised classification (Kazameini et al., 2020; Carducci et al., 2018) and extensive labeled datasets, SFA integrates semantic structure of word embeddings with empirical self-assessment data to recover and validate Big Five factor structure using an unsupervised methodology.

2.3 Environmental Sensitivity Framework

Environmental Sensitivity theory posits differential susceptibility to environmental influences (Aron and Aron, 1997). Lionetti et al. (2018) iden-

tified three sensitivity groups with distinct population distributions: Dandelions (low sensitivity, 29%), Tulips (medium, 40%), and Orchids (high, 31%). [Pluess et al. \(2024\)](#) introduced and validated the 18-item Highly Sensitive Person Revised Scale (HSP-R), subdividing sensory processing sensitivity into six distinct facets (See Appendix B). [Grimen and Diseth \(2016\)](#) demonstrated sensitivity correlates negatively with Extraversion and positively with Neuroticism, while [Homberg et al. \(2016\)](#) explored genetic underpinnings. The present work supports the HSP-R framework through empirically-derived LCA thresholds and comprehensive validation against Big Five dimensions.

3 Methodology

3.1 Semantic Factor Analysis (SFA)

SFA extracts personality structures from pre-trained word embeddings through the following procedure:

Embedding Extraction: The Google News Word2Vec model (300 dimensions, 3 million words, 100 billion token corpus) is utilized. For the IPIP-50 Big Five marker adjectives from [Goldberg \(1992\)](#)—comprising 10 items per dimension—300-dimensional embedding vectors \mathbf{v}_i are extracted directly from the pretrained model. The adjectives represent prototypical markers for each dimension (e.g., “creative”, “imaginative” for Openness; “talkative”, “energetic” for Extraversion).

Empirically-Weighted Semantic Space: Each adjective’s 300-dimensional vector is weighted by the mean participant response for that item’s corresponding IPIP-50 statement. Let \mathbf{w} denote the 50×1 vector of mean response scores (1-5 Likert scale) for each adjective. Weighted embeddings are constructed through element-wise multiplication:

$$\mathbf{V}_{\text{weighted}} = \mathbf{w} \odot \mathbf{V} \quad (1)$$

where \mathbf{V} is the 50×300 matrix of original embeddings and \odot denotes element-wise multiplication (broadcasting \mathbf{w} across all 300 dimensions). This operation incorporates empirical participant data into the semantic space, distinguishing SFA from purely distributional approaches. For example, if participants rated the statement corresponding to “anxious” with a mean of 3.2 on the 5-point scale, the adjective’s 300-dimensional Word2Vec vector would be multiplied by 3.2 across all dimensions,

amplifying its semantic representation proportionally to how strongly participants endorsed anxiety-related traits. This weighting reflects actual personality response patterns rather than pure distributional semantics alone.

Dimensionality Reduction via PCA: The weighted embeddings are standardized using z-score normalization, then Principal Component Analysis is applied to reduce dimensionality from 300 to 5 dimensions:

$$\mathbf{S} = \text{StandardScale}(\mathbf{V}_{\text{weighted}}) \cdot \mathbf{P} \quad (2)$$

where \mathbf{P} is the 300×5 PCA transformation matrix derived from the covariance structure of weighted embeddings, and \mathbf{S} is the resulting 50×5 factor score matrix representing each adjective in 5-dimensional semantic factor space.

Factor Assignment via K-means: K-means clustering ($k = 5$) is applied to the factor scores \mathbf{S} , assigning each adjective to its closest cluster centroid:

$$\text{cluster}_i = \underset{j \in \{1, \dots, 5\}}{\text{argmin}} \|\mathbf{s}_i - \mathbf{c}_j\|^2 \quad (3)$$

where \mathbf{s}_i is the factor score vector for adjective i and \mathbf{c}_j are the five cluster centroids. Factor purity is computed as the proportion of adjectives correctly clustering with their theoretically assigned Big Five dimension. Continuing the example of the adjective “anxious”, after PCA reduction and K-means clustering, this adjective’s weighted factor scores would ideally be assigned to the Neuroticism cluster alongside theoretically related adjectives like “nervous” and “worried”, with factor purity measuring how successfully such theoretically motivated groupings emerge from the semantic-empirical space.

Visualization: The PCA-reduced 5-dimensional space can be further visualized in 3D by retaining the first three principal components, preserving maximal variance while enabling visual interpretation of factor clustering patterns.

3.2 Baseline Analysis: Unweighted Word2Vec

To establish the value of empirical weighting, a baseline analysis used pure Word2Vec embeddings without participant responses, following the pilot studies’ word embeddings approach ([Müller and Degaetano-Ortlieb, 2025](#)).

Four IPIP adjectives were unavailable in the Word2Vec vocabulary and replaced with semantic equivalents verified through synonym analysis: *untalkative*→*silent*, *short-tempered*→*cranky*, *self-pitying*→*sad*, *uninquisitive*→*incurious* (see Appendix C). The baseline used identical analysis pipeline: Word2Vec extraction → PCA → K-means clustering → purity calculation, differing only in absence of empirical weighting. Silhouette analysis² identified optimal cluster count for both methods, enabling direct comparison of weighted versus unweighted approaches.

3.3 Traditional Confirmatory Factor Analysis

For direct comparison, standard CFA methodology is applied to empirical IPIP-50 data (Goldberg, 1999). The model includes five latent factors corresponding to Big Five dimensions with 10 observed indicators each and correlated errors permitted between theoretically related items (See Appendix A). The model is estimated using maximum likelihood in R package lavaan (Rosseel, 2012), evaluating fit through CFI (> 0.90), TLI (> 0.90), RMSEA (< 0.08), and SRMR (< 0.08). Standardized solutions are reported for interpretability.

3.4 Latent Class Analysis for Classification

Traditional personality classification often employs rather arbitrary cutoffs (e.g., median splits, percentile boundaries or via SDs). Instead, Gaussian Mixture Modeling (GMM) is applied to derive empirically-based thresholds. For K class solutions, GMM assumes data are generated from a mixture of K Gaussian distributions with probability density:

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \mu_k, \sigma_k^2) \quad (4)$$

where π_k are mixing coefficients ($\sum_{k=1}^K \pi_k = 1$), μ_k are class means, and σ_k^2 are class variances. Model selection uses information criteria balancing fit against complexity:

$$\text{BIC} = -2 \log \mathcal{L} + p \log(n) \quad (5)$$

²Silhouette analysis quantifies clustering quality by measuring how similar each data point is to its own cluster compared to other clusters, with scores ranging from -1 (misclassified) to +1 (well-clustered). Higher average silhouette scores indicate optimal cluster count by identifying the configuration with maximal within-cluster cohesion and between-cluster separation.

$$\text{AIC} = -2 \log \mathcal{L} + 2p \quad (6)$$

where \mathcal{L} is maximum likelihood, p is parameter count, and n is sample size. BIC penalizes complexity more heavily than AIC, favoring parsimonious solutions. Models with 2-5 classes are fit for each Big Five dimension and HSP-R total score, selecting optimal solutions through minimum BIC/AIC. Classification boundaries are extracted from fitted Gaussian means and variances, providing data-driven thresholds. For HSP-R, data are tested for support of the theorized trichotomous structure (Dandelions/Tulips/Orchids) or alternative configurations.

3.5 Participants and Procedure

Fifty-five participants (34 female, 21 male; $M_{age} = 24.3$ years, $SD = 4.2$, range 19-41) were recruited through the Saarland University network. All provided informed consent and received no compensation. Participants completed three online surveys on Qualtrics in controlled laboratory sessions: (1) IPIP-50 Big Five assessment (10 items per dimension, 5-point Likert scale); (2) HSP-R 18-item sensitivity scale (7-point Likert scale); (3) multimedia impact surveys including emotional response to controversial social media content (news reports) and product advertisement persuasiveness ratings. Session duration averaged 20 minutes for survey completion.

3.6 Measures and Scoring

IPIP-50: Raw scores were computed by summing item responses (reverse-scoring negatively keyed items), then doubling to yield 0-100 percentage scores per dimension following standard IPIP procedures (Goldberg, 1999).

HSP-R: Total sensitivity scores were calculated by averaging 18 items (range 1-7), then converting to percentages (score/7 × 100). The HSP-R comprises six subscales: Depth of Processing, Emotional Reactivity, Overstimulation, Sensitivity to Details, Sensitivity to Positive Experiences, and Social Sensitivity. Subscale scores were computed identically to facilitate correlational analyses.

Classification Thresholds: Based on LCA results (detailed in Results), empirically-derived cutoffs were established: Dandelions (<61.2% HSP-R), Tulips (61.2-76.5%), and Orchids (≥76.5%).

Multimedia Impact Surveys: To assess real-world behavioral manifestations of personality and sensitivity constructs, participants viewed

three video stimuli and rated their responses. Two videos featured controversial social media content selected for themes associated with increased negative emotional affect: (1) news report of missile strike on Ukrainian city; (2) news report of stabbing incident in Solingen, Germany. Content selection followed criteria established by [Bove et al. \(2024\)](#) and [Kusen and Strembeck \(2023\)](#) demonstrating differential emotional impact across personality profiles. The third video was a Coca-Cola product advertisement, selected to minimize gender-specific appeal. Participants rated emotional impact on 5-point scale after controversial content (1=Much worse, 5=Much better, baseline=3) and purchase likelihood after advertisement (1=Much less likely, 5=Much more likely, baseline=3). Responses were converted to percentage deviations from 50% baseline for analysis. Video presentation order was fixed (controversial content first, advertisement last) to prevent advertisement effects from influencing personality assessments, following experimental design principles minimizing survey fatigue and carryover effects ([Galesic and Bošnjak, 2009](#)).

4 Results & Discussion

4.1 LCA-Derived Classification Thresholds

Gaussian mixture modeling on HSP-R scores identified a three-class solution as optimal (BIC=312.4 vs. 2-class BIC=328.7, 4-class BIC=318.9), yielding thresholds: Dandelions (<61.2%, n=9, 16.4%), Tulips (61.2-76.5%, n=26, 47.3%), and Orchids (\geq 76.5%, n=20, 36.4%). Within-class homogeneity (entropy=0.84) confirmed distinct group membership. For Big Five dimensions, three-class solutions emerged as optimal for all traits with classification probabilities exceeding 0.80. These thresholds differed substantially from arbitrary median splits or tertile cutoffs, with several dimensions showing elevated boundaries reflecting actual score distributions.

4.2 Baseline vs. Weighted SFA Comparison

To establish the value of empirical weighting beyond pure semantic similarity, baseline analysis first examined unweighted Word2Vec embeddings. The baseline achieved 52.0% average cluster purity (range: 30.0%-60.0%), with optimal clustering at k=5 (silhouette score=0.476), validating that Big Five semantic foundation exists in distributional semantics. Both baseline and weighted

approaches identifying k=5 as optimal validates the Big Five structure: five personality dimensions emerge naturally from semantic patterns (c.f. [Müller and Degaetano-Ortlieb \(2025\)](#)). However, baseline performance varied dramatically, with Agreeableness showing particularly poor recovery (30.0% purity), demonstrating fundamental limitations of pure semantic similarity for socially-nuanced personality dimensions.

Empirically-weighted SFA achieved 62.5% average purity, representing substantial improvement over baseline. Critically, empirical weighting provided differential improvements: Agreeableness improved dramatically from 30.0% to 50.0%, while Openness improved from 60.0% to 85.7%. This pattern reveals that pure semantic similarity captures dimensions with clear lexical markers (Conscientiousness, Extraversion) but struggles with abstract constructs and socially-nuanced dimensions. Empirical weighting disambiguates context-dependent meanings through actual personality response patterns, substantially improving recovery for precisely those dimensions most challenging for distributional semantics alone.

4.3 Factor Structure Recovery: SFA vs. CFA

Table 1 presents the comparative analysis of factor structure recovery. Silhouette analysis indicated that a four-cluster solution (silhouette score=0.0433) provided the most coherent clustering of weighted adjective embeddings, achieving 66.7% average purity (see Figure 1). The optimal four-cluster solution included a highly pure Openness cluster (88.9% purity, n=9 adjectives), Conscientiousness cluster (72.7% purity, n=11 adjectives), Emotional Stability cluster (60% purity, n=10 adjectives), and a mixed cluster combining Extraversion, Agreeableness, and Emotional Stability items with 45% Extraversion dominance (n=20 adjectives).

In contrast, traditional CFA on IPIP-50 data (n=55) showed poor model fit across all indices: CFI=0.38 (target >0.90), TLI=0.35 (target >0.90), RMSEA=0.148 (target <0.08), SRMR=0.133 (target <0.08) (see Figure 2). Critical structural issues emerged: Openness and Extraversion factors collapsed, loading onto a common dimension (interfactor correlation $r=0.34$, the highest in the matrix). Despite excellent to good reliability for four dimensions—Extraversion ($\alpha=0.91$), Emotional

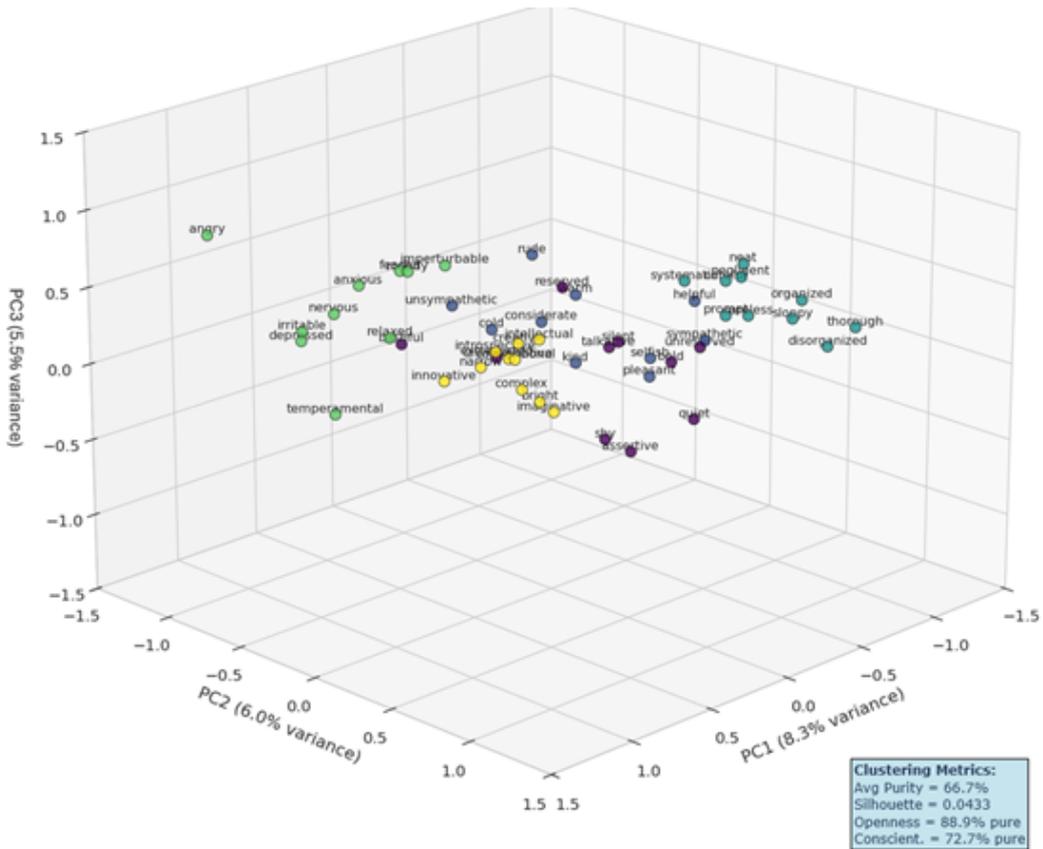


Figure 1: 3D Visualization of the Semantic Factor Analysis (SFA) results (Cyan=Conscientiousness, Green=Neuroticism, Purple=Extraversion, Blue=Agreeableness, Yellow=Openness to Experience)

Stability ($\alpha=0.86$), Conscientiousness ($\alpha=0.85$), and Agreeableness ($\alpha=0.79$)—the CFA failed to recover a coherent five-factor structure. Openness reliability fell to $\alpha=0.65$, well below the 0.70 acceptability threshold. Multiple items showed weak primary loadings (<0.40) and substantial cross-loadings. Modification indices suggested extensive model misspecification. While the small sample size ($n=55$) likely contributed strongly to CFA instability—conventional guidelines recommend $n>200$ for stable five-factor models (Comrey, 1973)—the results demonstrate SFA’s robustness to limited data when leveraging pre-trained semantic structures mediated by empirical participant data.

Both methods identified Conscientiousness as the most stable personality factor, with CFA showing good reliability ($\alpha=0.85$) and SFA achieving high cluster purity (72.7% in four-cluster solution). The most striking divergence concerned Openness to Experience: while CFA revealed complete factor collapse with items loading on Extraversion and reliability falling below acceptable thresholds—common in small-sample Big

Metric	SFA	CFA
<i>Factor Recovery (k=5 solution)</i>		
Openness	85.7%	Collapsed
Conscientiousness	60.0%	Recovered
Extraversion	50.0%	Collapsed
Agreeableness	50.0%	Recovered
Neuroticism	66.7%	Recovered
<i>Model Fit Indices</i>		
CFI	N/A	0.38
TLI	N/A	0.35
RMSEA	N/A	0.148
SRMR	N/A	0.133
<i>Baseline Comparison</i>		
Baseline purity	52.0%	N/A
Weighted purity	62.5%	N/A

Table 1: Comparison of SFA and CFA results.

Five assessments (Goldberg, 1992)—SFA identified Openness as the most coherent dimension with exceptional cluster purity (85.7% in five-cluster solution, 88.9% in four-cluster solution). This highlights the fundamental difference between empirical factor recovery in specific samples versus semantic structure validation using pre-trained language models.

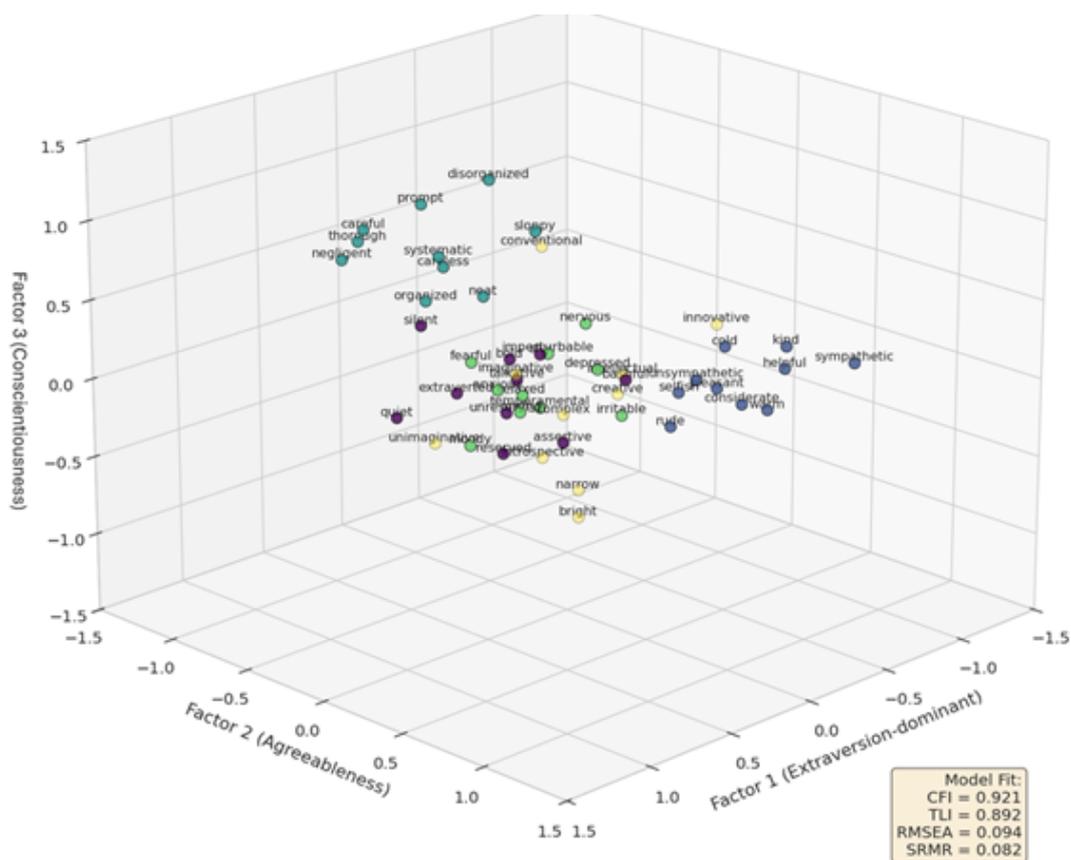


Figure 2: 3D Visualization of Confirmatory Factor Analysis (CFA) results

4.3.1 Higher-Order Factor Validation

Analysis of Big Five higher-order factors validated the hierarchical personality model (DeYoung et al., 2007). Stability (mean of Agreeableness, Conscientiousness, Emotional Stability) correlated strongly with constituent traits ($r=0.45-0.69$, all $p < 0.001$), while Plasticity (mean of Extraversion, Openness) showed very strong correlations ($r=0.76-0.91$, $p < 0.001$). Cross-loadings were minimal ($r_1 < 0.31$), supporting discriminant validity. The exceptionally high Extraversion-Openness interfactor correlation ($r=0.34$) likely reflects shared Plasticity variance and biological substrates (DeYoung et al., 2002), explaining why Openness items loaded on Extraversion in small-sample CFA despite forming distinct semantic clusters in SFA.

4.3.2 SFA vs CFA: Implications

Two implications emerge from the results. First, pre-trained embeddings capture personality-relevant semantic structures from natural language corpora at massive scale. When weighted by participant responses, these embeddings offer more robust representations than correlation matrices

from limited samples. Second, computational approaches may prove less sensitive to small-sample instabilities plaguing traditional factor analysis, as participant-weighted embeddings combine small-sample empirical data with semantic structures from billions of training tokens rather than relying solely on within-sample covariation. The factor purity metrics (50-88.9% range across solutions) validate that Word2Vec’s geometric organization aligns closely with theoretical personality constructs derived from decades of psychometric research. This convergence between computational linguistic structure and psychological theory strengthens the lexical hypothesis (Allport and Odbert, 1936): if personality is encoded in language, then distributional semantics—which learns from actual language use patterns—should naturally recover personality structures. The results provide strong empirical support for this theoretical connection.

4.4 Environmental Sensitivity Validation

4.4.1 HSP-R Subscale Structure and Big Five Correlations

Comprehensive analysis of the six HSP-R subscales revealed distinct patterns (M range: 60.5-79.6%) with Depth of Processing highest (M=79.6%, SD=14.9%) and Overstimulation lowest (M=60.5%, SD=22.7%). Subscales intercorrelated moderately ($r=0.130-0.543$), with strongest association between Social Sensitivity and Depth of Processing ($r=0.543$), supporting a higher-order sensitivity construct while maintaining discriminant validity.

Direct computation of HSP-R correlations with Big Five dimensions revealed a theoretically coherent pattern (see Appendix D). Overall HSP-R correlated most strongly with Conscientiousness ($r=0.442$, $p < 0.001$), followed by Neuroticism ($r=0.434$, $p < 0.001$), Agreeableness ($r=0.402$, $p < 0.001$), and Openness ($r=0.319$, $p < 0.05$), while Extraversion demonstrated essentially no relationship ($r=0.003$, n.s.). The Conscientiousness correlation magnitude was notably larger than anticipated, likely reflecting the HSP-R scale design incorporating Depth of Processing and Sensitivity to Details subscales that correlate strongly with conscientious functioning. Specifically, Depth of Processing correlated $r=0.455$ with Conscientiousness and $r=0.537$ with Openness ($p < 0.001$), suggesting shared variance in cognitive elaboration and reflective processing. Sensitivity to Details showed $r=0.473$ with Conscientiousness ($p < 0.001$), validating theoretical connections between sensory awareness and organized behavior.

The Neuroticism correlation ($r=0.434$) was primarily driven by Overstimulation ($r=0.478$, $p < 0.001$) and Emotional Reactivity ($r=0.479$, $p < 0.001$), demonstrating that sensitivity to environmental stressors and emotional responsiveness constitute core features of both constructs. The near-zero Extraversion correlation ($r=0.003$) emerged from opposing subscale effects: Overstimulation correlated negatively with Extraversion ($r=-0.218$, $p < 0.05$), reflecting social withdrawal under stimulation, while Sensitivity to Positive Experiences correlated positively ($r=0.213$, $p < 0.05$), reflecting openness to rewarding experiences. These opposing patterns largely canceled out in the overall correlation, supporting sensitivity as partially independent from Extraversion. Social Sensitivity showed strongest relationship with

Agreeableness ($r=0.514$, $p < 0.001$), validating that empathic awareness and emotional responsiveness constitute shared features between these constructs.

Multiple regression analysis revealed Big Five dimensions collectively explained 46.4% of HSP-R variance ($R^2=0.464$), remarkably consistent with large-scale validation findings of 44.0% despite substantially smaller sample size ($n=55$ vs. $n=400$) (Pluess et al., 2024). Conscientiousness ($\beta=0.295$) and Neuroticism ($\beta=0.315$) emerged as strongest unique predictors. The substantial unexplained variance (55%) demonstrates that Environmental Sensitivity represents a distinct phenomenon, with more than half of its variance independent from established personality dimensions. This supports Environmental Sensitivity as a construct separate from the Big Five rather than merely a combination of existing personality traits.

Gender analysis revealed females demonstrated higher overall HSP-R scores (M=75.0% vs. M=66.8%, difference=8.2%), with most pronounced difference in Overstimulation (females M=68.5% vs. males M=47.5%, difference=21.1%). See Appendix E for detailed gender distributions across dimensions.

4.5 Hypothesis Testing Outcomes

Eight sub-hypotheses examined relationships between personality traits, Environmental Sensitivity, and multimedia content responses—specifically controversial social media posts and product advertisements.

Binary comparisons contrasted high groups against combined low/medium groups for personality traits (H2-H5) and Orchids against combined Tulips+Dandelions for sensitivity (H6, H8). Four hypotheses received support: (H2) High Agreeableness associated with greater emotional reactivity (M=31.4% vs. 23.3%, Cohen's $d=0.76$); (H5) High Extraversion associated with increased purchase likelihood (M=2.8% vs. -0.8%); (H6) Orchids reported higher emotional impact (M=31.4% vs. 23.2%, $d=0.77$); (H9) Dandelions showed minimal purchase impact (M=-2.8%, within practical equivalence of baseline).

Unsupported hypotheses revealed important nuances: (H3) High Neuroticism showed minimal difference in emotional reactivity (M=27.1% vs. 25.6%), suggesting neuroticism involves self-focused rather than other-focused negative emo-

tions; (H4) High Openness showed advertising resistance rather than susceptibility ($M=-4.2\%$ vs. 2.3%); (H7) Dandelions demonstrated substantial emotional reactivity ($M=23.0\%$) rather than minimal response; (H8) Orchids showed no systematic pattern in purchase likelihood ($M=-2.5\%$ vs. -0.2%).

The pattern of results revealed important distinctions between emotional reactivity and consumer behavior domains. Environmental Sensitivity demonstrated clear predictive validity for emotional responses to controversial content, supporting core theoretical propositions about heightened environmental responsiveness. However, advertising susceptibility showed weaker patterns, suggesting that consumer behavior involves psychological mechanisms that are unlikely to be linked to the Environmental Sensitivity framework in any meaningful way. The finding that highly open individuals showed advertising resistance rather than susceptibility highlights the importance of empirical hypothesis testing.

5 Conclusion

The comparison between traditional CFA and the newly proposed SFA reveals a compelling case for methodological complementarity in personality structure research. Each approach provides unique insights that, when considered together, offer a more comprehensive understanding of personality organization than either method alone could achieve. SFA demonstrates that computational approaches integrating distributional semantics with empirical data can match or exceed traditional psychometric methods in personality structure recovery, particularly with small samples ($n=55$) where CFA shows factor collapse and poor model fit. The baseline comparison establishes that empirical weighting provides substantial improvement over pure semantic similarity, particularly for socially-nuanced dimensions. The successful recovery of personality structure through distributional semantics provides computational support for the lexical hypothesis.

Comprehensive HSP-R validation revealed robust correlations with Big Five dimensions ($R^2=0.464$), replicating large-scale findings ($R^2=0.44$, [Pluess et al. \(2024\)](#)) despite smaller sample. The substantial unexplained variance (55%) demonstrates Environmental Sensitivity as a distinct phenomenon, with more than half of

its variance independent from personality dimensions. The empirically-derived LCA thresholds (61.2%, 76.5%) provide data-driven boundaries validating the trichotomous model.

5.1 Future Work

Extending SFA to the Big Five Aspect Scale (BFAS) represents a critical methodological advance ([DeYoung et al., 2007](#)). The BFAS subdivides each Big Five dimension into two aspects (10 total), addressing the well-documented heterogeneity within broad Big Five dimensions. This finer-grained structure would enable more accurate personality assessment through SFA, as aspects are more homogeneous constructs than their parent dimensions. Individual-level prediction from natural language represents another critical extension, mapping text production to personality scores for social media analysis and workplace assessment. Cross-cultural and multilingual validation would establish generalizability beyond Western, English-language contexts. Application of SFA to domains beyond personality—including values, attitudes, and psychological constructs like depression or anxiety—would demonstrate broader utility of semantic factor analysis methodology.

6 Limitations

The study faces several limitations. The sample size ($n=55$), while adequate for validating SFA methodology, limits generalizability and contributed to CFA instability. Larger samples ($n>200$) would enable more rigorous comparison and formal mixture modeling, though results suggest embedding-based methods may be advantageous precisely when traditional approaches fail due to small samples. The Word2Vec training corpus (Google-News-300 model) may not optimally represent personality-relevant language, as it was trained on general news text rather than personality-relevant discourse. Domain-specific embeddings trained on psychological literature or conversational corpora could improve structure recovery.

The methodology required pre-specified adjective markers from [Goldberg \(1992\)](#), limiting flexibility in construct selection. SFA currently recovers dimensional structures but not individual-level scores, restricting application to group-level analyses. The approach depends on availability of val-

idated adjective inventories, which may not exist for all psychological constructs or languages.

Multimedia surveys suffered from single-item measures for emotional impact and purchase likelihood, potentially explaining null findings (H4). Age correlations were weak ($r=-0.09$ to $r=0.11$), reflecting restricted range (19-41 years) and trait stability in young adulthood (?). Only English-language assessment with mainly Western university students was tested, limiting cross-cultural generalizability.

Computational considerations include dependency on pre-trained embeddings quality and architecture. Results may vary with different embedding models (GloVe, FastText, BERT). PCA dimensionality reduction discards information (first 5 components explained only 30.1% variance). Alternative dimensionality reduction methods (t-SNE, UMAP) might improve structure recovery but require further investigation.

7 Ethics Statement

All participants provided informed consent. The study received institutional review board approval. Email addresses were stored separately from response data and deleted after results distribution.

Acknowledgments

Special thanks to Stefania Degaetano-Ortlieb for her ongoing support and guidance since our joint pilot study (Müller and Degaetano-Ortlieb, 2025).

References

- Gordon W Allport and Henry S Odbert. 1936. [Trait-names: A psycho-lexical study](#). *Psychological Monographs*, 47(1):i-171.
- Elaine N Aron and Arthur Aron. 1997. [Sensory-processing sensitivity and its relation to introversion and emotionality](#). *Journal of Personality and Social Psychology*, 73(2):345-368.
- Vincenzo Bove, Georgios Efthymoulou, Armine Ghazaryan, and Harry Pickard. 2024. [The emotional effect of terrorism](#). *Scientific Reports*, 14(1):26525.
- Giulio Carducci, Giuseppe Rizzo, Diego Monti, Enrico Palumbo, and Maurizio Morisio. 2018. [Twitpersonality: Computing personality traits from tweets using word embeddings and supervised learning](#). *Information*, 9(5):127.
- Andrew L Comrey. 1973. *A first course in factor analysis*. Academic Press, New York.
- Paul T Costa and Robert R McCrae. 1992. *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources, Odessa, FL.
- Colin G. DeYoung, Jordan B. Peterson, and Daniel M. Higgins. 2002. [Higher-order factors of the big five predict conformity: Are there social determinants of personality?](#) *Journal of Personality and Social Psychology*, 83(2):533-552.
- Colin G DeYoung, Lena C Quilty, and Jordan B Peterson. 2007. [Between facets and domains: 10 aspects of the big five](#). *Journal of Personality and Social Psychology*, 93(5):880-896.
- Mirta Galesic and Michael Bošnjak. 2009. [Effects of questionnaire length on participation and indicators of response quality in a web survey](#). *Public Opinion Quarterly*, 73(2):349-360.
- Lewis R Goldberg. 1992. [The development of markers for the big-five factor structure](#). *Psychological Assessment*, 4(1):26-42.
- Lewis R Goldberg. 1999. [A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models](#). *Personality Psychology in Europe*, 7(1):7-28.
- Listou H Grimen and Åse Diseth. 2016. [Sensory processing sensitivity: Factors of the highly sensitive person scale and their relationships to personality and subjective health complaints](#). *Perceptual and Motor Skills*, 123(3):637-653.
- Judith R Homberg, Daniel Schubert, Esther Asan, and Elaine N Aron. 2016. [Understanding the molecular mechanisms underlying environmental sensitivity](#). *Neuroscience & Biobehavioral Reviews*, 71:476-490.
- Amirmohammad Kazameini, Samin Fatehi, Yash Mehta, Sauleh Eetemadi, and Erik Cambria. 2020. [Personality Trait Detection Using Bagged SVM over BERT Word Embedding Ensembles](#). *arXiv preprint arXiv:2010.01309*.
- Emma Kusen and Mark Strembeck. 2023. [Investigating the impact of personality traits on emotional communication in social media: A study on twitter users during crisis events](#). *Computers in Human Behavior*, 141:107623.
- Francesca Lionetti, Elaine N Aron, Arthur Aron, Denise N Klein, and Michael Pluess. 2018. [Dandelions, tulips and orchids: Evidence for the existence of low-sensitive, medium-sensitive and high-sensitive individuals](#). *Translational Psychiatry*, 8(1):1-11.
- Robert R McCrae and Oliver P John. 1992. [An introduction to the five-factor model and its applications](#). *Journal of Personality*, 60(2):175-215.

- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). *arXiv preprint arXiv:1301.3781*.
- Oliver Müller and Stefania Degaetano-Ortlieb. 2025. [Embedded personalities: Word embeddings and the “big five” personality model](#). In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 205–215, Albuquerque, New Mexico. Association for Computational Linguistics.
- Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin E P Seligman. 2015. [Automatic personality assessment through social media language](#). *Journal of Personality and Social Psychology*, 108(6):934–952.
- Michael Pluess, Francesca Lionetti, Elaine N Aron, and Arthur Aron. 2024. [The highly sensitive person scale-revised: Psychometric properties and recommendation for use](#). *Assessment*, 31(2):289–304.
- Yves Rosseel. 2012. [lavaan: An r package for structural equation modeling](#). *Journal of Statistical Software*, 48(2):1–36.
- H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, and 1 others. 2013. [Personality, gender, and age in the language of social media: The open-vocabulary approach](#). *PloS one*, 8(9):e73791.
- Ernest C Tupes and Raymond E Christal. 1961. [Recurrent personality factors based on trait ratings](#). *Journal of Personality*, 60(2):225–251. Original work published 1961.

Appendix

A IPIP-50 Item List

The IPIP-50 comprises 10 items per Big Five dimension (50 total), rated on a 5-point scale (1=Very Inaccurate, 5=Very Accurate). Reverse-scored items marked with (R). Items were slightly modified to improve comprehension for German native speakers (e.g., "Seldom feel blue" → "I rarely feel sad"; "Shirk my duties" → "I neglect my duties"). See [Goldberg \(1999\)](#) for complete psychometric properties. The personal pronoun I was also added in front of each item.

Openness: 5. I have a rich vocabulary; 10. I have difficulty understanding abstract ideas (R); 15. I have a vivid imagination; 20. I am not interested in abstract ideas (R); 25. I have excellent ideas; 30. I do not have a good imagination (R); 35. I am quick to understand things; 40. I use difficult words; 45. I spend time reflecting on things; 50. I am full of ideas.

Conscientiousness: 3. I am always prepared; 8. I leave my belongings around (R); 13. I pay attention to details; 18. I make a mess of things (R); 23. I get chores done right away; 28. I often forget to put things back in their proper place (R); 33. I like order; 38. I neglect my duties (R); 43. I follow a schedule; 48. I am exacting in my work.

Extraversion: 1. I am the life of the party; 6. I don't talk a lot (R); 11. I feel comfortable around people; 16. I keep in the background (R); 21. I start conversations; 26. I have little to say (R); 31. I talk to a lot of different people at parties; 36. I don't like to draw attention to myself (R); 41. I don't mind being the center of attention; 46. I am quiet around strangers (R).

Agreeableness: 2. I feel little concern for others (R); 7. I am interested in people; 12. I insult people (R); 17. I sympathize with others' feelings; 22. I am not interested in other people's problems (R); 27. I have a soft heart; 32. I am not really interested in others (R); 37. I take time out for others; 42. I feel others' emotions; 47. I make people feel at ease.

Neuroticism: 4. I get stressed out easily; 9. I am relaxed most of the time (R); 14. I worry about things; 19. I rarely feel sad (R); 24. I am easily disturbed; 29. I get upset easily; 34. I change my mood a lot; 39. I have frequent mood swings; 44. I get irritated easily; 49. I often feel sad.

B HSP-R Scale Items

The Highly Sensitive Person Scale-Revised (HSP-R) comprises 18 items across six subscales, rated on 7-point scale (1=Not at all, 7=Extremely). See [Pluess et al. \(2024\)](#) for complete validation.

Depth of Processing: Do you tend to reflect on things deeply?; Do you like deep conversations?; Do you find yourself thinking about philosophical questions?

Emotional Reactivity: Are you easily affected by feedback (both negative and positive)?; Do you generally react strongly to your experiences, whether you show it or not?; Are you easily affected by the mood of people around you?

Overstimulation: Are you easily overwhelmed by things like bright lights, strong smells, coarse fabrics, or sirens close by?; Do you become unpleasantly aroused when a lot is going on around you?; Are you bothered by intense stimuli, like loud noises or chaotic scenes?

Sensitivity to Details: Do you notice when things have been moved around?; Do you tend to notice subtle signs of changing seasons?; Do you seem to notice changes in the weather more than others do?

Sensitivity to Positive Experiences: Are you deeply moved by the arts or music?; Do you notice and enjoy delicate or fine scents, tastes, sounds, works of art?; Do you tend to get deeply immersed in music?

Social Sensitivity: Do you easily recognize what others are feeling?; Are you good at anticipating how someone may feel about a situation?; Do other people tell you that you are good at understanding what they are feeling or thinking?

C 50 IPIP-50 Adjectives with Replacements

Complete list of 50 adjectives used in Semantic Factor Analysis, derived from IPIP-50 statements. Four adjectives were unavailable in Word2Vec vocabulary and replaced with semantic equivalents (marked with *).

Extraversion (10): extraverted, *silent**, unreserved, reserved, talkative, quiet, bold, bashful, assertive, shy

Agreeableness (10): selfish, warm, rude, sympathetic, unsympathetic, kind, cold, helpful, considerate, pleasant

Conscientiousness (10): organized, disorganized, thorough, careless, prompt, sloppy, neat,

negligent, systematic, careful

Neuroticism (10): anxious, relaxed, nervous, imperturbable, touchy, *cranky**, temperamental, moody, irritable, *sad**

Openness (10): intellectual, unintellectual, imaginative, *incurious**, innovative, unimaginative, bright, complex, introspective, creative

Synonym Replacements Used:

- Item S6 (Extraversion): *untalkative* → *silent*
- Item S29 (Neuroticism): *short-tempered* → *cranky*
- Item S49 (Neuroticism): *self-pitying* → *sad*
- Item S20 (Openness): *uninquisitive* → *incurious*

All replacements were verified to exist in the Google News Word2Vec vocabulary before analysis.

D HSP-R × Big Five Intercorrelations

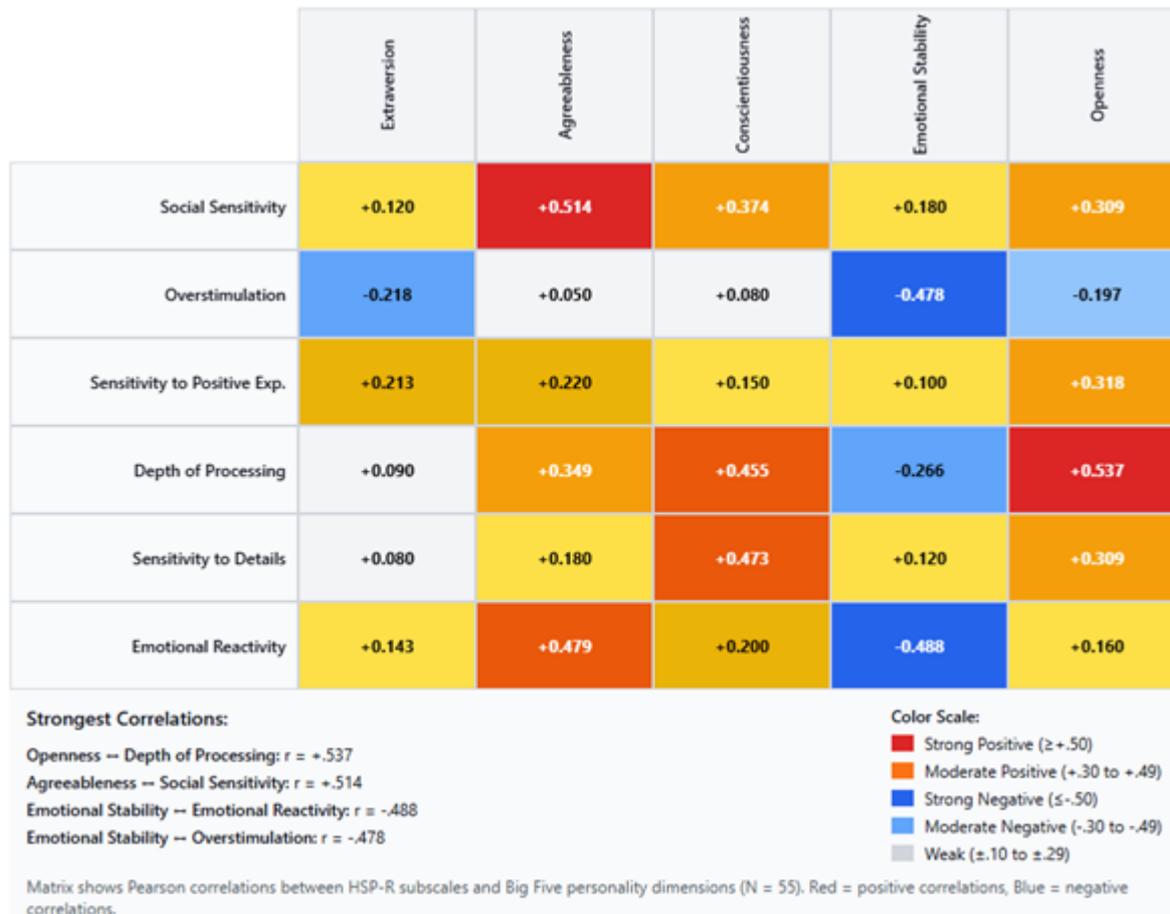


Figure 3: Intercorrelation heatmap between HSP-R subscales and Big Five personality dimensions (N=55)

E Gender Distributions

Dimension	Males (%)	Females (%)	Diff. (%)
<i>Big Five Personality Dimensions</i>			
Emotional Stability	63.1	45.4	-17.7
Extraversion	61.2	54.5	-6.8
Agreeableness	77.4	79.8	+2.3
Conscientiousness	70.6	71.0	+0.4
Openness	77.6	75.8	-1.8
<i>HSP-R Dimensions</i>			
Overall HSP-R	66.8	75.0	+8.2
Overstimulation	47.5	68.5	+21.1
Emotional Reactivity	69.4	77.4	+7.9
Sens. to Details	61.7	69.6	+7.9
Sens. to Pos. Exp.	72.8	80.6	+7.8
Depth of Processing	77.7	80.8	+3.1
Social Sensitivity	71.9	73.1	+1.2

Table 2: Gender differences Big Five & HSP-R dimensions. All scores in percentages. Negative differences indicate male advantage; positive differences indicate female advantage.

Quantitative Analysis of Rhyme and Metre in LLM-generated Translations of Poetry

Jan-Felix Klumpp

Department of Linguistics

University of Tübingen

jan-felix.klumpp@uni-tuebingen.de

Abstract

While machine translation systems have been applied to many tasks with remarkable success, machine poetry translation has remained a challenge. This study investigates the capabilities of generative Large Language Models (LLMs) in the translation of poetry (taking Shakespeare’s 154 sonnets as an example) from English to German. For this purpose, I define metrics that assess the reproduction of the rhyme scheme and the metre of the original in a quantitative way. The results indicate that LLMs still lag behind professional human translators (especially with regard to the reproduction of the rhyme scheme), but that their performance is significantly influenced by the chosen prompt strategy. In particular, iteratively refining the result emerges as a successful strategy in terms of the reproduction of the form, but this comes at the expense of other aspects such as grammaticality and the reproduction of the meaning.

1 Introduction

Poetry translation is a particularly difficult type of translation; it has even been claimed to be “one of the hardest possible tasks that can be considered in computational linguistics” (Genzel et al., 2010, p.1). The main reason for this is that the translation of a poem has to fulfill very different and sometimes conflicting expectations. One such conflict is that the translation has to resemble a poem in another language (in terms of content and style), but also constitutes a work of art in itself, which has led to the distinction between “free” vs. “faithful,” also called “domesticated” vs. “foreignized” translations (Schleiermacher and Bernofsky, 2021, Venuti, 2017, cited in Jones and Irvine, 2013). Another conflict is that the original poem has not only a specific meaning, but also a specific poetic form, whose reproduction imposes boundaries on rendering the meaning faithfully (Genzel et al., 2010). Which of these aspects are deemed more or less

important depends not only on the preferences of translators and readers, but also on aspects such as the author of the original (Gibbons, 1985), the target audience, and the context in which the translation is to be used (such that, e.g., a poem has to be translated differently if the translation is intended to be sung, Low, 2003).

Within just a few years, generative Large Language Models (LLMs) have permeated all areas of NLP applications. In the area of machine poetry translation, the flexibility of instruction-tuned LLMs to react to different prompt formulations may allow to produce translations more closely aligned to reader preferences. However, despite their impressive performance on many other NLP tasks, poetry translation still remains challenging for modern generative LLMs (Resende and Hadley, 2024).

The wide range of sometimes conflicting expectations also has consequences for the evaluation of translated poetry: A translation of a poem cannot be evaluated globally, but only with regard to specific aspects, which have to be defined clearly. So far, however, only little attention has been paid to analyzing individual aspects of LLM-translated poetry in a quantitative fashion. This study attempts to narrow this gap by conducting an analysis of LLM-generated translations of all 154 of Shakespeare’s sonnets into German, using numerical metrics that can transparently be mapped to well-defined aspects of poetry translation. This automated yet interpretable way of evaluation allows to compare different LLMs and prompting strategies on a relatively large corpus of poetry.

A quantitative analysis of LLM-generated translations of poetry requires focusing on aspects that can reliably be characterized numerically. This study focuses on three such aspects, namely, the *reproduction of the meaning* of the original (understood as the rather literal semantic content of the poem), the *reproduction of the metre*, and the *repro-*

duction of the rhyme scheme of the original. The reproduction of meaning and form (with the latter including metre and rhyme scheme) is often characterized as a tradeoff relation (Genzel et al., 2010), which makes these aspects particularly suitable for contrasting prompt strategies.¹

2 Related Literature

Research on the topic of machine poetry translation comes from two somewhat different directions. One strand of research approaches the task mainly as a technical problem and focuses on developing systems that reach a better overall performance. Since the 2010s, several approaches to neural machine poetry translation have been explored: Genzel et al. (2010) and Ghazvininejad et al. (2018) use decoding constraints that restrict the options to those fulfilling some manually defined formal criteria (e.g., rhyming words). Chakrabarty et al. (2021) fine-tune multilingual LLMs on translated poetry to improve their performance. To overcome the lack of large training corpora of translated poetry, Song et al. (2023) and Huynh and Bao (2024) employ a style injection paradigm, where the original poem is first translated into prose and then converted into poetry by a separate style injection module, which can be trained using target language poetry and synthetic prose equivalents generated via backtranslation. Finally, multi-step prompting strategies are used, where additional information retrieved from the language model itself (Wang et al., 2024) or a database (Chen et al., 2025) is given back to the model to improve the quality of the translation.

While the aforementioned literature has focused on developing systems that can translate poetry better, the availability of modern generative LLMs that are able to perform a variety of different tasks has led to another strand of research, which investigates the properties of translations generated by these LLMs. Using a combination of automated metrics and manual evaluation, Resende and Hadley (2024) analyse LLM-generated translations of four poems according to criteria such as the content, the form and the style (in a Spanish/Portuguese-to-English setting). In a similar vein, Karaban and Karaban (2024) compare LLM-generated English translations of twelve poems by the Ukrainian poet Ivan

Franko to those written by a human translator. Gao et al. (2024) compare translations of Chinese poetry by ChatGPT to those generated by conventional machine translation systems (DeepL and Google Translate). Other work that mostly uses manual close reading rather than automated quantitative metrics has investigated LLM-generated translations of individual poems from and into other (both high-resource and low-resource) languages, e.g., Kazakh to English (Manapbayeva et al., 2024), Arabic dialects to English (AlAfnan and Alshakhs, 2025), Spanish to Filipino and Spanish to English (Liwana et al., 2024), and English, French and Arabic to each other (Almaktary, 2025). However, most work that investigates specific aspects of translated poetry does so only on a small number of poems, which adversely affects a meaningful statistical analysis. Using quantitative yet transparently interpretable metrics, this study extends the analysis of LLM-generated translations of poetry to a larger number of poems and experimental conditions.

3 Data and Methods

3.1 Data

The original texts to be translated are William Shakespeare’s 154 sonnets, taken from a 1923 edition with modernized orthography, but without any other changes (Shakespeare, 2019). The LLM-generated translations are compared to five human-written German translations from the 19th and early 20th century: Dorothea Tieck’s translation (probably from 1825, published posthumously) (Jansohn, 1992), Gottlob Regis’ 1836 translation (Regis, 2004), Emil Wagner/Ludwig Walesrode’s 1840 translation (Walesrode, 2007), Max Joseph Wolff’s 1903 translation (Wolff, 1924), and Stefan George’s 1909 translation (George, 2004).

3.2 LLMs

Five LLMs are evaluated, three decoder-only Transformer models (OpenAI’s GPT-4o, Hurst et al., 2024, Anthropic’s Claude 3.5 Sonnet, Anthropic, 2024a, Anthropic, 2024b, and Google’s Gemini 1.5 Gemini Team, Google, 2024), one reasoning model (OpenAI’s OpenAI o4-mini, OpenAI, 2025b), and one unified model that flexibly uses both strategies (OpenAI’s GPT-5, OpenAI, 2025a).²

¹The code and other materials related to this study can be found at: <https://github.com/janfeliyklump/AnalyseLLMPoetryTranslation>.

²The API was used for gpt-5 and the analysis conditions of gpt-4o, and the web user interfaces (accessed in late 2024 and 2025) were used otherwise.

Beside human-written translations, the LLM-generated translations are also compared to translations produced by DeepL (DeepL SE, 2017), a standard machine translation tool based on a neural network architecture.

3.3 Prompting Strategies

I use zero-shot prompting with different instructions, both in single-step strategies that directly generate the translation and in multi-step strategies, where the output is reused in subsequent steps. Three single-step prompting strategies (which center different aspects) are evaluated for all LLMs:

- Plain: a simple instruction to translate a poem from English to German,
- PlainForm: like Plain, but with an explicit instruction to reproduce the form of the original,
- PlainMeaning: like Plain, but with an explicit instruction to reproduce the meaning of the original as closely as possible.

For GPT-4o, I evaluate additional multi-step strategies, which can be divided into two categories: First, following Chen et al. (2024), I test iterative strategies, where the later steps are used to improve on a previous translation attempt:

- IterativeForm: the output from PlainForm is given back to the LLM together with the original poem and the instruction to reproduce the form better,
- IterativeMeaning: the output from PlainMeaning is used likewise, but with the instruction to reproduce the meaning better,
- IterativeMixed: the output from PlainMeaning is used, but the instruction is to reproduce the form better,
- IterForm2Steps: the procedure used for IterativeForm is applied a second time to the output of that condition.

Second, I manually decompose the translation into a two-step procedure, where the LLM is instructed to analyse the form of the poem first and then to translate it in a second step. This can be regarded as a task-specific instance of the Generated Knowledge Prompting strategy introduced by Liu et al. (2022).

- AnalysisTranslate: in a first step, the form of the original is analysed, and this information is then provided to the LLM together with the original and the instruction to produce a translation that reproduces the form,
- AnalysisRewrite: in a first step, the form of the original is analysed; then, the output from PlainMeaning is adjusted to that form.

The exact prompts can be found in Appendix A.

3.4 Evaluation

3.4.1 Reproduction of form

Shakespeare’s sonnets almost all follow the prototypical form of the English (or Shakespearean) Sonnet, which is defined by its *metre* and its *rhyme scheme*. The metre is —for most English and German poetry— an abstract pattern of prototypically stressed and unstressed syllables in a line. The rhyme scheme characterizes which pairs of lines rhyme. In this study, a pair of lines is considered to rhyme if and only if the sequence of phonemes in one line from the last stressed vowel onward is also found at the end of the other line and the preceding consonant onset is different; however, in accordance with the conventions of much of German poetry (Peust, 2014), front rounded and front unrounded vowels are considered equivalent, as are long and short vowels.

These two aspects of the form are evaluated by comparing the translation to a reference metre and reference rhyme scheme. The reference metre for (almost all of) Shakespeare’s sonnets is the iambic pentameter $x/x/x/x/x/(x)$ (where “x” stands for an unstressed and “/” for a stressed position)³, and their reference rhyme scheme is ABABCDCDEFE-FGG (where all lines with the same letter rhyme with each other)⁴. Using the Phonemizer library for Python (Bernard and Titeux, 2021), I extract the sequence of stressed and unstressed syllables as well as the phoneme sequences at the end of each line from the translation. I define the following metrics to compare the phonetic properties of the translation to the reference metre and rhyme scheme:

1. The *metrical distance* is a line-wise weighted edit-distance (based on the Levenshtein distance, Levenshtein et al., 1966) between the

³except for Sonnet 145, where it is $x/x/x/x/(x)$

⁴except for Sonnet 99, where it is ABABACDCDEFEGG, and Sonnet 126, where it is AABBCDDDEEFF

sequence of stresses in the translation and the reference metre. It is a real number greater than or equal to 0 and is to be minimized. The costs for insertion, deletion and substitution operations are derived from stress values that are assigned to all syllables in both the translation and the reference metre. The cost of insertion or deletion is equal to the stress value of the inserted or deleted element, while the cost of substitution is the difference between the two stress values. Stressed syllables are assigned a stress value of 1.0, unstressed syllables are assigned 0.8, secondarily stressed syllables 0.9, and optional syllables 0.4. These stress values are chosen such that substituting a stressed for an unstressed syllable or vice versa is less costly than inserting or deleting a whole syllable. This corresponds to the fact that such substitutions are frequent even in Shakespeare’s original sonnets, while the number of syllables per line of the original always follows the metrical pattern (Heldner, 2013). If the translation and the reference differ in the number of lines, the shortest lines of the text with more lines are treated as if they were completely inserted or deleted (i.e., the corresponding metrical distance is the sum of stress values). The total metrical distance is divided by the number of lines in the reference metre to obtain the average metrical distance per line. Pseudo-code for the metrical distance algorithm can be found in Appendix B.

2. The *rhyme scheme similarity* is the harmonic mean (F_1 score) of precision and recall with respect to the overlap in rhyming pairs of lines in the reference rhyme scheme and the translation. It is a real number between 0 and 1 that is to be maximized. Each line is assigned a number (starting from the first line), and rhyming pairs are represented as pairs of these numbers. Such an evaluation based on individual pairs of lines allows to capture the internal structure of parts of the poem better than a global comparison of the rhyme scheme (as used by Resende and Hadley, 2024).

3.4.2 Reproduction of meaning

Previous studies on machine poetry translation have compared the machine-produced translations to human-written reference translations, using metrics such as BLEU (e.g., Genzel et al., 2010, Re-

sende and Hadley, 2024, Karaban and Karaban, 2024) or BERTScore (e.g., Song et al., 2023, Wang et al., 2024). However, human-written translations of poetry often consider or prioritize aspects other than the mere reproduction of the meaning and thus cannot be considered objective gold standard references for this purpose. For this reason, I use the CometKiwi metric (Rei et al., 2022), which is a quality-estimation metric that does not rely on reference translations but directly assigns a score based on the source text and the candidate translation.

4 Results

4.1 Plain prompting strategies

Fig. 1 shows the average CometKiwi scores, metrical distances per line and rhyme scheme similarities for the three plain conditions (Plain, PlainForm, PlainMeaning) for all five models used (GPT-4o, Claude 3.5 Sonnet, Gemini 1.5, OpenAI o4-mini, GPT-5), as well as the same metrics for the DeepL-generated translation and the five human translations. These results allow the following observations: First, there is a clear tradeoff relation between the reproduction of meaning (as approximated by the CometKiwi score) and the reproduction of form (as described by metrical distance and rhyme scheme similarity), with the DeepL translation prioritizing the reproduction of meaning, the human translations prioritizing the reproduction of form, and the LLM-generated translations in between. There is even some differentiation between human-written translations, such that, e.g., Gottlob Regis’ translation, which conforms to the form of the original less strictly than most other translations (Horstmann, 2002), performs lower on the two form-related metrics. LLM translations also are more spread out along the axes (i.e., less consistent in the weight they assign to the criteria), as can be seen from their higher standard deviations (see Appendix C).

Second, the chosen prompting strategy strongly influences whether the reproduction of the form or of the meaning is prioritized. PlainForm improves the reproduction of the form for all five models. The differences for metre and rhyme between PlainForm and Plain are clearly significant under a significance level of 5% (see Table 1).⁵ In contrast, PlainMeaning improves the reproduction

⁵All p -values in this study were obtained using paired t -tests.

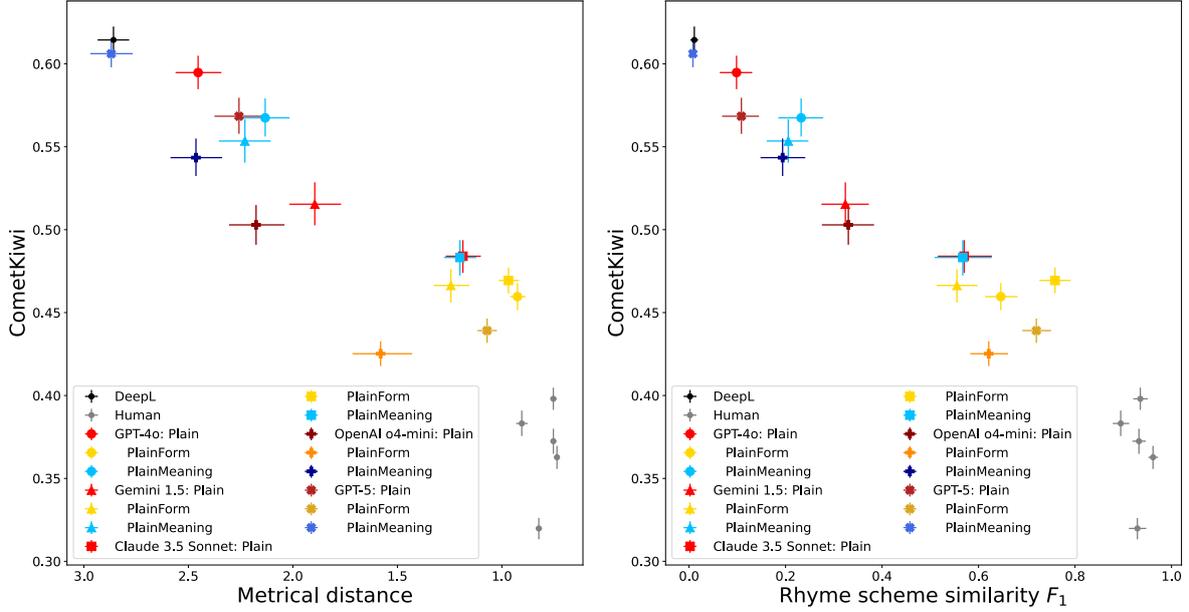


Figure 1: **Left:** Metrical distance per line (x-axis) and CometKiwi score (y-axis) for translations of Shakespeare’s sonnets into German by several LLMs under single-step prompt strategies. **Right:** Rhyme scheme similarity (x-axis) and CometKiwi score (y-axis) for the same conditions. The markers indicate the average values, the error bars indicate 95% bootstrapped confidence intervals.

of meaning only for Gemini 1.5, OpenAI o4-mini, and GPT-5, while it has the opposite effect for GPT-4o. For all models (except possibly Claude 3.5 Sonnet) the prompt formulations that lead to a better reproduction of the form reproduce the meaning less well.

Among the LLMs, Claude 3.5 Sonnet on average reproduces the rhyme scheme best, and tends to reproduce the form to a considerable extent even when the prompt does not explicitly mention this aspect. In contrast, the other models’ results are strongly prompt-dependent. OpenAI o4-mini on average performs worse than the other models (when only these three criteria are considered), with the CometKiwi scores being lower than those of otherwise comparable translations produced by other LLMs, particularly for PlainForm. This may be in part due to particularities of the resulting translations (see Section 4.3).

4.2 Multi-step prompting strategies

Fig. 2 shows the average CometKiwi scores, metrical distances per line and rhyme scheme similarities for the seven multi-step conditions (Iterative, IterativeForm, IterativeMeaning, IterativeMixed, IterForm2Steps, AnalysisTranslate, and AnalysisRewrite) for GPT-4o, as well as DeepL’s translation, the human-written translations and the plain conditions for comparison.

Model	metrical distance	rhyme scheme similarity F_1
GPT-4o	1.36×10^{-58}	1.70×10^{-52}
Claude 3.5 Sonnet	5.01×10^{-7}	1.08×10^{-8}
Gemini 1.5	2.59×10^{-17}	5.91×10^{-13}
OpenAI o4-mini	1.02×10^{-9}	4.95×10^{-16}
GPT-5	3.02×10^{-45}	4.24×10^{-55}

Table 1: p -values for the hypothesis that the explicit inclusion of the target criteria in the prompt leads to improvement on form-related metrics for different models (i.e., PlainForm compared to Plain).

These results indicate that iterative prompting strategies, i.e., IterativeForm, IterativeMixed and IterForm2Steps, succeed in reproducing the form better than any of the plain conditions. The differences to PlainForm for metre and rhyme scheme are clearly significant under a significance level of 5% (see Table 2). With these iterative strategies, the metre is reproduced similarly well as in the human translations, while in terms of the rhyme scheme, no strategy reaches human-like performance. On the other hand, IterativeMeaning and IterativeMixed do not reproduce the meaning better than Plain or PlainMeaning. This mirrors the pattern observed for the plain conditions, where

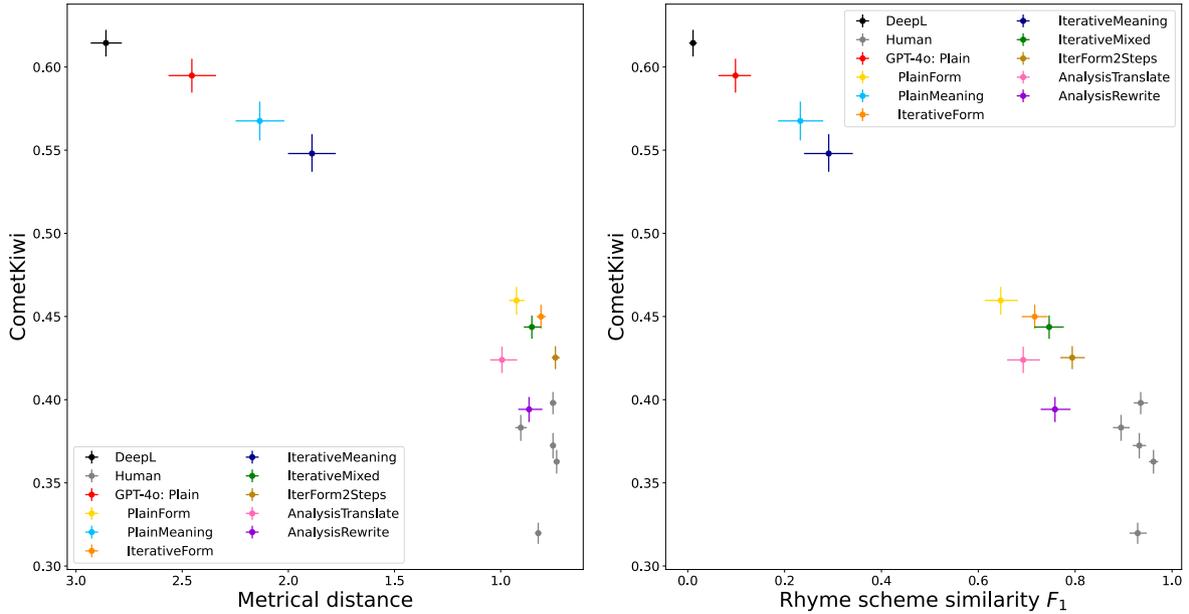


Figure 2: **Left:** Metrical distance per line (x-axis) and CometKiwi score (y-axis) for translations of Shakespeare’s sonnets into German by GPT-4o under different prompt strategies. **Right:** Rhyme scheme similarity (x-axis) and CometKiwi score (y-axis) for the same conditions. The markers indicate the average values, the error bars indicate 95% bootstrapped confidence intervals.

Condition	metrical distance	rhyme scheme sim. F_1
IterativeForm	6.71×10^{-10}	0.00057
IterativeMixed	0.0058	0.000056
IterForm2Steps	6.93×10^{-8}	0.0000045

Table 2: p -values for the hypothesis that additional iterative steps lead to improvement on form-related metrics for GPT-4o (compared to PlainForm).

PlainForm clearly led to a better reproduction of the form, but PlainMeaning did not do so for the meaning.

The results for the analysis prompting strategies are less clear. Overall, AnalysisTranslate and PlainForm lead to similar performance on the form-related metrics. AnalysisTranslate however is associated with a lower average CometKiwi score. Furthermore, the generated outputs for other conditions without a separate analysis step often already started with a brief analysis (e.g., by mentioning that the poem is a sonnet), which means that adding a separate step often did not provide much further information.

AnalysisRewrite reproduces the form better than PlainForm (but not better than some of the iterative conditions), but has a lower CometKiwi score than any other prompt condition. The differences to

PlainForm are significant under a 5% significance level, with $p = 0.041$ for metrical distance, $p = 6.98 \times 10^{-7}$ for rhyme scheme similarity, and $p = 1.50 \times 10^{-31}$ for CometKiwi.

4.3 Qualitative observations

The manual inspection of a few of the translated poems by the authors revealed some further particularities that do not fully manifest in the numerical metrics. Most importantly, those prompting strategies that lead to a closer adherence to the form of the original often also result in partially ungrammatical translations as well as divergent or even incomprehensible content (see Appendix D for an example). LLMs of the size of those used here do not usually make a lot of grammatical mistakes in text generation (neither in English nor German), nor were such mistakes produced at a high frequency in those translations that do not reproduce the form.

It is thus likely that the requirements of the form sometimes force the LLM to produce ungrammatical phrases. This would be an instance of a tradeoff relation not only between the reproduction of form and meaning, but also between poetic form and grammaticality. Furthermore, since the internal representations of an LLM are abstractions of patterns found in the training data, poetic licence (i.e., the fact that poets occasionally break grammatical

conventions in specific contexts or for a specific effect) in the training data may have contributed to the LLM learning patterns that allow deviations from regular grammar in poetry generally, even in cases or at a frequency that would be uncommon in human-written poetry. Finally, some of the grammar mistakes in the German translations can also be explained as interference from the English original.

Divergent or incomprehensible content cannot always be recognized from the CometKiwi score alone. Even in cases where it is still relatively high there may be completely incomprehensible or misleading lines. The reason might be that calculating one score for the whole poem obscures strong mismatches in the meaning of individual parts.

The qualitative analysis further revealed that an additional improvement step aiming at a closer reproduction of the form (as in IterativeForm, IterativeMixed, and IterForm2Steps) does usually not lead to whole parts of the previous attempt being copied. More often, a completely new translation is produced, which tends towards less literal word choices and a freer restructuring of the content (see [Appendix D](#) for an example).

Finally, LLM-produced translations sometimes follow a mostly consistent form that is not the one of the original (a strategy that is occasionally but rarely used by human translators as well). When this occurs, the alternative form usually employs metres and rhyme schemes common in German poetry (such as iambic hexameter and rhyming couplets). A notable exception are the reasoning model OpenAI o4-mini and the unified model GPT-5 (which depending on the prompt also uses reasoning) under the PlainForm strategy, which both produce a high number of identical rhymes (i.e., rhyme pairs where not only the phonemes from the nucleus of the last stressed syllable onward, but also that syllable’s onset consonants are identical). Even though identical rhyme is rare in German poetry and considered a stylistic mistake in most contexts ([Peust, 2014](#)), almost half (72) of OpenAI o4-mini’s and about two thirds (102) of GPT-5’s translations feature at least two such rhyme pairs. 28 of OpenAI o4-mini’s and 6 of GPT-5’s translations even contain five or more pairs of identical rhymes (i.e., at least 10 out of 14 lines).

5 Discussion

5.1 Comparison to human translations

Compared to the human translators, the LLMs more often produced translations that match the form of the original only partially. In particular, reproducing the rhyme scheme remains challenging, and when it was reproduced, this often was accompanied by a lower quality in terms of content reproduction, grammatical correctness, and style. On the other hand, the reproduction of the iambic pentameter was much less of a problem when the prompt clearly stated to reproduce the form.

Based on the CometKiwi scores alone, it may seem that on average the LLMs reproduce the meaning better than human translators. However, poetic meaning encompasses aspects other than the literal semantic content, and it is likely that the CometKiwi metric does not reflect all of its nuances equally well. Furthermore, the older and sometimes unconventional language of the human translations might have contributed to their lower CometKiwi scores. For example, Stefan George’s translation (which performs worst on this metric) is characterized by a nonstandard lowercase-only orthography and idiosyncratic usage of punctuation.

5.2 Comparison between LLMs

All five models tested here (GPT-4o, Claude 3.5 Sonnet, Gemini 1.5, OpenAI o4-mini, and GPT-5) showed the basic pattern of a tradeoff between the reproduction of form and meaning, and for all of them, the explicit instruction to reproduce the form, e.g., with the PlainForm prompt, led to a better performance on this criterion, and accordingly a less faithful reproduction of the meaning. While only three prompt formulations were tested for all models, these parallels in their behaviour suggest that different LLMs overall react to changes in the prompt similarly.

While the overall performance of all three standard LLMs (GPT-4o, Claude 3.5 Sonnet, Gemini 1.5), as well as the unified GPT-5, was similar, the reasoning model OpenAI o4-mini performed worse, at least in terms of the metrics used here.

A particularity of translations produced by OpenAI o4-mini and GPT-5 was the frequent occurrence of identical rhyme, which seemingly stems from the reasoning process used by these models. Since identical rhyme is rare in German poetry and more difficult to use than “canonical” rhyme, this can neither be explained by a high frequency in the

training data nor by the need to reconcile the reproduction of the form with other aspects. Given that the correct form would be expected to be produced more easily than the erroneously produced one, this is probably not a problem of the general model capabilities. Rather, it seems that somewhere during the multi-step reasoning procedure, the goal shifts from reproducing the form of the original to producing a specific, entirely different form.

5.3 Effect of prompting strategies

Comparing the results of several prompting strategies revealed that they differ in their effectiveness to align translations to a certain target criterion (i.e., reproduction of the form or the meaning).

If the goal is to prioritize the reproduction of the form, it improves the performance of all tested LLMs a lot if this is clearly stated in the prompt. This only holds for some models when there should be a focus on the meaning, or at least it does not consistently manifest in the resulting CometKiwi score. One explanation is that human translators—from whose work the LLM learns the relevant distinctions during training—usually consciously decide whether they want to reproduce the form or not, and sometimes explicitly state that they do so. In contrast, they do not usually make such a binary decision for the reproduction of the meaning.

To improve the reproduction of the form, it is beneficial to give the translation attempt back to the LLM and to instruct it to produce a better translation. On average, doing so more than one time further improves the reproduction of the form (compare `IterForm2Steps` to `IterativeForm`), but further research would be necessary to systematically inquire how exactly the properties of the input translation affect the output. In contrast, dividing the task into separate steps did not consistently lead to further improvement. These observations align with [Wu et al. \(2025\)](#)'s findings for a general, non-poetic translation task, where iterative refinement led to a better performance than task decomposition too. It may be the case that an explicit analysis would be more relevant for less common poetic forms. On the other hand, however, the quality of the analysis might be lower for rare poetic forms, as language models recognize more frequent forms (such as the sonnet) more reliably ([Walsh et al., 2024](#)).

Not including the original poem in the final translation step (in `AnalysisRewrite`) strongly impairs

the reproduction of the meaning, with the associated CometKiwi score being on average lower than in any other condition. This highlights the risks of splitting the process of poetry translation into two fully disconnected steps (literal translation and style injection). One of the challenges of poetry translation is that while it requires taking into account different aspects, the complex ways in which these aspects interact with each other make it difficult to divide the task into independent steps: Every part of the poem (be it a phrase, a line, or a longer section) influences the form and the meaning at the same time, and both form and meaning result from the interaction of several parts of the poem. Providing the original poem at every step prevents the LLM from moving too far away from the original meaning. Since the reasoning model OpenAI o4-mini presumably also divides the problem into subtasks, its lower performance and peculiar output might similarly stem from the impossibility to separate the task into isolated steps.

In general, LLM-generated translations that match the form of the original rather well often contain incomprehensible lines or grammatical mistakes. This suggests that a similar tradeoff relation as between the reproduction of the form and the meaning also exists between the reproduction of the form and other aspects (such as the adherence to the grammatical norms of the target language).

6 Conclusion

By using numerical and easily interpretable metrics describing the adherence of a translation of a poem to a specific form, this study confirms that the reproduction of the form of an English poem in its German translation remains challenging even for very large generative language models. However, the performance with regard to this aspect is strongly influenced by the chosen prompt strategy. In particular, giving the output back to the LLM together with the instruction to reproduce the form better helps to approximate the metre and rhyme scheme of the original. On the other hand, it is not possible to divide the task into fully independent steps, and the translations are on average better when the original poem is provided at each step of the translation process. These results might also be relevant for other domains where the output has to fulfill several criteria that interact in a complex way at the same time.

The results also confirm that there is a tradeoff

between different criteria. For both human translators and LLMs, reproducing the form better comes at the cost of having to compromise on other aspects (such as meaning, style and grammar). A translation that fulfills all criteria perfectly is impossible, but the influence of the chosen prompt strategy on the output suggests that—given further improvement with regard to the overall translation quality—LLMs can be used to tailor poetic translations to reader preferences more flexibly.

Limitations

While the restriction to a set of well-known poems that have often been translated into German allows a comparison to not only one but several human-written translations, it is likely that both the original poems and German translations are included in the training set of the LLMs, which may skew the results.

Furthermore, comparing the form of an original poem and its translation works best when the basic properties of this form are known to both the source and the target language tradition. In this case, the fact that the sonnet and the accentual-syllabic iambic pentameter are common in both English and German poetry allows a direct comparison, but further considerations about the transferability of the form as well as adjustments of the metrics could be necessary in situations where the form is specific to one of the languages or is defined differently in both of them.

More generally, this study considers only one author (William Shakespeare), one poetic form (the sonnet) and one language pair (English to German), and it is not certain to what extent the results hold for other kinds of poetry. In addition, the language of Shakespeare's sonnets (i.e., Early Modern English) differs from modern English, thus metrics such as CometKiwi may be less reliable here than they are for more recent texts.

Finally, the models tested here are closed-source models, which means that some of their exact specifications and settings are not known.

Acknowledgements

I would like to thank Michael Franke and Angelika Zirker as well as the anonymous reviewers for helpful feedback and advice. This work was supported by the Volkswagen Foundation through a Momentum grant.

References

- Mohammad Awad AlAfnan and Thana Alshakhs. 2025. Bridging linguistic and cultural nuances: A comparative study of human and AI translations of Arabic dialect poetry. *Advances in Artificial Intelligence and Machine Learning*, 5(1):186.
- Hussein Almakary. 2025. Reflections on translating poetry through ChatGPT: Insights from Arabic, English, and French. *Journal of English Studies in Arabia Felix*, 4(1):1–12.
- AI Anthropic. 2024a. The Claude 3 model family: Opus, Sonnet, Haiku. *Claude-3 Model Card*, 1:1.
- AI Anthropic. 2024b. Claude 3.5 Sonnet model card addendum. *Claude-3.5 Model Card*, 3(6).
- Mathieu Bernard and Hadrien Titeux. 2021. Phonemizer: Text to phones transcription for multiple languages in Python. *Journal of Open Source Software*, 6(68):3958.
- Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan. 2021. Don't go far off: An empirical study on neural poetry translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7253–7265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025. Benchmarking LLMs for translating classical Chinese poetry: Evaluating adequacy, fluency, and elegance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33019–33036, Suzhou, China. Association for Computational Linguistics.
- Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190, Sheffield, UK. European Association for Machine Translation (EAMT).
- DeepL SE. 2017. DeepL translate.
- Ruiyao Gao, Yumeng Lin, Nan Zhao, and Zhenguang G Cai. 2024. Machine translation of Chinese classical poetry: a comparison among ChatGPT, Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications*, 11(1):1–10.
- Gemini Team, Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Dmitriy Genzel, Jakob Uszkoreit, and Franz Och. 2010. “poetic” statistical machine translation: Rhyme and meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 158–166, Cambridge, MA. Association for Computational Linguistics.

- Stefan George. 2004. *Shakespeare. Sonette*. Zeno.org. (Original work published in 1909).
- Marjan Ghazvininejad, Yejin Choi, and Kevin Knight. 2018. **Neural poetry translation**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 67–71, New Orleans, Louisiana. Association for Computational Linguistics.
- Reginald Gibbons. 1985. Poetic form and the translator. *Critical inquiry*, 11(4):654–671.
- Christina Heldner. 2013. On rhythm in poetry translation: A new Swedish translation of Shakespeare’s sonnets.
- Gesa Horstmann. 2002. *Shakespeares Sonette in Deutschland. Zur Geschichte der Übersetzungen zwischen dem 18. Jahrhundert und den Übertragungen von Stefan George und Karl Kraus*. Doctoral dissertation, Technische Universität Berlin. Available at https://webdoc.sub.gwdg.de/ebook/diss/2003/tu-berlin/diss/1995/horstmann_gesa.pdf.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.
- Triet Minh Huynh and Quan Le Bao. 2024. Vietnamese poem generation & the prospect of cross-language poem-to-poem translation. *arXiv preprint arXiv:2401.01078*.
- Christa Jansohn, editor. 1992. *Shakespeares Sonette in der Übersetzung Dorothea Tiecks, kritisch herausgegeben von Christa Jansohn*. A. Francke Verlag Tübingen.
- Ruth Jones and Ann Irvine. 2013. The (un) faithful machine translator. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–101.
- Viacheslav Karaban and Anna Karaban. 2024. AI-translated poetry: Ivan Franko’s poems in GPT-3.5-driven machine and human-produced translations. In *Forum for Linguistic Studies*, volume 6.
- Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.
- Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. **Generated knowledge prompting for commonsense reasoning**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.
- Lois Mauri Anne Liwanag, Gabriel Luis Liwanag, and Leslie Anne Liwanag. 2024. AI in anthem: A comparative analysis of the English and Filipino ChatGPT 4 translations from the existing translations of the Philippine national anthem. *Recoletos Multidisciplinary Research Journal*, 12(2):91–102.
- Peter Low. 2003. Translating poetic songs: An attempt at a functional account of strategies. *Target. International journal of translation studies*, 15(1):91–110.
- Zhannura Manapbayeva, Gulkhadisha Zaurbekova, Karylgyash Ayazbekova, Aigul Kazezova, and Kunsulu Pirmanova. 2024. AI in literary translation: ChatGPT-4 vs. professional human translation of Abai’s poem ‘Spring’. *Procedia Computer Science*, 251:526–531.
- OpenAI. 2025a. **GPT-5 system card**.
- OpenAI. 2025b. **Introducing OpenAI o3 and o4-mini**.
- Carsten Peust. 2014. Towards a typology of poetic rhyme. In Eitan Grossman, Martin Haspelmath, and Tonio Sebastian Richter, editors, *Egyptian-Coptic linguistics in typological perspective*, volume 55, pages 341–385. Walter de Gruyter GmbH & Co KG.
- Gottlob Regis. 2004. *Shakespeare, William. Poetische Werke: Sonette*. Zeno.org. (Original work published in 1836).
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. **CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task**. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Natália Resende and James Hadley. 2024. **The translator’s canvas: Using LLMs to enhance poetry translation**. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 178–189, Chicago, USA. Association for Machine Translation in the Americas.
- Friedrich Schleiermacher and Susan Bernofsky. 2021. On the different methods of translating. In *The translation studies reader*, pages 51–71. Routledge.
- William Shakespeare. 2019. *Shakespeare’s Sonnets*. Wikisource. (Original work published in 1923).
- Wai Lei Song, Haoyun Xu, Derek F. Wong, Runzhe Zhan, Lidia S. Chao, and Shanshan Wang. 2023. **Towards zero-shot multilingual poetry translation**. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 324–335, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Lawrence Venuti. 2017. *The translator's invisibility: A history of translation*. Routledge.

Ludwig Reinhold Walesrode. 2007. *William Shakespeare's sämtliche Gedichte*. Wikisource. (Original work published in 1840).

Melanie Walsh, Anna Preus, and Maria Antoniak. 2024. *Sonnet or not, bot? Poetry evaluation for large models and datasets*. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15568–15603, Miami, Florida, USA. Association for Computational Linguistics.

Shanshan Wang, Derek Wong, Jingming Yao, and Lidia Chao. 2024. *What is the best way for ChatGPT to translate poetry?* In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14025–14043, Bangkok, Thailand. Association for Computational Linguistics.

Max J Wolff, editor. 1924. *William Shakespeare: Sonette*. Projekt Gutenberg-DE. (Original work published in 1925).

Di Wu, Seth Aycocock, and Christof Monz. 2025. Please translate again: Two simple experiments on whether human-like reasoning helps translation. *arXiv preprint arXiv:2506.04521*.

A Prompts

Plain:

Please translate the following poem to German:

<original poem>

PlainForm:

Please translate the following poem to German. Please make sure that your translation reproduces the form of the original (rhyme and meter):

<original poem>

PlainMeaning:

Please translate the following poem to German. Please make sure that your translation reproduces the meaning of the original as closely as possible:

<original poem>

IterativeForm:

You are provided with an English poem and an attempt at a German translation. Please suggest a translation to German that reproduces the form of the original (rhyme and meter) better.

Original:

<original poem>

Attempt at translation:

<translation from Condition B>

IterativeMeaning:

You are provided with an English poem and an attempt at a German translation. Please suggest a translation to German that reproduces the meaning of the original better.

Original:

<original poem>

Attempt at translation:

<translation from Condition C>

IterativeMixed:

You are provided with an English poem and an attempt at a German translation. Please suggest a translation to German that reproduces the form of the original (rhyme and meter) better.

Original:

<original poem>

Attempt at translation:

<translation from Condition C>

IterForm2Steps:

You are provided with an English poem and an attempt at a German translation. Please suggest a translation to German that reproduces the form of the original (rhyme and meter) better.

Original:

<original poem>

Attempt at translation:

<translation from Condition B>

AnalysisTranslate, first step:

Please tell me the meter and rhyme of the following poem (in the format: "Meter: [meter]; Rhyme scheme: [rhyme scheme]").

Poem:

<original poem>

AnalysisTranslate, second step:

Please translate the poem below to German. Please make sure to reproduce the meter and rhyme scheme of the original, making use of the given additional information.

Information on meter and rhyme scheme:

<output of the first step>

Poem:

<original poem>

AnalysisRewrite, first step:

Please tell me the meter and rhyme of the following poem (in the format: "Meter: [meter]; Rhyme scheme: [rhyme scheme]").

Poem:

<original poem>

AnalysisRewrite, second step:

Please translate the following to German:

<output of the first step>

AnalysisRewrite, third step:

Bitte schreiben Sie den folgenden Text in ein Gedicht mit den folgenden Eigenschaften um:

<output of the second step>

Text:

<translation from Condition C>

B Pseudocode for the metrical distance algorithm

```

metr_dist = 0;
while no_of_lines(transl) > no_of_lines(ref)
do
  metr_dist = metr_dist +
  min_x(sum_of_stress_values(transl_x));
  from transl remove
  transl_argmin_x(sum_of_stress_values(transl_x));
end
while no_of_lines(ref) > no_of_lines(transl)
do
  metr_dist = metr_dist +
  min_x(sum_of_stress_values(ref_x));
  from ref remove
  ref_argmin_x(sum_of_stress_values(ref_x));
end
metr_dist = metr_dist +
sum_{x in 1,...,no_of_lines(ref)} (ldist(transl_x, ref_x));
metr_dist = metr_dist ÷ no_of_lines(ref)

```

Algorithm 1: Metrical distance algorithm

$transl$ is the sequence of sequences (lines) of stress values extracted from the translation. ref is the sequence of reference line metres. $transl_x$ is the stress pattern of the x -th line of the translation, and ref_x is the reference line metre for the x -th line.

$ldist(transl, ref)$ is a standard edit distance with three operations (insertion, deletion, substitution), where the cost of insertion and deletion is the stress value and the cost of substitution is the difference in stress values.

C Means and standard deviations for all settings

Note: All results are given in the format *mean/standard deviation*.

	Translator, prompting strategy	CometKivi
1	Dorothea Tieck	0.363/0.045
2	Gottlob Regis	0.383/0.049
3	Ludwig Walesrode	0.372/0.048
4	Max Joseph Wolff	0.398/0.042
5	Stefan George	0.320/0.041
6	DeepL	0.614/0.051
7	GPT-4o, Plain	0.595/0.064
8	GPT-4o, PlainForm	0.460/0.052
9	GPT-4o, PlainMeaning	0.568/0.074
10	GPT-4o, IterativeForm	0.450/0.046
11	GPT-4o, IterativeMeaning	0.548/0.071
12	GPT-4o, IterativeMixed	0.444/0.045
13	GPT-4o, IterForm2Steps	0.425/0.044
14	GPT-4o, AnalysisTranslate	0.424/0.050
15	GPT-4o, AnalysisRewrite	0.394/0.048
16	Claude 3.5 Sonnet, Plain	0.484/0.062
17	Claude 3.5 Sonnet, PlainForm	0.469/0.049
18	Claude 3.5 Sonnet, PlainMeaning	0.483/0.067
19	Gemini 1.5, Plain	0.515/0.082
20	Gemini 1.5, PlainForm	0.466/0.064
21	Gemini 1.5, PlainMeaning	0.553/0.083
22	OpenAI o4-mini, Plain	0.503/0.076
23	OpenAI o4-mini, PlainForm	0.425/0.048
24	OpenAI o4-mini, PlainMeaning	0.544/0.072
25	GPT-5, Plain	0.568/0.070
26	GPT-5, PlainForm	0.439/0.046
27	GPT-5, PlainMeaning	0.606/0.052

	Metrical distance	Rhyme scheme similarity F_1
1	0.74/0.077	0.96/0.062
2	0.91/0.176	0.89/0.108
3	0.75/0.083	0.93/0.087
4	0.75/0.096	0.94/0.094
5	0.82/0.091	0.93/0.113
6	2.86/0.473	0.01/0.050
7	2.45/0.696	0.10/0.211
8	0.93/0.235	0.65/0.215
9	2.13/0.736	0.23/0.294
10	0.81/0.134	0.72/0.171
11	1.89/0.712	0.29/0.325
12	0.85/0.258	0.75/0.195
13	0.74/0.119	0.79/0.158
14	0.99/0.403	0.69/0.216
15	0.87/0.365	0.76/0.192
16	1.19/0.529	0.57/0.356
17	0.97/0.303	0.76/0.206
18	1.20/0.489	0.57/0.377
19	1.90/0.772	0.32/0.312
20	1.24/0.540	0.56/0.272
21	2.23/0.795	0.21/0.274
22	2.18/0.842	0.33/0.343
23	1.58/0.898	0.62/0.244
24	2.46/0.787	0.19/0.294
25	2.26/0.755	0.11/0.243
26	1.07/0.290	0.72/0.186
27	2.87/0.639	0.01/0.044

D Example of refinement over several steps

The following translations of Shakespeare’s Sonnet 29 by GPT-4o under the conditions PlainForm, IterativeForm, and IterForm2Steps serve as an example of the changes over several iterations. They are annotated for significant

content differences to the original as well as ungrammatical phrases. **Red** marks ungrammatical parts, **yellow** marks parts of the translations that strongly differ from the content of the original, and **green** marks changes in the order of contents.

William Shakespeare: Sonnet 29 (Shakespeare, 2019)

When, in disgrace with fortune and men's eyes,
I all alone beweep my outcast state,
And trouble deaf heaven with my bootless cries,
And look upon myself, and curse my fate,
Wishing me like to one more rich in hope,
Featur'd like him, like him with friends possess'd
Desiring this man's art, and that man's scope,
With what I most enjoy contented least;
Yet in these thoughts myself almost despising,
Haply I think on thee,—and then my state,
Like to the lark at break of day arising
From sullen earth, sings hymns at heaven's gate;
For thy sweet love remember'd such wealth brings
That then I scorn to change my state with kings.

Translation by GPT-4o, PlainForm

CometKiwi: 0.383, Metrical distance: 0.80, Rhyme scheme similarity F_1 : 0.42

Wenn ich, von Glück und **Menschens** Blick verbannt,

follows English rather than German morphology

Ganz einsam meinem ausgestoß'nen Stand
Nachweine, taubem Himmel Schmerz gesandt,
Verfluche mich und meiner Sterne Hand,
Und wünschte, **hoffnungsvoll wie jener Mann,**

Ansehnlich wie der Freundesreiche dann,

the phrase is missing a verb

Begehrte dies' Talent und jenen Plan,
Und schätzte nicht, was ich doch lieben kann;
Doch mitten in solch schmachvollem Verlangen
Denke ich, plötzlich, glücklich nur an dich.
Dann steigt mein Herz, wie Lerchen aufgesprungen,
Die singen, wenn der Tag das Dunkel bricht;

"from sullen earth" and "at heaven's gate" is not reproduced in this translation

Denn deines süßen **Denkens** reiche Gaben

Lässt mich auf Königsglück von Herzen verzichten.

number agreement mistake

Translation by GPT-4o, IterativeForm

CometKiwi: 0.427, Metrical distance: 0.81, Rhyme scheme similarity F_1 : 0.36

Wenn ich, verstoßen von des Schicksals Hand,
Ganz einsam **weine** meinen schalen Stand,

intransitive verb used with a direct object

Und **tauben Himmel** ruf mit leerem Schrei,

missing article

Mich anseh' selbst und meinem Schicksal groll',

Wünscht' ich, wie jener hoffnungsreich zu sein,
Gesellt wie er, mit Freunden wohl bedacht,
Begehr't' ich dies' Talent, des andern Schein,
Und das, was mir gefällt, schätz't' ich recht flach.
Doch mitten in solch fast verhasstem Trachten
Denk' ich an dich – und plötzlich strahlt mein Los;
Wie **Lerche**, die im Morgengrauen erwacht,

missing article

Die freudig singt, **als sei die Nacht nicht groß.**

"from sullen earth" and "at heaven's gate" is not reproduced in this translation

Denn süße Liebe, **die mir dich beschert,**
Macht, dass ich Königs Glück verschmäh' und wehr'.

Translation by GPT-4o, IterForm2Steps

CometKiwi: 0.438, Metrical distance: 0.66, Rhyme scheme similarity F_1 : 0.92

Wenn ich, verachtet, fern von Glück und Gunst,
Allein beweine' mein ausgestoßnes Los,
Und ruf' zum tauben Himmel meine **Kunst**,
Verfluchend mich und mein Verhängnis bloß;
Wenn ich mich wünsch', wie jener reich an Mut,
Und schön wie der, beliebt und hochverehrt,
Begehr't' dies' Mannes Kunst, des andern Gut,
Und was ich sonst geliebt, erscheint mir leer;

Doch denk' ich dann an dich in meinem Schmerz,

Da schwingt mein Geist sich auf in lichten Klang,

Gleich wie die Lerche, die mit frohem Herz

Vom düstren Acker steigt zum Morgensang.

"at heaven's gate" is not reproduced in this translation

Dein süßer Liebes **gruß** so reich mich macht,

Dass ich kein **Königs** königreich betracht'.

Note especially the third quartet (ll. 9-12) in the IterForm2Steps translation, where the contents of the original are compressed, expanded and reordered more freely than in the previous iterations to produce a more coherent and smoother translation. The ll. 9-12 can be translated back to English as "Yet if I think of you in my pain, / then my spirit arises into bright sound, / like the lark, who with a happy heart / rises from the dark fields for its morning song." Note also the less literal translation of words at the rhyming ends of lines, e.g., "Mut" (literally, "courage") for "hope", and "beliebt und hochverehrt" (literally, "popular and revered") for "with friends possess'd".

WikiLingDiv: a dataset for quantifying digital linguistic diversity using Wikipedia page views

Hannes Essfors

TU Wien Informatics,
Favoritenstraße 9-11,
1040 Vienna, Austria
hannes.essfors@tuwien.ac.at

Andreas Baumann

Faculty of Philological and Cultural Studies,
University of Vienna, Universitätsring 1,
1010 Vienna, Austria
andreas.baumann@univie.ac.at

Abstract

With the conflation of digital and non-digital spaces, and NLP technologies being integrated into an increasing number of aspects of daily life, linguistic diversity cannot be fully understood without considering language use in the digital space. To facilitate such research, we introduce WikiLingDiv, an openly accessible dataset for quantifying linguistic diversity in online knowledge retrieval using Wikipedia page views, corresponding to one dimension of digital linguistic diversity. Our dataset is based on yearly page views of 340 language editions of Wikipedia, aggregated across 239 countries and territories over 10 years (2015-2024). Using the dataset, we illustrate spatial and temporal patterns of digital linguistic diversity, suggesting that diversity has both increased and decreased across countries and regions, while highlighting country-specific dynamics in language usage. We release the dataset as an openly available and easily integrable data resource for researchers in computational linguistics, digital humanities, and the broader social sciences, enabling further work on linguistic variation and digital inequality.

1 Introduction

In the postdigital world, the lines between digital and non-digital have become blurred, with communication through digital media and tools moving from representing non-standard use of language to a non-optional part of daily interaction in large parts of the world. As a matter of fact, in December 2025, 73.2% of the global population were internet users and 68.7% were social media users (Statista, 2025a), while during Covid-19, digital working became the de facto standard for much of the knowledge-based part of the global economy (Marsh et al., 2022).

This raises the question whether traditional models of linguistic diversity as a function of the relative abundance of L1 speakers (Greenberg, 1956;

Ginsburgh and Weber, 2020; Grin and Fürst, 2022; Civico, 2025) remain sufficient—notably, the bias towards English in the digital space would suggest a large discrepancy between native language and the actual linguistic landscape (Benson et al., 2025). Instead, an operationalization of linguistic diversity based on language usage might be considered a more direct model of how language facilitates the tasks of, e.g., retrieving and producing information in the process of between-human interaction. While modelling language usage in the non-digital space intuitively appears difficult, it presents itself as fairly straightforward in the digital space due to the extensive datafication it has been subjected to (Mejias and Coudry, 2019; Flyverbom, 2019). Notably, such approaches towards linguistic diversity have been taken by, e.g., Magdy et al. (2014); Alshaabi et al. (2021); Hiippala et al. (2019); Väisänen et al. (2022) who assess digital linguistic diversity based on language production on social media, specifically Twitter/X and Instagram. These approaches come with three main caveats: Firstly, social media data is highly sensitive and thus restricted in its usage. For instance, since 2023, the Twitter API is no longer freely available for research (Blakey, 2024). Secondly, using social media data and similar artifacts to model digital linguistic diversity is crucially reliant on high accuracy in the NLP-task of language identification, which is notably challenging for multilingual documents and short texts (Jauhainen et al., 2019), both of which are common features of social media posts (Alshaabi et al., 2021). Thirdly, it has been shown that only a small fraction of tweets are geotagged (Pfeffer et al., 2023), severely limiting the possibility of analyzing linguistic diversity as a spatial phenomenon.

To address some of these shortcomings and facilitate research on digital linguistic diversity as an approach towards modelling human culture and society, we here introduce *WikiLingDiv*—a dataset

for quantifying digital linguistic diversity using Wikipedia page views. Our approach utilizes the fact that Wikipedia is a multilingual, widely used reference work, freely available in most of the world, and as of August 2025, the 7th most visited website globally (Statista, 2025b). Furthermore, the Wikimedia Foundation, which hosts Wikipedia, is committed to open data principles, making records and web activity associated with Wikipedia openly available through the Wikimedia Analytics API. This allows us to collect the number of page views associated with any given language edition of Wikipedia for 239 countries and territories over 10 years, which are made openly accessible through a CC BY-SA 4.0 license. Using this as a basis, we then normalize the page views to yield the proportion of attention given to each language across countries and years, based on which formal measures of diversity are derived.

Explicitly, our approach aims to shed light on digital linguistic diversity as a dynamic phenomenon arising when humans interact with—and through—digital technology. It follows that our work builds on two key assumptions: 1) Wikipedia page views approximate interest in retrieving knowledge in a given language, and 2), Patterns of Wikipedia access are impacted by socioeconomic factors such as internet access and level of education. Therefore, WikiLingDiv does not model digital linguistic diversity as a product of textual output, but rather as a proxy for language consumption. Thus, it complements work focusing on diversity in existing language technology infrastructure, such as the *Digital Language Equality Metric* (DLE) (Gaspari et al., 2022; Grütznert-Zahn and Rehm, 2022) of the European Language Equality (ELE) project, which assesses the availability of technology and support for different languages. Inherently, these two dimensions of digital linguistic diversity are linked: the language used depends on the available technological support for that language, and the available support partially reflects the linguistic demand. To understand the extent of this interaction, data on digital linguistic usage and the relative demand of digital technology is necessary, a gap which WikiLingDiv seeks to help to address.

The paper is structured as follows. First, we describe the process of compiling and structuring the dataset, as well as the measures taken in harmonizing the data to allow for integration with other datasets. Then, we provide a descriptive analysis

of the dataset. Finally, we showcase how WikiLingDiv can be utilized in a diversity analysis to address the research question of how digital linguistic diversity—from the perspective of language consumption in online knowledge retrieval—has changed over the last decade.

2 Creating the dataset

For the creation of the dataset and all subsequent analysis, the statistical programming language R (4.5.2) (R Core Team, 2025) was used.¹ As alluded to in the introduction, our empirical strategy utilizes the Wikimedia Analytics REST API to understand how users interact with the language-specific Wikipedia editions.² More specifically, we use the `httr` package (Wickham, 2023) to acquire a list of all available Wikimedia projects through the following API-request: <https://commons.wikimedia.org/w/api.php?action=\protect\@normalcr\relaxsitematrix&smttype=language&format=json>. We loop through the list and collect the 2 or 3-letter language codes used by Wikimedia to denote each project, e.g., `/en/` for English or `/ceb/` for Cebuano, together with its name in Latin script. For each language code, we then query the REST API to acquire the page views per country. The endpoint takes requests according to the following structure: `/pageviews/top-by-country/project/access/year/month`. We specify *project* using the respective language codes; we specify *access* to 'all-access' to capture both desktop and mobile users; and we specify the time frame as each year and month combination beginning with 2015, since the endpoint serves data from this year and onwards.

From each request to the API endpoint, we receive a value for the number of page views from each country for which at least one pageview exists in a given year and month. However, to protect user privacy, the Wikimedia Foundation does not publish the exact number of page views but instead ceils to the nearest 1000 (Wikimedia Foundation, 2024), meaning that the values are estimates. For example, if two language editions receive 1012 and 1756 views, respectively, they are both assigned

¹All code associated with the paper is made available at <https://github.com/Eszettfors/WikiLingDiv>. The dataset is published at <https://zenodo.org/records/18526766>

²For the exact API documentation provided by the Wikimedia Foundation, see <https://doc.wikimedia.org/generated-data-platform/aqs/analytics-api/>

the value 2000. It naturally follows that the more views an edition has, the smaller the rounding error proportionally is. We then aggregate the values for each year from 2015 to 2024 by summing the page views for each language edition and country across all 12 months. Finally, we restructure the dataset as a time series in long format such that each record consists of a uniquely defined year-country-language triplet specified by the number of page views.

Unfortunately, the page views by country endpoint does not allow specification of the nature of the agent, and accordingly, we cannot purposefully exclude self-identified bots from the statistics. While this naturally introduces some bias to the data, systematic scraping constitutes a fairly stable bias within countries—especially given the low granularity scope the data is aggregated at—meaning that the relative abundances we are seeking to compute is expected to remain robust. Furthermore, we would argue that a substantial fraction of automated information retrieval is indeed driven by human information needs. With this in mind, we suggest the aggregated page views per language and country be interpreted as a proxy for the relative interest in a particular language in that country.

2.1 Data harmonization

To harmonize and make the dataset interoperable, we address the language codes used by Wikipedia, which inconsistently employ both three- and two-letter language codes as well as multiple non-standard identifiers (Wikimedia Foundation, Inc., 2025). There are two major language catalogues that we intend to make the dataset easily mergeable with: Glottolog (Hammarström et al., 2025) for linguistic typology and Ethnologue (Eberhard et al., 2025) for speaker numbers. For this purpose we add both ISO 639-3 codes (ISO, 2023) and glottocodes. Since ISO-codes can denote both macrolanguages and individual languages, such as /ara/ being the macrolanguage identifier encompassing all Arabic varieties—multiple of which have their own Wikipedia language edition—we avoid using macrolanguage identifiers, unless they have a corresponding Glottocode. This is, for example, the case with Serbo-Croatian (hbs), which is recognized as an individual language in Glottolog with the respective standardized varieties Bosnian (bos), Serbian (srp) and Croatian (hrv), which are classified as dialects in Glottolog and thus have a

corresponding Glottocode. As a result, both Glottocodes and ISO639-3 function as unique language identifiers in the dataset, allowing for easy integration with, e.g., knowledge-bases used in multilingual NLP such as URIEL+ (Khan et al., 2025).

3 Descriptive analysis

All in all, the dataset comprises 295,510 year-language-country triplets covering a total of 340 languages and 239 countries/territories across 10 years (2015 - 2024). As seen in Figure 1a), the distribution of page views across countries remains fairly stable throughout the time period. While a noticeable increase in median and mean page views can be observed from 2015 (mean = $10^{7.38}$; median = $10^{7.37}$ to 2016 (mean = $10^{7.66}$; median = $10^{7.57}$), only small fluctuations are distinguished until 2022, after which it decreases. Furthermore, the geographical spread of the page views is by no means homogeneous: the boxplots are consistently layered according to continents across the time span, with countries in Oceania and Africa generally making up the first and second quartiles; North America the second and third; and Europe, Asia and South America the third and fourth. The number of countries also remains stable around 235, with a sharp decline to only 200 countries in 2024. This can reasonably be assumed to be a result of the Wikimedia Foundation’s data publication guidelines, as the missing countries for 2024 align with the list of countries and territories designated as protected due to the higher risk posed to individuals from their Wikipedia activity (Wikimedia Foundation, 2025). Because of this, the dataset only contains complete time series for 196 countries and territories, which, however, increases to 232 if the year 2024 is excluded.

Country	Page Views (%)	Cumulative
U.S.	22.50	22.50
Japan	7.06	29.56
Germany	6.35	35.91
U.K.	5.58	41.49
India	4.82	46.31
France	4.03	50.34
Italy	3.62	53.95
Russia	3.44	57.39
Canada	2.76	60.16
Spain	1.99	62.15

Table 1: Top 10 countries as a source of Wikipedia page views across the years 2015 - 2024.

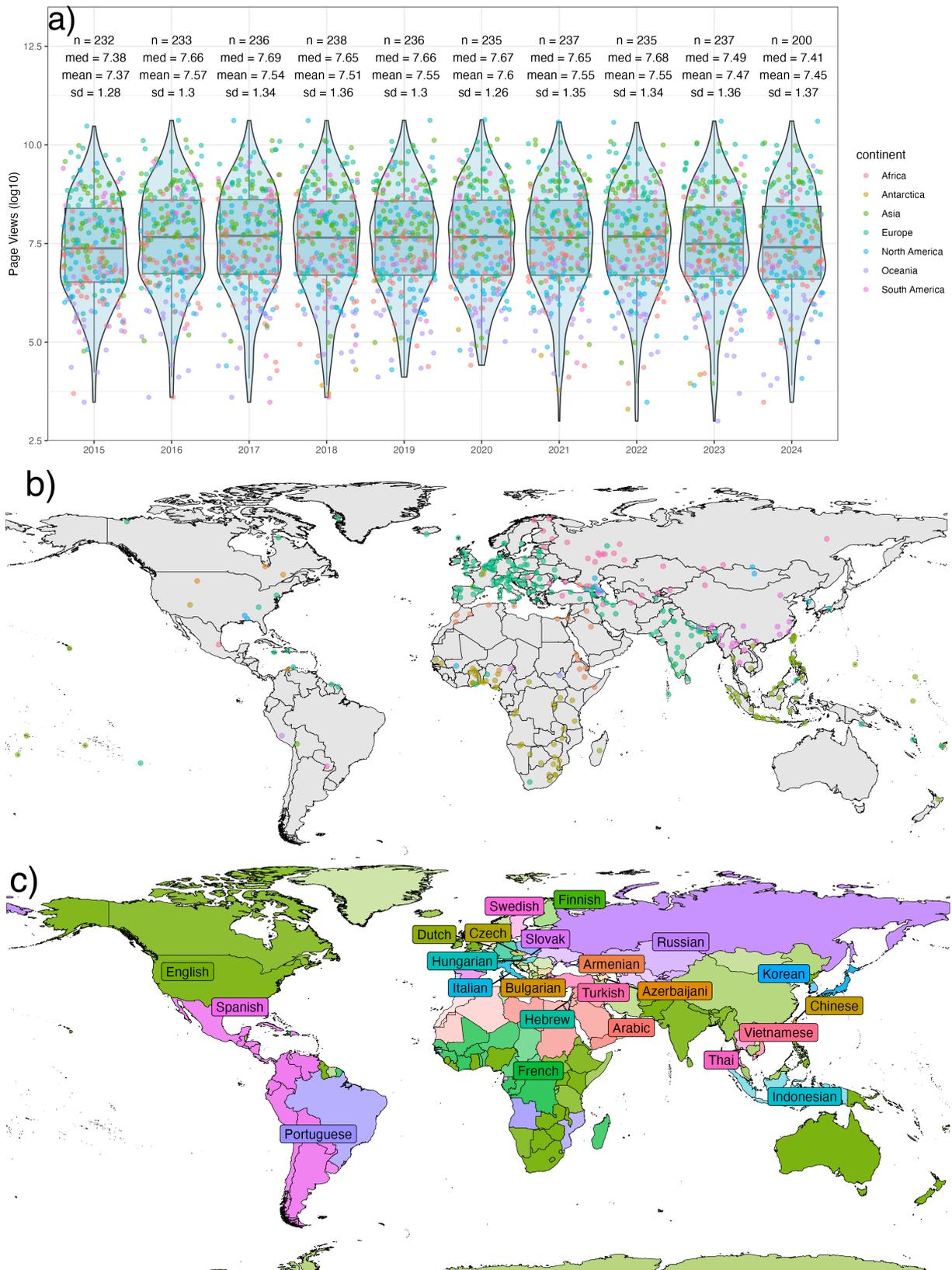


Figure 1: a) Boxplots overlaid by violinplots depicting the distribution of yearly total page views per country from 2015 to 2024, scaled using a base-10 logarithm. Each point corresponds to a country or territory, which is coloured according to its geographical continent of belonging. b) Geographical distribution of languages present in the dataset, coloured according to language family. The coordinates of each language were taken from Glottolog. c) World map with each country coloured according to the most visited language edition of Wikipedia between 2015 and 2024. Not all languages are labeled due to overlap. Colour intensity corresponds to the dominance of the language.

Naturally, Wikipedia is not used to the same extent across countries due to differences in, e.g., population size, digital access, information needs, and government restrictiveness on information access. This is evident from the fact that 22.5% of total page views across all years stems from the United States, with 62.15 % of all page views being sourced from 10—predominantly highly developed—countries as seen in Table 1. This difference is then partially reflected in which language editions receive attention, with 88.61 % of attention directed towards the 10 language editions with the most page views, as seen in Table 2, largely corresponding to the national language of the countries in Table 1. Although the dataset comprises 340 languages in total, almost 50 % of page views are concentrated into a single language, English, and looking at a family level, of the 32 language families present, 87.74 % of attention is directed towards the 137 Indo-European languages in the dataset, of which English is one. Indeed, as can be seen in Figure 1b), most languages are found across Eurasia, with a particularly high concentration of languages in Europe. However, quite a few languages from Central Asia, especially the Caucasus, are present. Also in Africa, notably the Atlantic-Congo (42 languages) and Afro-Asiatic (16 languages) language families are present. Languages from the Americas and Oceania, however, have a noticeably low presence in the data.

Language	Page Views (%)	Cumulative
English	49.77	49.77
Japanese	6.92	56.69
Spanish	6.55	63.24
German	5.96	69.21
Russian	5.07	74.28
French	4.66	78.94
Italian	3.38	82.32
Chinese	2.67	84.99
Portuguese	2.02	87.00
Polish	1.60	88.61

Table 2: Top 10 visited language editions of Wikipedia over 2015 and 2024.

On the other hand, looking at the country level and the dominant language of each country as seen in Figure 1c), only 24 languages dominate the 239 countries and territories in the data. While many instances of the national language being the most dominant language are found in Europe, such as for

Bulgaria, Sweden and Finland, a pattern of imperialistic legacy can be observed. For example, West Africa is dominated by French, while East Africa is dominated by English. Portuguese is dominant in the former colonies of Angola, Mozambique and Brazil, while the Russian Wikipedia is most popular in former Soviet states such as Ukraine and Kazakhstan. However, Russian is clearly more dominant in Russia (88% of page views) compared to Ukraine (57%) and Kazakhstan (66%), signaling that there is competition between Russian on the one hand, and the local national languages of Ukrainian (30%) and Kazakh (25%) and to a certain extent English (9.8% & 6.5%) on the other hand. At the same time, in other former Soviet states, such as Armenia (40% Armenian, 36 % Russian, 20% English) and Azerbaijan (42% Azerbaijani, 25% Russian, and 18% English), the local national languages are dominant, with the Russian Wikipedia being the second most used language. In the Baltic states, English dominates instead, followed by the local national language and Russian (Estonia: English 43%, Estonian 35%, Russian 16%; Latvia: English 37%, Russian 31%, Latvian 24%; Lithuania: English 43%, Lithuanian 41%, Russian 11%). Reasonably, these patterns occur as a result of different pressures acting on internet users, such as knowledge and familiarity of the languages in the population, the extent and usefulness of the Wikipedias (Kornai, 2013), and the political status and association with the languages, to name a few; all factors that suggest themselves to be accounted for in follow-up research. We are thus convinced that our dataset has great potential as a resource to model language competition and further our understanding of how technology interacts with—and shapes—both human society and cultural heritage.

4 Analyzing change in linguistic diversity

In this final part of the paper, we showcase how our dataset can be employed to answer research questions pertaining to linguistic diversity. Specifically, we aim to understand if digital retrieval of information has become more or less linguistically diverse over the last decade. Since the country coverage is substantially lower for 2024, we restrict the analysis to the years from 2015 to 2023, giving a time series of nine consecutive years. To increase the reliability of our estimates, we filter away countries that in any given year had fewer than 1,000,000

page views in total, which we consider a reasonable threshold. We then subset to countries and territories present across all nine years, yielding a complete time series with 199 countries. As a strategy to protect against sudden spikes in language richness due to scraping activities, we fixate the languages in each country to those present in 2015. Therefore, any changes in diversity will reflect a redistribution of attention among already established languages, rather than the introduction of new languages in the respective countries. While this strategy runs the risk of missing changes related to large-scale migration that explicitly took place after 2015, the actual impact on the measured diversity is likely limited within the relatively short timespan under study.

For each year, we calculate a relative abundance vector for each country by dividing the number of page views of each language by the total number of page views in that country. As such, we end up with 232×9 vectors, $p = [p_1, p_2, \dots, p_n]$ where p_n is the proportion of page view visits to the n :th Wikipedia, in a given country and year, such that

$$\sum_{i=1}^R p_i = 1, \quad (1)$$

where R is the length of the relative abundance vector.

We then define linguistic diversity according to the Leinster Cobbold Framework (Leinster and Cobbold, 2012)—suggested by Essfors (2025)—as the effective number of completely dissimilar and equally abundant languages. While we acknowledge that diversity is multivariate in its nature, we only employ one measure of diversity to maintain the example scope of the analysis. Specifically, we use a naive model of diversity, treating all languages as completely dissimilar and distinct from one another. To reduce the influence of rare languages that might have been introduced through automatic scraping and not necessarily purposeful information retrieval, we specify the sensitivity parameter q to 2, such that rare languages are given a lower weighting. In this set-up, our working definition of diversity D_{tc} in the country c at time t formally becomes

$$D_{tc} = \frac{1}{\sum_{i=1}^R p_i^q}, q = 2. \quad (2)$$

First, we consider the absolute change in diversity from 2015 to 2023 by calculating the log diversity

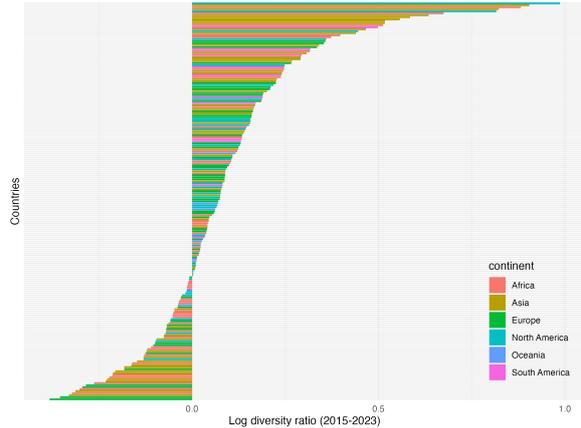


Figure 2: A barchart displaying the change in diversity across countries between 2015 and 2023. Each bar represents a country, with the length of the bar being proportional to the change in diversity. Bars directed to the left indicate a decrease in diversity, while bars directed to the right indicate an increase.

ratio between the years 2015 and 2023 as an effect size. Formally, we define the log diversity ratio ldr for a given country c as

$$ldr(c) = \log_2\left(\frac{D_{2023c}}{D_{2015c}}\right) \quad (3)$$

where D_{2023c} and D_{2015c} are the measured diversity of c in 2023 and 2015 respectively. It follows that a positive log-ratio indicates an increase in diversity, and a negative log-ratio a decrease. Furthermore, the measure is symmetric around 0 (i.e., no change), such that a doubling of diversity = 1, while a halving of diversity = -1.

By plotting the ldr of each country as seen in Figure 2, it is evident that more countries have seen an increase in diversity than a decrease. As a matter of fact, we note an increase in 137 countries, more than twice the number of countries observing a decrease (62), yielding a mean ldr of 0.095, 95% CI = [0.063, 0.13], suggesting that on average, linguistic diversity has increased by approximately 4.5%–9.4% from 2015 to 2023 due to a redistribution of attention away from the most dominant languages within each country. One of the most important insights from this result is that while indeed diversity is suggested to generally be increasing, the fact that the change is heterogeneous—with diversity decreasing for a substantial portion of countries—indicates that the underlying mechanisms are country-specific, and potentially culturally dependent. Considering the spatial distribution of ldr as seen in Figure 4, there

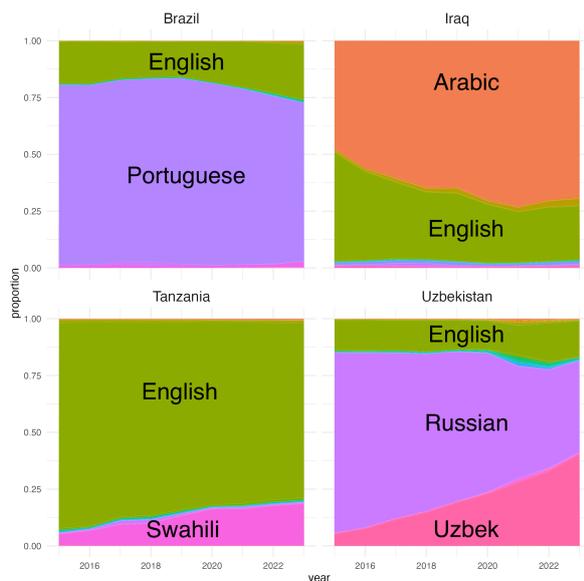


Figure 3: A streamgraph showing the proportional trajectories of dominant languages in Brazil, Iraq, Tanzania und Uzbekistan from 2015 to 2023.

is clearly some spatial dependence, with a substantial increase in diversity across most of Latin America, Central and South Asia, while the countries of the Middle East have seen a reduction. By investigating the relative abundance trajectories of the dominant languages within the respective countries, we can discern different patterns of interplay between the competing languages.

For example, across Latin America, exemplified by Brazil in Figure 3, we see a clear dominance of the national language—in the Brazil case, Portuguese—which has been declining in recent years, with increasing attention given to English, resulting in a diversification. This highlights an important aspect of diversity: what is perceived as diversification on the country level can be the result of a process yielding a concentration of languages on a global level. English is by far the most widely used language digitally, and as such an increase in its usage represents a decline in global diversity, while on a local level—as in Brazil—it is perceived as a diversification.

In Iraq and across the Middle East, on the other hand, the reverse pattern is observed: the English and standard Arabic wikis were on equal footing in 2015, after which increasing attention has been given to the Arabic one, making it dominant, thus resulting in a measured decline in diversity locally. This would, however, contribute to a global increase in diversity. In Tanzania, a similar pattern

is observed, with English in steady decline and attention being directed towards the Wikipedia in the local language of Swahili. And since English was heavily dominating in 2015, this resulted in an increase in measured diversity. In Uzbekistan, a third pattern is observed, with English remaining fairly constant across time, while the dominant Russian language loses ground to the national language of Uzbek, which only constituted 5.1% of page views in 2015, gradually increasing its share to 40.0% in 2023, just barely behind Russian with 40.1%.

While we can use WikiLingDiv to describe these patterns in language attention, the data in itself cannot explain why they occur. For this reason, other dimensions of digital linguistic diversity are equally important to consider, e.g., how the digital support for Arabic, Swahili and Uzbek has changed relative to English and Russian. The addition of such data could yield insights into how extending the technological resources of a language impacts its usage. For this purpose, intricate measures of language support, such as that of the DLE (Gaspari et al., 2022; Grütznert-Zahn and Rehm, 2022) for European languages, or Digital Language Support (DLS) of (Simons et al., 2022), building on (Kornai, 2013), for global estimates. To extend the analysis presented here, the most straightforward way would be to construct a metric quantifying the extent of the respective Wikipedia language edition, using, e.g., the number of Wikipedia articles and their sizes. However, equally important would be to also consider socioeconomic factors such as education level and second language acquisition, e.g., the increased use of the English Wikipedia in Latin America could very well be reflective of an increase in English proficiency. Further analyses could for example include the EF English Proficiency Index as a covariate to account for such heterogeneity.

5 Discussion and Conclusion

In this paper, we have introduced WikiLingDiv, a dataset for quantifying linguistic diversity in the digital space using Wikipedia page views as a proxy for interest in retrieving information in different languages. We have showcased its extensiveness with respect to spatial and temporal coverage, as well as linguistic breadth. In doing so, we illustrated how previously observed linguistic biases in available digital content—to a degree—manifest as patterns of digital knowledge retrieval. However, information retrieval and digital language consumption is

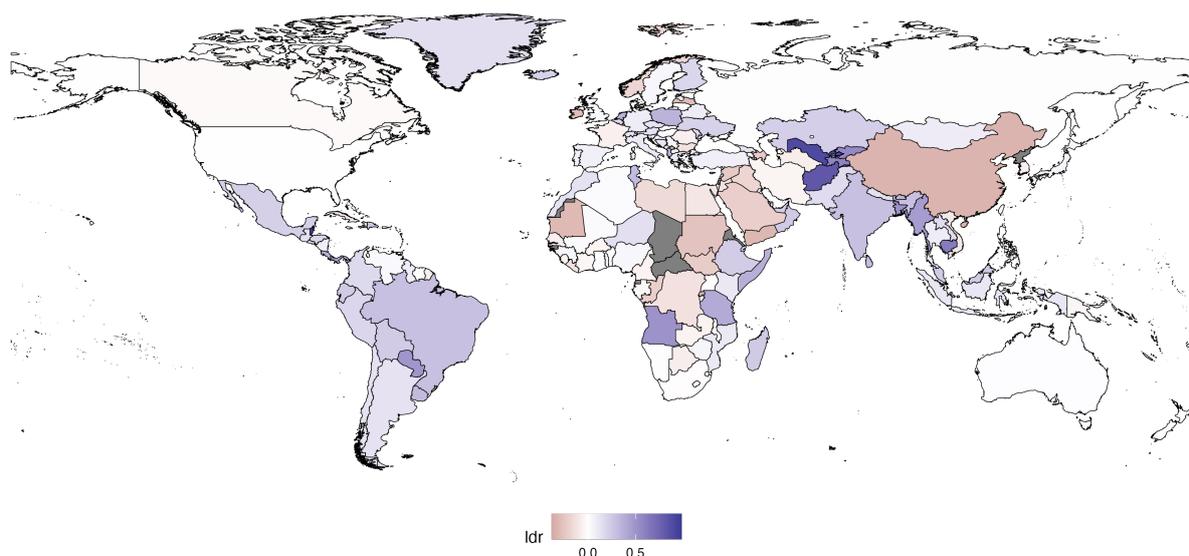


Figure 4: A choropleth map visualizing country-level change in the log diversity ratio (ldr) from 2015 to 2023. A blue color indicates an increase in diversity while a red color indicates a decrease, with more intense colors corresponding to a larger magnitude of change. An ldr of 1 is equivalent to a doubling of diversity. Grey countries indicate missing values.

only one dimension of digital linguistic diversity, and further research should aim at combining the dataset with other operationalizations, such as digital language support metrics.

While our approach is somewhat limited by the inability to differentiate between human and bot-generated page views, we have exemplified strategies to mitigate the impact of automatic scraping while still capturing signals of human linguistic preference. Furthermore, the dataset is potentially limited by the use of translation tools; it could very well be that traffic is directed towards the extensive English wikipedia, but then machine translated into a different language. This potentially yields a bias where the dominance of English is overestimated and the attention to languages with machine translation support is underestimated. In addition, while Wikipedia is one of the most widely used websites globally, its usage can be expected to vary within and across populations; e.g., men use Wikipedia to a larger extent than women (Johnson et al., 2021), there is a notorious rural-urban divide in internet access (International Telecommunication Union, 2024), and in some countries, access to Wikipedia has purposefully been restricted by the government (Zhang et al., 2017; Pan and Roberts, 2020; Yang and Roberts, 2021). Still, Wikipedia is arguably the closest approximation to a global and universally used digital encyclopedia there is.

Through our example analysis, we have demon-

strated how attention towards different language editions of Wikipedia varies overtime and across countries, resulting in cases of measurable increase as well as decrease in linguistic diversity. Most strikingly, however, we observe signs of an overall tendency towards *increasing* linguistic diversity, as per our operationalization, which to a certain extent challenges the widely-spread notion that linguistic diversity is steadily declining (Harmon and Loh, 2010; Bromham, 2023). However, since the increase is an average across countries, the observed diversification could result from English increasing its relative usage compared to already dominating language such as Spanish and Portuguese. Importantly, the longitudinal and spatial resolution of the data enables the identification of regional patterns of attention shifts from dominating global languages, such as English, to local languages—an interaction that ultimately plays a central role in shaping digital linguistic diversity. Explaining why these patterns emerge requires extending the analysis to inferential models, accounting for, e.g., the digital support of the languages, extent of internet access and digitalization across countries, and the linguistic diversity in the non-digital linguistic landscape, which could yield valuable insights into how diversity in digital language usage is shaped by the linguistic diversity of technology.

Ultimately, understanding the interaction between language, technology, and culture is cru-

cial as societies become increasingly digital, and NLP tools are further integrated into everyday life. While this endeavor is inherently complex, WikiLingDiv is a reusable and easy-to-integrate dataset that can support further research in areas of study where NLP, digital humanities, and linguistic diversity meet.

Acknowledgments

This research was funded by WWTF (grant number ICT23-012).

References

2023. [Iso 639:2023 — code for individual languages and language groups](#). Accessed: 2025-12-18.
- Thayer Alshaabi, David Rushing Dewhurst, Joshua R. Minot, Michael V. Arnold, Jane L. Adams, Christopher M. Danforth, and Peter Sheridan Dodds. 2021. [The growing amplification of social media: measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020](#). *EPJ Data Science*, 10(1):1–28.
- Juliane Benson, Katharina Zeh, Hannes Essfors, Hannes Fellner, Julia Neidhardt, and Andreas Baumann. 2025. [Linguistic diversity and digitalization: An ambivalent relationship](#). In *Digital Humanism: First Interdisciplinary Science and Research Conference, DIGHUM 2025, Vienna, Austria, November 20–21, 2025, Proceedings*, page 358–365, Berlin, Heidelberg, Springer-Verlag.
- Elizabeth Blakey. 2024. [The day data transparency died: How twitter/x cut off access for social research](#). *Contexts*, 23(2):30–35.
- Lindell Bromham. 2023. [Language endangerment: Using analytical methods from conservation biology to illuminate loss of linguistic diversity](#). *Cambridge Prisms: Extinction*, 1:e3.
- Marco Civico. 2025. [Measuring linguistic diversity: Limits and extensions of the greenberg index](#). *Journal of Quantitative Linguistics*, 0(0):1–28.
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025. [Ethnologue: Languages of the world](#).
- Hannes Essfors. 2025. [Global linguistic diversity: Adapting the leinster–cobbold framework from ecology for humanities research](#). In *Anthology of Computers and the Humanities*, volume 3, Austin, Texas, U.S.A. Association for Computers and the Humanities.
- Mikkel Flyverbom. 2019. [Digital and Datafied Spaces](#), page 25–38. Cambridge University Press, Cambridge.
- Federico Gaspari, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne, and Andy Way. 2022. [Introducing the digital language equality metric: Technological factors](#). In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, page 1–12, Marseille, France. European Language Resources Association.
- Victor Ginsburgh and Shlomo Weber. 2020. The economics of language. *Journal of Economic Literature*, 58(2):348–404.
- Joseph H. Greenberg. 1956. [The measurement of linguistic diversity](#). *Language*, 32(1):109–115.
- François Grin and Guillaume Fürst. 2022. [Measuring linguistic diversity: A multi-level metric](#). *Social Indicators Research*, 164(2):601–621.
- Annika Grütznher-Zahn and Georg Rehm. 2022. [Introducing the digital language equality metric: Contextual factors](#). In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, page 13–26, Marseille, France. European Language Resources Association.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. [Glottolog 5.2](#). Available online at <http://glottolog.org>. Accessed: 2025-12-18.
- David Harmon and Jonathan Loh. 2010. [The index of linguistic diversity: A new quantitative measure of trends in the status of the world’s languages](#). *Language Documentation & Conservation*, 4:97–151. Retrieved from ScholarSpace, University of Hawaii at Mānoa.
- Tuomo Hiippala, Anna Hausmann, Henriikki Tenkanen, and Tuuli Toivonen. 2019. [Exploring the linguistic landscape of geotagged social media content in urban environments](#). *Digital Scholarship in the Humanities*, 34(2):290–309.
- International Telecommunication Union. 2024. [Internet use in urban and rural areas](#). Facts and Figures 2024. Accessed: 9 February 2026.
- Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. [Automatic language identification in texts: A survey](#). *Journal of Artificial Intelligence Research*, 65:675–782.
- Isaac Johnson, Florian Lemmerich, Diego Sáez-Trumper, Robert West, Markus Strohmaier, and Leila Zia. 2021. [Global gender differences in wikipedia readership](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 15:254–265.
- Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. [URIEL+](#):

- Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.
- András Kornai. 2013. Digital language death. *PLOS ONE*, 8(10).
- Tom Leinster and Christina A. Cobbold. 2012. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489.
- Amr Magdy, Thanaa M. Ghanem, Mashaal Musleh, and Mohamed Mokbel. 2014. Exploiting geo-tagged tweets to understand localized language diversity. In *Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data*, GeoRich’14, New York, NY, USA. Association for Computing Machinery.
- Elizabeth Marsh, Elvira Perez Vallejos, and Alexa Spence. 2022. The digital workplace and its dark side: An integrative review. *Computers in Human Behavior*, 128:107118.
- Ulises A. Mejias and Nick Couldry. 2019. Datafication. *Internet Policy Review*, 8(4).
- Jennifer Pan and Margaret E. Roberts. 2020. Censorship’s effect on incidental exposure to information: Evidence from wikipedia. *SAGE Open*, 10(1).
- Jeffrey Pfeffer, Daniel Matter, Kokil Jaidka, Onur Varol, Anahita Mashhadi, Jonas Lasser, Dominik Assenmacher, Shuhan Wu, Dejing Yang, Christoph Brantner, Daniel M. Romero, Jan Otterbacher, Christian Schwemmer, Kiran Joseph, David Garcia, and Fred Morstatter. 2023. Just another day on twitter: A complete 24 hours of twitter data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1073–1081.
- R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Statista. 2025a. Internet and social media users in the world 2025. <https://www.statista.com/statistics/617136/digital-population-worldwide/>. Accessed: 2025-12-17.
- Statista. 2025b. Most visited websites worldwide. <https://www.statista.com/statistics/1201880/most-visited-websites-worldwide/>. Statistic on global website visits, based on data from November 2024; accessed 2025-20-12.
- Tuomas Väisänen, Olle Järv, Tuuli Toivonen, and Tuomo Hiippala. 2022. Mapping urban linguistic diversity with social media and population register data. *Computers, Environment and Urban Systems*, 97:101857.
- Hadley Wickham. 2023. *httr: Tools for Working with URLs and HTTP*. R package version 1.4.7.
- Wikimedia Foundation. 2024. Data platform/aqs/pageviews/pageviews per project. Accessed: 2025-12-18.
- Wikimedia Foundation. 2025. Legal: Wikimedia foundation country and territory protection list. Last edited 19 May 2025; accessed 23 December 2025.
- Wikimedia Foundation, Inc. 2025. List of wikipeidias. Meta-Wiki page, accessed 20 Dec 2025.
- Eddie Yang and Margaret E. Roberts. 2021. Censorship of online encyclopedias: Implications for nlp models. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, page 537–548, New York, NY, USA. Association for Computing Machinery.
- Ark Fangzhou Zhang, Danielle Livneh, Ceren Budak, Lionel Robert, and Daniel Romero. 2017. Shocking the crowd: The effect of censorship shocks on chinese wikipedia. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):367–376.

Modeling Linguistic Imprints of War Propaganda in a Russian Wikipedia Fork: A Comparative Analysis with the Original Wikipedia

Anastasiia Vestel

Saarland University

Saarbruecken, Germany

anastasiia.vestel@uni-saarland.de

Stefania Degaetano-Ortlieb

Saarland University

Saarbruecken, Germany

s.degaetano@mx.uni-saarland.de

Abstract

Although Wikipedia aspires to provide neutral information, alternative versions can be used for political manipulation. This paper analyzes how narratives about the Russo-Ukrainian War are linguistically reframed in a Russian Wikipedia Fork compared to the original Russian Wikipedia. Using Kullback-Leibler Divergence on a corpus of war-related edits in more than 13,000 articles, we identify key differences between the two versions. While the original Wikipedia features Ukrainian references and administrative details, direct war terminology, and Ukraine’s territorial designation, governance, and statehood, RWFork replaces or removes these elements, emphasizing reassignment of Ukrainian territories to Russia, favoring euphemistic war language, renaming locations, and recognizing Russia-backed DPR and LPR¹. These patterns closely align RWFork with demobilizational strategies observed in pro-Kremlin media.

1 Introduction and Related Work

Language, being a powerful tool, can be manipulated by malicious actors to shape public opinion. In this study, we examine how pro-Kremlin propaganda about the Russo-Ukrainian War is linguistically framed in a Russian Wikipedia Fork (RWFork) compared to the original Russian Wikipedia. RWFork, or Ruwiki, is a copy of Russian Wikipedia created in June 2023 and revised to comply with Russia’s legislation. It was launched by Vladimir Medeyko, the leader of Russian Wikipedia editors, who aspired to create a Kremlin-aligned alternative to the original Wikipedia (Cohen, 2023). Despite his denial about working for the Russian government, RWFork has

¹DPR (Donetsk People’s Republic) and LPR (Luhansk People’s Republic) are self-proclaimed separatist entities in Donbas that emerged in 2014, supported and recognized by Russia, which are internationally viewed as Ukrainian territory under Russian occupation.

been described as sponsored by the state and selectively edited to remove content unfavorable to the Kremlin (Roscoe, 2024).

Previous work on manipulative narratives in the context of the Russo-Ukrainian War has shown their evolving nature and adaptation to external events (Gerard et al., 2025a,b; Solopova et al., 2023). Therefore, RWFork presents a compelling case study for exploring the development of Russian propaganda by comparing two versions of Wikipedia separated by an important historical event: the start of Russia’s full-scale invasion of Ukraine.

As our method, we employ Kullback-Leibler Divergence (KLD; Kullback and Leibler, 1951), which, as opposed to transformer-based methods, offers an interpretable way to identify distinctive linguistic features. KLD has been established as a reliable method for detecting language variation and change, such as in studies by Bochkarev et al. (2014); Degaetano-Ortlieb and Teich (2019); Fankhauser et al. (2014); Hughes et al. (2012); Klingenstein et al. (2014); among others. Moreover, we have previously demonstrated that KLD is suitable for detecting divergences in propaganda strategies between traditional and social media covering the Russo-Ukrainian War (Vestel and Degaetano-Ortlieb, 2025). This study extends prior work by applying KLD to edits in RWFork, thus adding another text type (Wikipedia) and a diachronic dimension (the difference between the original and changed versions) to the analysis.

Propaganda studies on the Russo-Ukrainian War have mostly focused on media analysis (Akhynko et al., 2025; Hein, 2023; Vanetik et al., 2023, etc.). For instance, Alyukov et al. (2025) examined the differences between state-controlled traditional media (such as press and TV) and social media, showing that the former rely on demobilization by normalizing the war, while the latter have a more mobilizational character. A prominent

example of the normalization frame in traditional media is replacing direct war-related terminology with euphemisms such as *special military operation* (Alyukov et al., 2022; Vestel and Degaetano-Ortlieb, 2025). A similar trend was found in other media types, specifically in state-affiliated outlets as opposed to independent ones (Park et al., 2022) and in pro-Russian vs. pro-Ukrainian Telegram channels (Ustyianovych and Barbosa, 2024). More broadly, Gerard et al. (2025a) have demonstrated that Russia’s Telegram discourse seems to be detached from the realities of war, turning attention towards international responses and economic repercussions, which also aligns with a demobilizational strategy. Likewise, Russian bloggers more frequently focus on international actions instead of military developments, which is more characteristic of Ukrainian bloggers (Gerard et al., 2025b).

Wikipedia presents another text type, different from media: while aiming to be a neutral and objective source of information, it can also serve as a tool for knowledge manipulation, specifically in its alternative versions such as RWFork (Trokhymovych et al., 2025). Similarly to the above-mentioned studies on media propaganda, Trokhymovych et al. (2025) have shown that direct words like *war* and *invasion* are removed from RWFork. Moreover, RWFork replaces names of occupied Ukrainian regions with Kremlin-aligned terms and often removes Ukrainian references and administrative details (Trokhymovych et al., 2025). Both of these types of changes can be attributed to the normalization frame: the former aims to divert the readers’ attention from the war, whereas the latter tries to convince them that certain Ukrainian territories rightfully belong to Russia. Thus, RWFork seems to employ a demobilizational propaganda strategy, making it more similar to traditional media in contrast to social media.

The current study will test this hypothesis and not only investigate the divergences between the original and changed versions of Wikipedia but also compare the results with prior work on media analysis.

2 Data and Methodology

Our study uses Trokhymovych et al.’s (2025) RWFork dataset, containing edits from 1.9M page titles between May and September 2023. We filtered the data by selecting only the articles from the three categories of changes involving knowledge

manipulation linked to the 2022 Russian invasion of Ukraine, namely Territorial Claims Dispute, Terminology Changes Ukraine, and Sanctions Edit Adjustments (the category labels are assigned to each article; see the original paper for details). The resulting data comprised 13,048 articles (100,738 sentences in total). However, at first, we discovered that our data contained duplicated template sentences, requiring further pre-processing (refer to Appendix A for details). After removing them, the total number of sentences decreased to 92,629. Appendix B contains detailed statistics on the number of sentences, tokens, and analyzed words for each version before and after removing duplicates.

KLD is used to measure how much the language of RWFork diverges from Russian Wikipedia by quantifying the extra information needed to represent one probability distribution with the other, thereby identifying key linguistic features (e.g., words) that contribute to the linguistic differences. For instance, to calculate the KLD of the language of the *original* Wikipedia (O), given that of the *changed* RWFork (C), we use this formula:

$$D(O \parallel C) = \sum_i p(\text{feature}_i \mid O) \log_2 \frac{p(\text{feature}_i \mid O)}{p(\text{feature}_i \mid C)}$$

We applied KLD (Degaetano-Ortlieb and Teich, 2019) to the RWFork corpus, limiting the scope to the lemmas of nouns, proper nouns, verbs, adjectives, and adverbs, since these parts of speech (POS) are most likely to account for substantial differences in content words between the two versions of Wikipedia. The results, described in the next section, are based on the point-wise KLD, which measures the contribution of each feature to the overall KLD — high contribution reflects *distinctive* features. As opposed to other methods for detecting linguistic variation, such as the Jensen-Shannon Divergence, KLD is asymmetric, which makes it well-suited for capturing differences from distinct perspectives (Degaetano-Ortlieb and Teich, 2019).

Finally, we performed qualitative analysis by looking at the 50 most distinctive words for each version. In particular, we manually annotated each word with a topic it refers to. We drew inspiration from Trokhymovych et al.’s (2025) taxonomy of changes to classify the words according to topics, but the categories are more specific in our case, allowing for a more fine-grained analysis (see Appendix C for details). To find out the context around each word, we performed queries on the dataset us-

Topic	Original	Changed
DPR/LPR recognition	0.00%	10.00%
non-recognition of DPR/LPR	7.41%	0.00%
occupation/annexation (direct)	9.26%	0.00%
occupation/annexation (euphemism)	0.00%	4.00%
other	3.70%	10.00%
political figures	3.70%	0.00%
propaganda	1.85%	0.00%
renamed locations	0.00%	20.00%
sanctions	0.00%	8.00%
territorial designation & governance (Russia)	0.00%	34.00%
territorial designation & governance (Ukraine)	18.52%	0.00%
Ukraine's statehood	12.96%	0.00%
Ukrainian references & administrative details	22.22%	0.00%
war (direct)	20.37%	0.00%
war (euphemism)	0.00%	14.00%

Table 1: Percentages of words with the highest KLD values for the original and changed versions of Wikipedia grouped by topic.

ing Sketch Engine² as our corpus exploration tool. We used the parallel concordance feature, where the original and changed versions were aligned on the sentence level³, and examined the most frequent collocates for each word to inform our topic classification. In cases when a word could refer to more than one topic, it would receive multiple category labels.

3 Results

3.1 Overall Divergence

The overall KLD for the original Russian Wikipedia is close to 0.684, whereas for RWFork, the divergence is around 0.415, meaning that 0.269 additional bits are required to encode the former corpus with the latter. An explanation for this could be that RWFork is characterized by a more limited vocabulary, given that it contains many duplicated sentences (even after removing most of them, there were still a few template phrases left, as will be discussed in the next subsection).

3.2 Divergences between the Russian Wikipedia and RWFork

Figure 1 shows the most distinctive words for the Russian Wikipedia and RWFork (a list of these words, their translation into English, and their probabilities can be found in Appendix D). Additionally, Table 1 summarizes the percentage of words that belong to different topics in either version.

²<https://www.sketchengine.eu/>

³Note that only modified sentences were aligned; added and deleted sentences had empty entries in the Russian Wikipedia and RWFork, respectively.

About 22% of the words distinctive to the original version indicate Ukrainian references and administrative details in relation to certain locations occupied or claimed by Russia; they are often deleted from RWFork, implying that these territories are now part of Russia. Examples include *postal*, *telephone*, *code*, and *index*, the abbreviation *KOATUU* (which translates to "Classification of objects of the administrative-territorial system of Ukraine"), and links to the website of the Verkhovna Rada (Supreme Council), Ukraine's parliament (evidenced by the words *website*, *supreme*, *council*, and *Ukraine*).

The next most common category of words distinctive of the original Wikipedia is that of direct war terminology (around 20%), such as *invasion*, *war*, *aggression*, and *attack*, as well as their collocates in phrases like *full-scale invasion*, someone's *support* for the war, or *during* the war. These words and phrases are, for the most part, removed from RWFork or otherwise substituted with vague terms such as *conflict*, *military actions*, and *special military operation* (the words comprising these expressions are among the top KLD words for RWFork; see Figure 1 and Appendix D), as well as Russia's politics *in relation to* Ukraine. It is worth noting that 14% of the 50 most distinctive words for RWFork are war-related euphemisms or their collocates, making this category the third most common type of changed Wikipedia content. In a similar vein, direct words that refer to Russia's occupation of Ukrainian territories (e.g., *annexation*, *occupation*, *to occupy*, *to capture*) are distinctive for the Russian Wikipedia, comprising 9% of the top distinctive KLD words for this version; they are also mostly deleted or replaced with euphemisms such as *inclusion* and *entry* (into Russia).

Almost 19% of the words distinctive for the original version are those indicating the territorial designation and governance of Ukraine. For instance, the word for *council* (transliterated as *sovet*) and its collocate *rural council*, as a form of local government in Ukraine, are either removed or replaced with a corresponding form of local government in Russia (*okrug*). Another example is the word *autonomous*, which is among the top distinctive KLD words for the original Wikipedia and is often part of the phrase "Autonomous Republic of Crimea", as this is its official name according to the administrative division of Ukraine. After Russia annexed the peninsula in 2014, it renamed the territory to "Republic of Crimea", leaving out the word "au-



Figure 1: Top 50 words with the highest KLD values for the original (left) and changed versions (right). Size accounts for KLD values, and color indicates frequencies from higher (red) to lower (green).

onomous", which suggests its belonging to Russia. Therefore, this word was mostly deleted from the republic's name in RWFork.

A few words (approx. 13%) among those most distinctive for the Russian Wikipedia refer to the territorial statehood of Ukraine (such as *territorial*, *integrity*, *sovereignty*, and *independence*). They are mostly mentioned when talking about Russia's violation and disruption of Ukraine's statehood (hence, the words *violation* and *to disrupt* also have a high contribution to the language of the original version), and these phrases are almost always removed from RWFork.

More than one-third of the top distinctive KLD words for RWFork (34%) indicate the territorial designation and governance of Russia (including the so-called DPR and LPR, since they are backed by Russia). A lot of these words turned out to be a case of a few more template sentences that were duplicated in different articles and were not initially detected. These sentences state that before the year 2014 or 2022, a certain village was part of an occupied Ukrainian region; that since spring 2014, it has belonged to the DPR; or that since August 2022, it has been part of a city or municipal council of Kherson Oblast. These phrases contributed numerous words to the RWFork word cloud, such as *village*, *year*, *spring*, *August*, *municipal*, etc., implying that the specified territory now belongs to Russia. Such framing indicates territorial control over certain locations in the self-proclaimed DPR, as well as villages in Kherson Oblast captured by Russia during the full-scale invasion.

Furthermore, 20% of the top distinctive KLD words for RWFork are Ukrainian territories con-

trolled or claimed by Russia that have been renamed by the Russian government. This is often the case of administrative reorganization, when the borders of certain regions or districts are redrawn to their previous state during the Soviet regime, and their names are changed accordingly. A prominent example is *Krasnoarmeysk*, a former district in Donetsk Oblast partially occupied by Russia, which, according to the Ukrainian administrative division, is now part of Pokrovsk Raion (district). "Krasnoarmeysk" literally translates to "Red Army", and it was the Soviet name of the district (and its capital city) up until 2016, when it was renamed as Pokrovsk Raion by the Ukrainian government in the process of decommunization; in 2020, it was enlarged to include a few territories of other neighboring districts. By renaming certain locations in this way, the editors of RWFork aim to promote the narrative that they have inherently belonged to Russia.

Finally, the language of RWFork clearly indicates the recognition of the self-proclaimed DPR and LPR, as 10% of the most distinctive words for the changed version are related to those regions. These words are *LPR*, *Luhansk* (or *Lugansk*, if transliterated from Russian), *Donetsk*, *people's*, and *republic*. On the contrary, the original Wikipedia does not recognize these republics, which is evidenced by some words with high KLD values for this version. These include *self-proclaimed* and *separatist*, as well as *control* and *to be controlled* (in sentences like "[This territory] is controlled by the self-proclaimed DPR/LPR").

3.3 Discussion

By comparing our results with other studies on propaganda about the Russo-Ukrainian War, we have observed many similarities between the language of RWFork and that of traditional, state-affiliated, and pro-Russian media. The most remarkable example is that direct war-related terminology is often deleted (as was also noted by Trokhymovych et al., 2025) or replaced with vague euphemistic expressions, a pattern found by Alyukov et al. (2022); Park et al. (2022); Ustyianovych and Barbosa (2024); Vestel and Degaetano-Ortlieb (2025). This, and the fact that the words *war*, *invasion*, *aggression*, etc., are removed more often than they are substituted, points to a demobilizational strategy frequently employed by traditional media (Alyukov et al., 2025). Moreover, by renaming Ukrainian locations, deleting any references to Ukraine, and recognizing Russia-backed DPR and LPR, the authors of RWFork seek to convince the population that these territories have been historically part of Russia. These changes point to territorial control propaganda, which also reflects the normalization frame of the demobilizational approach, and the same trend was found to be distinctive for traditional media as opposed to social media in our previous study (Vestel and Degaetano-Ortlieb, 2025).

These findings reveal that studying propaganda in alternative versions of Wikipedia, such as RWFork, is equally important to media research. The encyclopedic style of RWFork gives it a certain degree of credibility as opposed to media sources, and the demobilizational strategies employed in it are a more subtle propaganda technique than mobilization. Therefore, ordinary users might rely on RWFork as an objective source of information without questioning its validity or suspecting possible manipulation. For example, while someone might not be searching for the news on purpose but stumbles upon an article about a certain town or village, they might see the location marked as "Russia" and not realize that this is Ukrainian territory occupied by Russia as a result of the invasion.

4 Conclusion and Future Work

By employing KLD to compare the two versions of Russian Wikipedia, we have identified key linguistic features that distinguish propagandistic rhetoric in war-related narratives. Specifically, the language of the original version is characterized by Ukrainian references and administrative details,

Ukraine's territorial designation, governance, and statehood, as well as direct terminology for the war and Russia's occupation or annexation of Ukrainian territories. In contrast, RWFork focuses on Russia's territorial designation and governance, renaming Ukrainian locations, euphemistic framing of the war, and recognizing the self-proclaimed DPR and LPR. In addition, we discovered that the language of RWFork aligns closely with pro-Kremlin propaganda spread in pro-Russian, state-affiliated, or traditional media, as opposed to the original version. This is evidenced by the efforts to downplay the war and the focus on territorial control, which point to a demobilizational strategy. Overall, our work sheds light on how propaganda is construed in political environments and enhances reproducible computational techniques for its detection.

In the future, we will expand our set of linguistic features (e.g., dependency relations) to investigate more subtle propagandistic framing. Furthermore, we aim to extend this study by applying other methods to the dataset, such as surprisal (Shannon, 1948), which models the (un)expectedness of words in particular contexts to capture more nuanced local linguistic changes. Finally, for a more fine-grained analysis and classification into different propaganda techniques, such as those proposed by Da San Martino et al. (2019), a comprehensive study of evaluative language is needed; this can be informed by word embeddings (Mikolov et al., 2013), which will allow us to model semantic divergences between the two versions of Wikipedia.

Acknowledgments

Funded by the European Union under grant 101119511. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

We sincerely thank Mykola Trokhymovych for his help with the RWFork dataset and Stefan Fischer for providing technical support.

References

Kateryna Akhynko, Oleksandr Kosovan, and Mykola Trokhymovych. 2025. *Hidden Persuasion: Detecting Manipulative Narratives on Social Media During the 2022 Russian Invasion of Ukraine*. In *Proceedings of the Fourth Ukrainian Natural Language Process-*

- ing Workshop (UNLP 2025), pages 194–202, Vienna, Austria (online). Association for Computational Linguistics.
- Maxim Alyukov, Maria Kunilovskaya, and Andrei Semenov. 2025. *Confuse and Normalise: Authoritarian Propaganda in a High-Choice Media Environment and Russia’s Invasion of Ukraine*. In Paul Goode, editor, *Russian Propaganda Today: Challenges, Effectiveness, and Resistance*, page in print. University of Michigan press, University of Manchester Press.
- Maxim Alyukov, Andrei Semenov, and Maria Kunilovskaya. 2022. Propaganda Setbacks and Appropriation of Anti-war language: “Special Military Operation” in Russian Mass Media and Social Networks (February-July 2022). Monitoring Report №1.
- Vladimir Bochkarev, Valery D. Solovyev, and Søren Wichmann. 2014. *Universals versus Historical Contingencies in Lexical Evolution*. *Journal of The Royal Society Interface*, 11(101):20140841.
- Noam Cohen. 2023. Russian Wikipedia’s Top Editor Leaves to Launch a Putin-Friendly Clone. *Bloomberg.com*.
- Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. *Fine-Grained Analysis of Propaganda in News Article*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.
- Stefania Degaetano-Ortlieb and Elke Teich. 2019. *Toward an Optimal Code for Communication: The Case of Scientific English*. *Corpus Linguistics and Linguistic Theory*, 18(1):175–207.
- Peter Fankhauser, Jörg Knappen, and Elke Teich. 2014. Exploring and Visualizing Variation in Language Resources. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 4125–4128, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Patrick Gerard, Svitlana Volkova, Louis Penafiel, Kristina Lerman, and Tim Weninger. 2025a. *Modeling Information Narrative Evolution on Telegram During the Russia-Ukraine War*. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:602–614.
- Patrick Gerard, Tim Weninger, and Kristina Lerman. 2025b. *Fear and Loathing on the Frontline: Decoding the Language of Othering by Russia-Ukraine War Bloggers*. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:615–635.
- Vitalij Hein. 2023. *Propaganda Detection in Russian and American News Coverage about the War in Ukraine through Text Classification*. Thesis, Technische Universität Wien.
- James M. Hughes, Nicholas J. Foti, David C. Krakauer, and Daniel N. Rockmore. 2012. *Quantitative Patterns of Stylistic Influence in the Evolution of Literature*. *Proceedings of the National Academy of Sciences*, 109(20):7682–7686.
- Sara Klingenstein, Tim Hitchcock, and Simon DeDeo. 2014. *The Civilizing Process in London’s Old Bailey*. *Proceedings of the National Academy of Sciences of the United States of America*, 111(26):9419–9424.
- Solomon Kullback and Richard A. Leibler. 1951. *On Information and Sufficiency*. *The Annals of Mathematical Statistics*, 22(1):79–86.
- Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.
- Chan Young Park, Julia Mendelsohn, Anjalie Field, and Yulia Tsvetkov. 2022. *Challenges and Opportunities in Information Manipulation Detection: An Examination of Wartime Russian Media*. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5209–5235, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jules Roscoe. 2024. Russia Clones Wikipedia, Censors It, Bans Original. <https://www.404media.co/russia-clones-wikipedia-censors-it-bans-original/>.
- Claude E. Shannon. 1948. *A Mathematical Theory of Communication*. *Bell System Technical Journal*, 27(3):379–423.
- Veronika Solopova, Christoph Benz Müller, and Tim Landgraf. 2023. *The Evolution of Pro-Kremlin Propaganda From a Machine Learning and Linguistics Perspective*. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 40–48, Dubrovnik, Croatia. Association for Computational Linguistics.
- Mykola Trokhymovych, Oleksandr Kosovan, Nathan Forrester, Pablo Aragón, Diego Saez-Trumper, and Ricardo Baeza-Yates. 2025. *Characterizing Knowledge Manipulation in a Russian Wikipedia Fork*. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:1924–1936.
- Taras Ustyianovych and Denilson Barbosa. 2024. Instant Messaging Platforms News Multi-Task Classification for Stance, Sentiment, and Discrimination Detection. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 30–40, Torino, Italia. ELRA and ICCL.
- Natalia Vanetik, Marina Litvak, Egor Reviakin, and Margarita Tyamanova. 2023. *Propaganda Detection in Russian Telegram Posts in the Scope of the Russian Invasion of Ukraine*. In *Proceedings of the Conference Recent Advances in Natural Language*

Anastasiia Vestel and Stefania Degaetano-Ortlieb. 2025. [From War to Special Military Operation: Interpretable Detection of Linguistic Propaganda Framing in Russian Media](#). *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*, 2025:50.

A Duplicated sentences

During the initial analysis, we observed that 52% of the top KLD words for RWFork came from only three template sentences used throughout various articles about different Ukrainian locations occupied or claimed by Russia. These sentences exhibited changes in territorial designation and governance, asserting that, according to an agreement between a certain Ukrainian region and the Russian Federation, the specified location belonged to Russia; that the Ukrainian authorities disputed this; and that Ukraine exercised de facto control over this territory. Examples of words that stem from these sentences and have a high contribution to the language of RWFork are *agreement*, *to dispute*, and *de facto* (see Figure 2). A possible explanation for the repetition of these template texts is that a significant share of RWFork’s editorial work may have been carried out by paid staff (Trokhymovych et al., 2025). If this is the case, the RWFork contributors might have received certain guidelines that prescribed inserting these sentences in multiple articles to unify RWFork’s content across different pages. This can be regarded as a specific type of information manipulation: for instance, a study by Da San Martino et al. (2019) showed that repetition was the third most common strategy within a fine-grained taxonomy of 18 propaganda techniques.

Since the duplicated sentences impacted our KLD results to such a great extent, we decided to remove them and rerun the experiment. The motivation behind this was the fact that we wanted to analyze how propaganda is framed linguistically compared to the original Wikipedia version, and the duplicated content accounted mostly for extralinguistic factors, therefore overshadowing more meaningful features that might point to certain manipulative strategies. After deleting the duplicates, we saw that 80% of the top 50 words distinctive for the original Wikipedia were the same as in the initial setup of the experiment, but the overlap between the 50 words with the highest KLD values

for RWFork with and without the repeated template sentences was only 58%.

B Descriptive statistics of the data

Table 2 shows the number of sentences, tokens, and POS-filtered target words used in our analysis for each version and in total, both before and after removing the duplicated template sentences discussed in Appendix A. The number of articles reported in Section 2 is the same for both versions.

C Classification of topics

Below is a detailed taxonomy of topics, expanded from the three initial categories of changes related to the Russo-Ukrainian War, which were proposed by Trokhymovych et al. (2025). Apart from making the three clusters more specific, we add more categories (under "Other"). In addition, we differentiate between a pro-Russian and a pro-Ukrainian stance by separating direct terminology from euphemisms, as well as discriminating between the language that does or does not recognize the DPR and LPR and the phrasing that supports Russia’s or Ukraine’s claim on a territory.

- Territorial Claims Dispute
 - DPR/LPR recognition
 - non-recognition of DPR/LPR
 - renamed locations
 - territorial designation and governance (Russia)
 - territorial designation and governance (Ukraine)
 - Ukrainian references and administrative details
- Terminology Changes Ukraine
 - occupation/annexation (direct)
 - occupation/annexation (euphemism)
 - war (direct)
 - war (euphemism)
- Sanctions Edit Adjustments
 - sanctions
- Other
 - political figures
 - propaganda
 - Ukraine’s statehood
 - other



Figure 2: Top 50 words with the highest KLD values from the original analysis for the Russian Wikipedia (left) and RWForK (right). Size accounts for KLD values, and color indicates frequencies from higher (red) to lower (green).

#	Before deduplication			After deduplication		
	Original	Changed	Total	Original	Changed	Total
Sentences	59,648	41,090	100,738	59,648	32,981	92,629
Tokens	859,234	632,051	1,491,285	859,234	477,274	1,336,508
Words analyzed	517,696	409,322	927,018	517,696	298,291	815,987

Table 2: Descriptive statistics of the data used for the analysis.

D Top 50 KLD words

Table 3 includes a list of 50 words with the highest KLD values, their translation into English, and their probabilities for the original and changed versions of Wikipedia. Note that to translate the names of Ukrainian locations, we did not use their official translations from Ukrainian into English but transliterated them from Russian into Latin script, since many of them are the ones that have been renamed (see Section 3.2); others, like *Zaporozh'ye* instead of *Zaporizhzhia*, have also been transliterated and not translated for consistency. However, *Luhansk* in the word list for the original Russian Wikipedia was translated in this way because this word was originally in Ukrainian (it appeared as part of an address), while its Russian-to-English transliteration, *Lugansk*, can be seen in the word list for RWForK, since in this version it was spelled in Russian. At the same time, most of the locations are adjectives used in the names of regions (e.g., *Kherson Oblast*), so we transliterated the city or town their names come from. The repeated word *Chaplynka* (name of a region) is a result of a parsing error when different word forms were not recognized as the same lemma. Finally, abbreviations that have not been mentioned in the paper are explained in parentheses.

words	translation	kld1	kld2	p1	p2
вторжение	invasion	0.08660169	-0.000542759	0.011774726	7.42601E-05
код	code	0.040392951	-0.001003039	0.007576064	0.000188129
коатуу	koatuu	0.032426735	-0.000178471	0.00432048	2.37791E-05
украина	ukraine	0.031051865	-0.018145495	0.040063383	0.023411474
индекс	index	0.020182594	-0.000198151	0.003025711	2.97062E-05
сайт	website	0.020110603	-0.001344424	0.005152738	0.000344468
рада	council	0.018103583	-0.001330889	0.004807351	0.000353414
верховный	supreme	0.016541175	-0.001630482	0.00494846	0.000487775
телефонный	telephone	0.015244526	-0.000608512	0.003280607	0.000130951
россия	russia	0.013438068	-0.007987493	0.017905223	0.010642738
почтовый	postal	0.01141654	-0.000860441	0.003060813	0.000230687
территориальный	territorial	0.010842433	-0.002122099	0.004607674	0.000901822
аннексия	annexation	0.009574253	-3.04384E-05	0.001153924	3.66855E-06
война	war	0.009325456	-0.002248677	0.004544352	0.001095794
самопровозглашённый	self-proclaimed	0.008389039	-0.000312659	0.00176766	6.58805E-05
совет	soviet	0.008257453	-0.003712976	0.007160961	0.003219936
деление	division	0.006969798	-0.000601835	0.001972392	0.000170314
луганська	luhansk	0.006099984	-1.9393E-05	0.000735193	2.33732E-06
российский	russian	0.005500208	-0.003841097	0.010618824	0.007415708
целостность	integrity	0.005355526	-0.000634959	0.001740909	0.000206404
украинской	ukrainian	0.005183336	-0.001723559	0.003263059	0.001085025
поддержка	support	0.004810887	-0.001659934	0.003133767	0.001081266
агрессия	aggression	0.004757816	-0.00040701	0.001341302	0.000114742
административный	administrative	0.004329221	-0.001950613	0.003763952	0.001695921
сельский	rural	0.00406735	-0.002547075	0.006023492	0.003772059
оккупировать	occupy	0.00378714	-0.000346861	0.001098143	0.000100578
население	population	0.003690449	-0.001834607	0.003659979	0.00181946
улице	street	0.003683599	-6.92844E-05	0.000642588	1.20864E-05
нарушение	violation	0.003490644	-0.000517941	0.00126811	0.000188162
окупация	occupation	0.003377183	-0.000146988	0.000746826	3.25047E-05
контроль	control	0.003344609	-0.001133892	0.002143218	0.000726596
российско	russo-	0.002988045	-0.001009512	0.001907661	0.000648847
классификация	classification	0.002898066	-0.000105284	0.000605943	2.20133E-05
время	time	0.002808515	-0.001581546	0.003899955	0.001908993
область	oblast	0.002346483	-0.002118394	0.01590522	0.014359159
автономный	autonomous	0.002212086	-0.001023394	0.001989203	0.00092028
подрывать	disrupt	0.0021831	-0.000176871	0.000602133	4.87838E-05
суверенитет	sovereignty	0.002001847	-0.000406664	0.000870586	0.000176855
обл	obl (oblast)	0.001888448	-0.001187279	0.002820506	0.001773269
так	so	0.001827927	-0.000800015	0.001533354	0.000671091
путин	putin	0.001819609	-0.001051689	0.002300623	0.001329704
независимость	independence	0.001639769	-0.000499208	0.000955699	0.000290951
контролировать	be controlled	0.001490477	-9.70816E-05	0.000378252	2.46373E-05
устройство	structure	0.001446351	-0.000227786	0.000542382	8.54199E-05
захватить	capture	0.001418415	-0.000274318	0.000598397	0.000115728
пропаганда	propaganda	0.001383671	-0.000224217	0.000526945	8.53708E-05
сепаратист	separatist	0.001332055	-0.000120349	0.00038406	3.46992E-05
полномасштабный	full-scale	0.001315102	-0.000150678	0.000420747	4.82071E-05
нападение	attack	0.001310835	-0.000208434	0.000494129	7.85708E-05
рф	rf (russian federation)	0.001277267	-0.000861606	0.002248867	0.001517017

words	translation	kld1	kld2	p1	p2
народный	people's	-0.006554873	0.038029752	0.002584231	0.014993069
республика	republic	-0.011293984	0.032744543	0.007354291	0.021322227
военный	military	-0.006373787	0.030521043	0.002820782	0.013507388
действие	action	-0.005416492	0.027148564	0.002329228	0.011674566
село	village	-0.012863659	0.024819316	0.013566939	0.026176233
год	year	-0.012912389	0.017643704	0.028668868	0.039173622
район	raion	-0.010701834	0.016946686	0.016137995	0.025555017
входить	be part of	-0.003767267	0.01007444	0.002654669	0.007099128
округ	okrug	-0.001425247	0.009956661	0.005082029	0.003550307
состав	composition	-0.003387915	0.006511743	0.003594048	0.00690794
херсонский	kherson	-0.003165557	0.006157423	0.003297911	0.006414869
донецкий	donetsk	-0.004173868	0.005775763	0.008906591	0.012324867
данным	data	-0.001425812	0.005676612	0.000715322	0.002847923
посёлок	village	-0.002631359	0.004569797	0.00330439	0.005738629
лнр	lpr	-0.001559755	0.004148277	0.001105274	0.002399554
запорожский	zaporozh'ye	-0.002244964	0.004000972	0.002692907	0.004799295
муниципальный	municipal	-0.000874695	0.003969679	0.000400837	0.001819142
городской	city	-0.002108014	0.003279932	0.003305228	0.00514272
луганский	lugansk	-0.002294771	0.003054	0.005565111	0.007406337
санкционный	sanctioned	-0.00167688	0.002443285	0.003087931	0.004499247
список	list	-0.00168797	0.00237826	0.003412674	0.004808279
начало	start	-0.001418093	0.00225028	0.002128781	0.003378025
тип	type	-0.001093008	0.002243665	0.001053681	0.002162695
вольнянском	volnyansk	-4.59238E-05	0.002031891	8.39951E-06	0.000371635
отношение	relation	-0.000930081	0.001721122	0.001047481	0.001938371
красноармейский	krasnoarmeysk	-0.000131614	0.001709402	3.56801E-05	0.000462113
мо	mo (municipal okrug)	-0.000115489	0.001702974	2.97482E-05	0.000438659
артёмовский	artemovsk	-0.000190333	0.001667124	6.07942E-05	0.000532497
км	km	-0.000998368	0.00160618	0.001455371	0.002341407
поселение	settlement	-0.001016529	0.001601508	0.001550107	0.002442143
специальный	special	-0.00050051	0.00151948	0.000312408	0.000948427
относится	belong	-0.000893753	0.001452661	0.001275423	0.002073009
краснолиманском	krasnyy liman	-1.51591E-05	0.001398687	2.32226E-06	0.000214268
август	august	-0.000725937	0.001321208	0.000840261	0.001529278
операция	operation	-0.000607059	0.001311385	0.000546318	0.00118017
численность	number	-0.000349115	0.00122901	0.000192273	0.000676871
присоединение	inclusion	-0.000545074	0.001178221	0.000490137	0.00105947
февраль	february	-0.000995856	0.001170766	0.004265949	0.005015209
александровский	aleksandrovka	-0.000475038	0.001097101	0.000393379	0.00090851
воладарский	volodarskoye	-9.98626E-05	0.001042632	2.9509E-05	0.000380893
чапльинский	chaplinsk	-1.89641E-06	0.001034915	2.08579E-07	0.000113827
сша	usa	-0.000751359	0.000877845	0.003347352	0.003910855
гуляйпольском	gulyaypole	-5.5183E-05	0.000863695	1.39062E-05	0.000217653
вхождение	entry	-0.000160475	0.000859995	6.6258E-05	0.00035508
куйбышевский	kuybyshevo	-0.000119403	0.000818604	4.29919E-05	0.000234745
весна	spring	-0.000341452	0.000808923	0.000275199	0.000650352
чапльинском	chaplinsk	-1.29288E-05	0.000795045	2.17503E-06	0.00013392
город	city	-0.00067337	0.000763929	0.003699068	0.004196538
конфликт	conflict	-0.000531337	0.00074783	0.001077582	0.001516642
значение	significance	-0.000396116	0.000733817	0.000445326	0.00082498

Table 3: The 50 most distinctive words for the original (left) and changed versions (right), together with their English translations and respective probabilities ("p"); "kld1" and "p1" are used for the Russian Wikipedia, while "kld2" and "p2" represent values for RWFork.

Stylometric Approach to AI-generated Texts. An Analysis of Contemporary French-Language Literature

Adam Pawłowski

University of Wrocław
pl. Uniwersytecki 1
50-137 Wrocław, Poland
adam.pawlowski@uwr.edu.pl

Tomasz Walkowiak

Wrocław University of Science and Technology
27 Wybrzeże Wyspiańskiego St.
50-370 Wrocław, Poland
tomasz.walkowiak@pwr.edu.pl

Abstract

The article focuses on a stylometric analysis of authentic literary texts and thematically related texts generated by large language models. The texts under study represent a fairly broad cross-section of twentieth-century French literature. Five models were used to generate the texts (ChatGPT 4-o, GPT 4-o mini, DeepSeek v.3, c4ai-command-r-plus, and c4ai-command-a). The original human-written stories of approximately 20,000 characters were summarized, and new narratives were then generated on the basis of these abstracts. In terms of plot and style, they were intended to resemble the originals. The research carried out with TF-IDF of the most frequent words showed that texts generated by specific LLMs and written by humans cluster relatively well as distinct groups. The experiments also showed that the "authorial" specificity of machine-generated texts partly matches the original clustering of human-written source texts.

1 Introduction

In today's communication landscape, machine-generated texts have secured a permanent place. Large Language Models (LLMs) are widely used to produce function texts on a variety of topics (formal letters, instructions, applications, etc.). LLMs translate, summarize, and assist humans in information retrieval and problem solving. They can also engage in quasi-natural conversations. One area where machine-generated texts have not yet reached a quality comparable to that of human-created ones is literary fiction. One might even say that literature has become the last shore defended against the expansion of machines, belonging to humans who reassure each other that there will never be an "electronic Aeschylus, Shakespeare, Molière or Balzac." Whether this will actually never happen remains unknown. However, it is certain that

some authors are already using LLMs as intelligent assistants to generate their literary texts. A comparative study of human- and machine-generated literary texts is therefore justified and necessary. On the one hand, such experiments serve as test of model performance; on the other hand, they reveal whether there still exists a boundary — and, if so, where it lies — separating human creativity from machine-generated AI outputs.

2 State-of-the-Art

Since the publication of the groundbreaking study by Mosteller and Wallace in 1964, it has been well established that reliable results in stylometric research are obtained by analyzing content-free function words rather than content-bearing lexemes (Mosteller and Wallace, 1964). This is because function words are resistant to conscious manipulation by the author and are largely independent of the subject matter of a text. A subsequent stage in the development of stylometric research involved the widespread adoption of advanced multivariate methods and the extension of the linguistic category of the word to include n-grams, defined as arbitrary sequences of characters.

Today, it is assumed that the aim of the interdisciplinary field of stylometry (Sara El Manar El Bouanani, 2014; Pöpcke et al., 2022; Eder et al., 2015) is to explore the relationship between the statistical features of texts and their meta-characteristics, such as authorship (Juola, 2006; Stamatatos, 2009; Kestemont, 2014) or the literary period of the work (Rabaev et al., 2023).

In recent years, increasing attention has been paid to the study of texts generated by artificial intelligence. The computational methods employed in this line of research do not differ substantially from those used during the period when only human-authored texts were under investigation. However, the principal challenge lies in the se-

lection of the textual material. Existing studies tend to rely on the most readily available corpora, which typically lack a clearly defined stylistic profile. Consequently, the conclusions drawn from such analyzes are of limited relevance to the study of creative or literary texts.

A representative example of this trend can be found in (Przystalski et al., 2026), where Wikipedia materials were processed without taking into account the fact that the encyclopedic style – particularly in the context of collaborative writing – does not exhibit a distinctive authorial stylistic signature and that Wikipedia contributors increasingly rely on AI-assisted tools when composing entries.

A more valuable contribution to the analysis of literary style in the context of creative writing is offered by (O’Sullivan, 2025). The author examined creative texts produced by both human writers and machines and concluded as follows: “The results reveal clear and consistent stylistic distinctions. Human-authored texts form broader, more heterogeneous clusters, reflecting the diversity of individual expression, writing ability, and interpretive engagement with the prompts. In contrast, LLM outputs, while fluent and coherent, display a higher degree of stylistic uniformity, clustering tightly by model.” (ibid.)

3 Goal of the Study

The experiment conducted aimed to apply selected stylometric measures to a corpus of twentieth-century literary fiction texts, which included samples written by humans (see the Appendix) and generated automatically by LLMs. Particular attention should be paid to the way the machine corpus was constructed. Our intention was to avoid comparing random text samples whose similarity or lack thereof would be difficult to interpret. To this end, we prepared a corpus of text pairs (human vs machine) related in terms of theme and style, which recount the same plots and employ the same narrative schemes.

The aim of the study was to verify how NLP and stylometric methods would cluster these texts, and to identify some hidden relations. This is certainly not an answer to the question of whether “a machine can write books” or whether AI-written narratives are as good as human literature; however, this measurement indirectly indicates the specificity of the literature produced by LLMs. The study was conducted on prose texts written in French by au-

thors from Europe (15 authors) and Africa (7 authors).

It is worth emphasizing that the task formulated here is demanding, as it goes beyond standard research aimed at distinguishing between machine-generated and human-written texts. Since the essence of literature is the creative fictionality based on the principle of mimesis, there is no truth criterion that could be applied to evaluate machine-generated texts. This radically distinguishes literary stylometry from studies of texts written under one’s own identity (e.g., letters, blogs, posts on social networks), as well as from functional texts (e.g., administrative or informational). The quality of literature has always been judged by the reader and his/her expectations rather than by its truthfulness or similarity to any specific pattern. There are numerous examples of works considered as worthy in one literary period but rejected by subsequent generations of readers. There are also works praised in one culture or language but dismissed in another. For this reason, a mixed “machine-human” corpus of texts linked by topics and (supposedly) style constitutes a valuable testing material for our research.

We formulated two hypotheses. The first posits that stylometric methods, which make it possible to identify the author of a literary text on the basis of style, will also prove effective in the analysis of machine-generated literature. The second hypothesis posits that texts generated by LLMs exhibit certain common features of a specific ‘collective identity’. This implies that, at a sufficiently high level of abstraction, human-written texts and texts produced by various LLMs will form distinct clusters.

4 Methods

The study was conducted in three stages. First, we compiled a corpus of contemporary French-language prose from the twentieth century (Pawłowski et al., 2025) (cf. Figure 1). From contemporary novels available in public resources (libraries and repositories) worldwide, we extracted samples of approximately 15,000 to 20,000 characters long. An average of ten samples were generated from each novel. In total, we produced 744 “human-authored” source samples that were further used to generate the corresponding machine-produced texts. The choice of twentieth-century texts, often protected by copyright and difficult to

access, was deliberate. Large language models are trained primarily on contemporary language, and we assumed that the resulting machine-generated texts would be of higher quality than those based on nineteenth-century literature, which is available in free-access repositories. In the second phase,

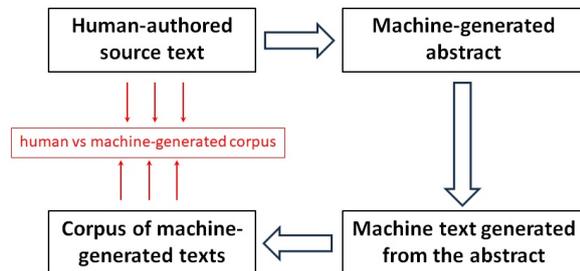


Figure 1: Data processing pipeline

we generated the texts corresponding to the source samples. Using each LLM, we produced 500-word summaries, and subsequently, on the basis of these summaries, full "literary" texts on the same topic were reproduced. An example of a summarization prompt for a novel by the Senegalese writer Fatou Diome was as follows:

Voici le texte de Fatou Diome, écrivaine sénégalaise d'expression française. Je voudrais obtenir un résumé de ce texte d'une longueur de 500 mots.

An example of a prompt used to generate an entire micro-story by the same author was as follows:

Voici le résumé de texte de Fatou Diome, écrivaine sénégalaise d'expression française. Sur la base de ce résumé, je voudrais obtenir un récit cohérent d'une longueur d'environ 2560 mots, rédigé dans le style de Fatou Diome.

It is worth noting that prompts defined in this way are highly elaborate and go beyond standard AI-based text generation methods. They comprise not only metadata (in the above query: the author's surname, gender, and nationality), but also a detail-rich content description embedded in the abstract.

Texts generated by the models rarely reached the expected length in the first generation cycle. A likely reason for this is a limitation inherent to LLMs, which have difficulty sustaining long-form narratives. Consequently, in many cases, we employed follow-up prompts explicitly requesting an increase in the length of the generated narrative. Since we used five LLMs, the final corpus comprised 3,999 samples: 744 human-authored texts and 3,255 machine-generated texts. Some API calls did not respond, which is why the number of machine-generated texts is lower than the ex-

pected $744 \times 5 = 3,720$. The models evaluated included two OpenAI models (GPT-4o and GPT-4o-mini), two models developed by Cohere Labs (c4ai-command-r-plus and c4ai-command-a), and one model from DeepSeek AI (DeepSeek-V3).

In the third phase, stylometric feature vectors were constructed for each text. We relied on a standard stylometric representation based on the frequencies of the most frequent words, weighted with TF-IDF (Salton G, 1988). For exploratory analysis, the resulting vectors were projected onto a two-dimensional space using the PaCMAP (Wang et al., 2021) dimensionality reduction technique. Next, using the classifier of the closest neighbors (k-NN) with $k = 5$, we evaluated three classification tasks: (i) detection of human text, distinguishing human-authored texts from AI-generated ones; (ii) detection of text origin, identifying whether a text was written by a human or by one of the language models used; and (iii) analysis of authorship, aimed at attributing texts to their respective authors. Finally, we quantified text similarity within the same author and source (human or specific LLM) by computing the average distance between each text's stylometric vector and its nearest neighbor from the same author and source.

5 Results

The results obtained should be considered valuable and, in some cases, even surprising. As expected, attribution of authorship for human authors yielded convincing results. However, relatively high performance was also achieved in the AI-author corpus (Table 1). Notably, reliable identification of AI-authored texts becomes possible only when at least 1,000 features are used, whereas in the human-authored corpus stylistic variation is more pronounced: high classification accuracy is observed even with a relatively small number of features. This indicates that human authors are more strongly individualized than the closely related group of AI authors.

Figure 2 shows that human texts can be effectively distinguished from machine-generated ones, despite the use of highly elaborate prompts (an instruction and a 500-word abstract). Figure 3 proved particularly striking. It revealed that LLMs exhibit their own stylistic identities: texts generated by different LLMs on the basis of identical prompts and abstracts formed distinct clusters. Using anthropocentric terminology, one would say that each

LLM possesses its own specific knowledge representation ("image of the world") and communicates through its own "idiolect". Texts generated by certain LLMs (e.g. GPT-4.0 and GPT-4.0-mini) cluster closely together and appear to have been trained on similar datasets, while others are positioned further apart.

Table 2 is also noteworthy. It shows that texts written by each human author are more similar to each other than texts written on the same topics by AI. LLMs seem to have some inherent constraints of their "individuality", understood as the ability to emulate the distinctive stylistic features of specific human authors. These constraints apparently impose limits to any nuanced stylistic imitation. LLMs thus appear to exhibit a form of dominant collective identity and a sort of collective authorial fingerprint. By contrast, human authors display greater stylistic coherence within samples of one author, while differing more markedly between authors.

A very interesting question is the division of machine-generated texts into two groups of equal size, as shown in Figures 2 and 3. Clearly explaining this phenomenon was not straightforward. The analysis of the data contained in both clusters convinced us that this division cannot be explained by semantic criteria (such as gender, year of publication, genre, European versus African origin, etc.). The explanation turned out to be entirely different. As noted above, some of the generated texts were too short, so LLMs were asked to produce new narratives of the required length. The division into two groups thus indicates a clustering of texts completed after the initial prompt and those generated in response to a second prompt. This secondary "urgency" prompt was in fact radically different from the first, as it incorporated the entire context — not only the original abstract but also the text already generated. This leads to an important methodological lesson. While we treated the "expanding" prompts as primary instructions, the LLMs interpreted them as composite prompts, taking into account as an intermediate product the previously generated text. As a result, we inadvertently violated the epistemological principle of conducting experimental research under identical conditions, known as *ceteris paribus* principle.

n	k-NN accuracy [%]				
	authorship				
	hum	cyber	all	hum	ai
5	82.72	45.04	9.95	21.10	9.95
10	91.35	62.39	16.65	32.26	14.65
20	94.57	74.07	27.48	56.99	21.44
50	96.52	81.12	32.93	66.26	25.93
100	96.85	81.77	37.08	73.39	29.43
200	96.37	81.55	41.31	76.61	33.70
500	95.27	78.87	50.19	80.91	43.47
1000	93.95	67.02	71.49	87.37	67.80
2000	90.60	55.71	80.47	88.58	78.13
3000	89.77	48.96	83.90	89.38	83.13
4000	89.45	45.59	86.25	89.52	85.19
5000	88.82	42.06	86.82	88.98	86.02

Table 1: k-NN classification accuracy (k = 5) for different numbers of the most frequent words (n) in three tasks: hum – human vs. LLM texts (2 classes), cyber – human vs. individual LLMs (6 classes), and authorship – author identification (22 classes). For authorship, results are reported for all texts (all), human-only (hum), and LLM-generated texts (ai).

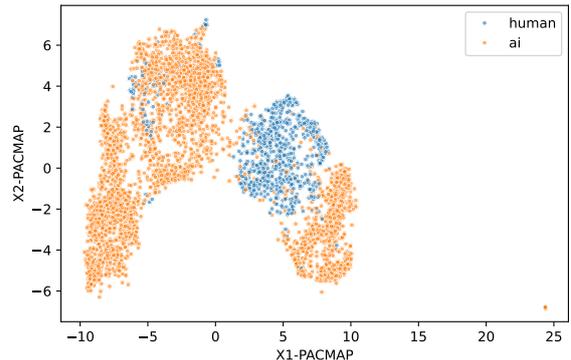


Figure 2: PaCMAP visualization based on TF-IDF vectors of the 100 most frequent words, illustrating a strong and consistent separation between human-authored texts and AI-generated outputs.

model	number of features			
	500	1000	3000	5000
c4ai-comm-a	0.425	0.486	0.598	0.659
c4ai-comm-r	0.402	0.465	0.590	0.654
deepseek-v3	0.486	0.549	0.679	0.748
gpt-4o	0.421	0.494	0.637	0.706
gpt-4o-mini	0.384	0.451	0.568	0.625
human	0.366	0.419	0.516	0.568

Table 2: Average distance between texts and its nearest neighbor from the same author and source (one of LLMs or human), evaluated for varying numbers of the most frequent words. Human-authored texts consistently show the lowest mean distances.

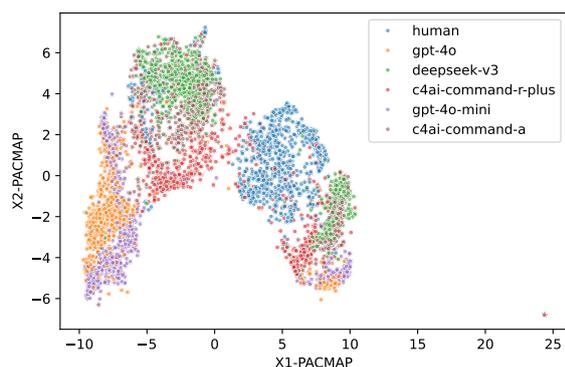


Figure 3: PaCMAP visualization using TF-IDF vectors of the 100 most frequent words, comparing multiple language models with human-authored texts. Outputs from each model separate into two clusters, while human texts remain grouped in one cohesive cluster.

6 Conclusions

The study provides a reliable extension of existing research on the relationship between human- and machine-generated texts in the field of literary fiction, particularly prose. The analysis was carried out on thematically and stylistically related human and machine-authored texts, representing narrative fiction. The study demonstrated that machine-generated texts, produced by various models, exhibit a sort of "collective character". Despite the use of elaborate prompts, the texts that mimicked different authors were more similar to each other than the original human-authored texts. In the case of LLMs, this appears as though a single mind were attempting to imitate multiple voices and writing styles. Unlike AI models, humans exhibit a more distinctive authorial voice and personality. The research also revealed that LLMs possess their own "identities," which include specific representations of knowledge and characteristic modes of conveying it, which could be called *idiolects*. In general, these findings suggest that despite the inherent limitations of the human species, authors of literary fiction remain more original and creative than LLMs.

Limitations

Literature, in its various forms of prose, poetry, and drama, is not intended for machines but for humans. Its purpose is to provide an esthetic and intellectual experience that leads to what has been termed *catharsis* since Aristotle. True humanists therefore argue that the ultimate measure of a literary work's value is human judgment, grounded in sen-

sitivity, knowledge, and culture. However, methods developed within natural language processing (NLP) are not compatible with readers' reception because algorithms do not "read" texts through human eyes. Consequently, an effective assessment of the similarity between machine-generated and human-authored texts should indeed be undertaken by humans. However, humans have a sluggish rate of processing textual information, despite the complexity and perceptiveness of the human mind. Therefore, it is difficult to imagine an efficient, technical evaluation of hundreds or thousands of literary samples in terms of quality, "humanness," "machineliness," or similarity to other texts when such an evaluation is carried out solely by human assessors. Consequently, only NLP tools can guarantee reliable and reproducible results when researching large volumes of human and machine-generated texts.

There are also other potential limitations to our inferences. Dimensionality reduction, which lies at the core of the methods we employ, is akin to a long journey – and the journey itself often transforms the traveler. It begins with data represented in, say, a 1,000-dimensional space and ends with a two-dimensional visualization. Therefore, dimensionality reduction entails an inevitable loss of information and a simplification of reality. This is further compounded by the profound subjectivity inherent in the creative literary process and its evaluation. Language is not a material substance, nor is it a collection of elementary particles to which methods validated in the study of inorganic matter can be straightforwardly applied. The meaning of words and sentences shifts depending on the context and communicative situation. Consequently, the results presented here do not provide strong evidence to confirm or refute the earlier formulated hypotheses. Rather, they serve as signposts or milestones that indicate the main orientation points within a new hybrid machine-human communicative space.

Appendix

In this study, we processed a corpus consisting of text samples by the following authors: Louis Aragon, Mariama Bâ, Georges Bernanos, Calixthe Beyla, Tanella Boni, Ken Bugul, Louis-Ferdinand Céline, Albert Cohen, Fatou Diome, Romain Gary (Émile Ajar), Jean Giono, Jean-Marie Gustave Le Clézio, Andreï Makine, Scholastique Mukasonga, Paul Pavlovitch, Georges Perec, Raymond Que-

neau, Aminata Sow Fall, Michel Tournier, Boris Vian, Marguerite Yourcenar.

Acknowledgements

The work was financed by CLARIN-PL: Common Language Resources and Technology Infrastructure (POIR.04.02.00-00C002/19, 2024/WK/01, FENG.02.04-IP.040004/24).

References

- Maciej Eder, Jan Rybicki, and Mike Kestemont. 2015. [Stylometry with r: A package for computational text analysis](#). *The R Journal*, 8:107–121.
- Patrick Juola. 2006. [Authorship attribution](#). *Found. Trends Inf. Retr.*, 1(3):233–334.
- Mike Kestemont. 2014. [Function words in authorship attribution. from black magic to theory?](#) In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66, Gothenburg, Sweden. Association for Computational Linguistics.
- Frederick Mosteller and David Wallace. 1964. *Inference and disputed authorship : The Federalist*. Addison-Wesley series in behavioral science : quantitative methods. Addison-Wesley, Reading Mass. [etc.
- James O’Sullivan. 2025. [Stylometric comparisons of human versus ai-generated creative writing](#). *Humanities and Social Sciences Communications*, 12.
- Adam Pawłowski, Ewa Kalinowska, and Tomasz Walkowiak. 2025. [Au-delà de l’attribution d’auteur : la stylométrie permet-elle d’identifier l’« identité collective » des textes littéraires et le sexe des auteurs ? analyse comparée de la fiction narrative d’europe et d’afrique](#). *Romanica Cracoviensia*, 25:203.
- Karol Przystalski, Jan K. Argasiński, Iwona Grabska-Gradzińska, and Jeremi K. Ochab. 2026. [Stylometry recognizes human and llm-generated texts in short samples](#). *Expert Systems with Applications*, 296:129001.
- Simon Pöpcke, Thomas Weitin, Katharina Herget, Anastasia Glawion, and Ulrik Brandes. 2022. [Stylometric similarity in literary corpora: Non-authorship clustering and Deutscher Novellenschatz](#). *Digital Scholarship in the Humanities*. Fqac039.
- Irina Rabaev, Marina Litvak, Vladimir Younkin, Ricardo Campos, Alípio Mário Jorge, and Adam Jatowt. 2023. [The competition on automatic classification of literary epochs](#). In *Proceedings of the IACT - The 1st International Workshop on Implicit Author Characterization from Texts for Search and Retrieval held in conjunction with the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023), Taipei, Taiwan, July 27, 2023*, volume 3477 of *CEUR Workshop Proceedings*, pages 49–56. CEUR-WS.org.
- Buckley C. Salton G. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.
- Ismail Kassou Sara El Manar El Bouanani. 2014. [Authorship analysis studies: A survey](#). *International Journal of Computer Applications*, 86(12):22–29.
- Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, 60(3):538–556.
- Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. 2021. [Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization](#). *Journal of Machine Learning Research*, 22(201):1–73.

Degree Zero of Translation: Using Interlinear Baselines to Quantify Translator Intervention

Maciej Rapacz
AGH University of Kraków
mrapacz@agh.edu.pl

Aleksander Smywiński-Pohl
AGH University of Kraków
apohllo@agh.edu.pl

Abstract

Literary translation is rarely a neutral act of linguistic transfer, but rather a continuous series of conscious interventions—restructuring, semantic shifts, and stylistic adaptations. While Translation Studies analyzes these shifts qualitatively, current computational methods focus primarily on quality evaluation (e.g., BLEU, COMET) or authorship attribution (e.g., stylometry), lacking a scalable metric to quantify the extent and character of the translator’s intervention. We propose a novel method to measure the translator’s signal by using Interlinear Translation – a strict word-for-word gloss – as a computational baseline representing translational “Degree Zero”, i.e. a neutral form of source text devoid of any stylistic adaptation. We define the Intervention Vector as the semantic difference between a literary translation and its interlinear counterpart in a high-dimensional vector space. We validate this approach on a multilingual corpus of the Greek New Testament translations comprising 5 interlinear baselines and 74 literary translations across 5 languages – English (16), French (14), Italian (12), Polish (16), and Spanish (16). Our results demonstrate that the magnitude of the Intervention Vector effectively ranks texts along a spectrum from literal to paraphrase, aligning with established theoretical categories. We find that this magnitude consistently distinguishes between translation strategies, yielding significantly longer vectors for dynamic and paraphrase strategies compared to literal and formal ones. This framework provides a quantitative method for analyzing translator agency without the need for a comprehensive corpus of reference translations.

1 Introduction

Translation, especially in a literary context, is rarely a neutral act of linguistic transfer, but rather an act of rewriting (Lefevere, 2016) or mediation (Munday, 2013). The translator makes a continuous series of conscious decisions to adapt the source text

to a new linguistic and cultural context, a specific audience, or a particular stylistic purpose. We propose a method to quantify these decisions, which manifest themselves in the collective set of syntactic restructuring, semantic shifts, and stylistic adaptations that distinguish a literary text from a mechanical, word-for-word gloss. Although this intervention is an inherent feature of human translation, quantifying it has lacked a reference-free, scalable method.

Current computational approaches do not capture this phenomenon. Natural Language Processing (NLP) has historically focused on Machine Translation (MT), where the objective is to optimize the output for correctness and fluency. Consequently, evaluation metrics – from n-gram matching such as BLEU (Papineni et al., 2002) to reference-free neural estimators such as COMET-QE (Rei et al., 2020) – are designed to correlate with human quality judgments (Fomicheva and Specia, 2016). These tools operate on an evaluative axis: they assign a scalar score to determine how “good” a translation is.

In contrast, Translation Studies (TS) originates from the analysis of human translation, where divergence from the source is viewed not as error but as a deliberate *strategy*. Theories such as those from the Manipulation School posit that “all translation is manipulation of the source text for a certain purpose” (Hermans, 1985). However, the field lacks the computational methodology to characterize this manipulation. Traditional approaches in Computational Stylometry, such as Burrows’ Delta (Burrows, 2002), are primarily used for *attribution*—identifying lexical “fingerprints” to determine *who* translated a text. Yet attribution metrics are inherently relative; they indicate that *Translation A* is statistically similar to *Translation B*, but they do not measure the shift from the source text. They capture lexical tendencies but miss the broader semantic and syntactic restructuring per-

formed by the translator.

We propose using the *Interlinear Translation* as a computational baseline. Unlike standard literary translations, an interlinear version strictly preserves the order of words in the source language, placing the target-language words directly under the corresponding source words (Shuttleworth and Cowie, 2014). Drawing on Barthes (1953)’s concept of Writing Degree Zero, a neutral mode of writing, we hypothesize that this extremely literal form of translation represents a neutral representation of the source text – or as close to it as possible – completing the necessary switch from one language to another while minimizing adaptation for style, audience, or fluency. By calculating the difference between vector representations of a literary translation under study and this baseline, we isolate the translator’s signal. We term this difference the *Intervention Vector*, and we posit that it captures both the character and the extent of the translator’s intervention.

We validate this method on a multilingual corpus of New Testament translations in five languages (English, French, Italian, Polish, and Spanish). We selected this domain because interlinear translations are primarily available as study resources for ancient and religious texts and because the rich history of translation of the New Testament offers a diverse range of strategies, which we use to test the sensitivity of the metric.

Our contributions¹ are as follows:

1. **Dataset and Validation Framework:** We construct a multilingual parallel corpus aligning literary New Testament translations with interlinear baselines across five languages. We use this dataset to establish a testbed for evaluating quantitative measures of translator intervention against theoretical categories.
2. **Spectrum Recovery:** We demonstrate that the magnitude of the Intervention Vector recovers the theoretical translation spectrum without supervision, effectively sorting translation strategies from literal to paraphrase.
3. **Pairwise Sensitivity:** We validate the metric’s granularity by showing that it consistently distinguishes intervention levels even in pair-wise comparisons of individual transla-

tions, proving it is sensitive enough to capture differences beyond broad strategy labels.

2 Related Work

Machine Translation Evaluation. Standard Machine Translation (MT) evaluation is based on reference-based metrics that penalize divergence. Neural metrics like COMET (Rei et al., 2020) improve on BLEU (Papineni et al., 2002) but still rely on some sort of reference, which might inadvertently introduce reference bias, penalizing valid creative variations Fomicheva and Specia (2016). Recent work has attempted to mitigate this by using multiple references (Wu et al., 2025), referenceless metrics or multi-agent LLM evaluators (Kim et al., 2025), but these approaches still stay within the evaluative paradigm, aiming to score the translation’s quality, rather than quantify the translator’s intervention.

Computational Stylometry. Stylometric research typically focuses on attribution and clustering. The most common method used here is Burrows’ Delta (Burrows, 2002), which distinguishes authors through frequent-word distributions. There have been multiple attempts to identify translators in the past, with varying success. It was found that, especially when dealing with translations of multiple works, clustering algorithms are more likely to detect the author than the translator (Rybicki, 2012). Identifying multiple translators within a single work resulted in greater success (Rybicki and Heydel, 2013). N-gram-based methods together with gradient boosting have been found to have been successful in translator attribution, even in multi-work, multi-translator scenarios (Mohamed et al., 2023). However, none of these measures quantifies the translator’s intervention. They rely either on relative distance (as with Burrows’ Delta) – which merely indicates that Translation A is statistically similar to Translation B, rather than defining its intrinsic character – or on surface-level features like n-grams. While the latter can describe stylistic tendencies within the target language, none of these methods capture the degree of departure from the source text, nor do they quantify the specific semantic shifts introduced by the translator.

Translationese and Corpus Linguistics. Corpus-based Translation Studies identify universal features of translated text, such as simplification and explicitation (Baker, 1993). Research in this domain typically quantifies “translationese” – the

¹The code and data are available at <https://github.com/mrapacz/sighum-interlinear-vector-baselines>

linguistic fingerprints distinguishing translated text from native writing – through surface-level features like type-token ratio, sentence length, or part-of-speech density (Volansky et al., 2015). While translationese is often viewed pejoratively as a mark of lower quality or lack of fluency (Wein and Schneider, 2024), we argue that in literary translation, the translator’s signal stems from deliberate strategy rather than linguistic interference. Consequently, our definition of intervention is broader than translationese: it encompasses valid, creative restructuring in the target language that would not be captured by metrics designed solely to detect non-native artifacts.

Vector Semantics in Digital Humanities. The use of embeddings to model semantic change is well-established in diachronic linguistics. Researchers use distributional models to measure semantic drift over time (Kutuzov et al., 2018; Hamilton et al., 2016). In literary studies, recent work has employed BERT-based embeddings to analyze narrative structures (Wegmann et al., 2022) or compared BERT and ELMo embeddings in the study of diachronic shifts (Kutuzov et al., 2018). We rely on vector representations to study semantic shift; here the shift is not diachronic but the distance between a full translation and the interlinear baseline.

Translator Intervention in a Quantitative Setting To our knowledge, our work is the first to propose a computational operationalization of translator intervention.

3 Methodology

Interlinear translation, often cited as the “archetype or ideal of all translation” (Benjamin, 1923), maps target-language words directly onto the source syntax without reordering for fluency (Shuttleworth and Cowie, 2014). To establish a baseline of non-intervention, we draw from Roland Barthes’ concept of Writing Degree Zero (Barthes, 1953), which describes a “colorless” and neutral style of writing that strips away all artistic flair to report reality as directly as possible. By analogy, we propose that interlinear translation is a “colorless” mode of translational practice that strips away all artistic flair to report the “reality” of the source text as directly as possible. Although no translation is entirely free of interpretation, even in extreme forms like the interlinear, we hypothesize that this state is as close to it as possible: a state where the compul-

sory language shift has occurred, but the specific adaptations for fluency, style, or target audience have not yet been applied.

3.1 Dataset Curation

The data consist of two text types: the translations under study and interlinear translations, which act as the point of reference.

The Interlinear Baseline. We scraped interlinear translations from five language-specific repositories: *BibleHub* (English), *Editeur BPC* (French), *Altevista* (Italian), *Oblubienica* (Polish) and *Bibliatodo* (Spanish). These texts preserve the syntactic order of the Ancient Greek source while mapping words to their literal target-language equivalents. We remove verse numbers, but retain original punctuation and capitalization.

The Literary Corpus. We utilized the *targum* corpus (Rapacz and Smywiński-Pohl, 2026), a multilingual collection containing 657 New Testament translations across our five target languages. To interpret the results effectively, we hand-picked a subset of 74 translations – English (16), French (14), Italian (12), Polish (16), and Spanish (16) – to have representatives in each of four orientational categories:

1. **Literal:** Strict word-for-word translation (e.g., Literal Standard Version).
2. **Formal:** Translations that adhere closely to source structure but permit necessary grammatical adjustments for fluency (e.g. New Revised Standard Version).
3. **Dynamic:** Translations that prioritize functional meaning, with substantial structural changes (e.g. New Living Translation).
4. **Paraphrase:** Texts with the highest allowance for adaptation, occasionally bordering on distinct literary works (e.g. The Message).

We emphasize that these labels are purely orientational and serve only to mark the varying degrees of allowance for adaptation; they do not imply any judgment on theological validity or fidelity.

Furthermore, we restricted ourselves to translations from 1900 onwards. We deliberately excluded earlier translations (e.g. from the 17th or 18th centuries) so as to compare translations synchronically and to avoid confounding effects from diachronic

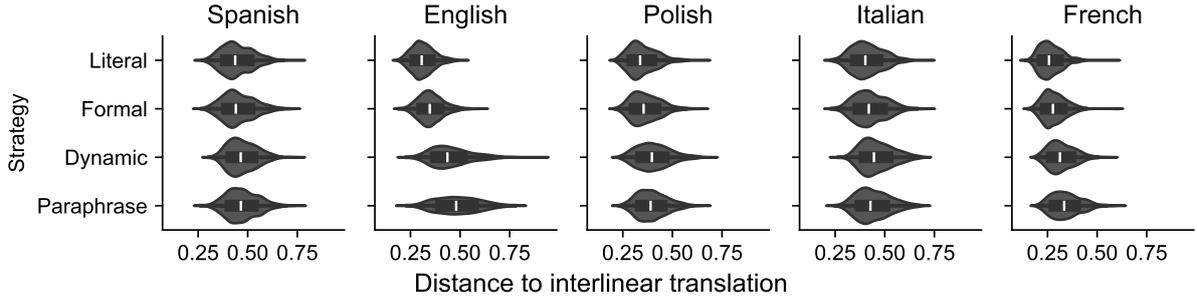


Figure 1: Violin plot showing distribution of intervention vector lengths for each chapter across languages and strategies.

language change. The complete list of translations is available in Appendix A.

3.2 Vectorization

We aggregate the text at the *Chapter Level* (260 units per translation). While verse-level segmentation offers finer granularity, it introduces alignment challenges (e.g., merged verses like “2-6a”) and noise from short verses (e.g., “Jesus wept”). Pericope-level segmentation, while semantically coherent, lacks a standardized schema across translations. We generate embeddings using the Qwen3-Embedding-8B model². We choose this model for its strong multilingual performance and the 32k token context window (Zhang et al., 2025), which fits full chapters of the New Testament without the need to segment or truncate the text.

3.3 The Intervention Vector

We define the *Intervention Vector* (V_{int}) as the difference between the vector representation of a literary translation (V_{trans}) and the Interlinear Baseline (V_{base}) for the same chapter:

$$V_{int} = V_{trans} - V_{base} \quad (1)$$

Using this vector, we can derive two straightforward metrics: magnitude (euclidean norm) and stability (standard deviation). Further, we also projected the vectors onto a plane using Principal Component Analysis (PCA) per language to visualize a simplified picture of the intervention characteristics.

4 Results

The Spectrum of Intervention. First, we evaluate whether the magnitude of the intervention

vector ($\|V_{int}\|$) accurately reflects the theoretical degree of translator intervention.

Figure 1 presents the distribution of intervention vector lengths. The violins show that more literal strategies cluster closer to the origin while dynamic and paraphrase versions extend further to the right.

We test whether chapter-level distances differ between strategy groups using Mann–Whitney U tests with Bonferroni correction (see Appendix B for full group-level and pairwise comparisons). In English and French the four strategies form a strict order: Literal < Formal < Dynamic < Paraphrase (all $p < 0.001$). In Italian and Polish, Literal and Formal lie closer to the baseline than Dynamic and Paraphrase (cross-group comparisons $p < 0.001$, or Formal < Paraphrase at $p < 0.01$); we do not find significant differences between Literal and Formal, or between Dynamic and Paraphrase. In Spanish, Literal is closer than all other groups ($p < 0.001$) and Formal closer than Dynamic ($p < 0.05$). At the level of individual translations, one of the two is significantly closer to the interlinear baseline in 90% of English pairs, 69.2% in French, 55.8% in Polish, 48.3% in Spanish, and 51.5% in Italian (Mann–Whitney with Bonferroni, $p < 0.001$).

The Range of Intervention. Median chapter-level distance to the interlinear baseline varies by language and the ranges differ (e.g. English 0.31–0.48, French 0.258 – 0.334, Polish 0.34–0.39, Italian 0.40–0.45, Spanish 0.44–0.66). That these ranges differ may reflect typological differences between each target language and Ancient Greek: where the language forces more structural adaptation, translations may consistently lie further from the interlinear baseline. We further check the relation between the mean and the standard deviation of the intervention vector and find that in the case of English and French there is a clear rela-

²<https://huggingface.co/Qwen/Qwen3-Embedding-8B>

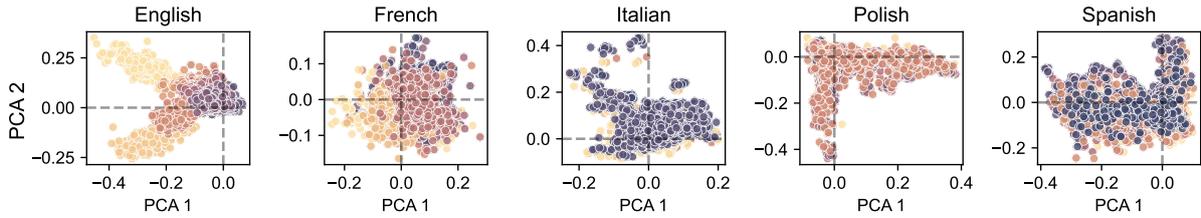


Figure 2: PCA projection of intervention vectors by language. PCA was calculated independently per language, so axes are not comparable. Dots represent chapters, colored by the translation’s mean intervention magnitude (brighter colors indicate higher magnitude).

tionship, 0.83 and 0.88, respectively. For the other languages we find a weak relationship for Polish (0.428), weak negative for Spanish (-0.511) and no relationship for Italian (0.068), though we cannot distinguish whether these values stem from typological differences between the languages or from better representation in English and French.

The Topology of Intervention. Finally, we investigate the geometric shape of the intervention space. While magnitude measures *how much* a translator intervenes, it does not indicate *how*. To assess whether the intervention vectors contain some characteristics of the translator’s strategy, we computed a PCA on the difference vectors (V_{int}) (independently for each language). We used PCA to further reduce dimensionality and to project it onto a 2D plane, as seen in Figure 2. We shift the coordinates so that the interlinear baseline is at the origin (0,0). In the figure, color intensity indicates mean distance from the interlinear baseline (brighter colors represent greater intervention); since PCA was performed separately for each language, the color scale is language-specific and not comparable across panels.

We observe distinct patterns in the intervention vectors for each language. For Polish, we find a clear, orthogonal v-shape. English³ and French, while lacking such a clear pattern, place translations with overall higher intervention further away from the origin. For Italian and Spanish, the patterns are more complex, and PCA did not identify anything that would align with the magnitude of the intervention.

5 Conclusion

We propose a novel framework for quantifying the translator intervention. By computing the intervention vector, the vector difference between the trans-

lation under study and its interlinear counterpart, we form a representation that captures characteristics of the translator’s intervention.

We demonstrate that the magnitude of the intervention vector can be used to rank translations according to their position on the translation spectrum, from literal to paraphrase. We also show that our method consistently yields longer vectors for higher-intervention strategies, both when comparing (literal, formal) vs. (dynamic, paraphrase) groups and when comparing individual translations; for the latter, we found a statistically significant difference in chapter-level distances in roughly half of the pairs (and in 90% of English pairs).

Finally, we show that the patterns in the intervention vectors differ from language to language. We observe different ranges of intervention vector lengths and different patterns when projecting them onto a plane using PCA. These patterns may arise from typological differences, representation quality, or sample composition.

Future work will extend this preliminary study in three directions. First, we would like to study how the metric behaves depending on parameters such as model choice (and model size) as well as the granularity of the text (chapter, verse, pericope). Second, in this work, we explicitly excluded the diachronic axis to isolate the method from drift; it would be useful to test how our method treats archaic texts. Third, we plan to validate this metric on distinct literary corpora beyond the New Testament to see how well these findings generalize to other domains.

Although we tested the method on a corpus of translations, the metric itself can be applied to a single translation. The concept of the interlinear baseline could also be applied in other domains, e.g. to isolate prosody from texts or musical interpretation from a very strict midi file.

³Erratum: English PCA results were not interpreted in the original submission due to an algorithmic error.

Limitations

First, we focused our analysis on the New Testament, because for this corpus the interlinear translations were readily available on-line. However, biblical texts are heavily represented in LLM training corpora, often within specific theological contexts. We do not know how this method performs on modern, secular literary works, which might exhibit different embedding behaviors. Future work is needed to determine if the stability we observed generalizes to domains less saturated in the model's training data.

Second, we treat the interlinear translation as a neutral reference point. In reality, rendering Ancient Greek into the target language requires selecting specific words, which is already an act of interpretation. Our results might be skewed by this fact; a translation could appear distant from the baseline not because the translator's intervention was substantial, but simply because their lexical choices differed from those of the interlinear author.

Third, we did not account for differences in the underlying source text. The New Testament exists in multiple critical editions (e.g., *Textus Receptus* vs. *Nestle-Aland*). If a translation follows a different critical edition than our interlinear baseline, the resulting vector distance captures textual variance—such as added or deleted verses—rather than the translator's intervention. Future work would need to align the critical edition of the baseline with that of the target text to better isolate the translator's style.

Fourth, for the purpose of this study, we labeled translations with broad categories such as Literal, Dynamic, and Paraphrase. We acknowledge that this is a coarse taxonomy intended only to approximate the general extent of adaptation, not the nuance of its character. These labels serve purely as auxiliary benchmarks to verify whether our methodology can distinguish fundamental approaches—separating literal strategies from dynamic ones—and do not constitute a qualitative judgment. For instance, designating a text as a “paraphrase” describes the magnitude of its structural divergence from the source, not its legitimacy as a translation.

Fifth, we relied on data scraped from third-party websites. These sources contained formatting errors, which we fixed manually wherever our validation logic spotted them, and our own processing

pipeline may have introduced others. Consequently, our metric might interpret these data artifacts as a translator intervention, even when such deviations are not present in the actual published text.

Acknowledgments

The research was supported by the National Science Centre, Poland under the project number 2025/57/N/HS2/04961. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2026/019145. The research presented in this article was partially supported by the funds of the Polish Ministry of Science and Higher Education assigned to the AGH University of Kraków.

References

- Mona Baker. 1993. Corpus Linguistics and Translation Studies: Implications and Applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In Honour of John Sinclair*, pages 233–250. John Benjamins.
- Roland Barthes. 1953. *Writing Degree Zero*. Hill and Wang, New York.
- Walter Benjamin. 1923. The Task of the Translator. In Lawrence Venuti, editor, *The Translation Studies Reader*, pages 15–25. Routledge, London.
- John Burrows. 2002. ‘Delta’: a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3):267–287.
- Marina Fomicheva and Lucia Specia. 2016. Reference Bias in Monolingual Machine Translation Evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 77–82, Berlin, Germany. Association for Computational Linguistics.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.
- Theo Hermans, editor. 1985. *The Manipulation of Literature: Studies in Literary Translation*. Croom Helm, London.
- Junghwan Kim, Kieun Park, Sohee Park, Hyunggug Kim, and Bongwon Suh. 2025. MAS-LitEval : Multi-Agent System for Literary Translation Quality Assessment. *arXiv preprint*. ArXiv:2506.14199 [cs] version: 1.

- Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. [Diachronic word embeddings and semantic shifts: a survey](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- André Lefevere. 2016. *Translation, rewriting, and the manipulation of literary fame*. Routledge.
- Emad Mohamed, Raheem Sarwar, and Sayed Mostafa. 2023. [Translator attribution for Arabic using machine learning](#). *Digital Scholarship in the Humanities*, 38(2):658–666.
- Jeremy Munday. 2013. *Style and Ideology in Translation: Latin American Writing in English*. Routledge, New York.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a Method for Automatic Evaluation of Machine Translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maciej Rapacz and Aleksander Smywiński-Pohl. 2026. [Targum – A Multilingual New Testament Translation Corpus](#). *arXiv preprint*. ArXiv:2602.09724 [cs].
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A Neural Framework for MT Evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Jan Rybicki. 2012. [The great mystery of the \(almost\) invisible translator: Stylometry in translation](#). In Michael P. Oakes and Meng Ji, editors, *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*, pages 231–248. John Benjamins Publishing Company.
- Jan Rybicki and Magda Heydel. 2013. [The stylistics and stylometry of collaborative translation: Woolf’s Night and Day in Polish](#). *Literary and Linguistic Computing*, 28(4):708–717.
- M. Shuttleworth and M. Cowie. 2014. *Dictionary of translation studies*. St. Jerome Publishing.
- Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. [On the features of translationese](#). *Digital Scholarship in the Humanities*, 30(1):98–118.
- Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. [Same Author or Just Same Topic? Towards Content-Independent Style Representations](#). In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics. TODO.
- Shira Wein and Nathan Schneider. 2024. [Lost in Translationese? Reducing Translation Effect Using Abstract Meaning Representation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 753–765, St. Julian’s, Malta. Association for Computational Linguistics.
- Si Wu, John Wieting, and David A. Smith. 2025. [Multiple References with Meaningful Variations Improve Literary Machine Translation](#). *arXiv preprint*. ArXiv:2412.18707 [cs].
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models](#). *arXiv preprint*. ArXiv:2506.05176 [cs].

A Complete list of translations used in the experiments

This appendix presents the complete list of translations used in the experiments across all five languages. For each translation, we provide its name, abbreviation, strategy classification (Literal, Formal, Dynamic, or Paraphrase), and year of publication. The abbreviations are used to label translations in the pairwise comparison figures in Appendix B.

Table 1: List of English translations included in the experiments.

Translation	Abbr.	Strategy	Year
Disciples’ Literal New Testament	DLNT	Literal	2011
Berean Literal Bible	BSB	Literal	2016
Literal Standard Version	LSV	Literal	2020
New American Standard Bible (2020)	NASB20	Formal	2020
NRSV Updated Edition	NRSVue	Formal	2021
English Standard Version	ESV	Formal	2025
New World Translation	NWT	Formal	2025
Bible in Basic English	BBE	Dynamic	1965
Complete Jewish Bible	CJB	Dynamic	1998
New Living Translation	NLT	Dynamic	2015
Good News Translation	GNT	Dynamic	2017
First Nations Version	FNV	Dynamic	2021
EasyEnglish Bible	EASY	Dynamic	2024
The New Testament in Modern English	Phillips	Paraphrase	1958
The Message	MSG	Paraphrase	2018
Orthodox Jewish Bible	OJB	Paraphrase	2023

Table 2: List of French translations included in the experiments.

Translation	Abbr.	Strategy	Year
Bible Bovet-Bonnet	BBO	Literal	1900
Louis Segond (1910)	LSG	Formal	1910
Bible Catholique Crampon	BCC	Formal	1923
Nouvelle Édition de Genève	NEG79	Formal	1979
Bible d’Ostervald	OST	Formal	1996
Bible King James Française	KJF	Formal	2006
Segond 21	S21	Formal	2007
La Bible des Peuples	BdP	Dynamic	1994
Bible en français courant	BFC	Dynamic	1997
La Bible du Semeur	BDS	Dynamic	2015
Parole Vivante	PVV	Paraphrase	2013
Parole de Vie	PDV	Paraphrase	2017
Traduction du monde nouveau	TMN	Paraphrase	2025

Table 3: List of Italian translations included in the experiments.

Translation	Abbr.	Strategy	Year
Riveduta (Luzzi)	RIV	Literal	1925
Nuova Riveduta	NR06	Literal	2006
Nuova Riveduta 2020	NR20	Literal	2020
Riveduta (1927)	RIV27	Literal	2020
Sacra Bibbia (Tintori)	TIN	Formal	1931
Sacra Bibbia (Ricciotti)	RIC	Formal	1940
Bibbia CEI (1974)	CEI74	Formal	1974
La Nuova Diodati	LND	Formal	1991
La Parola è Vita	PEV	Dynamic	2006
Parola del Signore (TILC)	TILC	Dynamic	2014
La Bibbia della Gioia	BDG	Paraphrase	2005
Traduzione del Nuovo Mondo	TNM	Paraphrase	2025

Table 4: List of Polish translations included in the experiments.

Translation	Abbr.	Strategy	Year
Nowy Testament (Szczepański)	SZCZ	Literal	1917
Przekład Dąbrowskiego	DAŁ	Literal	1961
Nowa Biblia Gdańska	NBG	Literal	2012
EIB Przekład Dosłowny	EIB-PD	Literal	2019
Biblia Jakuba Wujka	BJW	Formal	1923
Biblia Poznańska	BP	Formal	1975
Biblia Warszawska	BW	Formal	1975
Biblia Tysiąclecia (V wyd.)	BT5	Formal	2000
EIB Przekład Literacki	EIB	Formal	2016
Biblia Króla Jakuba	BKJ	Formal	2017
Uwspółcześniona Biblia Gdańska	UBG	Formal	2017
Biblia Warszawsko-Praska	BWP	Dynamic	1997
Nowy Przekład Dynamiczny	NPD	Dynamic	2021
Przekład Mariawicki	MAR	Paraphrase	1921
Słowo Życia	SŻ	Paraphrase	2016
Przekład Nowego Świata	PNŚ	Paraphrase	2025

Table 5: List of Spanish translations included in the experiments.

Translation	Abbr.	Strategy	Year
Versión Moderna (Pratt)	VM	Literal	1929
Nueva Biblia de las Américas	NBLA	Literal	1986
La Biblia de las Américas	LBLA	Literal	1997
Reina-Valera 1960	RVR60	Formal	1977
Biblia del Jubileo	JBS	Formal	2000
La Biblia de América	BAME	Dynamic	2010

Continued on next page

Table 5: List of Spanish translations included in the experiments.

Translation	Abbr.	Strategy	Year
La Palabra (España)	BLP	Dynamic	2010
Nueva Traducción Viviente	NTV	Dynamic	2010
Nueva Versión Internacional	NVI	Dynamic	2022
Dios Habla Hoy	DHH	Dynamic	2023
Biblia Jünemann (Septuaginta)	JUN	Paraphrase	1992
Biblia Latinoamericana	BL95	Paraphrase	1995
Traducción en Lenguaje Actual	TLA	Paraphrase	2002
Palabra de Dios para Todos	PDT	Paraphrase	2015
Biblia Peshitta en Español	PES	Paraphrase	2017
Traducción del Nuevo Mundo	TNM	Paraphrase	2025

B Pairwise Comparisons

This appendix presents pairwise Mann–Whitney U tests (one-sided, Bonferroni-corrected) on chapter-level distances to the interlinear baseline. Each heatmap cell indicates whether the row’s chapter-level distances are significantly smaller than the column’s: *** indicates $p < 0.001$, ** indicates $p < 0.01$, * indicates $p < 0.05$, and blank cells are not significant.

Figure 3 shows comparisons at the strategy-group level. Figures 4–8 show comparisons between individual translations, ordered by strategy category.

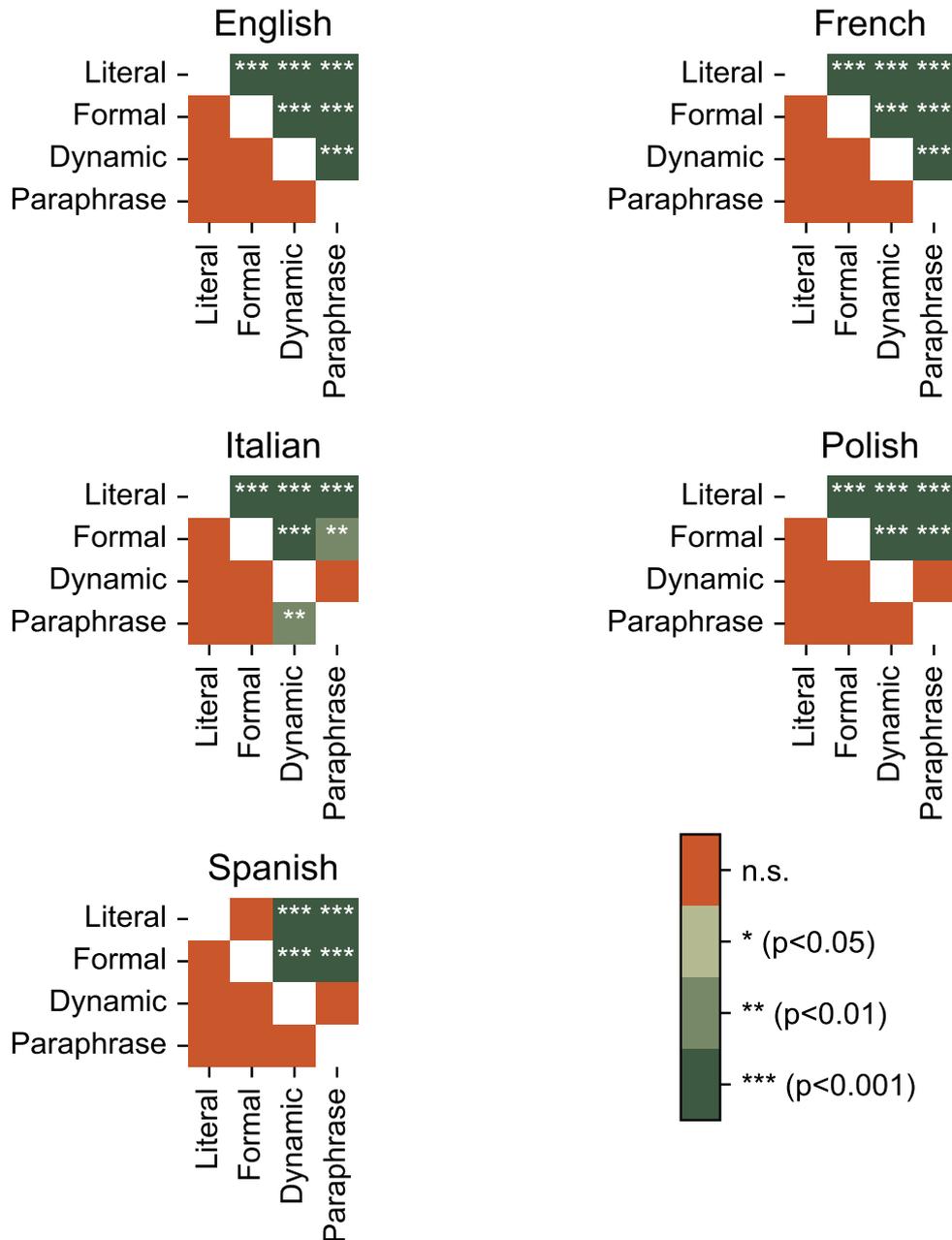


Figure 3: Strategy-group comparisons.

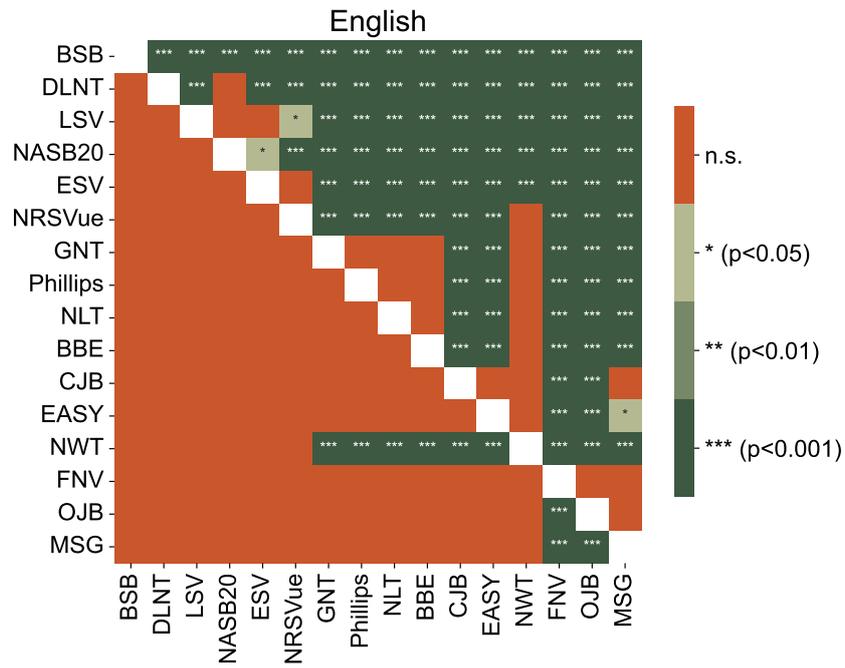


Figure 4: Pairwise comparisons of chapter intervention vector magnitudes for English translations.

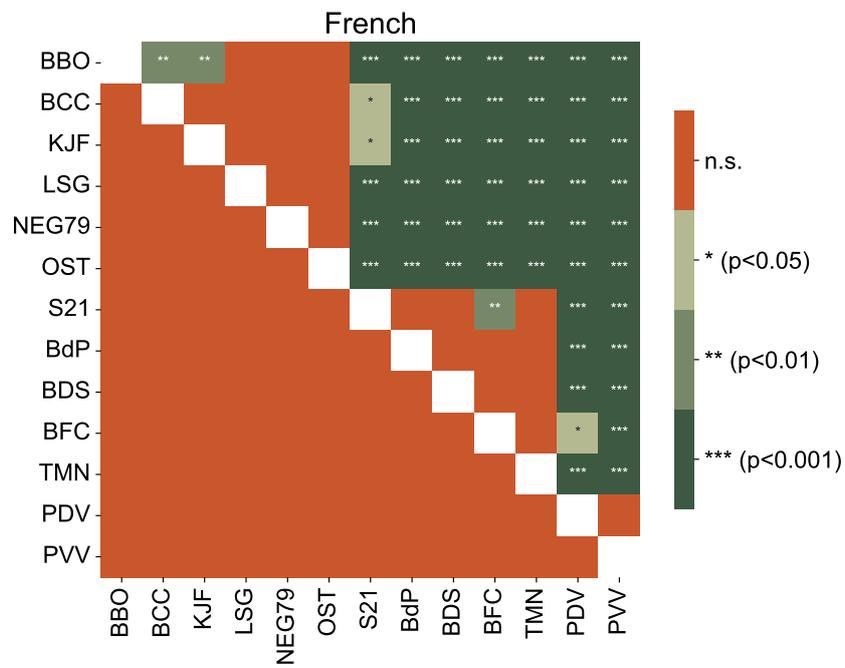


Figure 5: Pairwise comparisons of chapter intervention vector magnitudes for French translations.

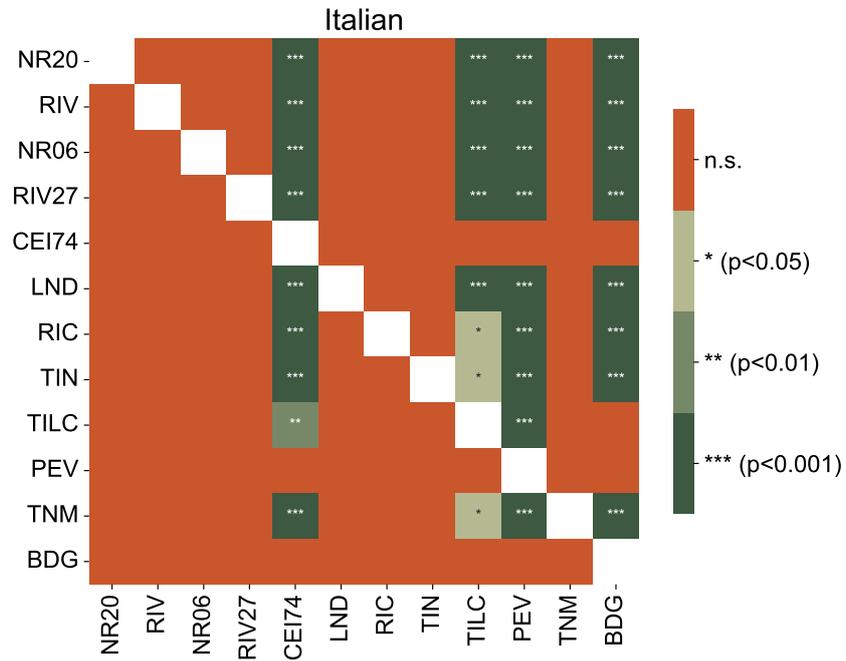


Figure 6: Pairwise comparisons of chapter intervention vector magnitudes for Italian translations.

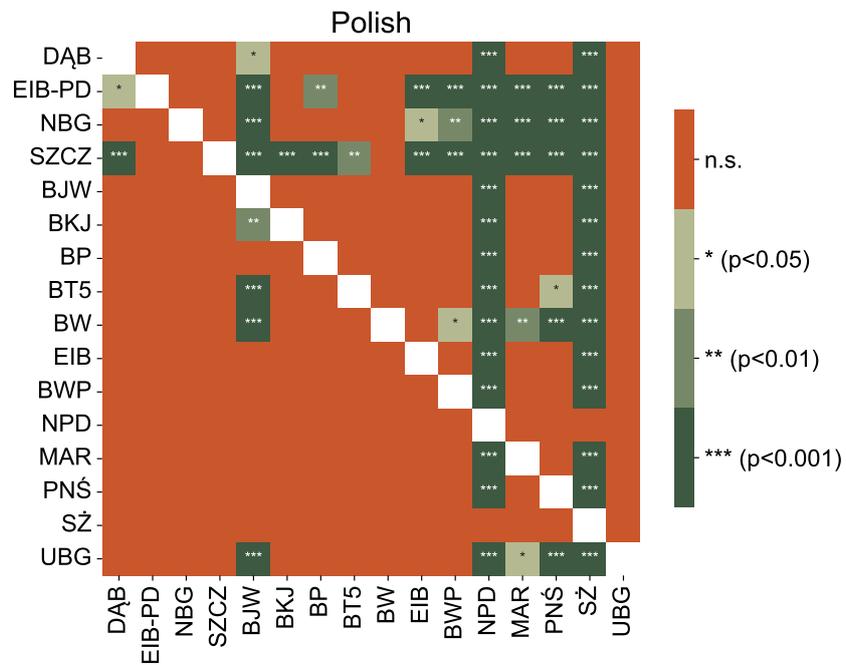


Figure 7: Pairwise comparisons of chapter intervention vector magnitudes for Polish translations.

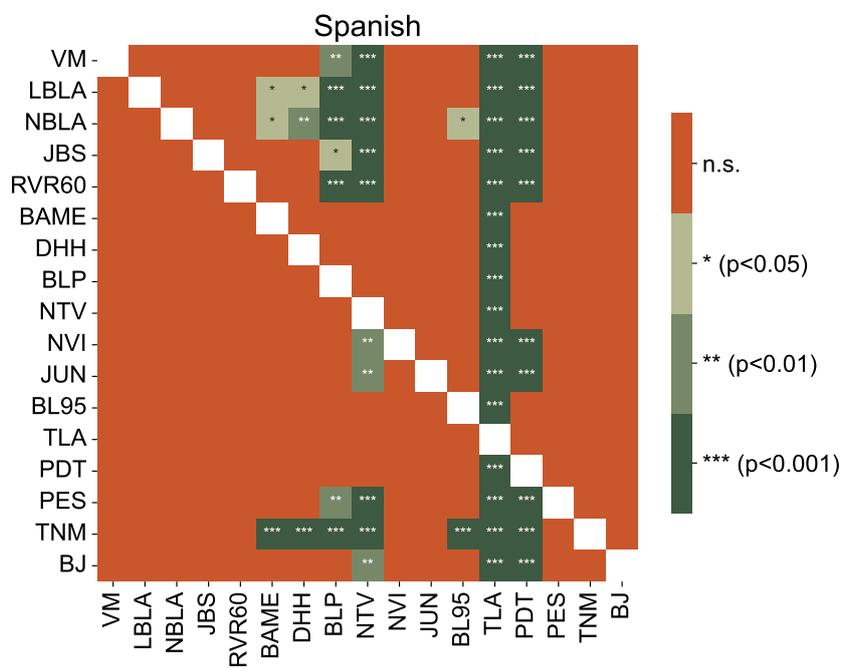


Figure 8: Pairwise comparisons of chapter intervention vector magnitudes for Spanish translations.

How to Efficiently Explore Noisy Historical Data? Leveraging Corpus Pre-Targeting to Enhance Graph-based RAG

Donghan Bian^{1,2} Marie Puren^{2,1} Florian Cafiero¹

¹Centre Jean-Mabillon, École nationale des chartes - PSL, Paris, France

²EPITA Research Laboratory, EPITA, Paris, France

{donghan.bian, florian.cafiero}@chartes.psl.eu, marie.puren@epita.fr

Abstract

Graph-based Retrieval-Augmented Generation (RAG) is increasingly used to explore long, heterogeneous, and weakly structured corpora, including historical archives. However, in such settings, naive full-corpus indexing is often computationally costly and sensitive to OCR noise, document redundancy, and topical dispersion. In this paper, we investigate corpus pre-targeting strategies as an intermediate layer to improve the efficiency and effectiveness of graph-based RAG for historical research.

We evaluate a set of pre-targeting heuristics tailored to single-hop and multi-hop of historical questions on HistoriQA-ThirdRepublic, a French question-answering dataset derived from parliamentary debates and contemporary newspapers. Our results show that appropriate pre-targeting strategies can improve retrieval recall by 3–5% while reducing token consumption by 32–37% compared to full-corpus indexing, without degrading coverage of relevant documents.

Beyond performance gains, this work highlights the importance of corpus-level optimization for applying RAG to large-scale historical collections, and provides practical insights for adapting graph-based RAG pipelines to the specific constraints of digitized archives.

1 Introduction

Retrieval-Augmented Generation (RAG) has become a central paradigm for grounding large language models in external textual evidence, enabling more reliable generation in knowledge-intensive settings (Lewis et al., 2020; Gao et al., 2023; Poibeau, 2025). Recent work has shown that RAG can scale to long documents and complex queries, particularly when combined with graph-based representations that support multi-hop reasoning and global semantic structure (Larson and Truitt, 2024; Xiang et al., 2025). As a result, RAG-based systems are increasingly considered for exploratory

access to large textual collections, including historical corpora. Beyond simple full-text search—often the default option for historians with limited digital experience—, RAG supports a more systematic, large-scale exploration of very large (and especially serial) sources. It also enables researchers to “dialogue” with their corpus, surfacing leads, aspects, and themes that were previously unknown, or less visible to domain experts, and that can subsequently be investigated through conventional methods of historical research.

At the same time, this promise comes with substantial practical challenges in archival settings. Historical archives constitute a particularly demanding testbed for such approaches. These documents are typically long, weakly structured, heterogeneous in genre, and are often affected by OCR noise (Piotrowski, 2012; Strange et al., 2014), a recurrent issue in DH corpora that often diverge from NLP benchmark assumptions (McGillivray et al., 2020). They also can exhibit strong redundancy and uneven information density, with large portions of text that are procedurally repetitive or only marginally relevant to a given research question. In this context, naive full-corpus indexing is not only computationally expensive, but can also amplify noise and degrade retrieval quality, especially in graph-based pipelines where indexing decisions directly affect graph construction costs.

While most existing work on RAG focuses on retrieval models, ranking strategies, or generation architectures (Gao et al., 2023), comparatively little attention has been paid to corpus-level decisions that precede retrieval, such as how much of a corpus should be indexed and under which constraints. This issue is particularly salient for historical research, where recent studies have shown both the potential and the limitations of applying RAG to digitized newspapers and parliamentary debates (Tran et al., 2024; Pellet et al., 2024). For large

historical collections, full-corpus indexing may be neither necessary nor desirable.

In this paper, we investigate corpus pre-targeting as an intermediate layer between raw historical archives and graph-based RAG pipelines. Pre-targeting consists in selectively reducing the indexed corpus based on query characteristics, with the goal of preserving relevant documents while filtering out material that is unlikely to contribute to answering a given class of questions. Building on recent proposals for pre-targeted RAG (Silvestre de Sacy et al., 2024), we hypothesize that different types of historical questions impose different retrieval constraints, and tailored pre-targeting strategies can improve the trade-off between retrieval effectiveness and computational cost.

We evaluate this hypothesis on HistoriQA-ThirdRepublic, a French-language question answering dataset derived from parliamentary debates and contemporary newspapers (Pellet et al., 2026). We distinguish between single-hop questions, which typically rely on localized factual information, and multi-hop questions, which require aggregating evidence across multiple documents. Our contributions are threefold: (1) a systematic evaluation of corpus pre-targeting strategies for graph-based RAG applied to historical archives; (2) empirical evidence that appropriate pre-targeting improves recall by 3–5% while reducing token consumption by 32–37%; and (3) practical insights for adapting RAG pipelines to the specific constraints of digitized historical collections.¹

2 Related Work

Our work connects three research threads: (i) institutional and scholarly needs for source-grounded exploration of large historical archives, (ii) emerging Digital Humanities uses of RAG on noisy, heterogeneous collections, and (iii) graph-based RAG pipelines and corpus-reduction strategies that trade indexing cost for retrieval effectiveness.

Archive exploration, transparency, and accountability. Beyond academic use, access to institutional archives is increasingly framed in terms of transparency and accountability: computational methods can enable traceable, source-grounded claims over large documentary collections, provided that provenance and citation granularity remain explicit (Jo and Gebru, 2020; Cafiero, 2023;

Cafiero et al., 2025). In that spirit, recent institutional initiatives experiment with question-answering interfaces over archival holdings, aiming to lower the barrier to exploration for non-specialists while preserving the ability to inspect underlying sources. For example, the National Library of Luxembourg has deployed a chatbot for exploring historical Luxembourgish newspapers (Bibliothèque nationale du Luxembourg, 2023), while the European Parliament Historical Archives provides an “Ask the EP archives” tool within a broader content-analysis environment for multilingual parliamentary material (European Parliament Historical Archives, 2024b,a).

RAG for historical research in Digital Humanities. RAG was initially introduced to ground generation on retrieved evidence for knowledge-intensive tasks (Lewis et al., 2020), and later work systematized a broad design space of retrieval, reranking, and generation modules (Gao et al., 2023). In Digital Humanities, RAG has been explored as a pragmatic approach to query long, heterogeneous, and OCR-noisy corpora where traditional search and close reading alone do not scale. As a result, RAG is increasingly being evaluated for examining large collections of historical documents (Nandula and Shenoy, 2024; Fan et al., 2025; Lee et al., 2025). Representative case studies include, in particular, historical newspapers (Tran et al., 2024) and reflections on RAG for parliamentary debates, with explicit discussion of methodological benefits (rapid exploration, cross-source synthesis) and risks (retrieval failures, over-trusting generated summaries) (Pellet et al., 2024). A recurring conclusion in this literature is that corpus characteristics—redundancy, uneven information density, OCR artifacts, and document structure—often dominate downstream performance, sometimes more than the choice of generator model.

Graph-based RAG and corpus reduction via pre-targeting. Graph-augmented RAG methods aim to improve evidence aggregation and multi-hop retrieval by exploiting explicit structure (entities/relations, neighborhoods, traversals) rather than relying solely on local semantic similarity. Survey work clarifies the main families of GraphRAG-like pipelines—graph construction, graph-guided retrieval, and graph-enhanced generation—and discusses when graphs help compared to simpler RAG stacks (Peng et al., 2025; Zhang et al., 2025a; Xiang et al., 2025). Concrete peer-reviewed

¹Scripts available at [Github](#)

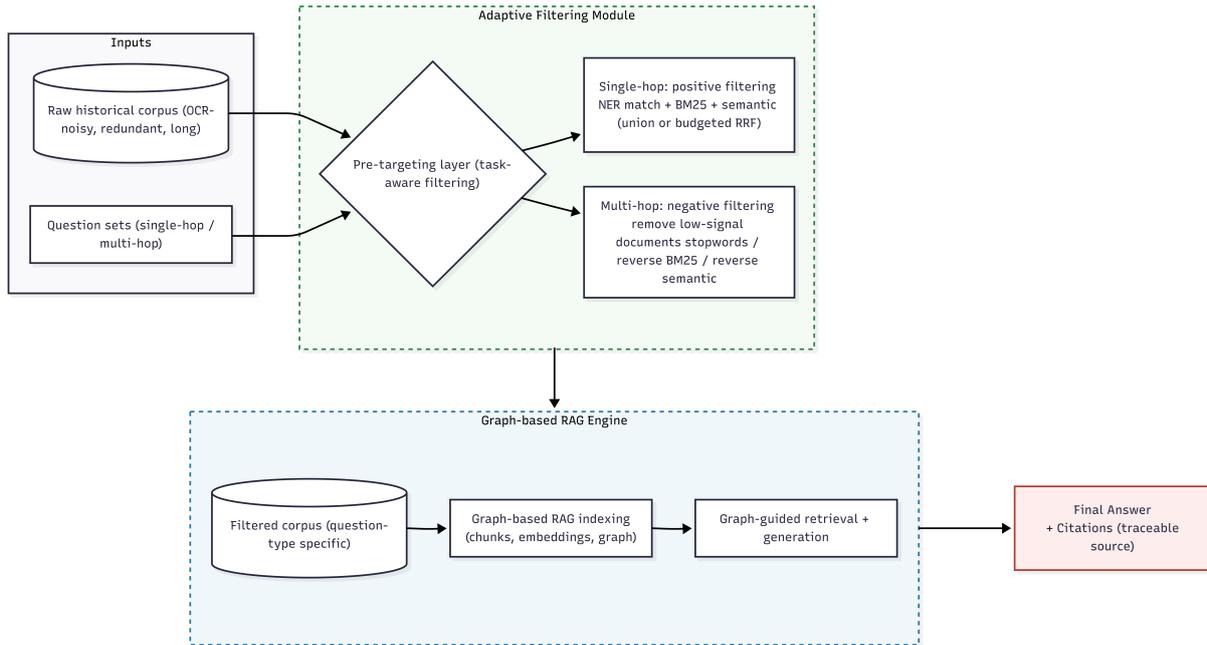


Figure 1: Overview of the pipeline. A task-aware pre-targeting layer reduces the indexed corpus differently for single-hop (positive filtering; union or budgeted Reciprocal Rank Fusion) and multi-hop questions (negative filtering). The filtered corpus feeds a shared Graph-based RAG indexing and QA stack, improving the recall–cost trade-off.

instantiations include HippoRAG, which combines LLM-driven graph induction with graph traversal mechanisms to improve multi-hop retrieval (Gutiérrez et al., 2024), and TRACE, which constructs knowledge-grounded reasoning chains over triples extracted from retrieved documents to integrate dispersed evidence (Fang et al., 2024).

However, graph construction can become the dominant cost on large archives, and indexing everything can amplify OCR noise and procedural redundancy. This motivates corpus reduction strategies that act before indexing, by filtering the corpus under query- or task-specific constraints while attempting to preserve gold evidence. Pre-targeted RAG formalizes this idea as a pre-filtering layer that reduces the indexed set prior to retrieval and generation (Silvestre de Sacy et al., 2024). We extend this line of work in a historical setting by (i) tailoring pre-targeting to question type (single-hop vs. multi-hop) and (ii) evaluating the resulting trade-offs specifically within a graph-based RAG pipeline on OCR-noisy parliamentary debates and newspapers.

3 Methodology

In this section, we outline our methodology. We begin by introducing the HistoriQA-ThirdRepublic dataset, and then present the strategies employed in

the pre-targeting process along with the evaluation protocol used to assess their effectiveness.

3.1 HistoriQA-ThirdRepublic

HistoriQA-ThirdRepublic is a French-language question answering corpus derived from parliamentary debates and newspapers of the French Third Republic (1870-1940) (Pellet et al., 2026). The dataset contains historical questions designed to evaluate RAG systems in domain-specific contexts, specifically focusing on complex reasoning patterns typical of historical inquiry.

The dataset is constructed from three primary sources, focusing specifically on texts from 1887: parliamentary debates from the Chamber of Deputies, and two contemporary newspapers, *Le Gaulois* and *L’Intransigeant*. For the parliamentary debate texts, content not directly related to debates, such as agendas, procedural notes, and vote lists, was removed in advance. After applying a specific segmentation strategy, the final text corpus comprises 78 documents from *Le Gaulois*, 79 documents from *L’Intransigeant*, and 3,227 documents from parliamentary debates, as shown in figure 2.

Questions were then generated through iterative refinement using prompt engineering and validated by a historian to ensure factual and contextual coherence. Single-hop questions were generated ex-



Parliamentary debate (Chamber of Deputies, 7 March 1887)



INTERPELLATION SUR LA SITUATION DE LA CORSE M. le président. L'ordre du jour appelle la discussion de l'interpellation de M. Cunéo d'Ornano sur le rôle de la magistrature en Corse et sur la situation actuelle de ce département. La parole est à M. Cuneo d'Ornano pour développer son interpellation. M. Cuneo d'Ornano. Messieurs, j'éprouve une très vive préoccupation en montant à cette tribune [...]

OCR text (recognition rate: 99.38%)



Newspaper (L'intransigeant, 9 March 1887)



L'interpellation de M. Cunéo d'Ornano sur le rôle de la magistrature en Corsica, nous rend, un peu de cette précieuse gaieté que la discussion de la loi sur les céréales semblait avoir tout jamais bannie. [...]

OCR text (recognition rate: 74.90%)

Figure 2: Corpus sources and OCR quality. The OCR texts are generated from scanned historical documents, and the recognition accuracy information is obtained from the corresponding Gallica pages.

clusively based on the content of parliamentary debates, whereas multi-hop questions were constructed following two strategies, newspaper-to-debate and newspaper-to-newspaper, resulting in three types of questions: follow-up questions (identifying opinions in debates and examining press reactions), comparative questions (contrasting differing viewpoints across sources), and bridge-entity questions (connecting sources through shared references).

To clarify the task distinction, we provide representative examples from HistoriQA-ThirdRepublic.

- **Single-hop example.** A question answerable from one debate document (localized factual evidence): *Quel est le rôle du bureau d'âge et comment est-il composé lors de l'ouverture de la session ordinaire de la Chambre des députés en 1887 ?* (What was the role of the Bureau of Age and how was it composed at the opening of the ordinary session of the Chamber of Deputies in 1887?)
- **Multi-hop example.** A question requiring evidence aggregation across at least two documents/sources (e.g., debate + newspaper, or

newspaper + newspaper): *En 1887, comment le baron de Mackau critique-t-il la politique éducative du gouvernement, et quelle analyse l'article de presse en donne-t-il ?* (In 1887, how did Baron de Mackau criticize the government's educational policy, and what analysis did the newspaper article provide?)

These examples reflect the operational criterion used in our experiments: single-hop targets local evidence, whereas multi-hop requires cross-document synthesis.

The final QA dataset comprises 897 single-hop questions and 885 multi-hop questions, the indexing data (document ID) required to answer the questions is also integrated into the QA dataset.

The corpus contains OCR noise as shown in Figure 2. We adopted a fast and deliberately conservative heuristic to estimate the word error rate (WER), treating as errors words that exceed 20 characters in length, contain interleaved digits and letters, or include non-French characters. This approach yields an estimated WER of approximately 2.73%, which should be regarded as a lower bound, as the OCR error rates reported by Gallica, the digital library

of the Bibliothèque nationale de France, for these sources are generally higher.²

3.2 Pre-targeting strategies regarding different type of questions

Overall, our approach is inspired by the methodology proposed by Silvestre de Sacy et al. (2024). The core objective is to define constraints for pre-filtering the corpus based on the characteristics of the query under investigation, thereby reducing the amount of text to be indexed as much as possible while preserving the target documents (gold documents) required to answer the questions. This strategy aims to lower the cost of graph construction and indexing time, partially mitigate the impact of OCR noise, and ultimately improve the performance of RAG systems during the question-answering stage.

With respect to question types, existing RAG approaches generally perform better on single-hop questions than on multi-hop questions (Larson and Truitt, 2024; Gutiérrez et al., 2025). This discrepancy stems from the fundamentally different nature of the two question types. Single-hop questions can be answered using information contained within a single document and are typically tied to concrete facts. In most cases, answers can be directly extracted from the original text. In contrast, multi-hop questions require aggregating and synthesizing information across multiple related documents, which places higher demands on the robustness of the retrieval stage as well as the long-context processing capabilities of large language models.

These observations directly motivate our experimental design. We evaluate the effectiveness of pre-targeting separately for single-hop and multi-hop questions in order to identify the most suitable strategies for each case. For each question type, we randomly sample ten sets of fifty questions from the HistoriQA-ThirdRepublic QA dataset without replacement and report the average performance of different pre-targeting strategies across the ten sets.

For single-hop questions, we assume that precise semantic matching is sufficient to rapidly localize the documents required to answer a given question. Accordingly, we adopt three complementary strategies. First, we apply GLiNER (Zaratiana et al., 2024), an advanced named entity recognition model, to detect four types of entities, including person, location, date, and organization, in the

query, and subsequently search for these entities in the text corpus to retrieve the corresponding documents. Second, after tokenizing the query, we perform a bag-of-words retrieval over the entire corpus using BM25 (Robertson et al., 2009) and retain the top 500 ranked documents. Third, we segment the corpus into chunks of 1,000 tokens with an overlap of 200 tokens, compute the cosine similarity between the query and each chunk using the Qwen3-Embedding-0.6B model (Zhang et al., 2025b), and treat the document containing a chunk as a candidate document whenever the similarity exceeds 0.7. We further evaluate the union of the results produced by these three strategies to assess whether their combination yields additional improvements in retrieval performance.

In contrast to single-hop questions, multi-hop questions exhibit weaker direct semantic alignment with individual documents. Consequently, the objective of pre-targeting shifts from directly locating relevant texts to filtering out documents with low information density or those potentially unrelated to the question-answering task. We again employ three strategies. First, we define a stop word list and consider a document to be low in informational content if stop words account for more than 20% of its text, in which case the document is filtered out.³ Second, we rank all documents using BM25 for each question and exclude any document that appears among the bottom 500 documents for at least one question. Finally, we compute cosine similarity using the same embedding model, but with a different criterion: if a document’s similarity score with respect to all questions does not exceed 0.6, it is considered as unrelated documents. These strategies yield three lists of excluded documents, and we further explore the union of these lists to identify the most effective combination of filtering strategies.

The thresholds and cutoffs used in pre-targeting (e.g., cosine thresholds, BM25 rank cutoffs, stop-word ratio) are operational hyperparameters selected for this corpus and task configuration, not universal constants. Our objective is to evaluate whether task-aware corpus reduction can improve the recall–cost trade-off under realistic archival constraints, rather than to claim corpus-independent optimal values. In practice, these parameters should be re-calibrated when corpus characteristics

²OCR error information is available on the document’s Gallica page, here is an [example](#).

³The stop-word list is available in our [GitHub repository](#) (block 19).

change (OCR quality, document length distribution, genre balance, or language), and we therefore interpret the reported gains as evidence of the approach’s usefulness in this setting, with external transfer requiring additional validation.

3.3 Retrieval performance

Based on the selection of the optimal pre-targeting strategies, we choose three filtered corpora and compare their performance against full-corpus indexing on the corresponding question sets. We adopt HippoRAG 2, one of the state-of-the-art RAG frameworks (Gutiérrez et al., 2025; Xiang et al., 2025), for retrieval and generation. For embedding and generation models, we use Qwen3-Embedding-0.6B and DeepSeek-V3.2 (Liu et al., 2025), respectively. To support the former, we implement additional scripts to enable compatibility between HippoRAG 2 and the Qwen 3 model architecture, while the latter is accessed via an API. Finally, we report both the recall performance of the different configurations and their token consumption. Figure 3 illustrates an example of a knowledge graph generated under full-corpus indexing.

4 Results

4.1 Pre-targeting stage results

We first evaluate the performance of different filtering strategies in the pre-targeting stage. Table 1 summarizes the results using three key metrics: recall percentage, which measures the proportion of gold documents retained relative to the total set of gold documents; precision percentage, representing the proportion of gold documents in filtered corpus; document count, indicating the number of documents remaining after filtering; and document coverage, defined as the percentage of documents preserved from the original corpus.

For single-hop questions, individual filtering strategies showed varying performance characteristics. BM25 Filtering achieved the highest recall at 97.00% while maintaining relatively low document coverage (34.93%, 1182.8 documents). In contrast, NER Filtering demonstrated 84.00% recall but required substantially more documents (1891.7 documents, 55.87% coverage), and Cosine Similarity Filtering achieved only 62.80% recall with the smallest document set (120.5 documents, 3.56% coverage).

When combining strategies, BM25 + Cosine Similarity (Union) emerged as the optimal approach, achieving 97.20% recall with 1185.4 documents (35.01% coverage). The motivation for adopting the union of the results is to account for the complementary information captured by bag-of-words models and dense vector representations. In addition, we apply Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) to combine these two lists of results by weighting documents according to their positions in each list, while constraining the size of the fused list to the longer of the two inputs to prevent an excessive expansion of the candidate set, thereby assessing their relative importance under the two ranking schemes. We observe that RRF attains the same recall as the union strategy, suggesting that it successfully fuses the complementary features captured by the two retrieval strategies in an efficient manner. Although the union of all three strategies (NER + BM25 + Cosine) reached the highest recall at 98.80%, it came at the cost of substantially increased document coverage (64.75%), making it less efficient for practical deployment.

For multi-hop questions, which require reasoning across multiple documents, reverse filtering strategies proved efficient. Among individual strategies, Reverse Cosine Similarity Filtering achieved the best balance with 98.64% recall using 1828.6 documents (54.00% coverage), outperforming both Reverse BM25 Filtering (98.15% recall, 61.52% coverage) and Reverse Stopwords Filtering (84.95% recall, 72.15% coverage). Combined strategies further improved performance, with Reverse BM25 + Cosine (Union) reaching 99.39% recall at 74.96% coverage. The most comprehensive approach, combining all three reverse strategies, achieved near-perfect recall at 99.52% but required processing 98.74% of the document collection (3343.3 documents), offering minimal practical advantage over simpler combinations.

Based on these results, we selected BM25 + Cosine Similarity (RRF) for single-hop questions and Reverse Cosine Similarity Filtering for multi-hop questions as our optimal strategies for the subsequent RAG pipeline (Figure 1), as they provide the best balance between high recall and computational efficiency.

These results should be read as corpus- and setup-specific: they show that simple task-aware heuristics can be effective on HistoriQA-ThirdRepublic, but they do not imply that the same thresholds will transfer unchanged to other histori-

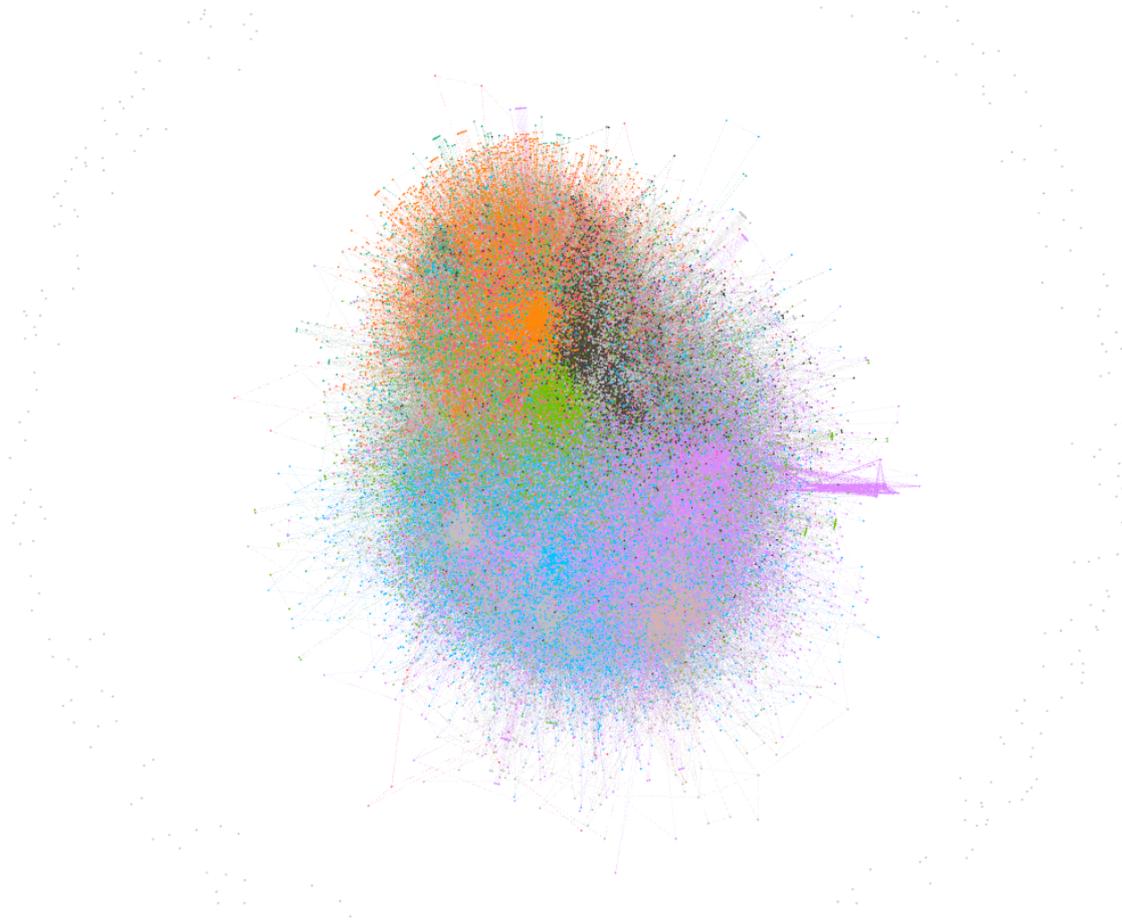


Figure 3: Knowledge graph generated from the HistoriQA-ThirdRepublic dataset under full-corpus indexing. The network comprises 30,514 nodes and 648,792 edges. We employed the Louvain algorithm (Blondel et al., 2008) for community detection, identifying 332 distinct communities with a modularity of 0.619, indicating a robust community structure. For visualization, the nodes are color-coded by their community membership, and the network layout is optimized using the Yifan Hu multilevel force-directed algorithm (Hu, 2005) to enhance interpretability.

cal collections. The main transferable claim is the principle of differentiated pre-targeting for single-hop versus multi-hop questions.

4.2 RAG performance

We then compared the performance of default full indexing against the pre-targeting approach across different retrieval depths (top-3, top-5, and top-10) for both question types. Table 2 presents the recall rates and token consumption for each strategy, with numbers in brackets representing the total tokens used across all three experimental runs.

For single-hop questions, the pre-targeting strategy demonstrated consistent improvements over full indexing across all retrieval depths. Pre-targeting achieved average recall rates of 55.33% at top-3, 62.00% at top-5, and 70.67% at top-10, representing improvements of 4.66, 2.67, and 2.00 percentage points respectively over full indexing (50.67%, 59.33%, and 68.67%). These

improvements were achieved while using substantially fewer tokens on average: 12,156,845.7 tokens compared to 19,434,237 for full indexing, representing a 37.5% reduction in computational cost. The standard deviations across runs (± 6.80 , ± 7.12 , and ± 6.60 for top-3, top-5, and top-10 respectively) indicate reasonable stability in performance.

The advantages of pre-targeting were more pronounced for multi-hop questions. Pre-targeting achieved recall rates of 43.00%, 52.67%, and 62.67% at top-3, top-5, and top-10 respectively, outperforming full indexing by 5.00, 4.34, and 2.00 percentage points. The performance gains were accompanied by a 32.1% reduction in average token usage (13,194,585.3 tokens versus 19,434,237 for full indexing). The lower standard deviations for multi-hop questions (± 5.66 , ± 4.71 , and ± 2.87) suggest more stable performance compared to single-hop questions, particularly at higher retrieval depths where the standard deviation

Strategy	Recall (%)	Precision (%)	Doc Count	Doc Coverage (%)
Single-Hop Questions				
<i>Individual Strategies:</i>				
NER Filtering	84.00	2.28	1891.7	55.87
BM25 Filtering	97.00	4.10	1182.8	34.93
Cosine Similarity Filtering	62.80	26.69	120.5	3.56
<i>Combined Strategies:</i>				
BM25 + Cosine Similarity (Union)	97.20	4.10	1185.4	35.01
BM25 + Cosine Similarity (RRF)	97.20	4.11	1182.8	34.93
NER + BM25 + Cosine (Union)	98.80	2.29	2192.3	64.75
Multi-Hop Questions				
<i>Individual Strategies:</i>				
Reverse Stopwords Filtering	84.95	2.79	2443.0	72.15
Reverse BM25 Filtering	98.15	3.79	2083.1	61.52
Reverse Cosine Similarity Filtering	98.64	4.35	1828.6	54.00
<i>Combined Strategies:</i>				
Reverse BM25 + Cosine (Union)	99.39	3.14	2538.3	74.96
Reverse Stopwords + BM25 + Cosine (Union)	99.52	2.39	3343.3	98.74

Table 1: Average Performance of Individual and Combined Filtering Strategies (10 sets). The bold rows indicate our selected optimal strategies for the subsequent RAG pipeline. The low precision is a result of sparsity of gold documents relative to the scale of the corpus.

Strategy	Recall@3 (%)	Recall@5 (%)	Recall@10 (%)	Token Consumption (total)
<i>Single-Hop</i>				
Full Indexing	50.67	59.33	68.67	19,434,237
Pre-targeting (std.)	55.33 (± 6.80)	62.00 (± 7.12)	70.67 (± 6.60)	12,156,845.7 (36,470,537)
<i>Multi-Hop</i>				
Full Indexing	38.00	48.33	60.67	19,434,237
Pre-targeting (std.)	43.00 (± 5.66)	52.67 (± 4.71)	62.67 (± 2.87)	13,194,585.3 (39,583,756)

Table 2: Comparison of Full Indexing and Pre-targeting Performance. The information of token consumption is obtained from the official API platform.

dropped to ± 2.87 . Additionally, it is noteworthy that despite the substantial disparity in the total number of documents between the single-hop and multi-hop text corpora, their token consumption remains relatively close. This phenomenon can be attributed to the significant variation in token counts across texts from different sources in the original dataset (Pellet et al., 2026).

The pre-targeting approach demonstrated substantial computational efficiency gains while improving retrieval performance across both question types. For single-hop questions, the strategy reduced token usage by 37.5% while consistently improving recall rates. For multi-hop questions, despite requiring more tokens than single-hop pre-targeting due to the inherent complexity of multi-hop reasoning and also the different filtering strategies, pre-targeting still achieved a 32.1% reduction in computational cost compared to full indexing. These results indicate that pre-targeting not only enhances retrieval quality but also provides mean-

ingful computational benefits, making it a more practical solution for large-scale RAG applications.

5 Discussion

Our results demonstrate that corpus pre-targeting is a viable strategy for enhancing graph-based RAG, achieving 3–5% higher recall while reducing token consumption by 32–37% in our experimental setting. This suggests that optimizing the document preprocessing pipeline can complement ongoing improvements in model architectures and retrieval algorithms, particularly for domain-specific applications where computational resources are limited. It should be noted, however, that practical deployment requires corpus-specific calibration (e.g., threshold tuning and rank cutoffs), since the heuristics evaluated here are not expected to generalize unchanged across collections with different OCR quality, textual genres, or language distributions.

Our approach assumes effective document identification through lexical-semantic similarity, which may not generalize to all query types. The upfront filtering cost (12-13M tokens) remains substantial for single queries, though this amortizes across multiple queries on the same corpus. Additionally, our evaluation on a single historical dataset requires validation across diverse domains and languages.

Promising directions include developing lightweight query routers for automatic strategy selection, adaptive filtering with dynamic thresholds, iterative pre-targeting to recover initially missed documents, and extending to multi-lingual scenarios. Incorporating OCR noise evaluation and the impact of the level of noise on performance could help refine our approaches depending on the corpus treated.

Acknowledgements

We would like to express our sincere gratitude to Aurélien Pellet for his invaluable support and assistance with the dataset. This work has been supported by PSL’s CultureLab project.

References

Bibliothèque nationale du Luxembourg. 2023. [L’intelligence artificielle au service du patrimoine imprimé luxembourgeois](#). Press release on the launch of the chatbot.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Florian Cafiero. 2023. [Datafying diplomacy: How to enable the computational analysis and support of international negotiations](#). *Journal of Computational Science*, 71:102056.

Florian Cafiero, Jean-Philippe Cointet, and Grégoire Mallard. 2025. [Digital accountability can re-legitimate multilateralism](#). Preprint / working paper, HAL.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.

European Parliament Historical Archives. 2024a. [Archives content analysis dashboard and “ask the ep archives” \(project page\)](#). Project web page (no peer-reviewed publication identified for this citation key).

European Parliament Historical Archives. 2024b. [The ep archives unit launches its first generative ai tool](#). Web page describing the “Ask the EP archives” tool within the Archives Content Analysis Dashboard.

Yang Fan, Zhang Qi, Xing Wenqian, Liu Chang, and Liu Liu. 2025. [Research on graph-retrieval augmented generation based on historical text knowledge graphs](#). Preprint, arXiv:2506.15241.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. 2024. [Trace the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8472–8494. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, and Haofen Wang. 2023. [Retrieval-augmented generation for large language models: A survey](#). *arXiv preprint arXiv:2312.10997*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. [Hipporag: neurobiologically inspired long-term memory for large language models](#). In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS ’24*, Red Hook, NY, USA. Curran Associates Inc.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. [From RAG to memory: Non-parametric continual learning for large language models](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 21497–21515. PMLR.

Yifan Hu. 2005. Efficient, high-quality force-directed graph drawing. *Math J*, 10:37.

Eun Seo Jo and Timnit Gebru. 2020. [Lessons from archives: strategies for collecting sociocultural data in machine learning](#). In *FAT* ’20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 306–316. ACM.

Jonathan Larson and Steven Truitt. 2024. [Graphrag: Unlocking llm discovery on narrative private data](#).

Jeong Ha Lee, Ghazanfar Ali, and Jae-In Hwang. 2025. [A retrieval-augmented generation system for accurate and contextual historical analysis: Ai-agent for the annals of the joseon dynasty](#). *Computer Animation and Virtual Worlds*, 36(4):e70048.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

- Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. [DeepSeek-V3.2: Pushing the frontier of open large language models](#). *arXiv preprint arXiv:2512.02556*.
- Barbara McGillivray, Thierry Poibeau, and Pablo Ruiz. 2020. [Digital Humanities and Natural Language Processing: “Je t’aime... Moi non plus”](#). *Digital Humanities Quarterly*, 14(2).
- Srianusha Nandula and Saachi Shenoy. 2024. [Enhancing historical understanding with retrieval augmented generation](#). Retrieved 2026-01-05.
- Aurélien Pellet, Julien Perez, and Marie Puren. 2024. [Generative artificial intelligence and historical research: Challenges, potentials, and limitations. application of RAG to french parliamentary debates of the third republic \(1881–1940\)](#). In *A Conversation between AI and the Humanities*.
- Aurélien Pellet, Marie Puren, and Julien Perez. 2026. [HistoriQA-ThirdRepublic: Multi-Hop Question Answering Corpus for Historical Research, Parliamentary Debates from the](#). Working paper or preprint.
- Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2025. [Graph retrieval-augmented generation: A survey](#). *ACM Trans. Inf. Syst.*, 44(2).
- Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- Thierry Poibeau. 2025. *Understanding Conversational AI*. Ubiquity Press, London.
- Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Foundations and trends® in information retrieval*, 3(4):333–389.
- Antoine Silvestre de Sacy, Adam Faci, Stéphane Pouyllau, and Léa Maronet. 2024. [Pre-targeted-RAG: Retrieval Augmented Generation sur des groupes préciblés de communautés d’articles de recherche](#).
- Carolyn Strange, Daniel McNamara, Josh Wodak, and Ian Wood. 2014. [Mining for the meanings of a murder: The impact of ocr quality on the use of digitized historical newspapers](#). *Digital Humanities Quarterly*, 8(1).
- The Trung Tran, Carlos-Emiliano González-Gallardo, and Antoine Doucet. 2024. [Retrieval augmented generation for historical newspapers](#). In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, pages 1–5.
- Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jinsong Su. 2025. [When to use graphs in RAG: A comprehensive analysis for graph retrieval-augmented generation](#). *arXiv preprint arXiv:2506.05690*.
- Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. [Gliner: Generalist model for named entity recognition using bidirectional transformer](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376.
- Qinggang Zhang, Chuanjie Wu, Zhishang Xiang, and 1 others. 2025a. [A survey of graph retrieval-augmented generation for customized large language models](#). *arXiv preprint arXiv:2501.13958*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025b. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *arXiv preprint arXiv:2506.05176*.

Detecting reported speech as a token classification task: an application to Classical Latin?

Agustin Dei

Sorbonne Université - Paris, France

EDITTA (Edition, Interprétation et traduction des textes anciens)

CERES (Centre d'expérimentation en méthodes numériques
pour les recherches en Sciences Humaines et Sociales)

agustin.dei@sorbonne-universite.fr

Abstract

This paper presents the first application of an automatic token-classification approach for detecting reported speech spans in Classical Latin using transformer-based neural architectures. Focusing on Seneca the Elder's Declamatory Anthology, the study addresses the text's highly polyphonic nature, resulting from the use of reported speech. Instead of relying exclusively on sentence-level syntactic information, the proposed approach treats reported speech detection as a token-level sequence labeling problem. This enables the identification of reported speech spans extending across multiple sentences. We fine-tune three Latin neural language models —LatinBERT, LaBERTa, and PhilBERTa— for binary token-level classification and conduct experiments both with and without punctuation. The results show that RoBERTa-based models effectively identify reported speech, with LaBERTa achieving the best performance (F1 scores above 0.90).

1 Introduction

Reported speech detection in ancient classical languages (Ancient Greek and Latin) remains a task not only unsolved but also unexplored.

This paper presents a work in progress on the first method for reported speech detection in Classical Latin, developed through an automatic approach based on the fine-tuning of encoder-only Latin pre-trained language models¹. At this stage, we base our method on a use case drawn from the declamatory anthology of Seneca the Elder (henceforth Seneca), father of the renowned philosopher. The text was written around 30 AD. This anthology is the oldest surviving record of Latin declamatory practice and brings together thousands of quotations from the practice of the *Controversiae* (mock judicial speeches) and the *Suasoriae* (mock

speeches of private advice), which were delivered in pedagogical and public contexts. In this way, the text is distinguished by its highly polyphonic nature, resulting from the use of reported speech in different modalities. Up to 116 speakers cited by Seneca have been identified by scholars (Echavarren, 2007). Due to the text's manuscript tradition, the only complete books that have been transmitted to us are the following: *Controversiae* 1, 2, 7, 9, 10, and the book of *Suasoriae*. The whole corpus contains 96,465 tokens, with punctuation excluded.

2 State of the art and Seneca's corpus

Reported speech is a powerful linguistic tool. By choosing how to report someone's words, it is possible to reshape the original meaning of a text and influence how others perceive the cited speaker. Seneca writes his anthology of declamations with a cultural and a historical purpose: he seeks to preserve the memory of the speakers he claims to have heard by transcribing their speeches. The resulting text therefore constitutes an important stylistic testimony. The speakers are cited and integrated in a network of imitation and competition: each declaimer is positioned in relation to others, who also treat the same subjects. Seneca's own words and the speakers utterances are then mixed in one text through the implementation of reported speech. In this way, quotations in both direct and indirect speech reveal the relationships among the different declaimers and this cultural and literary activity.

The anthology is structured in two macro-sections: the *sententiae*, where long sequences of the declaimers' speeches are cited and the *diuisiones et colores*, which gather Seneca's critical comments alongside other quotations (mostly from the same speakers). These quotations may be either long or short and introduced as direct or indirect reported speech. The work also includes a preface to each book, where the author mainly focuses on

¹The code and dataset are available at https://github.com/agu-oli/latin_reported_speech_detection.git.

the presentation of a speaker.

In Latin, reported speech has been the object of linguistic studies (see Bolkestein, 1996; Shalev, 2005; Rosén, 2013; Charles et al., 2020). These studies have focused on the use of non-finite verbal constructions and subordinate clauses, the use of markers of reported speech in Latin prose, speech verbs (*uerba dicendi*), as well as deictic elements such as the adverbs *hic* and *ecce*, personal pronouns, and the use of the vocative case, (see also Dalbera, 2024).

In Seneca, an important indicator of reported speech can be the presence of a syntagm or lemmas that not only announce the reported speech cataphorically but also introduce a metalinguistic comment on the cited text; in such cases, a speech verb may not be governing the sentence. For instance, Seneca frequently uses the lemma *color*, which in this context refers to the perspective from which the declamator develops the argumentation, for introducing reported speech with a critical motivation. The lemma *color* functions often as a complement of the verb *utor*, which is the root of the phrase, as in the example *Albucius hoc colore usus est*² (*Controuersiae* 2,1,31; 7,6,14; 7,7,13; 7,7,15; 9,5,13), but sometimes the verb is elided, see *Controuersiae* 1,1,17; 10,5,17.

Also, Latin's free word order affects the structure of reported speech at the phrase level: reported speech spans can be discontinuous and therefore interrupted by governing verbs or other phrase elements that do not belong to the reported span.

Finally, as it also happens in Seneca the Elder's anthology, reported speech in Ancient Greek and Latin can include sudden transitions from indirect to direct reported speech or vice versa. Syntactically this implies that, for instance, non-finite verb constructions can be juxtaposed with finite verbal clauses³. Seneca's texts present examples of this reported-speech modality, as in *Controuersiae* 2,2,7 where the speech of Marullus is first cited as indirect speech through a non-finite verb construction, followed by a sudden transition to direct reported speech: *Marullus praeceptor noster licenter uerbo usus est [...] cum diceret uxorem intellexisse mariti mendacium: et ipsa aduersus temerarios mariti iocos relusit*⁴.

²"Albucius made use of this colour".

³This phenomenon that has been explained as a mixed quotation style by Maier (2012).

⁴"Marullus, our teacher, made a bold use of a word [...] when he said that the woman had understood the husband's

In the field of Computational Linguistics applied to Latin, several resources have been developed including Treebanks and neural models such as LatinBERT, PhilBERTa, LaBERTa, and morphosyntactic parsers (Sprugnoli et al., 2024). There have also been advances in Named Entity Recognition task (Erdmann et al., 2016; Beersmans et al., 2023). Nevertheless, no attention has been paid to the automatic extraction of reported speech in Classical Latin. In modern literary studies, we can cite the development of a tagger framework including recognizers for detecting four different types of reported speech in German (Brunner et al., 2020).

Since v.2.10, the Universal Dependencies (UD) introduced changes for reported speech, which has become a subtype of clausal complements such as *csubj* and *ccomp*, instead of a subtype of the *parataxis* relation.

3 Challenges

Within this context, some challenges need to be addressed in order to detect reported speech in Latin as a token classification task:

- **UD reported subtype:** The reported subtype, already integrated in UD v. 2.10, is mostly absent in Latin treebanks. Even if already included in the gold standard of the EvaLatin dependency parsing 2024 campaign, among the five Latin treebanks only the Dante treebank includes some examples of reported speech as a subtype of clausal complements.
- **Reported speech as a within-sentence or span-level problem:** if we approached reported speech as an exclusively syntactic problem, one possible solution would be to enrich the Treebanks with the reported speech subtype in order to improve current parsers. However, such an approach would limit the analysis to the sentence level and the results would be inadequate: reported speech sequences can indeed cover multiple sentences as is the case of Seneca's quotations.
- **Punctuation-based approaches:** Punctuation marks could help identify reported speech since they can function as speech type delimiters. However, a punctuation-based approach would not be suitable for Latin or other ancient languages. The manuscript transmission

lie: 'she joked back to her husband's jests'".

of these texts is not consistent with respect to punctuation, as punctuation conventions have evolved over time. Moreover, the punctuation found in Ancient Greek and Latin texts has largely been added by modern editors and may differ between editions.

4 Dataset

In this paper, we will focus on the reported speech in the *diuisiones et colores* section of the text, since it presents a wide variety of reported speech modalities, including different reporting markers (see *supra* 1 and 2). Moreover, because the *diuisiones et colores* section also includes the author’s critical voice on the declaimers being cited and on the practice of declamation itself, the identification of reported speech spans in this context constitutes a fundamental first step towards analyzing quotation modalities within the text as a whole, as well as for later stylometric analysis of the cited speeches.

For this first experiment, we prepared a small gold standard of 6,203 tokens, including punctuation, as well as a version of the same dataset without punctuation, consisting of 5,068 tokens. The gold standard contains extracts primarily from the *diuisiones et colores* sections of Books 1, 2 and 9 of the *Controuersiae*, as well as an extract from the preface to Book 2 and from the *Suasoriae* (1,4), in order to obtain a more balanced sample of the work as a whole. The text corresponds to the A. Kiessling edition, which is in the public domain⁵.

The dataset contains the following features: the word form as in the sentence, the lemma, the part-of-speech (PoS) tag for each word form, and a customized binary feature indicating whether the form corresponds to a non-finite verb (NFV). Under UD, participles and infinitives are classified as VERB, together with finite forms. However, in the case of reported speech, non-finite verb (infinitives) constructions are one of the possible ways of expressing reported speech in Latin. Finally, each token in the gold dataset is annotated with a binary label indicating reported or non-reported speech, which is then the target to be predicted by the developed model⁶.

Although reported speech is not a homogeneous category—at a minimum, direct, indirect, and free

⁵Seneca, L. A. (1922). *Annaei Senecae oratorum et rhetorum sententiae, diuisiones, colores* (A. Kiessling, ed.). Leipzig: B. G. Teubner.

⁶Discontinuous spans (e.g. speech verb interrupting a reported sequence) are handled by labeling only the quoted material and treating interruptions as non-reported tokens.

indirect forms can be distinguished—the present study focuses on a binary distinction between reported and non-reported macro-categories, so that future work can build on this baseline and address the internal differentiation of reported speech types.

Form	Lemma	PoS	NFV	Label
Belle	belle	ADV	0	0
de	de	ADP	0	0
hoc	hic	DET	0	0
uitio	uitium	NOUN	0	0
illius	ille	DET	0	0
Scaurus	scaurus	PROPN	0	0
aiebat	aio	VERB	0	0
,	,	PUNCT	0	0
illum	ille	DET	0	1
acta	ago	VERB	1	1
in	in	ADP	0	1
aurem	auris	NOUN	0	1
legere	lego	VERB	1	1

Table 1: Excerpt of token-level annotation for reported speech from *Controuersiae* 2,1,39 including linguistic features and reported speech label from two different sentences. Labels indicate reported speech (1) and non-reported speech (0). The two sentences are not split.

5 Method

We approach reported-speech detection in Latin as a token-level sequence labeling problem, a perspective that enables the identification of spans that cover multiple sentences. This entails labeling tokens as reported or non-reported speech. In this manner, our approach can exploit features available at the sentence and token levels, such as lemmas, part-of-speech tags, and the presence of non-finite verbal forms, while analyzing them at the span-level in order to determine whether each token belongs to reported or non-reported speech.

This first experiment consisted of fine-tuning three encoder-only Latin neural models, Latin BERT (Bamman and Burns, 2020) and LaBERTa, which are monolingual, and the multilingual model PhilBERTa (Riemenschneider and Frank, 2023). All three were fine-tuned for binary token-level classification of textual sequences as reported or non-reported speech in Seneca the Elder’s anthology, using token-level features (lemma, PoS, NFV) encoded as embeddings⁷ and concatenated with the model’s contextual token representations.

⁷Lemma embeddings were initialized from the pretrained model’s vectors.

Model	Punctuation	Acc	Prec	Rec	F1	PR AUC
LatinBERT	with punct.	0.744	0.769	0.892	0.826	0.803
LatinBERT	without punct.	0.681	0.737	0.855	0.791	0.822
LaBERTa	with punct.	0.877	0.921	0.897	0.909	0.949
LaBERTa	without punct.	0.873	0.935	0.883	0.908	0.933
PhilBERTa	with punct.	0.849	0.891	0.886	0.889	0.927
PhilBERTa	without punct.	0.847	0.910	0.870	0.890	0.937

Table 2: Performance of three fine-tuned models with and without punctuation on the test set.

We fine-tuned all three models with a learning rate of $1e-5$, batch size 8, and weight decay 0.01. All experiments were run with a fixed random seed (42) (source code link in footnote 1). Checkpoints were selected by validation loss. Evaluation is performed at the token level after alignment of word-level labels to subword tokens.

Preprocessing choices: we conducted two experiments per model on two preprocessing choices on the same dataset, with and without punctuation.

Dataset splitting strategy: the span-level approach presented a new challenge for splitting the dataset in a training, validation, and test sets. Since the dataset consists of sequences of text containing reported speech spans which can extend across more than one sentence, it is not possible to adopt a sentence-level splitting strategy. Such a split would not be appropriate and would prevent the model from learning to detect sequences of tokens belonging to the same speech type across multiple sentences. To address this issue, the dataset was then organised into 18 sections, which do not necessarily coincide with the book and chapter divisions in Seneca’s work. Each section contains both labels; section boundaries were defined during annotation by distinguishing coherent units.

In short, the dataset was split at the level of these sections. We randomly assign 70% of the sections to training, 20% to validation, and 10% to test sets using a fixed random seed (42). No section appears in more than one split.

6 Results

The results show (see Table 2) that transformer-based Latin models are able to identify reported speech at the token level even after a first experiment in a small data setting and without relying on sentence-level segmentation.

Among the three models, LaBERTa consistently achieves the best performance, with F1 scores above 0.90 and the highest PR AUC value, in-

dicating robust discrimination between reported and non-reported tokens despite the limited size of the dataset. LatinBERT exhibits higher recall than precision in the test set, indicating sensitivity to reported speech spans, a behavior that can be desirable for literary and stylometric analysis because it can be refined via a post-processing step. However, LatinBERT shows clearly lower performance than RoBERTa-based models, which could reflect differences in tokenization strategies (see Bamman and Burns, 2020). In contrast, both LaBERTa and PhilBERTa (see Riemenschneider and Frank, 2023), which share the same tokenization method, demonstrate more stable performance, indicating better generalization.

LaBERTa’s superior performance could be due to its pretraining data, which is limited to the *Corpus Corporum*. In future work, it should be investigated whether this training corpus provides more stable lexical and contextual representations for Latin. By contrast, PhilBERTa achieves very similar performance to LaBERTa, even though it is a multilingual Greek–Latin model and was pre-trained on more heterogeneous data: the *Open Greek & Latin project*, the *CLARIN corpus of Greek Medieval Texts*, the *Patrologia Graeca*, the *Corpus Corporum* and also Project Gutenberg texts.

The comparison between punctuation-aware and punctuation-free datasets does not reveal considerable differences between LaBERTa and PhilBERTa performances. While the presence of punctuation slightly coincides with a recall improvement, LaBERTa and PhilBERTa models trained without punctuation remain highly competitive, which suggests that lexical and contextual information play a more important role than punctuation for reported speech detection in Latin. This is particularly relevant for the future of this study on a larger dataset and for the deployment of a reported speech detection tool on Latin corpora given the inconsistency

and instability of punctuation in the manuscript tradition and editorial practices.

7 Error analysis

At the multi-word level, subordinate clauses whose structure closely resembles that of reported speech remain difficult to distinguish from genuinely reported sequences. All models appear to overfit to reported clauses introduced by the Latin particle and subordinating conjunction *an*, which are frequent in the Seneca corpus because they are used to indicate which points are addressed by the speakers (the *diuisio* of the speech). Clauses introduced by *an* are also used by Seneca in his own authorial voice to analyze how this is addressed by the various speakers⁸.

The proximity of tokens to a speech verb may also lead to false positives across all models, as in *Controversia* 1,1,15⁹, where the speech verb is followed by a direct object and then by a clausal modifier of that object which constitutes the proper reported speech span. The object is incorrectly tagged as part of the reported segment. This example illustrates a construction that is particularly frequent in Seneca, in which an object or oblique nominal is followed by a clausal modifier that introduces the reported-speech sequence. It functions as a transition between the author's critical comments on the speakers and the strictly reported text. The question of how best to handle these two constructions will be addressed in future work.

In both punctuation and non-punctuation training settings, the models tend to tag adverbs such as *deinde* ("then"), from Seneca's voice, which interrupt the quotations, as reported speech. By contrast, while LaBERTa and PhilBERTa models trained on a dataset containing punctuation correctly identify the speech verb *adicio* ("to add") when it interrupts a reported sequence, the models trained without punctuation and LatinBERT fail to recognise it in this function.

Furthermore, models trained without punctuation show a tendency to produce isolated false-negative tokens within an otherwise correctly identified reported span. Models trained with punctuation, on the other hand, exhibit errors on punctua-

⁸*An abdicari debeat per haec quaesiit: an...* "On the question 'should he be disinherited', he examined..." (*Controversia*. 1,1,13)

⁹*et sua figura dixit omnia propter quae uelle deberet. 'Quare ergo abdicas?'*"and through his figure he said all the reasons why he should want it: 'Why do you disinherit me?'"

tion marks both within reported spans and at span boundaries. Both types of error can be addressed in a post-processing step.

8 Conclusions and future work

These initial results support the feasibility of token-level reported speech detection in Latin, as well as the advantages of a span-based segmentation strategy, despite the small size of the dataset used in this first experiment. A span-level approach based on fine-tuning transformer-based Latin models, to detect reported speech sequences at the token level, enables the effective extraction of reported speech sequences that extend beyond sentence boundaries.

Future work will focus on enlarging the dataset and further examining how linguistic features and contextual cues affect the performance of token-level reported speech detection, with the aim of improving model robustness. It will also involve testing the approach on the internal types of reported speech and conducting further error analysis in order to develop a post-processing strategy that could include rule-based approaches or syntax-based heuristics.

Just as reported speech is a powerful linguistic tool in Seneca's text, advances in its detection can lead to new contributions to information extraction from Latin texts—including non-literary ones—in which other sources are cited, and may also provide inspiration for comparable frameworks in other languages.

Acknowledgments

This research was funded by the Émergence project, at Sorbonne University and Rouen Normandie University (France), "SenecAI: Artificial Intelligence for the Stylistic and Authorial Analysis of Seneca the Elder's Declamatory Anthology"¹⁰.

References

- David Bamman and Patrick J. Burns. 2020. *Latin BERT: A contextual language model for classical philology*. *Computing Research Repository*, arXiv:2009.10053.
- Marijke Beersmans, Evelien de Graaf, Tim Van de Cruys, and Margherita Fantoli. 2023. Training and evaluation of named entity recognition models for classical latin. In *Proceedings of the Ancient Language Processing Workshop*, pages 1–12, Varna, Bulgaria. INCOMA Ltd.

¹⁰<https://editta.hypotheses.org/senecai>

- A. Machtelt Bolkestein. 1996. [Reported speech in Latin](#). In Theo Janssen and Wim van der Wurff, editors, *Reported Speech: Forms and Functions of the Verb*, pages 121–140. John Benjamins Publishing Company, Amsterdam.
- Ann Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. 2020. [To BERT or not to BERT: Comparing contextual embeddings in a deep learning architecture for the automatic recognition of four types of speech, thought and writing representation](#). In *Proceedings of SwissText/KONVENS*.
- Lise Charles, Frédérique Fleck, and Lyliane Sznajder. 2020. [Les formes de discours rapporté : repérage et interprétations](#). *Lalies*. Langues et littérature : actes des sessions de linguistique et de littérature.
- Joseph Dalbera. 2024. [Verbes introducteurs et stratégies d'introduction du discours direct dans la narration romanesque latine \(le Satyricon de Pétrone et les Métamorphoses d'Apulée\)](#). In Concepción Cabrillana, editor, *Recent Trends and Findings in Latin Linguistics*, pages 503–520. De Gruyter, Berlin, Boston.
- Arturo Echavarren. 2007. *Nombres y personas en Séneca el Viejo*. Ediciones Universidad de Navarra.
- Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. 2016. Challenges and solutions for latin named entity recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 85–93, Osaka, Japan. The COLING 2016 Organizing Committee.
- Emar Maier. 2012. [Switches between direct and indirect speech in ancient greek](#). *Journal of Greek Linguistics*, 12(1):118–139.
- Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.
- Hannah Rosén. 2013. [About non-direct discourse: Another look at its parameters in latin](#). *Journal of Latin Linguistics*, 12(2):231–268.
- Donna Shalev. 2005. [Action nouns in reports of speech acts](#). *Journal of Latin Linguistics*, 9(2):719–730.
- Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. 2024. [Overview of the evalatin 2024 evaluation campaign](#). In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 190–197, Turin, Italy. ELRA and ICCL.

Annex: inter-annotator agreement

100% of the data were tagged independently by two experts¹¹, leading to a Cohen's kappa of 0.92. Disagreements were resolved through joint review to produce the final gold standard.

¹¹The author of this work and Pierre Belenfant.

Narrative in Short German Prose: A Multi-Phenomenon Dataset for Computational Literary Analysis

Hans Ole Hatzel^{*1}, Haimo Stierner^{*2},
Evelyn Gius², Chris Biemann¹

University of Hamburg¹, TU Darmstadt²

{hans.ole.hatzel, chris.biemann}@uni-hamburg.de,

{haimo.stierner, evelyn.gius}@tu-darmstadt.de

Abstract

We present the novel dataset GermAnProse, an annotated corpus consisting of four German short prose texts accompanied by an extensive set of narrative-focused annotations. As part of this dataset, we contribute an annotation scheme for mentions, speech, and character agency: Characters in Action (ChiA). GermAnProse also contains information on narrative phenomena: narrativity, semantic verb classes, and plot keyness. Moreover, we include reader reception data in the form of timing information for audiobook performances, indicating pauses between sentences and the time taken to read a specific sentence in a performance. We release the dataset, which contains more than 18,000 manually created standoff annotations in JSON format, enabling researchers to utilize this resource for further exploratory applications.

1 Introduction

In recent years, the computational analysis of literature has gained increasing attention, as evidenced by the emerging field of Computational Literary Studies (CLS). Our novel corpus GermAnProse (short for **German annotated prose**), contains annotations of a series of narrative phenomena for four literary texts and is meant to serve the field. Figure 1 showcases our annotations on a toy example. Specifically, our dataset contains annotations for narrativity (Vauth et al., 2021), character mentions and agency, semantic verb classes (Hatzel et al., 2025), plot keyness (Hatzel et al., 2023), scenes (Zehe et al., 2021), and audiobook timing information. With this publication, we pursue the goal of providing a dataset with rich annotations for performing explorative studies of narrative phenomena with audiobook information, in the form of the time taken to read a sentence and the length of pauses between sentences, serving as a reader-reception-focused extension. In the established CLS process

^{*}Equal contribution.

Keyness: 0.5
Reading Speed: 1.2

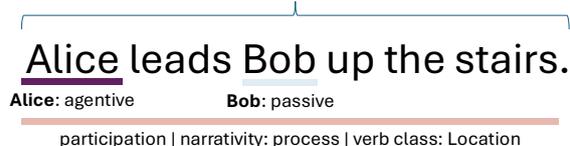


Figure 1: In our dataset, phenomena are annotated at a sentence level (plot keyness and reading speed) or at a phrase level (agency, narrativity, and verb classes). The roles of characters in a given agency annotation are identified on a span level; our scene annotations are not shown here. The example shows an interaction of *Alice* and *Bob* in which *Alice* is the active participant.

of operationalization, annotation, automation, and exploration (Pichler and Reiter, 2020), we envision our dataset to serve as an initial exploration resource. Researchers can use the annotated texts to identify interesting phenomena and explore their interaction, guiding their subsequent research on larger corpora. Moreover, the dataset can serve as ground truth for the development and evaluation of automated approaches. We see this work as part of a larger trend with drastic improvements in natural language processing, allowing the field to increasingly shift its focus from low-level local linguistic phenomena to higher-level narrative and semantic phenomena.

The dataset is provided in JSON format with text and corresponding character-level standoff annotations, allowing for a low barrier of entry and easy exploration in any scripting environment.¹

2 Related Work

Here, we will only discuss prior work that does not integrate into our dataset. Our dataset contains a range of previously released annotations that we merge into one unified dataset, see Section 3.1 for details on this data.

¹<https://github.com/forTEXT/GermAnProse>

CLS has a strong tradition of reusable datasets; largely, these datasets focus on providing collections of plain texts or texts in TEI-XML with formatting information in conjunction with limited metadata (Gius et al., 2021a; Schöch et al., 2021, e.g.). The DraCor initiative (Fischer et al., 2019), for example, focuses on drama and provides the texts in TEI-XML with metadata, such as publication year, and limited annotations, such as speaker identification. Often CLS researchers publish annotations for a single phenomenon of interest on a limited set of project-specific texts (Underwood, 2018; Piper and Bagga, 2022; Coll Ardanuy et al., 2020, e.g.). Hagedorn and Darányi (2022) set out to create a dataset of folk tales that is extensible by other researchers.

Audiobooks have received limited attention as a resource in NLP research. Largely, approaches focused on audiobooks as a speech data source (Székely et al., 2012; Boeffard et al., 2012). Stiemer et al. (2025) introduced the idea of using audiobook information as a means of analysis for the original text to the CLS community. The public domain audiobooks on Librivox² have long attracted the interest of the research community in the humanities and natural language processing alike (Weber, 2021; Beilharz et al., 2020).

3 Corpus & Annotations

In our selection of texts, we largely follow the selection by Vauth et al. (2021), including representation of modern German literature while relying only on out-of-copyright texts (see Table 4 for details). This selection enables us to build upon their published annotations for our own work. Further, for resource reasons, we limit our selection to short narratives with the longest text consisting of just over 22,000 tokens. We annotate the texts in their entirety, rather than excerpts of long texts like some prior work (Bamman et al., 2020, e.g.); we want to, thereby, enable the analysis of entire self-contained narratives rather than focusing on local linguistic features in service of future automation. Here, we depart from the text selection by Vauth et al. (2021) in removing the annotated excerpt of the much longer text *Effi Briest*, including instead the full text *Der blonde Eckbert*; as a result, all our texts are fully annotated. The annotated texts originate either from the TextGrid corpus (TextGrid, 2021) or the d-Prose 1870-1920 corpus (Gius et al., 2021a). In Table 1

²<https://librivox.org>

we provide a full list of the works with relevant metadata.

Table 2 provides an overview of the source and quality of the annotations we discuss in detail below.

3.1 Pre-existing Data

In this section, we describe pre-existing published annotations that are included in our dataset and give a rationale for including each set of annotations.

3.1.1 Narrativity

Narrativity or eventfulness refers to the quantification of “how much” is happening in a narrative text passage. Our dataset contains narrativity annotations as operationalized in the EvENT concept by Gius and Vauth (2022). In the approach each finite verb in the texts is annotated as expressing one of four *event* types: change of state, process, state, or non-event. The event types, in the order listed above, are associated with decreasing narrativity (Vauth et al., 2021). Events are not derived from semantic units – as the concept is typically understood in natural language processing (see e.g. Doddington et al., 2004) – but are instead informed by their textual representation. As a result, the concept can be understood as the first step in a bottom-up approach to event semantics, ranging from simple and text surface-based phenomena to more complex semantic phenomena. Change of state events mark the most eventful moments in a text, while non-events indicate eventlessness (since they do not denote actual events in the fictional world but rather descriptive passages, questions, or counterfactual statements).

3.1.2 Semantic Verb Classes

The verb class data were generated to provide a coarse-grained semantic profile of individual literary texts based on the meanings of their verbs (Hatzel et al., 2025). Following the narrativity annotations (Vauth and Gius, 2021), verb phrases were used as the annotation unit, with each annotation semantically describing the finite verb in a phrase. Building on this event-centered segmentation, each main verb was classified according to the semantic field of its lemma in GermaNet, resulting in one of fifteen verb classes such as *communication*, *emotion*, *location*, *change*, or *perception*. Ambiguous or missing verb senses were resolved through contextual interpretation to maintain internal consistency across each annotated text. The classes represent broad categories of actions and processes

Text	Author	Published	Short Name
Der blonde Eckbert	Ludwig Tieck	1796	Eckbert
Das Erdbeben in Chili	Heinrich von Kleist	1807	Erdbeben
Krambambuli	Marie von Ebner-Eschenbach	1896	Krambambuli
Die Verwandlung	Franz Kafka	1915	Verwandlung

Table 1: We list the four texts in our dataset with their respective publication dates and original author names. For reference, we also supply short-form names that are used to identify the texts in some of our tables.

Annotations	Publication	Quality
Verbclasses	Hatzel et al. (2025)	
Keyness	Hatzel et al. (2023)	
Scenes	Ours	
Speech Time	Ours	
Agency ^{ChiA}	Ours	
Mentions ^{ChiA}	Ours	
Narrativity	Vauth et al. (2021)	

Table 2: We list the annotation source and quality across the different phenomena annotated in our dataset.  refers to gold data,  refers to automated annotations, and  refers to single annotator annotations. The ChiA superscript denotes data that is part of our novel annotation scheme.

and were selected to capture the central event types that shape narrative meaning.

For each literary work, Hatzel et al. (2025) aggregated the resulting annotations to produce text-specific verb class distributions. The resulting frequency profiles are used as interpretable summaries of a text’s semantic structure, highlighting tendencies such as an emphasis on emotional, social, or physical processes.

We expect that the verb classes can be used as a first order approximation of action semantics in exploratory text analysis.

3.1.3 Plot Keyness

With “plot keyness” we refer to a given sentence’s relevance to the plot of the overall story. For example, a passage describing the natural beauty of the scenery will, in most cases, not contain the most important actions of the story. The data captures plot keyness by using a corpus of summaries for measuring how many of them reference a given passage of text (Hatzel et al., 2023). Conceptually, this data builds on the ideas of the Pyramid method (Nenkova and Passonneau, 2004), asserting that the most relevant events are mentioned in the greatest number of summaries. For the plot keyness data, we use text passages instead of atomic event units. In

the downstream analysis of our dataset, we expect this to be used to identify low-level textual features for relevance to the overall story.

Specifically, our plot keyness measure is the fraction of summaries in our dataset that reference a given sentence. A value of 1.0 means all summaries refer to a specific sentence, whereas a value of 0.5 means only half of all summaries reference the events in the sentence.

The plot keyness and its relation to concepts like narrativity is a phenomenon of interest for CLS. We showcase the exploration of this relationship on our data (see Section 4).

3.2 Novel Annotations

In this section, we describe new annotations we contribute in this publication that are based on existing annotation schemes. The newly created annotations are marked as *ours* in Table 2.

3.3 Scenes

To enable the aggregation and contextual analysis of the data compiled in this dataset across selected text segments, we divided the corpus texts into scenes. The guidelines applied for this purpose (Gius et al., 2021b) are based on the narratological understanding of a scene as a narrative segment that constitutes a coherent unit of time, space, action, and character constellation—roughly corresponding to a film shot in terms of its continuity.

The guidelines specify how scene boundaries can be identified by examining changes in these four aspects. They also provide clear criteria for distinguishing between narrative scenes and non-scenic passages such as summaries, reflections, or general descriptions. To ensure reliable and consistent annotations, the guidelines assign different weights to each aspect, prioritizing the coherence of events and character constellations over the temporal or spatial continuity.

In addition, the guidelines define lexical and structural indicators of scene changes (e.g., “suddenly”, “the next morning”), recommend the use

of the film-shot analogy as an intuitive test, and explain how to handle transitional or ambiguous passages. The resulting annotations enable the segmentation of literary texts into meaningful narrative units, thereby allowing for a systematic analysis of plot structure and narrative rhythm across different genres and styles. Scenes are one of the aspects of our work that can potentially be extended to large corpora.

3.4 Audiobook Timing

Audiobook information serves as additional reception-based data in our dataset, allowing us to connect purely narrative phenomena to a simple form of reception representation. Potentially, readers’ interpretations of a text can hint at deeper phenomena, and analyzing the relationship between narrative phenomena and reader presentation is a worthwhile endeavor. We obtain our data from amateur readers (from the platform Librivox or read by student assistants) and professional readers in commercial audiobooks. Each text and its audio performances are put through a forced alignment pipeline to obtain token-level timing information.³ We break this information down to the sentence level and provide a reading speed comparison as measured against a text-to-speech (TTS) system that is based on public training data (Müller and Kreutz, 2021).⁴ While the speech synthesis has dealt with explicitly modeling speech speed given phonemes (Ni et al., 2019), we take a comparatively simple approach in relying on the duration of a single speech generation. Our approach is potentially sensitive to individual terms that the TTS system fails to reproduce accurately on a story-analysis level. If such a term is central to a given text and, for example, co-occurs with major events at an increased rate, it is conceivable that the approach could adversely affect downstream analysis.

3.5 Mentions, Direct Speech and Character Agency

We developed annotation guidelines for systematically capturing how characters act and interact as agents within narrative texts; the guidelines are referred to as characters in actions (ChiA). To maintain compatibility with the narrativity and semantic verb class data, we follow the two annotation schemes in regard to the annotation unit. That is to

say, for each verbal phrase, the character actions associated with the verb are modeled. Our guidelines specify that these actions encompass all forms of activity within the fictional world, including mental and emotional processes, speech, and physically manifested actions.

Our annotations chiefly focus on different degrees of agency, describing the extent to which a character actively shapes an event: ranging from intentional and controlled actions (agentive) to involuntary or imagined actions (low-agency), and further to situations in which the character is merely affected by or influenced by events (passive). By distinguishing between these levels, the guidelines enable the tracing of narrative dynamics and the influence of characters throughout the text. We refer to these annotations that capture co-presence and the agentivity of characters on a verbal phrase level as the **agency** annotations.

In addition to character agency, we annotate character mentions in texts. For the mention extents, we adopt an existing annotation scheme for character mention span extents (Krug, 2020). Beyond that, we differ from prior work in that we only annotate individual characters (potentially multiple for one span) rather than annotating collective mentions on a group level (e.g. “the three friends” would refer to three separate characters in our annotation scheme). Further, we introduce two special characters *background character* and *crowd* that are used for non-specific or non-reoccurring characters and unspecified groups of people, respectively. Figure 1 illustrates a toy example of an active and a passive character. For real world example from our dataset see Section 3.7 As illustrated, character mentions are also annotated with a canonical name for the corresponding character, as the mention’s surface form does not always unambiguously identify a character.

In addition to character mentions and agency, direct speech is also annotated, and text inside speech acts is not considered for agency annotations. All three tags are linked to their corresponding characters; in the case of agency, the particular roles (agentive, low-agency, passive) of each character are also annotated. Taken together, these annotations provide a structured representation of how characters appear, act, and speak, thereby enabling a systematic analysis of their narrative functions and their contribution to plot development.

³<https://github.com/feldberlin/timething>

⁴<https://github.com/OHF-Voice/piper1-gpl>

3.5.1 Agreement

For the purposes of calculating inter-annotator agreement, we treat the identification of agentive, low-agency, and passive participants as a binary classification problem for each character in each annotation unit. In terms of Krippendorff’s Alpha, in a weighted average setup, we achieve agreements of up to 0.6 in identifying characters in the agentive class. For the low-agency and passive classes, by contrast, we observe much lower agreement, ranging from 0.35 to 0.51 (see Table 3 for a breakdown by class and text). Qualitatively speaking, we can attribute many of the errors to the special *background character* and *crowd* characters; we observe that it can be difficult to discern, especially for larger groups of people, who, besides the main characters, are included in an action.

For the mutual accuracy metric, we calculate the accuracy across all pairs of span-wise matching annotations, macro-averaging across annotator pairs. This results in a perhaps more interpretable measure: showing that annotators agree on the exact set of characters in approximately 75% of cases.

3.6 Annotation Workflow

Both the scene and ChiA annotations were manually created using the web-based tool CATMA (Gius et al., 2025) in conjunction with the GitMA package (Vauth et al., 2022) for programmatic processing. All annotators were students of literary studies and employed as student assistants or were full-time researchers and co-authors of this paper.

For the ChiA guidelines, extensive pilot annotations accompanied by annotator training were performed, iteratively improving the guidelines. At least two annotators annotated each text, and for all but one text, at least two annotators annotated the text in its entirety, in addition to another annotator creating the gold standard. Regular coordination meetings were held to discuss and resolve ambiguous annotation cases, which were additionally documented using a dedicated uncertainty tag. Based on the resulting annotations, gold-standard annotations were produced by an additional annotator who reconciled inconsistencies. To facilitate this process, we developed a semi-automatic adjudication pipeline that inserted annotations on which both annotators agreed (on both the spans and properties of an annotation) into a new collection in CATMA. All annotations automatically created in this manner were subsequently carefully reviewed. We found

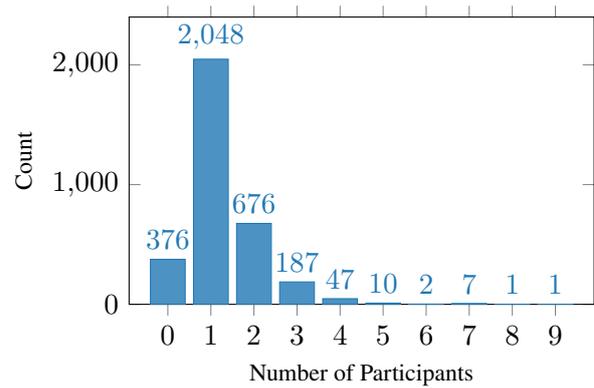


Figure 2: The distribution of the number of participants in each agency annotation in our data is shown. Each annotation includes an average of 1.27 participants. Very few annotations have more than three participants.

the approach to vastly speed up the creation of the gold labels.

We have around 3,400 gold agency annotations across the four texts in our dataset. For each agency annotation in our gold data, 0.40 agentive characters, 0.57 low-agency characters, and 0.30 passive characters are annotated on average; this results in a total of 1.27 characters per annotation, see Figure 2 for the distribution. The maximum number of characters participating in one agency annotated phrase in our dataset is nine, occurring in a sentence “When **they** arrived in the Dominicans’ church” from *Das Erdbeben in Chili*, in which a large portion of the story’s cast is involved.

In terms of co-occurrence, there is a sizable number (376) of agency annotations with no characters involved, as each verbal phrase is an annotation unit regardless of its semantics. For example, the phrase “The weather is nice.” has no participants.

3.7 Example

Figure 3 shows an example of real annotations from our dataset. Note that the single sentences consist of multiple verbal phrases, each with its own narrativity and participation annotations. Further, the example illustrates that not every participant must have a surface representation in the phrase in which they appear. The phrase “ließ aber eine Flasche [...] bringen”, for example has the character *Hopp* perform an action. In practice, unlike in the visualization, our annotators label character identities rather than their spans. However, since all character mentions are separately annotated, the corresponding text mention spans—as depicted in our graphics—can still be retrieved.

	Mutual Accuracy			Krippendorff’s Alpha		
	Erdbeben	Verwandlung	Krambambuli	Erdbeben	Verwandlung	Krambambuli
passive	0.722	0.762	0.731	0.381	0.444	0.509
low-agency	0.684	0.653	0.674	0.354	0.443	0.409
agentive	0.730	0.785	0.732	0.510	0.603	0.590

Table 3: The agreements for annotating character agency as measured by mutual accuracy and Krippendorff’s Alpha, show that generally speaking, the agentive class is annotated most consistently.

Text	Number of						
	Tokens	Sentences	Scenes	Agency	Mentions	Speech Acts	Narrativity
Der blonde Eckbert	7508	314	14	259	1205	20	930
Das Erdbeben in Chili	6499	224	14	627	1064	39	712
Krambambuli	4573	226	16	477	628	39	571
Die Verwandlung	22299	850	26	1992	2884	106	2363
Total	40879	1614	70	3355	5781	204	4576

Table 4: Annotation Overview. All annotation schemes cover each of the four texts; we omit some of the annotation schemes for brevity.

4 Data Exploration

After laying out the components of our dataset, we perform a preliminary analysis of the data. Table 4 shows a range of dataset statistics.

Narrativity and Plot Keyness For narrativity, as measured by event scores (Vauth et al., 2021), we find a very weak correlation with plot keyness ($r=.067$, $p<0.05$). The results indicate some connection between the two phenomena, but, as captured by our data, there is no clear discriminative value for either variable in predicting the other. Narratologically speaking, we expect to find a limited correlation as keyness captures plot relevance, where narrativity only captures narration at a local level; yet, one might expect non-narrative passages to only rarely constitute key passages. Hatzel et al. (2023) came to a similar conclusion, based on comparable data.

Semantic Verb Classes and Narrativity In Figure 4, we illustrate the distribution of event types in ascending order of narrativity with each semantic verb class. We find that some classes, particularly *Stative Verbs*, *Cognition*, and *Emotion*, are associated with lower narrativity. Perhaps most notably, the classes *Location*, *Communication*, and *Contact* are associated with higher narrativity.

Pauses, Reading Speed, and Plot Keyness We examine the phenomenon of pauses and reading speed in the audiobook timing information contained in ChiA-Seeds. The idea was initially ex-

plored by Stiemer et al. (2025), and we expand on their analysis using additional texts and introduce the concept of reading speed to the analysis.

In our data, professionals make longer pauses than amateur readers, with an average length of 0.87 and 0.64 seconds, respectively. The two groups read at an equal pace when considering only the sentences themselves, without taking into account the pauses (1.25 times faster than the TTS system for the professional readers and 1.26 times faster for the amateurs). For all readers in our data, we observe a very small but significant correlation between the pause after a sentence and its plot keyness ($r=.05$, $p<0.05$). The same effect can be observed for the correlation of reading speed and plot keyness, but it does not rise to statistical significance. Sentences that contain direct speech are also read slightly faster in our dataset ($r=.04$, $p<0.05$). Further work is required to identify potential differences in presentations by amateur readers compared to professional readers.

Character Interaction As our character agency annotations capture the co-presence of characters in verbal phrases, we can use the annotations to build a novel kind of character network. In conjunction with the scene annotations, the data can also be used to create dynamic character networks, changing with each scene. We show an instance of a character network for a subset of the scenes in *Das Erdbeben in Chili* in Figure 5. So far, high-quality data for performing character network visualization on prose texts has not been available, with many

Herr Hopp sagte vorerst kein Wort von dem Wohlgefallen, das er an dem Hunde gefunden hatte,

participation | Communication | non-event

Hobb: passive

participation | Cognition | stative event

Hobb: agentive

Krambambuli: passive

ließ aber eine Flasche von dem guten Danziger Kirschbranntwein bringen,

participation | Social | process | Hopp: agentive

den der Löwenwirt damals führte, und schenkte dem Vazierenden fleißig ein.

participation | Social | process

Wirt: low agentive

participation | Change | process | Hopp: agentive

Forstgehilfe: passive

Figure 3: Example annotations for a single sentence from *Krambambuli*. The sentence has a keyness of 0.25 and a reading speed of 1.36 in our data. Translation: “Mr. Hopp said for the time being not a word of the pleasure that he had found in the dog, but had a bottle of the good Danzig cherry brandy brought, which the Lion-innkeeper carried at that time, and poured for the wanderer diligently.”

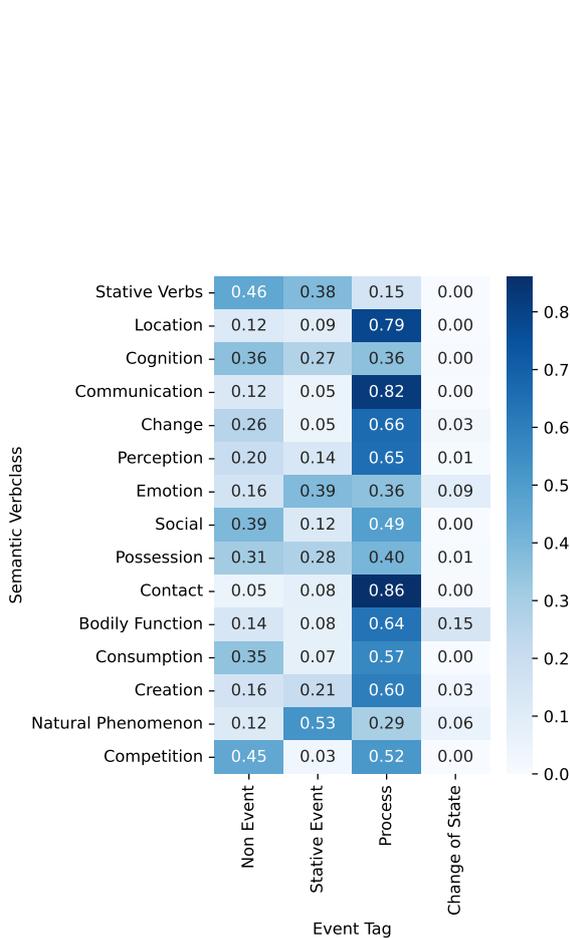


Figure 4: The event tags are listed in ascending order of narrativity. For each verb class, we show the relative distribution of event tags within its occurrences.

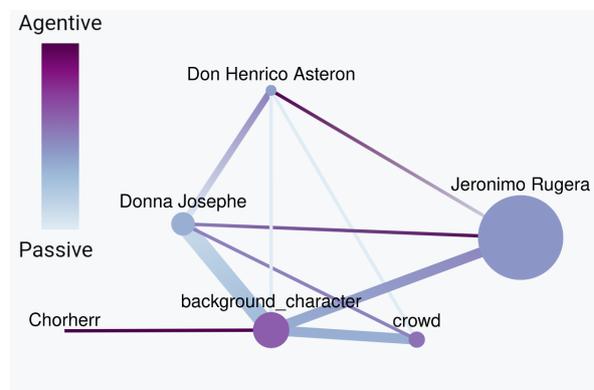


Figure 5: In our character networks with agency data, the color of nodes describes character agency, while the colors at the ends of each describe each character’s agency in only that specific interaction. Here, we show the character network for the first five scenes in *Das Erdbeben in Chili*.

prior approaches falling back on using only specific mentions (e.g., explicit names) and their co-occurrence (Ketschik, 2025) or using hybrid models that include coreference resolution with name clustering or additional rules (Perri et al., 2022, e.g.). Ketschik (2025) also points out that many prior approaches fail to account for differences in character mentions, a problem that our data, at least partially, resolves, as direct speech is not included, and our agency data captures character interactions (or, in the case of low-agency, co-presence).

5 Conclusion

In this work, we introduce GermAnProse, a novel dataset for exploring narrative phenomena in short narratives. Our extensive annotations cover a total of over 40,000 tokens, with more than 18,000 manually created span annotations in various categories,

out of which more than 9,000 are newly created for this work. In addition, we provide automated sentence timing information for audiobooks as aligned with the original texts, allowing a form of reader studies to be performed on our data.

In our data exploration, we show initial examples of identifying phenomena that warrant further investigation. We provide a novel dataset allowing for the exploration of narrative texts.

Despite its focus on narrative exploration, we believe that our dataset will also prove useful for established computational approaches such as character coreference resolution (Roesiger et al., 2018) or the attribution of direct speech (Petersen-Frey and Biemann, 2024). Likewise, we expect that our dataset can aid in the creation of automation for the character agency annotations. We recommend that all such approaches adopt text-level leave-one-out cross-validation schemes to verify their generalizability.

6 Limitations

We see the main limitation of our work in the limited size of the dataset. Resource-wise, creating a much larger dataset would not be feasible. In our view, the dataset is sufficiently large to enable the identification of potentially interesting effects, but might not support robust quantitative conclusions in most cases.

As mentioned, our audio speech assessments are potentially sensitive to systematic errors in the speech synthesis system we use. Further analysis, comparing across readers and potentially with other speech synthesis approaches, would be beneficial.

7 Ethical Considerations

In terms of legal considerations, all primary annotated works are out of copyright. The audiobooks we incorporated into the analysis are, in part, commercial products. We only report derived data (i.e., timing information) intended for analysis that does not allow a reconstruction of the original performance.

In terms of gender representation in our dataset, one of four texts is written by a female author (*Krambambuli* by Marie Ebner von Eschenbach).

References

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An Annotated Dataset of Coreference in English Lit-

erature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Benjamin Beilharz, Xin Sun, Sariya Karimova, and Stefan Riezler. 2020. LibriVoxDeEn: A Corpus for German-to-English Speech Translation and German Speech Recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3590–3594, Marseille, France. European Language Resources Association.

Olivier Boeffard, Laure Charonnat, Sébastien Le Maguer, and Damien Lolive. 2012. Towards Fully Automatic Annotation of Audio Books for TTS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 975–980, Istanbul, Turkey. European Language Resources Association (ELRA).

Mariona Coll Ardanuy, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson, and Barbara McGillivray. 2020. *Living Machines: A study of atypical animacy*. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4534–4545, Barcelona, Spain (Online). International Committee on Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Frank Fischer, Ingo Börner, Mathias Göbel, Angelika Hechtel, Christopher Kittel, Carsten Milling, and Peer Trilcke. 2019. *Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama*. In *Digital Humanities 2019: "Complexities" (DH2019)*, Utrecht. Zenodo.

Evelyn Gius, Svenja Guhr, and Benedikt Adelman. 2021a. *d-Prose 1870-1920*.

Evelyn Gius, Jan Christoph Meister, Malte Meister, Marco Petris, Dominik Gerstorfer, Mari Akazawa, and Stefanie Messner. 2025. *CATMA*.

Evelyn Gius, Carla Sökefeld, Lea Dümpelmann, Lucas Kaufmann, Annekea Schreiber, Svenja Guhr, Nathalie Wiedmer, and Fotis Jannidis. 2021b. *Guidelines for detection of scenes*.

Evelyn Gius and Michael Vauth. 2022. *Towards an event based plot model. a computational narratology approach*. *Journal of Computational Literary Studies*, 1(1).

- Joshua Hagedorn and Sándor Darányi. 2022. [Bearing a Bag-of-Tales: An Open Corpus of Annotated Folktales for Reproducible Research](#). *Journal of Open Humanities Data*.
- Hans Ole Hatzel, Evelyn Gius, Haimo Stierner, and Chris Biemann. 2023. [Narrativität und Handlung: Zum Verhältnis von Handlungszusammenfassungen und relevanten Ereignissen](#). In *DHd2023: Open Humanities, Open Culture*, Trier, Germany/Belval, Luxembourg. Zenodo.
- Hans Ole Hatzel, Haimo Stierner, Evelyn Gius, and Chris Biemann. 2025. Scalable Verb-Based Literary Semantics. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2025)*.
- Nora Ketschik. 2025. [Netzwerkanalysen narrativer Texte - ein Vorgehensmodell](#). In *DHd 2025: Under Construction*, Bielefeld, Germany. Zenodo.
- Markus Krug. 2020. [Techniques for the automatic extraction of character networks in german historic novels](#). Released Online via OPUS Würzburg, accessed 2025-10-11.
- Thorsten Müller and Dominik Kreutz. 2021. [Thorsten - open german voice \(neutral\) dataset](#).
- Ani Nenkova and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Jinfu Ni, Yoshinori Shiga, and Hisashi Kawai. 2019. [Duration Modeling with Global Phoneme-Duration Vectors](#). In *Proc. Interspeech 2019*, pages 4465–4469, Graz, Switzerland.
- Vincenzo Perri, Lisi Qarkaxhija, Albin Zehe, Andreas Hotho, and Ingo Scholtes. 2022. One Graph to Rule them All: Using NLP and Graph Neural Networks to analyse Tolkien’s Legendarium. In *Proceedings of the Computational Humanities Research Conference 2022*, volume 3290 of *CEUR Workshop Proceedings*, pages 291–317, Antwerp, Belgium. CEUR.
- Fynn Petersen-Frey and Chris Biemann. 2024. Fine-grained quotation detection and attribution in German news articles. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 196–208, Vienna, Austria. Association for Computational Linguistics.
- Axel Pichler and Nils Reiter. 2020. [Reflektierte Textanalyse](#). In *Reflektierte algorithmische Textanalyse: Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*, pages 43–60. De Gruyter.
- Andrew Piper and Sunyam Bagga. 2022. Toward a Data-Driven Theory of Narrativity. *New Literary History*, 54(1):879–901.
- Ina Roesiger, Sarah Schulz, and Nils Reiter. 2018. Towards Coreference for Literary Text: Analyzing Domain-Specific Phenomena. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 129–138, Santa Fe, New Mexico. Association for Computational Linguistics.
- Christof Schöch, Roxana Patras, Tomaž Erjavec, and Diana Santos. 2021. [Creating the European Literary Text Collection \(ELTeC\): Challenges and Perspectives](#). *Modern Languages Open*, 0(1):25–44.
- Haimo Stierner, Hans Ole Hatzel, Chris Biemann, and Evelyn Gius. 2025. [Pause im Text. Zur Exploration semantisch konditionierter Sprechpausen in Hörbüchern](#). In *Book of Abstracts - DHd 2025*, Bielefeld, Germany.
- Éva Székely, Joao Paulo Cabral, Mohamed Abou-Zleikha, Peter Cahill, and Julie Carson-Berndsen. 2012. Evaluating expressive speech synthesis from audiobook corpora for conversational phrases. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 3335–3339, Istanbul, Turkey. European Language Resources Association (ELRA).
- TextGrid. 2021. Die digitale bibliothek bei textgrid. <https://textgrid.de/de/digitale-bibliothek>.
- Ted Underwood. 2018. Why Literary Time is Measured in Minutes. *ELH*, 85(2):341–365.
- Michael Vauth and Evelyn Gius. 2021. [Richtlinien für die Annotation narratologischer Ereigniskonzepte](#). Zenodo.
- Michael Vauth, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann. 2021. Automated Event Annotation in Literary Texts. In *CHR 2021: Computational Humanities Research Conference*, volume 2989 of *CEUR Workshop Proceedings*, pages 333–345, Amsterdam, The Netherlands.
- Michael Vauth, Malte Meister, Hans Ole Hatzel, Dominik Gerstorfer, and Evelyn Gius. 2022. [GitMA](#).
- Millicent Weber. 2021. “Reading” the Public Domain: Narrating and Listening to Librivox Audiobooks. *Book History*, 24(1):209–243.
- Albin Zehe, Leonard Konle, Lea Katharina Dümpelmann, Evelyn Gius, Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, Annekea Schreiber, and Nathalie Wiedmer. 2021. Detecting Scenes in Fiction: A new Segmentation Task. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3167–3177, Online. Association for Computational Linguistics.

Sense-Based Annotation of Geographical Nouns in Ancient Greek and Latin: A Diachronic Study with LLMs

Andrea Farina¹, Michele Ciletti², Barbara McGillivray¹, Andrea Ballatore¹

¹Department of Digital Humanities, King’s College London, United Kingdom

²Department of Humanities, University of Foggia, Italy

Correspondence: andrea.farina@kcl.ac.uk

Abstract

This paper investigates the lexicalisation of geographical nouns in Latin and Ancient Greek using a diachronic, multi-genre corpus (8th cent. BCE – 2nd cent. CE) and Large Language Models for Word Sense Disambiguation. We focus on two main aspects: the onomasiological question of which words encode core geographical concepts, and the semasiological distribution of senses across lemmas. Across both languages, city-related concepts are the most frequently expressed, but Greek shows a stronger focus on maritime terms, whereas Latin favours concepts related to land. Semasiologically, Latin shows clearer evidence of semantic change over time (e.g., *civitas* ‘citizenship’ → ‘city’, *aequor* ‘flat surface’ → ‘sea’), while Greek displays more gradual or distributed shifts. These results show that computational annotation enables cross-linguistic and diachronic analysis of geographical semantics, allowing us to compare the frequency of concepts across languages, genres, and periods, and to track when semantic change occurs and how core concepts evolve over time.

1 Introduction

Language technologies have opened new avenues for studying cultural heritage texts, enabling large-scale analyses of phenomena traditionally explored qualitatively. One such phenomenon is the lexicalisation of geographical concepts, which is central to understanding how ancient societies conceptualised space. This paper investigates the distribution and semantic disambiguation of common nouns referring to geographical entities in Latin and Ancient Greek using computational approaches. Our work adopts an onomasiological perspective, asking: given a concept, which words express it, and how does this set of words evolve across time and genres? We leverage Large Language Models (LLMs) to annotate geographical expressions and perform Word Sense Disambiguation (WSD),

testing whether temporal metadata improves performance. With a comparison between Latin and Ancient Greek, we aim to gain insights into cross-linguistic tendencies in geographical semantics and contribute to the broader study of diachronic conceptual change.

2 Related Work

Research on historical lexical semantics distinguishes two complementary perspectives: semasiology (from words to concepts) and onomasiology (from concepts to words). Semasiology asks, given a word, which meanings it expresses – for example, Latin *civitas* can mean ‘citizenship’, ‘the citizens’, or ‘a city’. Onomasiology asks, given a concept, which words encode it – for instance, <CITY> in Latin may be *urbs*, *oppidum*, *civitas*, or *metropolis*. These relations vary across genres, registers, and time, shaped by cognitive, social, and contextual factors (Meillet, 1958; Blank, 2001). Semasiological studies focus on polysemy and sense-frequency trajectories (Haspelmath, 1999; Geeraerts, 1997), while onomasiological analyses track concept lexicalisation and synonym evolution (Bréal, 1897; Lehrer, 1985). While qualitative studies have documented such phenomena for Latin and Greek (Haverling, 2012; Moussy, 1965; Fruyt, 1994; Torrego, 2020; Williams, 1976; Lehrer, 1985; Stolova, 2015; Adams, 2013), with a few exceptions (Farina, 2023) large-scale quantitative analyses are missing due to the lack of sense-annotated diachronic corpora.

Recent NLP research has introduced automatic methods for semantic change detection based on distributional semantics (Hamilton et al., 2016; Kutuzov et al., 2022). These approaches model words as vectors in high-dimensional spaces and track their shifts across time-defined subcorpora. While effective for modern languages, they typically treat lexemes as atomic units, ignoring sense-level vari-

ation and genre effects. For historical languages, preliminary experiments on WSD using contextual embeddings (Latin BERT) achieved promising results (Bamman and Burns, 2020; Lendvai and Wick, 2022; Ghinassi et al., 2024; Mercelis et al., 2025), but coverage and temporal sensitivity remain limited. Lendvai and Wick (2022) fine-tuned Latin BERT on data from the *Thesaurus Linguae Latinae*, achieving substantial gains over static embeddings. Ghinassi et al. (2024) proposed a language-pivoting approach that propagates sense annotations from English to Latin via parallel corpora. Kaše et al. (2025) explored transformer-based embeddings and attention mechanisms for WSD using dictionary senses.

On the onomasiological side, resources such as WordNet (WN) (Fellbaum, 1998) and its Latin and Ancient Greek counterparts (Minozzi, 2010) provide structured sense inventories, yet they have not been systematically linked to diachronic corpora. This prevents quantitative analyses of how concepts were lexicalised over time. Existing diachronic corpora annotated with WN senses for other languages (e.g., Bulgarian, Dutch, Slovenian) were generated automatically without manual curation and cover only a few centuries (Martelli et al., 2022).

In this context, WN-style sense inventories provide a principled framework for operationalising an onomasiological perspective by anchoring lexical items to shared conceptual units (Fellbaum, 1998) (Section 3). WN has been successfully employed for – often cross-linguistic – semantic and lexical studies both on modern (Gonzalo et al., 1998; Burgun and Bodenreider, 2001; Apidianaki and Sagot, 2012; Rudnicka et al., 2022; Klimczak et al., 2024) and historical (Biagetti et al., 2023; Khan et al., 2023; Brigada Villa et al., 2025) languages, and it therefore offers a solid basis for constructing comparable sense inventories for historical languages.

Our study focuses on geographical nouns (GNs), intended as common nouns denoting geographical entities (as opposed to toponyms), a lexically and conceptually salient domain that is central to both narrative structure and cultural conceptualisation in Ancient Greek and Latin. Annotating GNs in historical texts is challenging because of diachronic change and sparse data, yet it is essential for linking lexical mentions to conceptual and real-world geographical domains (Erdmann et al., 2016; Beersmans et al., 2023; Kenyon et al., 2023). GNs attract attention in ethnophysiology and cognitive linguistics as evidence of how cultures

conceptualise the natural and human-made environments, and are generally characterised by high cross-linguistic and cross-cultural variability (Burenhult and Levinson, 2008; Mark et al., 2011; Falcinelli et al., 2024). Compared to previous studies, this is the first computational study on two historical languages combining both the semasiological and the onomasiological perspectives, and employing LLMs, specifically targeted towards GNs.

3 Research Questions

Our study integrates sense-annotated diachronic resources for historical languages by combining onomasiological and semasiological perspectives with recent advances in LLMs. To operationalise this approach, geographical meanings are anchored in WN sense inventories, where each synset (synonym set) represents a concept shared across lexical items. Linking lexical forms to synsets allows us to analyse both how concepts are lexicalised across words and how words map onto multiple senses in a sense-aware way. This framework also allows for a principled integration of computational annotation: LLMs are applied to automatically disambiguate senses in a diachronically diverse corpus including multiple literary genres (Section 4), providing both linguistic and quantitative data on the lexicalisation of geographical concepts. Building on this framework, we focus on two main research questions (RQs):

RQ1: LLM-based annotation of geographical common nouns How accurately can LLMs annotate geographical senses of common nouns in Ancient Greek and Latin texts? How does model performance vary across languages, centuries, and literary genres?

RQ2: Onomasiological patterns of geographical lexicalisation How are geographical concepts (e.g., <CITY>, <SEA>) lexicalised through common nouns in Ancient Greek and Latin? How do the distributions of lemmas associated with the same geographical sense change diachronically and across literary genres? To what extent do these languages exhibit convergent/divergent onomasiological patterns?

4 Corpus Overview

The REMOVE Base Corpus (Farina, 2026) is a diachronic and multi-genre collection of Ancient Greek and Latin texts originally developed for a

manually-curated comparative analysis of prefixed verbs of motion. The corpus was explicitly designed to enable cross-linguistic comparison by ensuring balanced coverage across time periods, genres, and authors in both languages. While originally constructed for the study of motion verbs, its comparable design and systematic sampling make it equally suitable for investigating GNs and other geographical expressions (Farina et al., 2025). As is the case for all historical languages, the fragmentary survival of texts and their uneven transmission prevent the construction of a fully representative corpus (McEnery and Wilson, 2001). The corpus is therefore conceived as a controlled and purpose-oriented sample, structured to support comparative investigations of lexical and semantic phenomena.

The corpus builds on a subset of the Ancient Greek and Latin Dependency Treebank (AGLDT) (Bamman and Crane, 2007, 2011), supplemented with additional works to improve chronological continuity and genre coverage. It comprises 35 texts (19 Ancient Greek, 16 Latin) from the 8th cent. BCE to the 2nd cent. CE, covering major linguistic phases – Archaic, Classical, Hellenistic, and Imperial Greek (Charalambakis, 1996; Babiniotis, 2002; Ralli, 2012) and Early, Classical, and Post-Classical Latin (Cuzzolin and Haverling, 2009). Texts span historiography, poetry, philosophy, theatre, oratory, and the novel (labels from Biagetti et al., 2021), allowing us to evaluate whether the semantic behavior of place nouns varies across registers, including figurative or freer-word-order contexts.

The full corpus contains 541,620 tokens (punctuation excluded), divided almost evenly between Ancient Greek (275,913 tokens; 50.9%) and Latin (265,707 tokens; 49.1%). Historiography is the most represented genre (approximately 44% of Greek tokens and 36% of Latin tokens), followed by poetry (around 29% in both languages). For a more detailed breakdown of authors, works, centuries, and genres we refer to Appendix A and Farina (2026).

5 Annotation Pipeline

The selection of target GNs is grounded in the place vocabulary by Ballatore (2015). This is a structured collection of nouns describing both natural (e.g., rivers, mountains) and human-made (e.g., cities, roads) places. In addition to lexical coverage, this

resource associates each entry with GeoNames¹, DBpedia² and WN (Bond and Bond, 2019). Although the resource is English-based, it offers a theoretically informed and systematically curated starting point for defining a cross-linguistic set of place nouns and for grounding their semantic interpretation.

Starting from this English inventory, candidate Latin equivalents were obtained through a combination of dictionary-based translation for well-attested and semantically central terms, and automated generation for less frequent or contextually complex ones. Primary lexicographic support for direct translation was provided by a suite of classic English-Latin dictionaries,³ including Smith and Hall (1871), Döderlein (1874), Lewis and Short (1879), and Ogilvie (1901), which together cover a broad range of core vocabulary and synonymic nuances in Latin. Candidate lemmas were then manually reviewed, discarding entries that were semantically inappropriate, exhibited excessive ambiguity with respect to the intended geographical meanings, or were anachronistic (e.g., Eng. *airport*).

This procedure resulted in an initial list of 227 Latin lemmas. To ensure cross-linguistic comparability, these lemmas were then mapped to their Ancient Greek counterparts using the digitised version of Schrevel (1832). Latin entries for which no suitable Ancient Greek equivalent could be identified were excluded. The final inventory comprises 227 place-noun lemmas attested in Latin and 264 lemmas attested in Ancient Greek, which form the basis for the automatic annotation of this study.

Our data was obtained by extracting each token corresponding to an identified Latin or Ancient Greek place-noun lemma from the REMOVE Base Corpus. The source texts were lemmatised with latinCy (Burns, 2023) and greCy (Myerston, 2024). Then, for each matching token, we stored its lemma, its context (the sentence it occurs in), additional metadata (author, language, passage reference), and gave it a unique identifier. The lemmatisation process accounted for dialect variation (e.g., *thálassa* vs *thálatta*) by linking variants to a stable form.

¹<https://www.geonames.org/>.

²<https://www.dbpedia.org/>.

³<https://latinitium.com/latin-dictionaries/>.

6 GNs in the REMOVE Base Corpus

Lemma Frequencies Before considering the LLM-based annotation, it is informative to examine the raw distribution of place-noun lemmas in the REMOVE Base Corpus. The extraction procedure (Section 5) yielded 16,681 occurrences (Table 1), of which 252 had already been manually annotated and serve as our validation set (Section 8). Of the remaining 16,429 occurrences, 8,555 are from Greek and 7,874 from Latin. This roughly balanced distribution provides a solid foundation for cross-linguistic comparisons of geographical terminology and contextualises the frequency patterns of lemmas prior to any automated sense annotation.

Table 1: Dataset distribution

	Latin	Ancient Greek	Total
Unlabelled data	7,874	8,555	16,429
Validation set	106	146	252
Total	7,980	8,701	16,681

Table 2 presents the ten most frequent lemmas for each language. The frequency data for Greek and Latin lemmas shows a strong focus on urban, territorial, and natural features, though with some language-specific emphases. In Greek, the most frequent lemmas are *pólis* ‘city’ (1,537 occurrences), *thálassa* ‘sea’ (384), *archē* ‘(ruled) territory’⁴ (373), *nēsos* ‘island’ (338), and *teíkhos* ‘wall’ (304). The least frequent Greek lemmas include specialised or context-specific terms such as *bórboros* ‘mud’, *domē* ‘building’, *kamára* ‘vaulted chamber’, *kapēleíon* ‘tavern’, and *xenōn* ‘guest chamber’. Latin shows a broadly similar pattern with the most frequent lemmas being *urbs* ‘city’ (721), *domus* ‘house’ (482), *terra* ‘land’ (310), *civitas* ‘community, city’ (271), and *ager* ‘field’ (261), while the least frequent terms include *Oceanus* ‘Ocean’, *ager publicus* ‘public land’, *apotheca* ‘storehouse’, *canalis* ‘channel’, and *zona* ‘zone’, reflecting more specialised contexts.

Groupings Examining groups of GNs reveals interesting parallels and differences between Greek and Latin (Table 3). In both languages, urban centres (Gr. *pólis*, *ástu*, Lat. *urbs*, *civitas*) dominate frequency counts, followed by natural features such

⁴The main meanings of *archē* are ‘origin’ and ‘power’, which accounts for its high overall frequency. The semantic annotation (Section 8) was used to isolate those instances in which *archē* functions as a GN.

Greek		Latin	
Lemma	#	Lemma	#
<i>pólis</i> ‘city’	1,537	<i>urbs</i> ‘city’	721
<i>thálassa</i> ‘sea’	384	<i>domus</i> ‘house’	482
<i>archē</i> ‘(ruled) area’	373	<i>terra</i> ‘land’	310
<i>nēsos</i> ‘island’	338	<i>civitas</i> ‘city’	271
<i>teíkhos</i> ‘wall’	304	<i>ager</i> ‘land’	261
<i>khôra</i> ‘region’	299	<i>regnum</i> ‘kingdom’	252
<i>nomós</i> ‘district’	259	<i>via</i> ‘street’	210
<i>potamós</i> ‘river’	239	<i>finis</i> ‘border’	206
<i>khōrion</i> ‘region’	213	<i>aedes</i> ‘house, temple’	200
<i>oikos</i> ‘house’	202	<i>provincia</i> ‘province’	181

Table 2: Most frequent GNs in Greek and Latin.

as seas and rivers (Gr. *thálassa*, *pélagos*, *póntos*, *háls*, *potamós*, Lat. *mare*, *pelagus*, *flumen*), mountains (Gr. *óros*, Lat. *mons*), and fields or land (Gr. *pedíon*, *khthôn*, Lat. *ager*). Greek texts show a relatively higher occurrence of maritime terms (*thálassa*, *pélagos*, *háls*, *póntos* sea’, *nēsos* island’), which is consistent with the historically central role of the sea in Greek civilisation, both for trade (Reed, 2003), colonisation (Nash, 2018), and cultural exchange (Lindenlauf, 2004; Beaulieu, 2016). Latin, by contrast, emphasises inland features, administrative regions, and cultivated land (*terra* ‘earth, land’, *ager* ‘land’, *provincia* ‘province’, *regnum* ‘kingdom’), reflecting the Roman focus on land (Roselaar, 2010; Zuiderhoek, 2015).

Diachrony When normalising for total tokens per century, the distribution of place-related vocabulary in Greek and Latin texts becomes clearer. In Greek, the most evident difference lies in the chronological gap between the Homeric lexicon (8th cent. BCE) and Classical Greek (5th cent. BCE). The most frequent place name in Homer is *dōma* ‘house’, dominating with roughly 22.8 occurrences per 10,000 tokens, followed by *póntos* ‘sea’ (17.4) and *oikos* ‘house’ (13.7). This distribution also reflects the composition of the corpus: the Homeric material for the 8th cent. BCE includes a portion of the *Odyssey* (Appendix A), which features many scenes set around houses and numerous references to the sea due to Odysseus’ journey.

In the 5th cent. BCE, *pólis* ‘city’ becomes extremely frequent at 98.7 per 10,000 tokens, followed by *teíkhos* ‘wall’ and *thálassa* ‘sea’ at only 23.3 and 19.8 per 10,000 tokens. Notably, the terminology for ‘sea’ changes by this period: *thálassa* is now preferred over *póntos* (Section 9.2). As for *pólis*, it remains consistently among the top three most frequent words up until the 2nd cent. CE.

Semantic Group	Greek (occurrences)	Latin (occurrences)
City / Town	<i>pólis</i> ‘city’ (1537), <i>ástu</i> ‘town’ (111), <i>agorá</i> ‘marketplace’ (108), <i>oikía</i> ‘house’ (80), <i>hierón</i> ‘temple’ (91)	<i>urbs</i> ‘city’ (721), <i>civitas</i> ‘community, city’ (271), <i>oppidum</i> ‘town’ (131), <i>vicus</i> ‘village’ (50), <i>colonia</i> ‘colony’ (46)
Sea / River	<i>thálassa</i> ‘sea’ (384), <i>potamós</i> ‘river’ (239), <i>pélagos</i> ‘sea’ (75), <i>póntos</i> ‘sea’, <i>háls</i> ‘sea’ (10), <i>ōkeanós</i> ‘ocean’ (24)	<i>mare</i> ‘sea’ (178), <i>flumen</i> ‘river’ (154) <i>pelagus</i> ‘sea’ (49), <i>pontus</i> ‘sea’ (66), <i>oceanus</i> ‘ocean’ (31)
Mountain / Hill	<i>óros</i> ‘mountain’ (140), <i>hóros</i> ‘mountain’ (140), <i>lóphos</i> ‘hill’ (53), <i>líthos</i> ‘rock’ (52), <i>skópelos</i> ‘peak’ (13)	<i>mons</i> ‘mountain’ (160), <i>saxum</i> ‘rock’ (116), <i>collis</i> ‘hill’ (44), <i>rupes</i> ‘peak’ (31), <i>scopulus</i> ‘peak’ (45)
Land / Territory	<i>khthōn</i> ‘land’ (102), <i>pedíon</i> ‘plain’ (87), <i>agrós</i> ‘field’ (37), <i>erēmía</i> ‘desert’ (64), <i>khorós</i> ‘region’ (299)	<i>terra</i> ‘earth,land’ (310), <i>ager</i> ‘field’ (261), <i>campus</i> ‘plain’ (133), <i>regio</i> ‘region’ (83), <i>patria</i> ‘territory’ (141)
Wall / Fort	<i>teĩkhos</i> ‘wall’ (304), <i>phulakē</i> ‘garrison’ (128), <i>púrgos</i> ‘tower’ (49)	<i>murus</i> ‘wall’ (148), <i>moenia</i> ‘wall’ (144), <i>praesidium</i> ‘garrison’ (103), <i>castellum</i> ‘fort’ (16)

Table 3: Frequency of major semantic groups for GNs in Greek and Latin.

Latin shows a similar pattern. In the 3rd cent. BCE, *aedes* ‘house, temple’ and *domus* ‘house’ are the most frequent, with 44.4 and 23.4 per 10,000 tokens respectively. Across all centuries, *urbs* ‘city’ and *domus* ‘house’ remain extremely frequent. These distributions are also shaped by genre (see below): for instance, the early Latin texts in our corpus are dominated by comedies, which often emphasise domestic settings, explaining the prominence of house-related vocabulary.

Genre In Greek, *pólis* ‘city’ largely dominates historiography with 101.2 occurrences per 10,000 tokens, far surpassing any Latin counterpart, where the top word *aedes* ‘house, temple’ reaches only 44.4 per 10,000 tokens in the 3rd cent. BCE. This prominence of *pólis* may reflect the centrality of the concept of <CITY> in the Greek history: city-states (e.g., Athens, Sparta) were fundamental for history of Greece, so historiographical narratives often revolve around individual cities, their governance, and their interactions (Rood, 2001; Azevedo, 2010; Stefanovski and Čavoški, 2023; Harris and Lewis, 2024). By contrast, in Roman historiography, *urbs* primarily refers to Rome itself (Witcher, 2005), while other words such as *oppidum*, *castellum*, or *vicus* denote more specific types of settlements (Becker, 2008), resulting in a more distributed lexical pattern for urban terms.

Other Greek words show strong genre specialisation: *nomós* ‘district’ is highly frequent in oratory (78.9 per 10,000 tokens), *thálassa* ‘sea’ and *teĩkhos* ‘wall’ in historiography (22.7–23.1 per 10,000 tokens), and *dōma* ‘house’ and *póntos* ‘sea’ in poetry (14.4–15.3 per 10,000 tokens), whereas philosophy prefers terms like *arkhē* (41.8 per 10,000 tokens) – often employed with its primary sense of ‘beginning, authority’ – and *pólis* (39.0 per 10,000

tokens). In Latin, the picture is broadly similar but less extreme: *aedes* and *domus* dominate across genres, *urbs* ‘city’ appears frequently in oratory (66.2 per 10,000 tokens) and historiography (42.2 per 10,000 tokens), and *civitas* ‘community’ is most frequent in historiography (22.9 per 10,000 tokens).

Some correspondences emerge: both Greek and Latin historiography privileges words for ‘city’ and ‘walls’ (Gr. *pólis*, Lat. *urbs*; Gr. *teĩkhos*, Lat. *murus*), while poetry in both languages seems to favour natural and domestic space (Gr. *dōma*, *póntos*, *nēsos*; Lat. *domus*, *terra*, *saxum*). Greek exhibits higher lexical peaks and more pronounced genre-specific distributions, suggesting greater stylistic and lexical flexibility, whereas Latin shows more even frequencies across genres.

Specialised terms, such as Gr. *stratōpedon* ‘camp’, *diéxodos* ‘passage’, or *múlē* and Lat. *praesidium* ‘camp’, *palus* ‘marsh’, or *balneum* ‘bath’, are concentrated in historiography or poetry. This pattern can be explained on two grounds: in historiography, the more technical or administrative nature of the discourse naturally calls for precise terminology, so specialised terms appear frequently (Ampolo et al., 2004; Cardinali, 2017); in poetry, the use of rare words may reflect particular imagery or expressive nuance, meter constraints, or a desire to avoid repetition (Ferri, 2011).

7 Experimental Setup

During our automatic annotation procedure, we tasked an LLM with assigning a WN synset to each token (Section 3). We selected GPT-5.2 (with medium reasoning effort) as our annotator. This is the most advanced iteration of the GPT-5 family and demonstrated strong potential for Latin and An-

cient Greek semantic annotation among 13 model families in Farina and Ciletti (2026). Sample tests confirmed its competence in handling GNs, and we avoided models marked as deprecated by their creators (such as GPT-5) to ensure reproducibility.

For each Latin token, we retrieved the WN synsets associated with its lemma via the Latin WordNet (LWN) API⁵. Additionally, we retrieved all synsets associated with its English counterpart(s) in the Open English WordNet (OEWN) (McCrae et al., 2020) based on our initial mapping, to ensure maximum semantic coverage. For Ancient Greek, whose WN is still under construction (Marchesi et al., 2025), we followed a similar procedure: each token was mapped to its corresponding Latin and English lemma(s), and candidate synsets were stored from both the LWN and OEWN. The model input included the target token, its lemma, context, and the list of candidate synsets. The synsets were presented with glosses and customised IDs in a randomised order, stripped of any information regarding their source language. The prompt also included detailed instructions (Appendix B) and up to five few-shot examples of the same lemma from the validation set. The LLM was instructed to return the most appropriate synset, or to explicitly reject all candidates if none were accurate, along with a binary classification on whether the meaning was literal (yes) or metaphorical/metonymic (no), and a confidence score (0–1).

8 Results

We evaluated the model against our gold standard of 252 annotated tokens (Section 6). Across all tokens, including cases where the model abstained (NA), it achieved a precision of 0.583, recall of 0.966, and F1 of 0.727, reflecting the high frequency of predictions but the challenge of fine-grained sense disambiguation (see below). The model predicted a sense for 98.0% of tokens, and on these committed predictions, the precision was 0.583. As this metric considers only tokens with predictions, precision, recall, and F1 are identical, reflecting accuracy when the model commits. The mean confidence score given by the model itself was 0.85 – high even for incorrect predictions, but proportionally higher for correct annotations –, and 94.7% of tokens were tagged as literal, reflecting their status as GNs.

⁵<https://latinwordnet.exeter.ac.uk/api>.

Languages Breaking down performance by language (Table 4), Latin tokens achieved slightly higher precision on committed predictions (0.594), while Greek tokens scored lower (0.574). Inter-annotator agreement between human and the model’s annotation, measured with unweighted Cohen’s Kappa, also varied slightly: Latin achieved 0.562 and Greek 0.542, indicating moderate agreement in both languages.

These results suggest that the model is slightly more conservative in Latin, avoiding some false positives, while Greek tokens – despite higher coverage and more examples per token – show marginally lower precision. Overall, literal GNs are reliably annotated, whereas metaphorical or less frequent senses remain challenging, highlighting the limits of current LLM-based WSD for fine-grained WN synsets.

Language	Precision
Greek	0.574
Latin	0.594
All	0.583

Table 4: Model precision by language and overall.

Previous Latin WSD studies report higher F1 scores on standard benchmarks (Ghinassi et al., 2024: 0.618; Lendvai and Wick, 2022: 0.794; Kaše et al., 2025: 0.68). We compare our precision on committed predictions to their F1 because, in the committed subset, precision, recall, and F1 are identical. Comparable LLM-based approaches to semantic analysis using English meanings rather than WordNet synsets achieved F1 scores of 0.596 for Latin and 0.666 for Greek (Farina and Ciletti, 2026).

Centuries and Genres We also examined model performance across centuries and macro-genres (prose/poetry) to assess temporal and stylistic variation (Table 5). Overall, Latin achieves higher precision values across most periods and shows a relatively homogeneous pattern, although data for the 3rd century BCE and 2nd century CE are very limited. Greek is also fairly homogeneous, with peaks in the 8th and 5th centuries BCE; these likely reflect the higher volume of texts from these periods rather than systematic differences in model performance.

Macro-genres were used because coverage is highly uneven across more specific genres. In Latin, poetry exhibits higher precision than prose (0.692 vs. 0.563), whereas Greek is more homogeneous

across poetry and prose (0.588 vs. 0.570). This suggests that the model handles Greek texts relatively consistently across genres, while in Latin the differences between poetry and prose may reflect the particular types of GNs present in poetry within the REMOVE Base Corpus.

Lang.	Genre	Precision
Greek	Poetry	0.588
Greek	Prose	0.570
Latin	Poetry	0.692
Latin	Prose	0.563

Table 5: Model performance by macro-genre.

Qualitative Insights An analysis of the synsets with the highest precision reveals clearer patterns. After excluding single-occurrence synsets, 12 achieved a perfect precision (1.0), while five others fell within a moderate range of 0.625–0.928. All other synsets failed to score higher than 0.375. The highest-performing synsets generally benefit from distinct glosses and low semantic ambiguity. For instance, 02901994-n “a structure that allows people or vehicles to cross an obstacle such as a river or canal or railway etc.”, was unequivocally assigned to tokens linked to Eng. *bridge* (Lat. *pons*). In general, high-scoring synsets are related to the following GNs: human-made places (‘camp’, ‘temple’, ‘house’), water masses (‘lake’, ‘river’), and land masses (‘land’).

Most discrepancies appeared to stem from issues of granularity (Burgun and Bodenreider, 2001; Bond and Piasecki, 2017; Kafe, 2017; McGillivray et al., 2023). WN synsets can be highly specific, and multiple similar senses may exist for a single lemma. Consequently, many annotation divergences reflected slightly different interpretations rather than outright errors. Consider, for instance, synsets 08542298-n “a large and densely populated urban area; may include several independent administrative districts” (Eng. *city*) and 08683242-n “an urban area with a fixed boundary that is smaller than a city” (Eng. *town*). Precision for 08542298-n was high (0.815), while for 08683242-n it dropped to 0.375. A closer examination shows that in all misclassified cases the model predicted 08542298-n instead of 08683242-n. This reflects the fact that, in some passages, Lat. *urbs* and Gr. *pólis* – normally denoting larger urban centres – actually refer to smaller settlements, but the model defaulted to the more common, prototypical sense for these lemmas.

9 Discussion

In this section, we outline general tendencies in the semantic development of place-related lemmas in Latin and Greek, before turning to two case studies in Sections 9.1 and 9.2. Although the overall precision of the automatic annotation is moderate (Section 8), the case studies focus on high-frequency concepts for which the model achieves a precision of 0.660 (i.e., <CITY> and <SEA>). A closer examination of the 18 errors shows that in 13 cases, apparent mistakes stem solely from the fine granularity of WN synsets, with only 5 representing genuine misannotations. Some of these apparent errors involve, for example, synsets for towns being annotated as cities and vice versa (Section 8); since these still denote closely related concepts, they do not affect our analysis, which examines how these concepts are lexicalised and whether some lemmas undergo genuine semantic shifts across distinct meanings.

In Latin, the data show that certain lexical items undergo noticeable shifts in dominant meaning over time, often reflecting cultural and functional changes in the environment. For instance, *aedes* initially denotes a dwelling (42 per 10,000 occurrences in the 3rd cent. BCE) but gradually acquires a prominent secondary meaning as a temple or place of worship (0.94 per 10,000 occurrences in the 3rd cent. BCE; 4.8 per 10,000 occurrences in the 1st cent. BCE; 15.4 per 10,000 occurrences in the 2nd cent. CE), illustrating a narrowing of sense toward religious buildings. Similarly, *moenia*, originally a general term for architectural partitions, increasingly refers specifically to defensive city walls, while *villa* shows a shift from extensive rural estates to domestic dwellings, with residual usage for large properties.

Greek place-related lemmas, by contrast, tend to display more gradual and distributed changes across coexisting senses, rather than abrupt reorganisations of dominant meaning. Semantic evolution often manifests as moderate redistributions of lexical prominence, suggesting slower or more incremental shifts in sense hierarchy. While the Greek data do not show the sharp transitions observed in Latin, this apparent stability may partly result from the temporal resolution of the available corpus (8th cent. BCE – 2nd cent. CE): longer diachronic windows might reveal more substantial reorganisation of sense dominance (Section 10).

9.1 The concept of <CITY>

The concept of <CITY> provides a productive case study for examining semasiological organisation and diachronic distribution in Greek and Latin. Quantitatively, Greek exhibits a higher overall lexical dispersion for the concept, with 1,585 tokens distributed across several lemmas, compared to 1,057 tokens in Latin. In Greek, the lexicalisation of <CITY> is strongly dominated by *pólis* (Section 6), which alone accounts for 1,457 tokens, while *ástu* (111 tokens) and *políteuma* (17 tokens) play marginal and more specialised roles. Latin, by contrast, shows a more balanced lexical system: although *urbs* is the most frequent term (693 tokens), it is complemented by *civitas* (222 tokens) and *oppidum* (121 tokens), with minor contributions from *patria* (14 tokens) and *municipium* (7 tokens).

This distribution reflects semantic differentiation rather than lexical competition. In Greek, *pólis* is the unmarked term for the city, *ástu* denotes the citadel or built core, and *políteuma* refers to the civic body rather than urban space. Latin shows a similar but less polarised pattern: *urbs* is the general urban term, *civitas* overlaps with ‘city’ and ‘citizenship’, and *oppidum*, *patria*, and *municipium* mark more specific urban notions (McGillivray and Nowak, 2025).

From a semantic perspective, both languages cover comparable conceptual domains, but distribute them differently across their lexicons. In Greek, 08542298-n “large and densely populated urban area” (897 tokens) and 08185877-n “politically organized body of people under a single government” (593 tokens) are both largely subsumed under *pólis*. Conversely, Latin displays a sharper semantic partitioning: while 08542298-n is predominantly encoded by *urbs* (731 tokens), the political sense is primarily associated with *civitas* (202 tokens), and smaller settlements (08683242-n “an urban area with a fixed boundary that is smaller than a city”) are lexicalised with *oppidum* (124 tokens). Genre distribution reveals a contrast: in both languages historiography dominates quantitatively, but Greek <CITY> terminology remains overwhelmingly centred on a single lemma across genres, while Latin exhibits greater lexical and semantic diversification.

Semasiologically, the Latin lemma *civitas* emerges as a particularly revealing example of semantic shift within a system that explicitly differentiates between civic community (‘citizenship’)

and urban space (‘city’). In the present corpus, this semantic shift becomes more evident from the 1st cent. BCE onward, when the initially dominant sense of *civitas* as a politically organised body of citizens progressively redistributes toward the sense of a densely populated urban area (Figure 1). This development is a case of metonymic extension, whereby the collective entity of citizens comes to stand for the physical space they inhabit.

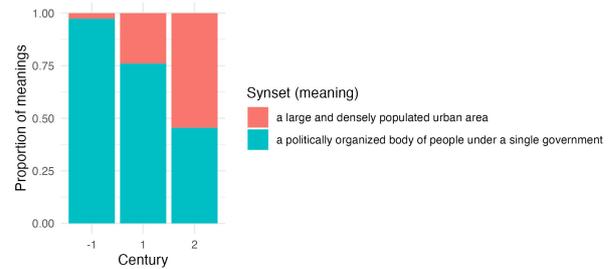


Figure 1: Semantic shift of Lat. *civitas*.

9.2 The concept of <SEA>

The concept of <SEA> is lexicalised in both Greek and Latin by four etymologically distinct lemmas. In Greek, it is expressed through *thálassa*, *póntos*, *pélagos*, and *háls*,⁶ collectively covering two similar synsets: 09368829-n “any very large body of (salt) water” (288 tokens) and 09449666-n “a division of an ocean or a large body of salt water partially enclosed by land” (225 tokens). *Thálassa* dominates the corpus with 307 tokens (ca. 60% of all occurrences), followed by *póntos* (139 tokens), *pélagos* (60), and *háls* (7).

At the genre level, *thálassa* is mostly employed in historiography (273 tokens), while *póntos* in epic poetry, probably reflecting their etymologies. This preference is statistically confirmed by a Chi-squared test ($\chi^2 = 727.83$, $p < .001$). The preference for *póntos* in epic does not imply that *thálassa* is absent from Homeric poetry. Rather, qualitative inspection of the contexts shows a systematic semantic contrast. *Thálassa* tends to denote the sea as an unmarked geographical setting, without an explicit focus on movement or traversal, as in (1), where it simply designates the maritime landscape. By contrast, *póntos* typically occurs in contexts of motion, journeying, or crossing, foregrounding the sea as a route or medium of transport, as in (2).

⁶The lemmas differ etymologically (Beekes, 2009): *thálassa* is the unmarked term for the sea; *póntos* denotes the sea as a route or crossing; *pélagos* refers to the open or high sea; *háls*, etymologically ‘salt’, denotes the sea by its substance (metonymy).

- (1) *Tēlémakhos d' apáneuthen iōn epì thîna thalássēs*
 ‘Telemachus went far off to **the shore of the sea.**’ (Od. 2.260)
- (2) *Plēōn epì oínopa pōnton ep' allothrōous anthrōpous*
 ‘**Sailing on the wine-dark sea** to men of another language.’ (Od. 1.183)

In Latin, <SEA> is lexicalised through *mare*, *pontus*, *aequor*, and *pelagus*,⁷ covering the same two synsets as Greek: 09368829-n (214 tokens) and 09449666-n (73 tokens). Here, *mare* and *aequor* are the dominant forms, with poetry accounting for the majority of occurrences (190 tokens), followed by historiography (56 tokens) and theatre (32 tokens). *Aequor* also provides an example of semantic change within the corpus: its primary sense of ‘flat surface’ diminishes over time, while the sense of ‘sea’ rises in the 1st cent. BCE and becomes dominant, reflecting a metonymic shift from a general flat expanse to the specific surface of the sea (Figure 2).

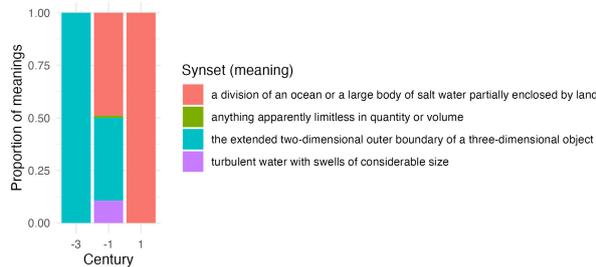


Figure 2: Semantic shift of Lat. *aequor*.

10 Conclusion and Future Work

This study demonstrates that combining LLM-based WSD with an onomasiological and semasiological framework may provide new insights into the lexicalisation of geographical concepts in Latin and Ancient Greek. Greek shows higher lexical dispersion and semantic conflation, while Latin exhibits sharper semantic partitioning and gradual sense shifts. Genre and diachrony modulate lexical patterns, highlighting the interaction between linguistic, cultural, and textual factors. Due to space

⁷The lemmas differ etymologically (de Vaan, 2008): *mare* is the general term for the sea; *pontus* and *pelagus* are borrowings from Greek (*póntos*, *pélagos*) and retain their original nuances; *aequor* derives from Latin *aequus* ‘level, even’ and originally denotes a flat expanse, later extending metonymically to bodies of water.

constraints, we focused on two case studies, while the full annotated dataset is made available on Zenodo (Ciletti et al., 2026).

Although the model proved useful for comparative onomasiological analyses, its overall accuracy was moderate, reflecting challenges in disambiguating fine-grained WN synsets. Future work should explore fine-tuning alternative models and systematically investigating the best strategies for conducting onomasiological and semasiological analyses. Preliminary studies (Farina and Ciletti, 2026) show that fine-tuned open models can reach F1 scores up to 0.85 when assigning English meanings to Latin and Ancient Greek verbal prefixes, doubling the scores of their base versions. It still remains to be tested whether this strategy is feasible for assigning meanings via WN synsets. Moreover, expanding the corpus diachronically and by adding further texts may provide a clearer picture of semantic change, enabling more reliable identification of when meanings shift and when concepts come to be expressed by different lexical items, and contributing to the development of historical lexical resources.

Acknowledgments

This work is supported by the UKRI under the Horizon Europe Guarantee (grant number UKRI947) for the project COALA (Computational Corpus Annotation for Quantitative Analysis of Latin Lexical Semantics) successfully evaluated by the ERC, and by King’s College London’s AHRS Research Grant (Research & Scholarship Development Stream) for the project “Mapping meaning with Large Language Models”.

Author Contributions

AF: Conceptualisation; Data curation; Formal analysis; Investigation; Methodology; Visualisation; Writing – original draft (Sections 3–6, 8–10); Writing – review & editing.

MC: Investigation; Software; Writing – original draft (Sections 5, 7, 8); Writing – review & editing.
 BMcG: Conceptualization; Funding acquisition; Supervision; Writing – original draft (Sections 1, 2); Writing – review & editing.

AB: Conceptualization; Data curation; Writing – original draft (Section 2).

References

- J. N. Adams. 2013. *Social Variation and the Latin Language*. Cambridge University Press, Cambridge.
- Carmine Ampolo, Ugo Fantasia, and Leone Porciani. 2004. *Lexicon historiographicum Graecum et Latinum (LHG & L)*. Number 1 in *Lexicon historiographicum Graecum et Latinum (LHG & L)*. Ed. della Normale, Pisa.
- Marianna Apidianaki and Benoît Sagot. 2012. [Applying cross-lingual WSD to wordnet development](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 833–840, Istanbul, Turkey. European Language Resources Association (ELRA).
- Flávia Lemos Mota de Azevedo. 2010. [The polis and the marvelous in the narrative of herodotus](#). *Revista Archaï*, (1):33–42.
- Georgios Babiniotis. 2002. *Lexikó Tis Néas Ellinikís Glóssas [Dictionary of Modern Greek]*. Athens: Kentro Lexicologias.
- Andrea Ballatore. 2015. [Place Vocabulary](#). GitHub.
- David Bamman and Patrick Burns. 2020. [Latin bert: A contextual language model for classical philology](#). arXiv preprint arXiv:2009.10053.
- David Bamman and Gregory Crane. 2007. The Latin Dependency Treebank in a cultural heritage digital library. *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*. Prague (Czech Republic), pages 33–40.
- David Bamman and Gregory Crane. 2011. The Ancient Greek and Latin Dependency Treebanks. In *Language Technology for Cultural Heritage*, pages 79–98, Berlin/Heidelberg. Springer.
- Marie-Claire Beaulieu. 2016. *The sea in the Greek imagination*. PENN/University of Pennsylvania Press, Philadelphia.
- Hilary Becker. 2008. Urbs, oppidum, castellum, vicus: settlement differentiation and landscape nomenclature in Etruria. *La Città Murata in Etruria. Atti del XXV Convegno di Studi Etruschi ed Italici*, pages 73–80.
- Robert Stephen Paul Beekes. 2009. *Etymological dictionary of Greek*. Number 10 in Leiden Indo-European etymological dictionary series. Brill, Leiden.
- Marijke Beersmans, Evelien de Graaf, Tim Van de Cruys, and Margherita Fantoli. 2023. Training and evaluation of named entity recognition models for classical Latin. In *Proceedings of the Ancient Language Processing Workshop*, pages 1–12.
- Erica Biagetti, Chiara Zanchi, and Silvia Luraghi. 2023. [Linking the Sanskrit WordNet to the Vedic dependency treebank: a pilot study](#). In *Proceedings of the 12th Global Wordnet Conference*, pages 77–83, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.
- Erica Biagetti, Chiara Zanchi, and William M. Short. 2021. Toward the creation of wordnets for ancient indo-european languages. In *Proceedings of the 11th Global Wordnet Conference*, pages 258–266. Global Wordnet Association.
- A. Blank. 2001. Words and concepts in time: Towards diachronic cognitive onomasiology. In *Words in Time: Diachronic Semantics from Different Points of View*, pages 37–65.
- Francis Bond and Arthur Bond. 2019. [GeoNames Wordnet \(geown\): extracting wordnets from GeoNames](#). In *Proceedings of the 10th Global Wordnet Conference*, pages 387–393, Wroclaw, Poland. Global Wordnet Association.
- Francis Bond and Maciej Piasecki. 2017. Proceedings of the Workshop on Challenges for Wordnets. *Language, Data and Knowledge Conference. Galway, Ireland, 18 June 2017*.
- Michel Bréal. 1897. *Essai de sémantique*. Paris: Hachette.
- Luca Brigada Villa, Andrea Farina, and Chiara Zanchi. 2025. [Formulaic Networks as Prototypical Categories: Combining the Ancient Greek Dependency Treebank with the Ancient Greek WordNet for a Pilot Study on the Iliad](#). In Jesús De La Villa, Araceli Striano, and Rodrigo Verano, editors, *Advances in Ancient Greek Linguistics*, pages 737–758. De Gruyter.
- Niclas Burenhult and Stephen C Levinson. 2008. Language and Landscape: A Cross-linguistic Perspective. *Language sciences*, 30(2-3):135–150.
- Anita Burgun and Olivier Bodenreider. 2001. Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. *Proceedings of NAACL'2001 Workshop, "WordNet and Other Lexical Resources: Applications, Extensions and Customizations"*, pages 77–82.
- Patrick J Burns. 2023. [Latincy: Synthetic trained pipelines for latin nlp](#). arXiv preprint arXiv:2305.04365.
- Luca Cardinali. 2017. Alcune osservazioni sulla lingua degli storiografi arcaici. *Pallas. Études de linguistique latine* II(103):171–179.
- Christophoros Charalambakis. 1996. Istoría tis ellinikís glóssas [history of the greek language]. *Encyclopedia Papyros-Larousse-Britannica: Greece*.
- Michele Ciletti, Andrea Farina, Barbara McGillivray, and Andrea Ballatore. 2026. [A diachronic dataset of semantically-annotated geographical nouns in ancient greek and latin](#).

- Pierluigi Cuzzolin and Gerd V. M. Haverling. 2009. Syntax, sociolinguistics, and literary genres. In *New perspectives on historical Latin syntax*, pages 16–63.
- Michiel de Vaan. 2008. *Etymological Dictionary of Latin*. Brill, Leiden / Boston.
- Ludwig Döderlein. 1874. *Handbook of Latin Synonyms*. W. F. Draper, Andover.
- Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. 2016. Challenges and solutions for Latin named entity recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 85–93.
- Ilenia Falcinelli, Chiara Fini, Claudia Mazzuca, and Anna M Borghi. 2024. The geo domain: a review on the conceptualization of geographical and geopolitical entities. *Frontiers in Psychology*, 15:1389581.
- Andrea Farina. 2023. [Lost at sea: A dataset of 25+ sea words morpho-semantically annotated in ancient greek and latin](#). *Journal of Open Humanities Data*, 9(1):24.
- Andrea Farina. 2026. *A comparative study on Ancient Greek and Latin preverbed motion verbs. The creation of PrevNet as an expandable online resource for preverbs and preverbation*. King’s College London: PhD Thesis.
- Andrea Farina, Andrea Ballatore, and Barbara McGillivray. 2025. Mapping meaning in latin with large language models: A multi-task evaluation of preverbed motion verbs and spatial relation detection in llms. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*.
- Andrea Farina and Michele Ciletti. 2026. [Probing preverbs: Evaluating large language models on Latin and Ancient Greek preverbed motion verbs](#). Forthcoming. In press.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Rolando Ferri. 2011. The Language of Latin Epic and Lyric Poetry. In James Clackson, editor, *A companion to the Latin language*, pages 344–366. Wiley-Blackwell, Malden, MA.
- Michèle Fruyt. 1994. Typologie des cas de synonymie. In Claude Moussy, editor, *Les problèmes de la synonymie en latin*, pages 25–46. PUPS.
- Dirk Geeraerts. 1997. *Diachronic Prototype Semantics: A Contribution to Historical Lexicography*. Oxford: Clarendon Press.
- Iacopo Ghinassi, Simone Tedeschi, Paola Marongiu, Roberto Navigli, and Barbara McGillivray. 2024. [Language pivoting from parallel corpora for word sense disambiguation of historical languages: A case study on Latin](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10073–10084, Torino, Italia. ELRA and ICCL.
- Julio Gonzalo, Felisa Verdejo, Carol Peters, and Nicoletta Calzolari. 1998. [Applying EuroWordNet to Cross-Language Text Retrieval](#). *Computers and the Humanities*, 32(2-3):185–207.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1489–1501.
- Edward M. Harris and David M. Lewis. 2024. [City and country in the Greek world](#). In Miko Flohr and Arjan Zuiderhoek, editors, *A Companion to Cities in the Greco-Roman World*, 1 edition, pages 306–320. Wiley.
- Martin Haspelmath. 1999. Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft*, 18(2):180–205.
- Gerd Haverling. 2012. Literary late latin and the development of the spoken language. In Martti Leiwo, Hilla Halla-aho, and Marja Vierros, editors, *Variation and Change in Greek and Latin*. Foundation of the Finnish Institute at Athens.
- Eric Kafe. 2017. [How Stable are WordNet Synsets?](#) *Proceedings of the Workshop on Challenges for Wordnets*, pages 1–12.
- Vojtěch Kaše, Sarah Lang, and Petr Pavlas. 2025. [Embedded in the labyrinth: Investigating latin word senses through transformer-based contextual embeddings and attention](#). *Anthology of Computers and the Humanities*, 3:498–512.
- Jeremy Kenyon, Jason W. Karl, and Bruce Godfrey. 2023. Evaluation of placename geoparsers. *Journal of Map & Geography Libraries*, 19(3):185–197.
- Fahad Khan, John P. McCrae, Francisco Javier Minaya Gómez, Rafael Cruz González, and Javier E. Díaz-Vera. 2023. [Some considerations in the construction of a historical language WordNet](#). In *Proceedings of the 12th Global Wordnet Conference*, pages 101–105, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.
- Karol Marek Klimczak, Jan Makary Fryczak, Dominika Hadro, and Justyna Fijałkowska. 2024. [Equivalent sentiment measures for cross-language analysis of corporate communications](#). *MethodsX*, 12:102745.
- A. Kutuzov, E. Velldal, and L. Øvrelid. 2022. Contextualized embeddings for semantic change detection: Lessons learned. *Northern European Journal of Language Technology*, 8.

- A. Lehrer. 1985. The influence of semantic fields on semantic change. In Jacek Fisiak, editor, *Historical Semantics—Historical Word Formation*, pages 283–296. Mouton de Gruyter.
- Piroska Lendvai and Claudia Wick. 2022. Finetuning latin bert for word sense disambiguation on the thesaurus linguae latinae. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 37–41.
- Charlton T. Lewis and Charles Short. 1879. *A Latin Dictionary. Founded on Andrews' edition of Freund's Latin dictionary. revised, enlarged, and in great part rewritten*. Clarendon Press, Oxford: Clarendon Press.
- Astrid Lindenlauf. 2004. [The sea as a place of no return in ancient Greece](#). *World Archaeology*, 35(3):416–433.
- Beatrice Marchesi, Annachiara Clementelli, Andrea Maurizio Mammarella, Silvia Zampetta, Erica Biagetti, Luca Brigada Villa, Virginia Mastellari, Riccardo Ginevra, Claudia Roberta Combei, and Chiara Zanchi. 2025. Towards the semi-automated population of the ancient greek wordnet. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 647–658.
- David M Mark, Andrew G Turk, Niclas Burenhult, and David Stea. 2011. *Landscape in Language: Transdisciplinary Perspectives*, volume 4. John Benjamins Publishing.
- F. Martelli, M. Maru, C. Campagnano, R. Navigli, P. Velardi, R.J. Ureña-Ruiz, F. Frontini, V. Quochi, J. Kallas, K. Koppel, and M. Langemets. 2022. [D3.8 lexical-semantic analytics for nlp](#). Technical Report.
- John P. McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. [English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology](#). In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 14–19, Marseille, France. The European Language Resources Association (ELRA).
- Tony McEnery and Andrew Wilson. 2001. *Corpus Linguistics. An Introduction. Second edition*. Edinburgh: Edinburgh University Press.
- Barbara McGillivray, Iacopo Ghinassi, and Paola Marongiu. 2023. [The challenges of sense granularity in word sense disambiguation for latin \(poster\)](#).
- Barbara McGillivray and Krzysztof Nowak. 2025. Tracing the semantic change of socio-political terms from classical to early medieval latin with computational methods. In *Varietate delectamur: Multifarious Approaches to Synchronic and Diachronic Variation in Latin*, pages 357–570. Brepols Online.
- Antoine Meillet. 1958. Comment les mots changent de sens. In *Linguistique historique et linguistique générale*, pages 230–280. Champion.
- W. Mercelis, T. Van Hal, A. Keersmaekers, J. de la Villa, A. Striano, and R. Verano. 2025. [Tongue, language or noise? word sense disambiguation in ancient greek with corpus-based methods](#). In *Advances in Ancient Greek Linguistics: Vol. 5.1*, pages 813–828. De Gruyter, Berlin/Boston.
- Stefano Minozzi. 2010. The latin wordnet project. In *Latin Linguistics Today: Akten des 15. Internationalen Kolloquiums zur Lateinischen Linguistik*, pages 707–716, Innsbruck.
- Claude Moussy. 1965. Gratus et iucundus. *Revue des Études latines*, 42:389–400.
- Jacobo Myerston. 2024. [greCy: Ancient Greek language models for spaCy](#).
- John M. Nash. 2018. *Rulers of the Sea – Maritime Strategy and Sea Power in Ancient Greece 550-321 BC*. The Australian National University: PhD Thesis.
- Robert Ogilvie. 1901. *Horae Latinae: Studies in Synonyms and Syntax*. Longmans, Green, and Co., London.
- Angela Ralli. 2012. Greek. *Revue Belge de Philologie et d'Histoire*, 90(3):939–966.
- C. M. Reed. 2003. *Maritime traders in the ancient Greek world*. Cambridge University Press, Cambridge, UK ; New York.
- Tim Rood. 2001. [Thucydides on the polis](#). *The Classical Review*, 51(2):241–242.
- Saskia T. Roselaar. 2010. *Public Land in the Roman Republic: A Social and Economic History of Ager Publicus in Italy, 396-89 BC*. Oxford Studies in Roman Society and Law Ser. Oxford University Press, Incorporated, Oxford.
- Ewa Rudnicka, Łukasz Grabowski, Maciej Piasecki, and Tomasz Naskręt. 2022. [In Search of Gaps between Languages and Wordnets: the Case of Polish-English WordNet](#). *International Journal of Lexicography*, 36(1):68–92. _eprint: <https://academic.oup.com/ijl/article-pdf/36/1/68/50049909/ecac005.pdf>.
- Cornelis Schrevel. 1832. *Lexicon manuale graecolatium et latino-graecum*. Collins and Hannay, Philadelphia.
- William Smith and Theophilus Hall. 1871. *A Copious and Critical English-Latin Dictionary*. Harper & brothers, New York.
- Mirjana Stefanovski and Kosta Čavoški. 2023. [Polis, Loimos, Stasis: Thucydides about Disintegration of the Political System](#). *Conatus*, 8(2):629–656.
- N. I. Stolova. 2015. *Cognitive Linguistics and Lexical Change: Motion Verbs from Latin to Romance*. John Benjamins, Amsterdam/Philadelphia.

Esther Torrego. 2020. The expression of knowledge in Latin: *cognosco, nosco, scio, nescio* and *ignoro*. In L. van Gils, C. Kroon, and R. Risselada, editors, *Lemmata Linguistica Latina, Volume II: Clause and Discourse*, pages 20–47. De Gruyter.

Joseph Williams. 1976. Synaesthetic adjectives: a possible law of semantic change. *Language*, 52(2):461–478.

Rob Witcher. 2005. *The extended metropolis: Urbs, suburbium and population*. *Journal of Roman Archaeology*, 18:120–138.

Arjan Zuiderhoek. 2015. *Introduction: Land and Natural Resources in the Roman World in Historiographical and Theoretical Perspective*. In Paul Erdkamp, Koenraad Verboven, and Arjan Zuiderhoek, editors, *Ownership and Exploitation of Land and Natural Resources in the Roman World*, pages 1–17. Oxford University Press.

A Corpus Composition

This appendix documents the textual composition of the *Anonymised Corpus*, providing an explicit overview of the Ancient Greek and Latin works included. The selection spans multiple centuries and literary traditions in both languages, allowing for controlled comparison across historical periods and genres.

Table 6 reports the authors and texts included in the corpus, together with their chronological attribution, literary genre, and token counts. Works that are fully or partially annotated in the Ancient Greek and Latin Dependency Treebanks (AGDT/LDT) are indicated by an asterisk (*) or a degree symbol (°), respectively. The corpus combines diachronic breadth with genre diversity while taking advantage of existing syntactic resources, and it is intended to support both its original focus on preverbed motion verbs and its reuse for related studies, such as the analysis outlined in this paper. For a deeper discussion about the creation of the corpus, we refer to [Farina \(2026\)](#).

B System Prompt

This appendix presents the complete prompt template used for the automatic annotation. The complete data of this study are available on Zenodo ([Ciletti et al., 2026](#)).

System Prompt:

You are an expert in Latin and Ancient Greek semantics, historical linguistics, and semantic annotation.

Your task for each query:

- You are given a TOKEN, its LEMMA, LANGUAGE, a SENTENCE providing context, and a list of candidate noun synsets.

- Each candidate synset has:

- a small integer ID (1, 2, 3, ...)

- an English gloss.

You may also be given gold-annotated example sentences with the same lemma, each with one gold sense gloss.

Use these examples only as guidance; they may involve different senses than the current sentence.

Your job:

1. Choose at most one synset ID whose gloss best matches the meaning of the TOKEN as used in the given SENTENCE.

2. If none of the synsets plausibly matches, select "n/a" instead of any numeric ID.

3. Decide whether the TOKEN is used literally (compositional, non-figurative) or metaphorically/figuratively in context.

4. Provide a confidence score from 0 to 1 for your overall decision.

Important:

- Focus strictly on the actual usage in the SENTENCE, not all possible meanings of the lemma.

- Prefer precise matches over loose associations.

- Treat metonymic or metaphorical readings as "not literal".

Output:

Return ONLY a single JSON object matching the provided schema.

Table 6: The REMOVE Base Corpus: Authors, Texts, Centuries, Literary Genres, and Token Counts. Texts fully or partially annotated in the AGDT/LDT are marked with * and ° respectively.

Author	Text	Century	Literary Genre	Tokens
Ancient Greek				
Homer	<i>Odyssey</i> 1–12*	8th BCE	Poetry, epic	46,021
Pindar	<i>Olympians</i>	5th BCE	Poetry, lyric	5,953
Aristophanes	<i>Thesmophoriazusae</i>	5th BCE	Theatre, comedy	7,894
Aeschylus	<i>Agamemnon</i> *	5th BCE	Theatre, tragedy	8,495
	<i>Prometheus Bound</i> *	5th BCE	Theatre, tragedy	6,155
Sophocles	<i>Antigone</i> *	5th BCE	Theatre, tragedy	7,658
Euripides	<i>Medea</i>	5th BCE	Theatre, tragedy	8,295
Herodotus	<i>Historiae</i> 1	5th BCE	Historiography	29,081
Lysias	<i>De caede Eratosthenis</i> *	5th BCE	Oratory	2,478
	<i>In Alcibiadem</i>	5th BCE	Oratory	2,591
Plato	<i>Euthyphro</i> *	5th BCE	Philosophy	5,417
Aristotle	<i>Athenian Constitution</i>	4th BCE	Philosophy	16,112
Thucydides	<i>Histories</i> 1*	5th BCE	Historiography	22,308
Apollonius Rhodius	<i>Argonautica</i> 1–3	3rd BCE	Poetry, epic	27,066
Polybius	<i>Histories</i> 1*	2nd BCE	Historiography	25,658
Diodorus Siculus	<i>Bibliotheca Historica</i> 11*	1st BCE	Historiography	23,336
Plutarch	<i>Alcibiades</i> *	2nd CE	Historiography, biography	10,249
	<i>Lycurgus</i> *	2nd CE	Historiography, biography	9,662
Lucian of Samosata	<i>Vera Historia</i>	2nd CE	Novel	11,484
Latin				
Ennius	<i>Annales</i>	3rd BCE	Poetry, epic	1,194
Plautus	<i>Amphitruo</i>	3rd BCE	Theatre, comedy	9,988
	<i>Mostellaria</i>	3rd BCE	Theatre, comedy	9,780
Caesar	<i>De bello Gallico</i> 1–4°	1st BCE	Historiography	20,498
Cicero	<i>In Catilinam</i> 1–3°	1st BCE	Oratory	11,625
	<i>De amicitia</i> °	1st BCE	Philosophy	9,471
Sallust	<i>Bellum Catilinae</i> *	1st BCE	Historiography	10,655
Livy	<i>Ab Urbe condita</i> 1–2	1st BCE	Historiography	39,913
Vergil	<i>Aeneid</i> °	1st BCE	Poetry, epic	63,719
Propertius	<i>Elegiae</i> 1.1–1.22*	1st BCE	Poetry, elegy	4,384
Horace	<i>Satires</i> 1	1st BCE	Poetry, satire	7,048
Seneca	<i>De ira</i>	1st CE	Philosophy	22,614
	<i>Medea</i>	1st CE	Theatre, tragedy	5,693
Tacitus	<i>Historiae</i> 1°	1st–2nd CE	Historiography	11,852
Suetonius	<i>Life of Augustus</i> °	2nd CE	Historiography, biography	13,915
Apuleius	<i>Metamorphoses</i> 1–5	2nd CE	Novel	23,358

Evaluating Humanities Theory Alignment in Large Language Models: Incremental Prompting and Statistical Assessment

Axel Pichler

Department of German Studies
University of Vienna
axel.pichler@univie.ac.at

Janis Pagel

Department of Digital Humanities
University of Cologne
janis.pagel@uni-koeln.de

Abstract

We propose a method to evaluate the extent to which an LLM’s observable input–output behavior aligns with established theories in the humanities and cultural studies. We instantiate the framework on three humanities theories—Davidson’s truth-conditional semantics, Lewis’s truth in fiction, and Iser’s concept of textual gaps—using a top-down, theory-driven black-box framework. Core assumptions of these theories are reconstructed into testable behavioral rules and assessed via controlled classification tasks with systematic prompt comparisons and significance testing. Our experiments show that theory-uninformed classification prompts generally outperform theory-enriched prompts in Lewis and Iser settings, while theory-informed prompts help in the Davidson task. Gemini Flash consistently achieves the highest scores across tasks and corpora, while the Iser gap detection task remains substantially harder than binary truth-conditional judgments. Statistical tests confirm robust prompt effects and the failure of basic prompts. However, model behavior under incremental theory exposure is unstable and architecture-dependent.

1 Introduction

In the Digital Humanities in general, and in Computational Literary Studies in particular, the question of how theories and concepts from the humanities and cultural studies can be modeled and operationalized has long played a central role (Moretti, 2014; Flanders and Jannidis, 2018). This question revolves around the problem of how cultural and textual phenomena can be datafied without losing core specificities. Here, we understand “datafication” as the translation of textual properties into formal, machine-readable representations, and “operationalization” as the development of procedures that, in multiple steps, map the extension of field-specific concepts onto indicators in texts, thereby

enabling subsequent measurement (Pichler and Reiter, 2022; Jacke, 2025). Answering this question is complicated by the fact that there are no widely accepted conceptualizations of theory in the humanities and cultural studies in general, nor in literary studies in particular. While the notion of “concept” in literary studies appears to share the same facets as in the general philosophical debate on concepts (Margolis and Laurence, 2023), the notion of “theory” is characterized by nuances that repeatedly depart from an understanding of theory as explicit, ordered, and logically consistent systems of categories designed to describe and explain the phenomena within a given domain. For modeling literary theories and operationalizing their guiding concepts, it follows that reconstruction procedures are often necessary; their selection is guided by pragmatic criteria and geared toward the further algorithmic processing of their results.

With the rise of generative large language models (LLMs) and the so-called prompt-and-predict paradigm that characterizes their use (Liu et al., 2023), the impression arises that at least one of the central problems of modeling textual data and operationalizing text-analytical concepts can be partially addressed: Now that one can interact with the models via string manipulation, it appears less necessary to clarify in detail how one’s data model relates to texts or how computational measurement methods relate to traditional concepts. One might simply articulate this relationship verbally (Halterman and Keith, 2025) without having to formulate formal-technical mapping rules—by which we mean explicit, transparent mappings between theoretical categories and algorithmic operations.

Tempting as this hope may sound, it is gradually becoming clear that it holds only to a limited extent, given that LLMs are not ideology-agnostic models but, depending on training methods and training data, incorporate a wide variety of biases. In this regard, we can roughly distinguish between

(a) socio-cultural biases, (b) methodological or architectural biases, and (c) theoretical priors that arise from the distribution of the training data. It remains unclear to what extent such biases can be defined away through mere instruction in the course of prompt-and-predict usage. This raises, particularly for Computational Literary Studies, questions of interpretive validity and the reproducibility of LLM-based results.

Against this backdrop, we pose the following two questions:

RQ1. How can we test whether, and to what degree, a particular theory adheres to an LLM—that is, the extent to which an LLM acts in accordance with this theory without being explicitly prompted to do so? This question targets the theoretical bias of large language models and its empirical detectability.

RQ2. To what extent does an LLM’s behavior change when it is provided—step by step—with increasing amounts of information about a theory? Here, we are interested in the models’ conditionability and the stability of their outputs under incremental theory exposure.

In what follows, we propose a method to evaluate how closely an LLM’s observable input–output behavior aligns with established theories in the humanities and cultural studies. For this purpose, we have selected three “theories” that we take to be operationalizable once they have been explicated to the level of precision required for the tests to be conducted: Donald Davidson’s conception of meaning, David Lewis’s conceptualization of truth in fiction, and Wolfgang Iser’s *Leerstelle* (“gap”).¹ Drawing on NLP work in behavioral evaluation, we adopt a theory-driven, top-down approach. Starting from a given theory, we reconstruct its testable claims and decision criteria to derive observable indicators of theory-conforming behavior. Based on these indicators, we design tasks and test datasets, then assess the degree to which an LLM’s outputs match the expected decisions of a scholar following the reference theory. Our primary aim is to demonstrate a general pathway for answering our

¹As the debate on the operationalization of literary-theoretical concepts has shown, such explications may, on the one hand, risk narrowing the original concepts, but, on the other hand, if carried out with sufficient precision, they allow the resulting measurements to be reintegrated into humanities discourse.

two research questions: derive testable indicators from suitably explicated theories and use them to evaluate LLM behavior. We selected Davidson, and Lewis first and foremost because their core claims admit precise operationalization into falsifiable decision criteria; second, because questions about states of affairs in narrated worlds (truth conditions) are central to literary interpretation and thus to addressing interpretive problems.

2 Related Work

Behavioral Analysis of LLMs In line with prior work, we adopt the NLP tradition of behavioral testing (Beizer, 1995), which assesses system capabilities by validating input–output behavior without access to internal structure (i.e., black-box evaluation).

There are some attempts to test the behavior of LLMs and compare them to human behavior in similar settings. Wang et al. (2024) compare different values of self-attention heads and feed-forward layers of LLMs with human eye-tracking measures and find some correlations that are not present in non-transformer language models. Akata et al. (2025) let LLMs interact in settings derived from game theory and find LLMs to perform well in games where self-interest is involved and less well in games where coordination is required. Pichler and Pagel (2025) test LLMs via prompts with increasing amount of information on a re-conceptualization of a theory of focalization and find that the systems do not deviate from their pre-conceived concepts.

Use of LLMs in Computational Literary Studies There are multiple recent studies who have used LLMs in order to solve classification and/or information extraction tasks from Computational Literary Studies (Bamman et al., 2024; Hicke and Mimno, 2024; Konle et al., 2024; Pagel et al., 2024; Wu et al., 2024; Bamman et al., 2025; Gius et al., 2025; Graciotti et al., 2025; Guhr et al., 2025; Hicke et al., 2025b; Hicke and Mimno, 2025; Irfan and Ali, 2025; Jannidis et al., 2025; Klähn et al., 2025; Majumdar et al., 2025; Michel et al., 2025; Pichler et al., 2025; Tudor et al., 2025; Werner and Reiter, 2025).

Hicke et al. (2025a) use LLMs to carry out focalization annotations in a selection of novels by Stephen King. Interestingly, they found the models to be not very receptive to prompt variations and in particular, one of the best performing prompts

was one that did not contain any theoretical notions about the target concept (i.e. focalization) at all and they theorize that GPT-4o had utilized pre-existing notions of focalization from its pre-training data (see [Hicke et al., 2025a](#), p. 745).

3 Workflow

In the following, we describe the workflow that can be used to check whether an LLM produces outputs consistent with the decision criteria of a researcher working within a specific theory in the humanities or cultural studies. The workflow implements a form of behavioral alignment testing. Behavioral alignment testing is a method that compares the behavior of an LLM with that of an informed human being. Therefore, the three theories are converted into a format that allows their core assumptions to be tested. Specifically, we derive testable behavioral rules from the core assumption(s) of each respective theory. These behavioral rules are formulated as conditionals, predicting specific behaviors that align with the antecedent conditions, i.e. the conditions of the literary text. The individual steps of the workflow we use are as follows:

Identification of Core Assumptions The first step takes as its starting point the characteristics of humanities theories already outlined in the introduction—especially the fact that these are often not explicit, ordered, and logically consistent systems of categories. It involves the systematic identification of core assumptions or foundational principles underlying a given theoretical framework. These assumptions serve as the theoretical premises from which testable behavioral rules are derived. They define the central characteristics of the subject area and specify how these characteristics or statements about them are to be linked.

Formation of Behavioral Rules Based on the identified core assumptions, testable behavioral rules are formulated as conditional statements. We operationalize the consequence relation between a core assumption and its predicted consequence as logical entailment (\models). These rules predict how an individual is expected to behave under the given theoretical framework. They define the expected behavioral outcomes and specify the conditions under which deviations from these predictions may occur. This step ensures that predicted behavior is directly grounded in the theoretical premises and can be empirically tested.

Experimental Design An experimental setting is developed to test manually created sentences derived from the behavioral rules. LLMs are presented with these sentences as controlled input, followed by a classification task or a sentence completion task. In the classification task, we evaluate model predictions over a constrained label set (e.g., True, False, Undetermined). In the sentence completion task, the model is prompted to continue a given sentence, with completion evaluated via the same constrained label mapping. We compare three setups: (i) sentence completion task, (ii) theory-uninformed classification, and (iii) theory-enriched classification. The design isolates the influence of specific variables on the LLMs’ interpretive or classificatory outputs and evaluates whether their responses align with the predictions derived from the theoretical premises.

4 Reference Theories and Operationalization

4.1 Davidson’s Radical Interpretation

The first theory we reconstruct is Donald Davidson’s theory of meaning as he developed it in particular in *Radical Interpretation* ([Davidson, 1973](#)). Davidson develops his theory based on the following questions: What knowledge could enable a listener to understand the utterances of a speaker of a language unknown to the listener? How could the listener acquire this knowledge?

Davidson’s answer is guided by his central hypothesis “that a theory of truth, modified to apply to a natural language, can be used as a theory of interpretation” ([Davidson, 1973](#), p. 189). This aligns with our aim of operational reconstruction insofar as Davidson’s program already suggests a procedure: derive a compositional truth theory licenses interpretive competence. We adapt this procedure to account for LLM-specific constraints (non-agency, lack of beliefs), treating LLM outputs as simulations of a competent speaker’s judgments.

Core Principle (Truth-Conditional Test):

Given a factual text T that represents a situation K , and an assertion φ (attributed to a speaker S at time Z), we evaluate truth *relative to* T over a two-way label space by restricting items to cases where $K \models \varphi$ or $K \models \neg\varphi$. K contains only information textually licensed by T :

$$\text{Truth}_T(\varphi) = \begin{cases} \text{True} & \text{if } K \models \varphi, \\ \text{False} & \text{if } K \models \neg\varphi. \end{cases}$$

By construction, items for which T does not fix the truth value of φ are excluded from the dataset.²

Interpretive Constraint (Charity/Coherence):

Before evaluating $\text{Truth}_T(\varphi)$, fix an interpretation function I (reference assignment, disambiguation, coreference, quantifier scope) that *maximizes coherence and charity* subject to T and publicly available evidence (Davidson, 1973). Charity constrains the choice of I ; it is *not* itself a condition for truth. We then assess φ under I relative to T .

4.2 Lewis’ Theory of Fiction

David Lewis, in his 1978 essay *Truth in Fiction* (Lewis, 1983), made the classical contribution to addressing the question of what holds true in the fictional world of a literary text. The core idea of his approach is that the fictional truth of a story depends on identifying the possible worlds where the events of the story take place (Köppe, 2014). Building on this idea, Lewis proposed the following three principles for determining fictional truth in stories:

1. **Principle 1:** What is explicitly described or narrated in the story is true in the fictional world.
2. **Principle 2:** What corresponds, in its factuality, to our “real world” is also true in the fictional world (the *reality principle*).
3. **Principle 3:** What corresponds, in its factuality, to collective belief worlds at the time the fictional story was created is also true in the fictional world (the *mutual belief principle*).

The principles³ have been subject to criticism from various theoretical perspectives. Regarding

²In view of the reconstructed Core Principle, several limitations must be noted in relation to Davidson’s theory of meaning. First, LLMs are not agents, they do not possess beliefs, and consequently cannot hold propositions to be true; we therefore assess simulated judgments. Second, the LLM–human interaction lacks the real-world situational and causal framework presupposed by radical interpretation (Davidson, 2001). Accordingly, our evaluation is conducted relative to the factual text T : we restrict the dataset to cases where $K \models \varphi$ or $K \models \neg\varphi$ and use a two-way label space {True, False}. Charity and coherence serve as constraints on interpretation (choice of reference, disambiguation) rather than as components of truth conditions. Thus, the test set examines to what extent LLM outputs approximate the communicative judgments of a competent speaker in Davidson’s sense, relative to T .

³For Principles 2 and 3, we operationalized our unique closest world assumption by grounding Principle 2 in currently valid knowledge, while grounding Principle 3 in knowledge that was valid at the time of the publication of the literary texts.

the first principle, it has been argued that it is insufficient because it does not account for implicit truths. With respect to the second principle (the reality principle), Currie (1990) argued that it results in: (a) statements that are irrelevant to the plot but true in reality being considered true in the fictional world, and (b) in cases of unreliable narration, a bias toward interpretations that align with our reality. Following Lewis (1983), we operationalize three core principles of truth in fiction, each paired with a two-way behavioral labeling rule. The full formal definitions and interpretive constraints are provided in Appendix A.

4.3 Iser’s Concept of the *Leerstelle*

Alongside Hans Robert Jauss, Wolfgang Iser is a central representative of reception theory in the German-speaking world. This approach places an idealized construct of the reader at the center of its literary-theoretical considerations. Drawing on Roman Ingarden’s phenomenological theory, Iser understands literary comprehension as a dynamic process in which the reader continually draws on what has already been read while simultaneously forming expectations about what is to come (Sneis, 2018, pp. 144–161). Because texts never fully specify all aspects of a represented world, the reader supplies missing information—from visual completion, to the establishment of logical coherence. Central to this is the concept of the *gap* (German: *Leerstelle*).

As recent scholarship has emphasized (Willand, 2015), Iser’s notion is formulated with considerable openness, which necessitates explication for practical application. In this paper, we use the following: A “gap” is a textually grounded indeterminacy—namely, a missing relation between explicitly articulated partial units (e.g., propositions, perspectives, or events) that prompts the reader to establish coherence through inferential reasoning and expectation formation, or to process provisional incoherences.

Core Principle (Iser/Gaps) Given a text T that represents a situation K , and an assertion φ (attributed to a speaker S at time Z), under a fixed interpretation I :

- φ is *True* in T if $K \models \varphi$,
- φ is *False* in T if $K \models \neg\varphi$,
- φ is *Not articulated (Gap)* if neither $K \models \varphi$ nor $K \models \neg\varphi$ holds, because T leaves a requi-

site relation between explicit units unspecified under I .

Behavioral Rule (Three-way labeling) Under the fixed interpretation I , assign:

- True if $K \models \varphi$,
- False if $K \models \neg\varphi$,
- Not articulated (*Gap*) if neither entailment holds due to an indeterminacy (a missing relation among explicit units in T).

5 Experimental Setup

5.1 Data Annotation

We evaluate our approach using manually created test data with respect to the narrative worlds of two literary reference texts—Arthur Schnitzler’s *Amerika* and Wolfgang Borchert’s *Das Brot*—as well as selected excerpts from newspaper articles. The test instances consist of sentences specifically designed for the three theoretical frameworks and, like the reference texts, are formulated in German. In the case of Davidson and Lewis, each sentence is assigned a truth value (true/false), while for Iser an additional option, “not fully articulated,” is available (for examples see Appendix C). The test datasets for Davidson and Lewis were created by one of the authors of this paper and annotated by another, whereas the test dataset for Iser was created collaboratively. We retained only instances with full annotator consensus to ensure data quality. Table 1 provides an overview of the resulting test datasets.⁴

5.2 Models

We evaluate four state-of-the-art large language models via the OpenRouter API: Gemini 2.0 Flash, Claude Sonnet 4.5, Qwen and Ministral. This diverse selection spans different model families, sizes, architectures, and training paradigms, allowing us to assess the generalizability of our findings across the current LLM landscape.

5.3 Prompt Design

We systematically compare multiple prompt templates varying in structure and theoretical grounding. Baseline prompts require the model to com-

⁴All manually created reference data, the Jupyter notebooks used for experimentation, and the resulting evaluation tables are publicly available at the project’s GitHub repository: <https://github.com/AxPic/LLM-Theory-Alignment>.

File	Prin.	κ	N
Davidson	–	0.709	170
Lewis-Bor	P1	0.971	69
Lewis-Bor	P2	0.971	69
Lewis-Bor	P3	0.941	68
Lewis-Sch	P1	0.850	74
Lewis-Sch	P2	0.774	71
Lewis-Sch	P3	0.822	73
Iser	–	–	60

Table 1: Overview of the manually created test datasets, including inter-annotator agreement (Cohen’s κ) and the number of rows with full annotator consensus.

plete the input’s last sentence directly with the label; classification prompts explicitly present the labels but not the underlying theory; theory-informed prompts incorporate explicit theoretical principles relevant to the classification task. All prompts include full textual context and request binary or, in the case of Iser, multiclass classification outputs; exemplary prompts for the Lewis framework are provided in Appendix D. As an example of a full-length context, Arthur Schnitzler’s short story “Amerika” is included in Appendix B as a machine translation produced with GPT-5.2.

5.4 Evaluation Setup

Experiments comprise the Davidson truth-conditional task (10 short newspaper articles), six evaluations of Lewis’s principles of truth in fiction across two texts (Schnitzler and Borchert, 3 principles each), and two Iser narrative gap detection tasks (Schnitzler and Borchert).

For each model-prompt-dataset combination, we conduct up to three independent runs – 3 for the Davidson and Iser, 2 for the Lewis data – with temperature 0.1 to balance consistency with minimal stochasticity while capturing any residual variation in model responses.

Classification Metrics For the Lewis and Davidson experiments (binary classification: true/false), we report F1-score (weighted), precision, recall, and accuracy. For the Iser experiment (3-class classification: true/false/not formulated), we additionally report F1-macro and F1-micro scores, alongside per-class precision, recall, and F1-scores to assess performance on the theoretically critical “not formulated” category representing Iser’s textual gaps (*Leerstellen*).

Statistical Significance Testing To assess statistical significance, we employ a multi-tiered approach adapted to each experiment’s classification

structure and number of runs: For binary classification (Lewis, Davidson), we use McNemar’s test to compare prompt pairs on individual predictions, accounting for the dependency structure of identical test instances. For multi-class classification (Iser), we employ Cochran’s Q test, which tests whether different prompts achieve systematically different success rates across the three classes.

All statistical tests use a significance level of $\alpha = 0.05$. This framework enables both granular instance-level comparisons and robust assessment of run-level consistency, while appropriately accounting for the paired nature of our experimental design and the distinct classification structures of each dataset.

6 Results

We report results from nine classification experiments across four large language models. We present overall performance (§6.1), prompt engineering effects (§6.2), and statistical significance (§6.3).

6.1 Overall Performance

Across all nine experiments and prompts, the models exhibited a stable performance ranking with only minimal variations, though absolute scores varied substantially by task type. Table 2 summarizes best performance across all experiments: Models achieved F1-scores ranging from 0.752 to 1.000, with substantial variation across tasks and prompt types. The F1 scores of the majority baseline are added for comparison, showing that on average, models easily outperform this simple baseline.

Gemini 3 Flash consistently outperformed all competitors and achieved the highest observed peak scores. Its performance remained robust across corpora and task formulations, indicating strong task-general capabilities.

Qwen 3 Next and *Mistral Ministral 14B* occupied intermediate positions across experiments. While neither model surpassed Gemini in any condition, both achieved moderate peak performance on selected tasks, with Qwen generally outperforming Mistral except in one of the Iser-settings.

Claude Sonnet 4.5 showed the weakest overall performance. Its best observed result was achieved on the Lewis-Borchert task with the theory-uninformed prompt (peak F1=0.869), while highlighting substantial brittleness with respect to

task formalization and prompt style.

The following patterns emerged with regard to the individual tasks:

Binary Classification Tasks (Davidson, Lewis)

The binary classification experiments achieved best F1-scores between 0.841 and 1.000, demonstrating high overall performance. The Davidson experiment, requiring truth-conditional evaluation against single contexts, reached F1=0.953 with Gemini 3 Flash using theory-enriched prompting. The Lewis experiments, evaluating three different principles of truth in fiction, showed solid performance: Principle 1 achieved perfect classification on Schnitzler (F1=1.000), though the reality principle (0.916) and the mutual belief principle (0.905) proved more challenging. On both test sets best performances were achieved in relation to Principle 1, which relies exclusively on statements explicitly articulated in the text, suggesting that the models – regardless of the theoretical enrichment of the prompts – have an easier time with this principle over the others.

Multi-Class Classification (Iser) The Iser experiments, which require three-way classification (*true/false/not formulated*), proved substantially more challenging than the binary settings. Across models, best performance reached $F1_{\text{weighted}} = 0.752$ on Schnitzler and 0.783 on Borchert, corresponding to an average deficit of approximately 0.15 F1 points relative to the binary tasks. These results suggest that Iser’s concept of *Leerstellen* poses a fundamental challenge for current LLMs.

6.2 Prompt Engineering Effects

Figure 1 presents prompt effectiveness across all experiments. Contrary to initial hypotheses, theory-uninformed classification prompts consistently outperformed theory-enriched prompts in Lewis experiments (mean advantage: +0.16 F1 points) and Iser experiments (Schnitzler: +0.25, Borchert: +0.20), while theory prompts excelled only in Davidson contexts.

Prompt effectiveness varied dramatically in Davidson: `prompt_theory` achieved mean F1=0.778, `prompt_class` F1=0.833, while `prompt_basic` achieved rather low scores in comparison (F1=0.373).

Across all six Lewis evaluations (3 principles \times 2 corpora), theory-uninformed classification prompts dominated: Schnitzler class=0.822 vs. theory=0.674 (+0.148); Borchert class=0.828 vs.

Experiment	Task Type	Best Model	Best Prompt	Peak F1	Maj. F1	N
Davidson	Binary	Gemini 3 Flash	theory	0.953	0.373	170
Lewis-Sch P1	Binary	Gemini 3 Flash	theory_1	1.000	0.379	74
Lewis-Sch P2	Binary	Gemini 3 Flash	class	0.916	0.423	71
Lewis-Sch P3	Binary	Gemini 3 Flash	theory_3	0.905	0.437	73
Lewis-Bor P1	Binary	Gemini 3 Flash	class	0.986	0.408	69
Lewis-Bor P2	Binary	Gemini 3 Flash	class	0.841	0.425	69
Lewis-Bor P3	Binary	Gemini 3 Flash	class	0.868	0.436	68
Iser-Sch (3-class)	Multi-class	Gemini 3 Flash	theory	0.752	0.176	60
Iser-Bor (3-class)	Multi-class	Gemini 3 Flash	theory	0.783	0.176	60
Mean (Binary)				0.918		
Mean (Multi-class)				0.768		

Table 2: Best performance per experiment. F1-scores represent weighted F1-scores.

theory=0.656 (+0.172). This pattern held consistently across principles with only one exception: Schnitzler Principle 1, where prompt_theory_1 achieved perfect F1=1.000 (vs. class 0.879). Theory prompts showed high inter-model variance (SD=0.335–0.390 vs. 0.075–0.092 for theory-uninformed classification prompts), indicating sensitivity to model architecture and suggesting that formal notation does not universally improve classification.

For the Iser task, prompt effectiveness diverges sharply by model architecture rather than exhibiting a uniform advantage for either prompting strategy. Across both corpora, Claude consistently failed under theory-enriched prompting (Macro-F1 ≤ 0.02 on Schnitzler; 0.00 on Borchert) while performing substantially better with theory-uninformed classification prompts (0.41 and 0.37, respectively). Gemini 3 Flash exhibited the opposite pattern, achieving its strongest results under theory-enriched prompting on both Schnitzler (Macro-F1 = 0.75 vs. 0.54) and Borchert (0.78 vs. 0.64). Mistral and Qwen occupied an intermediate regime: both benefited from classification prompts, but theory-enriched prompting degraded collapsed performance (Schnitzler: Mistral 0.13, Qwen 0.42; Borchert: Mistral 0.00, Qwen 0.46). These results contradict the hypothesis that Iser-specific *Leerstellen* terminology would uniformly aid gap recognition.

Relevance of logical notation To test whether logical notation itself causes confusion, we evaluated a variant (prompt_theory_no_symbol) only on the Davidson-data in which symbols were replaced by natural-language equivalents (“ $K \models \varphi$ ” → “situation logically supports φ ”, “iff” → “if and only if”). Table 3 shows divergent responses to symbol removal. Contrary to expectations, re-

Model	With Symbols	No Symbols	Δ	Signif.
Gemini 3 Flash	0.945	0.943	-0.002	p=1.00
Qwen 3 Next	0.840	0.827	-0.013	p=0.50
Mistral 14B	0.822	0.763	-0.059	p<0.001***
Claude Sonnet	0.507	0.386	-0.120	p=0.001***
Average	0.778	0.730	-0.049	—

Table 3: Symbol removal effects per model (Davidson theory prompts). Negative Δ indicates performance loss without symbols. Significance from McNemar tests comparing predictions.

moving symbols *decreased* performance (F1 = 0.730 vs. 0.778), with significant degradation for Claude (-0.120 , $p = 0.001$) and Mistral (-0.059 , $p < 0.001$), while Gemini remained essentially unaffected (-0.002 , $p = 1.0$). This counterintuitive result indicates that formal symbols provide *structural scaffolding* that is beneficial even to models struggling with formal reasoning: Claude’s performance, though poor with symbols (0.507), heavily lowers without them (0.386).

Model-Prompt Interactions Table 4 shows model-specific prompt preferences averaged across all nine experiments. Gemini favored formal prompts marginally in Davidson but theory-uninformed classification prompts in Lewis/Iser, yielding near-parity overall (+0.01). Claude showed dramatic preference for theory-uninformed classification instructions (+0.20), while Qwen and Mistral favored theory-uninformed classification prompts moderately (+0.08, +0.09).

6.3 Statistical Significance

We assess statistical significance at the instance-level differences by comparing model predictions across prompts, using McNemar tests for binary tasks and Cochran’s Q tests for multi-class Iser tasks. McNemar tests on binary classification revealed significant prompt differences in 71%

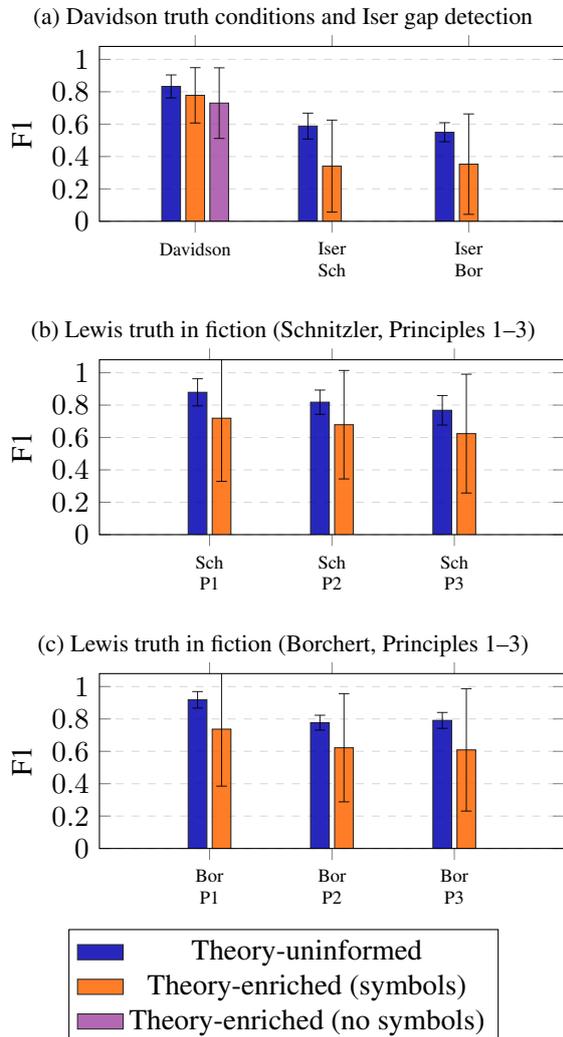


Figure 1: F1-scores by prompt type across tasks, split by theoretical framework and corpus. Error bars indicate the standard deviation across models and runs, with some bars extending beyond the theoretical F1 maximum of 1.0 due to high inter-model variance. Blue bars denote theory-uninformed classification prompts; orange bars denote theory-enriched prompts with formal notation; violet bars denote theory-enriched prompts without notation (Davidson only). Large error bars for theory-enriched prompts (e.g., Lewis-Sch P1) demonstrate architecture-dependent effectiveness, with Gemini achieving near-perfect scores while Claude fails completely on the same prompts.

Model	Theory	Classification	Δ
Gemini 3 Flash	0.850	0.861	+0.011
Claude Sonnet	0.365	0.565	+0.200
Qwen 3 Next	0.694	0.772	+0.078
Mistral 14B	0.619	0.708	+0.089

Table 4: Average F1-scores for theory-enriched vs. theory-uninformed prompts across all experiments. Δ indicates preference strength (positive values favor role-based prompts). Theory-enriched prompts substantially harm Claude (-0.20).

Experiment	Signif.	Total	Rate
Davidson	9	12	75%
Lewis-Sch P1	8	12	67%
Lewis-Sch P2	8	12	67%
Lewis-Sch P3	9	12	75%
Lewis-Bor P1	8	12	67%
Lewis-Bor P2	8	12	67%
Lewis-Bor P3	8	12	67%
Iser-Sch	8	12	67%
Iser-Bor	9	12	75%
Overall	75	108	70%

Table 5: Statistical significance summary. “Signif.” = comparisons with $p < 0.05$.

of model-experiment combinations. For multi-class Iser experiments, Cochran’s Q tests showed 67% significance (Schnitzler) and 82% significance (Borchert), with Borchert’s higher rate reflecting greater ambiguity and thus larger prompt-driven performance differences (Table 5).

Consistent patterns emerged: basic vs. theory-uninformed/theory-enriched comparisons achieved 100% significance (all $p < .001$), confirming the failure of basic prompts. Critically, theory vs. theory-uninformed classification comparisons showed significance in only 17% of cases, indicating theory-enriched prompting rarely yields statistically distinguishable improvements over simple theory-uninformed classification prompts.

7 Conclusion

Our experiments across three theoretical frameworks (Davidson truth conditions, Lewis’s three principles of truth in fiction, and Iser’s gaps) and multiple texts yield four main findings. First, theory-uninformed prompting consistently outperforms theory-enriched prompting across all tasks and models, indicating that the explicit incorporation of theoretical constraints does not reliably improve performance under the evaluated conditions. Second, formal notation provides structural

scaffolding: removing logical symbols from theory prompts leads to a universal performance decrease ($\Delta = -0.049$), with particularly severe effects for Claude ($\Delta = -0.120$), suggesting that symbolic structure anchors model behavior even when it is not semantically transparent. Third, Gemini 3 Flash performs best overall (rank #1 in 9/9 experiments) while exhibiting near-complete notation-agnosticism ($\Delta = -0.002$), pointing to architectural capacity for abstraction beyond surface form. Fourth, multi-class classification in the Iser gap task remains fundamentally challenging (best mean F1 = 0.768 vs. 0.924 for binary tasks, $\Delta = -0.16$), with uniformly low per-class performance (F1 = 0.27) indicating limited conceptual grounding.

Taken together, these results suggest that the extent to which a literary theory adheres to an LLM cannot be inferred from theory-enriched prompting alone, but is more reliably assessed through controlled comparisons with theory-uninformed baselines. Across tasks, models do not consistently act in accordance with the target theory when left unprompted; instead, apparent theory alignment emerges primarily when task structure aligns with strong surface regularities, such as binary truth-conditional judgments in the Davidson or Lewis settings. By contrast, tasks requiring abstract or meta-textual distinctions, such as Iser gap detection, show no evidence of latent theoretical bias, with overall weaker performance despite partial gains from theory-enriched prompting.

Incremental exposure to theoretical information does not lead to stable or monotonic improvements in model behavior. Rather, LLMs differ markedly in their conditionability: while some models (notably Gemini 3 Flash) remain robust under increasing theoretical and notational complexity, others (notably Claude) exhibit degradation when confronted with formalized theory, especially when logical notation is involved. The consistent performance drop observed after removing formal symbols further suggests that models might rely on notation as structural scaffolding rather than on semantic understanding of the underlying theory. Overall, theory prompting yields unstable and architecture-dependent effects, undermining assumptions of smooth theoretical internalization through stepwise exposure.

For the use of LLMs in CLS, these findings highlight the need for precise operationalizations of theoretical concepts and for annotation guidelines and reference data derived from these operational-

izations. Such resources are essential for systematically evaluating model behavior on tasks grounded in literary theory. Only with such resources can the construct validity of model outputs be ensured. Moreover, by evaluating outputs against these reference data, we can determine a model’s suitability for a given task irrespective of whether it is genuinely theory-aligned or merely exhibits surface-level conformity.

Declaration on Generative AI

We used GPT-5 and Claude Sonnet 4.5 for proofreading and code generation; all outputs were reviewed by the authors, who take full responsibility for the final content.

Limitations

Our study faces several limitations. First, we identify and operationalize only a single core principle per theory; ideally, multiple such core principles would be delineated and tested to probe theoretical coverage and robustness. Second, because our evaluation relies on black-box behavioral testing, we cannot determine whether a model is genuinely theory-aligned or merely exhibits surface-level conformity. Third, we evaluate on compact, manually constructed datasets (e.g., 60 Iser items per corpus; 68–74 per Lewis setting; 170 for Davidson), which constrains external validity and risks item-design artifacts. Fourth, reproducibility is limited: each model–prompt–dataset condition was run only two to three times (temperature 0.1), and results are explicitly characterized as indicative rather than conclusive. Fifth, the notation-manipulation result (performance drop after symbol removal) was tested only in the Davidson setting, limiting generalization of the “notation as scaffolding” interpretation to Lewis and Iser. Sixth, with respect to the basic prompts, which were designed to elicit a labeled continuation of an explicitly marked output sentence, we observe that instruction-tuned models no longer behave in line with the classical next-token prediction paradigm. Accordingly, this setup yields consistently lower scores across all evaluated metrics. Seventh, findings depend on model family and serving configuration—four instruction-tuned models were accessed via the OpenRouter API, and several effects were architecture-dependent and unstable under theory prompts. Finally, our task and prompt space is narrow: we compare basic, theory-uninformed classification, and theory-

enriched prompts (plus a notation variant), leaving broader process-oriented protocols outside the present scope.

Acknowledgments

We thank Janina Jacke, with whom the foundational groundwork for this study regarding the operationalization of the Davidson and Lewis theories has been collaboratively developed. We also thank Dominik Gerstorfer and Jonas Kuhn for their collaboration on an early version of the Davidson experiments. The second author has carried out this work under funding of the German Research Foundation (DFG, grant number 508319395).

References

- Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2025. [Playing repeated games with large language models](#). *Nature Human Behaviour*, 9(7):1380–1390.
- David Bamman, Sabrina Baur, Mackenzie Hanh Cramer, Anna Ho, and Tom McEnaney. 2025. [Measuring the Stories in Contemporary Songs](#). *Anthology of Computers and the Humanities*, 3:820–844.
- David Bamman, Kent K. Chang, Li Lucy, and Naitian Zhou. 2024. [On Classification with Large Language Models in Cultural Analytics](#). In *CHR2024*, pages 494–527, Aarhus.
- Boris Beizer. 1995. *Black Box Testing: Techniques for Functional Testing of Software and Systems*. John Wiley, New York.
- Gregory Currie. 1990. *The Nature of Fiction*, 1 edition. Cambridge University Press.
- Donald Davidson. 1973. [Radical interpretation](#). *Dialectica*, 27(3/4):313–328.
- Donald Davidson. 2001. *Epistemology Externalized*, 1 edition, pages 193–204. Oxford University Press Oxford.
- Julia Flanders and Fotis Jannidis, editors. 2018. *The Shape of Data in the Digital Humanities: Modeling Texts and Text-based Resources*, 1 edition. Routledge, Abingdon, Oxon ; New York, NY : Routledge, 2019. | Series: Digital research in the arts and humanities.
- Evelyn Gius, Stefanie Messner, and Axel Pichler. 2025. [How are Literary Histories written? An LLM-based Analysis of Objects and Perspectives in German Literary History](#). *Anthology of Computers and the Humanities*, 3:1090–1107.
- Arianna Graciotti, Franziska Pannach, Valentina Pretutti, and Federico Pianzola. 2025. [Llamas Don't Understand Fiction: Application and Evaluation of Large Language Models for Knowledge Extraction from Short Stories in English](#). *Anthology of Computers and the Humanities*, 3:4–32.
- Svenja Guhr, Huijun Mao, and Fengyi Lin. 2025. [Rethinking Scene Segmentation. Advancing Automated Detection of Scene Changes in Literary Texts](#). In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 79–86, Albuquerque, New Mexico. Association for Computational Linguistics.
- Andrew Halterman and Katherine A. Keith. 2025. [What is a protest anyway? Codebook conceptualization is still a first-order concern in LLM-era classification](#). *arXiv preprint*. ArXiv:2510.03541 [cs].
- Rebecca Hicke and David Mimno. 2024. [\[Lions: 1\] and \[Tigers: 2\] and \[Bears: 3\], Oh My! Literary Coreference Annotation with LLMs](#). In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 270–277, St. Julians, Malta. Association for Computational Linguistics.
- Rebecca M. M. Hicke, Yuri Bizzoni, Pascale Feldkamp, and Ross Deans Kristensen-McLachlan. 2025a. [Says Who? Effective Zero-Shot Annotation of Focalization](#). *Anthology of Computers and the Humanities*, 3:739–755.
- Rebecca M. M. Hicke, Brian W. Haggard, Mia Ferrante, Rayhan Khanna, and David Mimno. 2025b. [Are You There God? Lightweight Narrative Annotation of Christian Fiction with LMs](#). *Anthology of Computers and the Humanities*, 3:1012–1035.
- Rebecca M. M. Hicke and David Mimno. 2025. [Looking for the inner music : Probing LLMs' understanding of literary style](#). *Computational Humanities Research*, 1:e3.
- Saniya Irfan and Syed Juned Ali. 2025. [QaLLM: An LLM-based NER Dataset Curation, Annotation and Evaluation in Historical Urdu Elegies](#). *Anthology of Computers and the Humanities*, 3:922–937.
- Janina Jacke. 2025. [Operationalization and interpretation dependence in computational literary studies](#). *JCLS*, 4(1).
- Fotis Jannidis, Rabea Kleymann, Julian Schröter, and Heike Zinsmeister. 2025. [Do Large Language Models Understand Literature? Case Studies and Probing Experiments on German Poetry](#). *JCLS*, 4(1).
- Jannis Klähn, Janos Borst-Graetz, and Manuel Burghardt. 2025. [From dictionaries to LLMs – an evaluation of sentiment analysis techniques for German language data](#). *Computational Humanities Research*, 1:e4.

- Leonard Konle, Merten Kröncke, Fotis Jannidis, and Simone Winko. 2024. [On the Unity of Literary Change. The Development of Emotions in German Poetry, Prose, and Drama between 1850 and 1920 as a Test Case](#). In *CHR2024*, pages 282–300, Aarhus.
- Tilman Köppe. 2014. 8. [Fiktive Tatsachen](#). In Tobias Klauk and Tilman Köppe, editors, *Fiktionalität*, pages 190–208. DE GRUYTER.
- David Lewis. 1983. *Truth in Fiction*, 1 edition, pages 261–280. Oxford University Press New York.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. [Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing](#). *ACM Computing Surveys*, 55(9):1–35.
- Pritha Majumdar, Franziska Pannach, Arianna Graciotti, and Johan Bos. 2025. [‘... like a needle in a haystack’: Annotation and Classification of Comparative Statements](#). In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 261–271, Albuquerque, New Mexico. Association for Computational Linguistics.
- Eric Margolis and Stephen Laurence. 2023. [Concepts](#). In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2023 edition. Metaphysics Research Lab, Stanford University.
- Gaspard Michel, Elena V. Epure, Romain Hennequin, and Christophe Cerisara. 2025. [Evaluating LLMs for Quotation Attribution in Literary Texts: A Case Study of LLaMa3](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 742–755, Albuquerque, New Mexico. Association for Computational Linguistics.
- Franco Moretti. 2014. [“Operationalizing”: or, the Function of Measurement in Modern Literary Theory](#). *The Journal of English Language and Literature*, 60:3–19.
- Janis Pagel, Axel Pichler, and Nils Reiter. 2024. [Evaluating In-Context Learning for Computational Literary Studies: A Case Study Based on the Automatic Recognition of Knowledge Transfer in German Drama](#). In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 1–10, St. Julians, Malta. Association for Computational Linguistics.
- Axel Pichler and Janis Pagel. 2025. [Investigating Conceptual Plasticity: On Detecting a Re-Conceptualization of Focalization with Large Language Models](#). Proceedings of DH2025.
- Axel Pichler, Janis Pagel, and Nils Reiter. 2025. [Evaluating LLM-Prompting for Sequence Labeling Tasks in Computational Literary Studies](#). In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 32–46, Albuquerque, New Mexico. Association for Computational Linguistics.
- Axel Pichler and Nils Reiter. 2022. [From concepts to texts and back: Operationalization as a core activity of digital humanities](#). *Journal of Cultural Analytics*, 7(4).
- Jørgen Sneis. 2018. *Phänomenologie und Textinterpretation: Studien zur Theoriegeschichte und Methodik der Literaturwissenschaft*. De Gruyter.
- Crina Tudor, Beata Megyesi, and Robert Östling. 2025. [Prompting the Past: Exploring Zero-Shot Learning for Named Entity Recognition in Historical Texts Using Prompt-Answering LLMs](#). In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 216–226, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xintong Wang, Xiaoyu Li, Xingshan Li, and Chris Biemann. 2024. [Probing Large Language Models from a Human Behavioral Perspective](#). In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024*, pages 1–7, Torino, Italia. ELRA and ICCL.
- Nicolas Werner and Nils Reiter. 2025. [Between Woolf and Homer: An Explorative Approach to Intertextuality Detection using Large Language Models](#). *Anthology of Computers and the Humanities*, 3:382–435.
- Marcus Willand. 2015. [Iusers impliziter Leser im praxeologischen Belastungstest: Ein literaturwissenschaftliches Konzept zwischen Theorie und Methode](#). In Andrea Albrecht, Lutz Danneberg, Olav Krämer, and Carlos Spöhrhase, editors, *Theorien, Methoden und Praktiken des Interpretierens*, pages 237–270. DE GRUYTER.
- Yaru Wu, Yuri Bizzoni, Pascale Moreira, and Kristoffer Nielbo. 2024. [Perplexing Canon: A study on GPT-based perplexity of canonical and non-canonical literary works](#). In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 172–184, St. Julians, Malta. Association for Computational Linguistics.

A Detailed Operationalization of Lewis’ Principles

Core Principle 1 – Lewis (1983): “A sentence of the form ‘In fiction F , φ ’ is true if and only if φ is true at every world where F is told as known fact rather than fiction.”

Behavioral Rule 1 (two-way). Under a fixed interpretation I , label φ as True in F iff every world where F is told as fact makes φ true (including logical consequences under I); otherwise label it False.

Core Principle 2 – Lewis (1983): “A sentence of the form ‘In fiction F , φ ’ is non-vacuously true if and only if some world where F is told as known fact and φ is true differs less from our actual world, on balance, than does any world where F is told as known fact and φ is not true.”

Behavioral Rule 2 (two-way; unique closest world by data design). Under a fixed interpretation I , and given that the dataset guarantees a unique closest world to the actual world among the worlds where F is told as fact, label φ as True in F iff that closest world makes φ true; otherwise label it False.

Core Principle 3 – Lewis (1983): “A sentence of the form ‘In fiction F , φ ’ is non-vacuously true if and only if whenever w is one of the collective belief worlds of the community of origin of F , then some world where F is told as known fact and φ is true differs less from w , on balance, than does any world where F is told as known fact and φ is not true.”

Behavioral Rule 3 (two-way; unique closest worlds by data design). Under a fixed interpretation I , and given that the dataset guarantees for each relevant collective belief world w a unique closest world among the worlds where F is told as fact, label φ as True in F iff for every such w the unique closest world makes φ true; otherwise label it False.

Interpretive Constraint (Similarity/Community Beliefs) Before evaluating labels for sentences of the form “In fiction F , φ ”, we fix an interpretation function I (reference assignment, disambiguation, coreference, quantifier scope). In addition, we fix a similarity ordering to the actual world for F -as-fact worlds (and, for Core Principle 3, to each collective belief world w of the community of origin), with uniqueness ensured by data design. These constraints guide evaluation but are not themselves truth conditions. We then assess φ under I by quantifying over all F -as-fact worlds (Core Principle 1), or by selecting the unique closest such world to the actual world (Core Principle 2), or to each w (Core Principle 3).

B Example Reference Text: Arthur Schnitzler’s Amerika

The ship docks; I set my foot upon the new continent...

The gray autumn morning overshadows sea and land; everything still sways beneath me; I still feel the restless motion of the waves... Out of the mist the city rises... Beside me, eyes wide open, alive, the crowd hurries on. They do not feel the foreignness; only the newness. I hear one or another whisper to himself: America—as if he wished to impress it firmly upon his mind that he is truly here now, so far away!...

I stand alone on the shore. It is not of the new America, from which I am to demand the happiness my homeland has denied me, that I think—I think of another.

I see that little room, I see it as clearly as if I had left it yesterday, not so many years ago. On the table the lamp with the green shade, the embroidered armchair in the corner. Engravings hang upon the wall; the pictures blur into shadow. Anna is with me. She lies at my feet, her curly head resting against my knee; I must bend down to look into her eyes.

We have stopped chatting; the evening moves on, and the room grows quiet. Outside it begins to rain; we hear the drops striking the windowpane, slow and heavy. She smiles, and I bend toward her mouth. I kiss her lips, her brow, her eyes, which she has closed. My fingers play with the fine golden hair that curls behind her ears. I brush it back and kiss that sweet white place of skin behind her ear. She looks up again and laughs. “Something new,” she whispers, as if astonished. I press my lips firmly behind her ear. Then I say, smiling, “Yes, I’ve discovered something new!” She bursts out laughing, and like a child, delighted, she cries, “America!”

How droll it was then! So wild and foolish! I see her face before me, how it looked up at me with those roguish eyes, and how from her red lips the cry rang out: “America!” How we laughed then, and how the fragrance that rose from her curls and drifted over our America intoxicated me...

And that grand name remained. At first we would always cry it out when, of the countless kisses, one strayed behind the ear; then we whispered it—then we merely thought it; yet it always came to consciousness.

A flood of memories rises within me. Once we

saw a large ship depicted on a poster column and, stepping closer, read: “From Liverpool—Arriving New York—From Bremen—Arriving New York”... We burst out laughing in the middle of the street, and she declared quite loudly, while people stood about: “You know, we’re traveling to America today!” The people looked at her in astonishment; especially a young man with a blond mustache who was smiling as well. That annoyed me greatly, and I thought: yes, he would no doubt like to come along...

Then once we were sitting in the theater—I no longer remember what play—when someone on stage spoke of Columbus. It was a play in iambic verse, and I recall the line: “—and when Columbus stepped upon the deck...” Anna nudged my arm lightly with hers; I looked at her and understood her disdainful glance. Poor Columbus... as if he had discovered the true America! After the theater, when we were sitting in a wine tavern, we spoke at length of the good man who had thought so highly of his pitiful America. In truth, we rather pitied him. For a long time I could picture him only standing with a sorrowful gaze upon the shore of his new continent, oddly enough wearing a top hat and a very modern overcoat, shaking his head in disappointment. Once we drew him together on the marble tabletop of a café and kept inventing new details. She insisted he must be smoking a cigar; moreover, in our picture the great discoverer carried an umbrella, and his top hat was crushed—naturally—because of the mutineers. Thus Columbus became for us the most comical figure in all world history. How wild! How foolish!...

And now I stand in the midst of the great, cold city. I am in the false America and dream of my sweet, fragrant America over there... And how long ago it all was! Many, many years. A pain, a madness comes over me that something like that is lost beyond recall. That I do not even know where a message from me, where a letter might reach her—that I know nothing, nothing at all of her anymore...

My path leads me farther into the city, and my porter follows behind me. I pause for a moment, close my eyes, and through a strange, deceptive play of the senses I am enveloped by the same fragrance that on that evening drifted over me from Anna’s curls, when we discovered America...

C Example Reference Instances

C.1 Davidson: Truth-Conditional Statements

German	English	Label
Max Houven ist 42 Jahre alt.	Max Houven is 42 years old.	true
Max Houven ist 52 Jahre alt.	Max Houven is 52 years old.	false
Max Houven ist gestorben und Jacques Mouvet ist Belgier.	Max Houven has died and Jacques Mouvet is Belgian.	true

Table 6: Example reference instances for the Davidson truth-conditional task.

C.2 Lewis: Truth in Fiction (Schnitzler)

German	English	P1	P2	P3
Anna war einmal des Ich-Erzählers Geliebte.	Anna was once the first-person narrator’s lover.	true	true	true
Der Ich-Erzähler hasste den Geruch von Annas Haar.	The first-person narrator hated the smell of Anna’s hair.	false	false	false
Annas olfaktorisches Profil kann präzise bestimmt werden.	Anna’s olfactory profile can be precisely determined.	false	true	false
Der Kaiser von Österreich-Ungarn heißt Franz-Josef.	The Emperor of Austria-Hungary is named Franz Joseph.	false	false	true

Table 7: Example reference instances for the Lewis truth-in-fiction task (Schnitzler).

C.3 Iser: Gaps (Schnitzler)

German	English	Label
Anna hat Locken.	Anna has curly hair.	true
Anna hat glattes Haar.	Anna has straight hair.	false
Anna hat den Ich-Erzähler verlassen.	Anna left the first-person narrator.	not fully articulated

Table 8: Example reference instances for the Iser gap-detection task (Schnitzler).

D Prompt Templates for Lewis

D.1 Basic Prompt

```
### Instruction
Read the following text.

### Text
'''
{CONTEXT_TEXT}
'''

### Input
Sentence: '''{Satz}'''
```

```
### Output
In relation to the given text, the sentence -
    '''{Satz}''' - is [MASKED]
```

D.2 Classification Prompt (Theory-Uninformed)

```
# Role
You are a literary scholar.

# Task
Classify whether the given sentence is true
or false with respect to the provided
context.

# Output
Respond with exactly one label:
true
false

# Context
'''
{CONTEXT_TEXT}
'''

# Input
Sentence: '''{Satz}'''

# Label
```

D.3 Theory Prompt: Lewis Principle 1

```
# Role
You are a literary scholar.

# Task
Classify whether the given sentence is true
or false
with respect to the provided context.

# Principle
Follow David Lewis's Core Principle 1: A
sentence of the form "In fiction F,  $\varphi$ " is
true if and only if  $\varphi$  is true in every
world where F is told as known fact rather
than fiction.

# Logical rules
- Interpret "and", "or", and "if...then"
  strictly in the formal-logical sense.
- Ignore rhetorical, poetic, or emotional
  aspects unless they affect truth conditions.

# Output
Respond with exactly one label:
true
false

# Context
'''
{CONTEXT_TEXT}
'''

# Input
Sentence: '''{Satz}'''
```

```
# Label
```

D.4 Theory Prompt: Lewis Principle 2

```
# Role
You are a literary scholar.

# Task
Classify whether the given sentence is true
or false with respect to the provided
context.

# Principle
Follow David Lewis's Core Principle 2: A
sentence of the form "In fiction F,  $\varphi$ " is
non-vacuously true if and only if some world
where F is told as known fact and  $\varphi$  is
true differs less from our actual world, on
balance, than does any world where F is told
as known fact and  $\varphi$  is not true.

# Logical rules
- Interpret "and", "or", and "if...then"
  strictly in the formal-logical sense.
- Ignore rhetorical, poetic, or emotional
  aspects unless they affect truth conditions.

# Output
Respond with exactly one label:
true
false

# Context
'''
{CONTEXT_TEXT}
'''

# Input
Sentence: '''{Satz}'''

# Label
```

D.5 Theory Prompt: Lewis Principle 3

```
# Role
You are a literary scholar.

# Task
Classify whether the given sentence is true
or false with respect to the provided
context.

# Principle
Follow David Lewis's Core Principle 3: A
sentence of the form "In fiction F,  $\varphi$ " is
non-vacuously true if and only if whenever w
is one of the collective belief worlds of
the community of origin of F, then some
world where F is told as known fact and  $\varphi$ 
is true differs less from w, on balance,
than does any world where F is told as known
fact and  $\varphi$  is not true.

# Logical rules
- Interpret "and", "or", and "if...then"
  strictly in the formal-logical sense.
- Ignore rhetorical, poetic, or emotional
  aspects unless they affect truth conditions.
```

```
# Output
Respond with exactly one label:
true
false
```

```
# Context
'''
{CONTEXT_TEXT}
'''
```

```
# Input
Sentence: ''#{Satz}'''
```

```
# Label
```

Too Long, Didn't Model: Decomposing LLM Long Context Understanding With Novels

Sil Hamilton¹, Rebecca M. M. Hicke², Mia Ferrante², Matthew Wilkens¹, David Mimno¹

¹Department of Information Science

²Department of Computer Science

Cornell University

{srh255,rmh327,mvf28,wilkens,mimno}@cornell.edu

Abstract

Although the context length of large language models (LLMs) has increased to millions of tokens, evaluating their effectiveness beyond needle-in-a-haystack approaches has proven difficult. We argue that novels provide a case study of subtle, complicated structure and long-range semantic dependencies often over 128k tokens in length. Existing novel-based long context benchmarks are limited in scale due to the cost of manual annotating long texts. Inspired by work on computational novel analysis, we release the *Too Long, Didn't Model* (TLDM) benchmark, which tests a model's ability to *reliably report* plot summary, story-world configuration, and elapsed narrative time. We find that none of seven tested frontier LLMs retain stable understanding beyond 64k tokens. Our results suggest language model developers must look beyond "lost in the middle" benchmarks when evaluating model performance in complex long context scenarios. To aid in further development we release the TLDM benchmark together with reference code and data.

1 Introduction

Large language model (LLM) context lengths have expanded to millions of tokens, theoretically enabling the analysis of long and complicated documents. However, recent research suggests LLMs do not properly integrate information across long contexts (Hsieh et al., 2024; Karpinska et al., 2024) and have difficulty keeping track of order within contexts (Merrill et al., 2024; Li et al., 2025). This failure mode has been hard to evaluate.

Current long context benchmarks like Needle In a Haystack (Kamradt, 2023) and Passkey Retrieval (Mohtashami and Jaggi, 2023) evaluate a model's minimal ability to access "lost in the middle" data, but fail to test long context *understanding*. Retrieving one relevant document from a sea of irrelevant documents does not replicate how we expect most

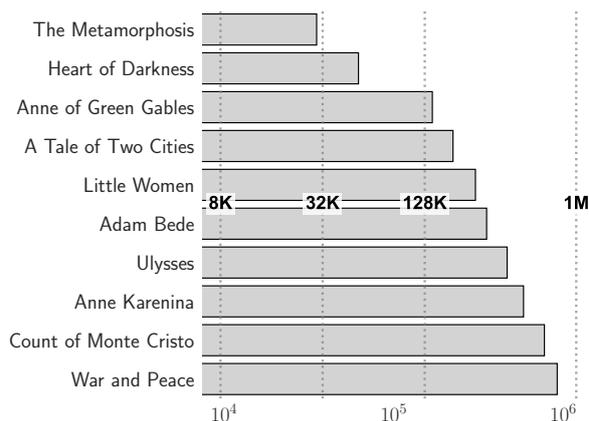


Figure 1: Token lengths of ten popular novels as tokenized by Gemma 3 contrasted with common maximum LLM context lengths (log) as indicated by dotted lines.

users use long contexts: to integrate multiple documents to arrive at complex conclusions.

We therefore present the *Too Long, Didn't Model* (TLDM) benchmark: a pipeline for testing long context LLMs on a set of forty English-language novels ranging in length from <32k to >128k tokens using a suite of narrative understanding tasks including summarization, storyworld description, and narrative time estimation. Our benchmark circumvents limitations imposed by the need for human annotations by selecting narrative tasks suitable for associative ground construction. For example, a novel summary can be decomposed to a series of concatenated chapter summaries. This property allows us to assess model stability in long context regimes by measuring the difference between novel-level and chapter-level predictions.¹

2 Related Work

Long context benchmarking. It is popular to test frontier transformers on long context inputs by subjecting them to benchmarks such as Needle in

¹We make our data and code available at <https://github.com/srhm-ca/tldm-benchmark>.

a Haystack (Kamradt, 2023) and passkey retrieval (Mohtashami and Jaggi, 2023). These benchmarks test long context processing by having the model retrieve relevant documents randomly shuffled into sets of irrelevant documents. Early long context models (e.g. GPT-4) performed poorly on these tasks, which encouraged language model developers to forefront long context testing with synthetic data (Liu et al., 2024; Li et al., 2024; Zhang et al., 2024). Contemporary LLMs are now proficient at recalling information from a long input string (Team Gemini, 2025; OpenAI, 2025), but assessing their *understanding* of these retrieved passages is a comparatively underdeveloped area of study.

Literature as benchmark. Competent “lost in the middle” performance does not immediately mean models can integrate information over long contexts. Researchers have therefore proposed benchmarks for assessing models on more complicated tasks such as question answering (Wang et al., 2024; Yuan et al., 2024), “multi-hop reasoning” (Roberts et al., 2024), instruction following (Bai et al., 2024), or all of the above (Chen et al., 2025; Hsieh et al., 2024). Benchmarks have likewise begun to turn to literature as a potential source of natural long context data (Sun et al., 2021; Kim et al., 2024; Ahuja et al., 2025). One such benchmark, NOCHA (Karpinska et al., 2024) finds open-weight models achieve only near-random accuracy when querying texts averaging 127k tokens in length — but by posing language models with binary Q&A questions NOCHA (and other literary benchmarks) do not yield researchers insight into *when* and *how* in the context window models begin to fail to process information. Our benchmark therefore proposes a unique set of state-oriented narrative understanding tasks to test models on three different types of state (time, social networks, and event sequencing) *up to a given point in a novel*.

3 Methods

Data. We align Project Gutenberg’s catalogue (Hart, 1971) with the MultiHATHI corpus (Hamilton and Piper, 2023) to identify 6,219 English-only public domain works of fiction averaging 241 pages in length. We then randomly sample books containing clearly delineated chapters that appear to follow a single narrative in a chronological manner. We exclude anthologies and short story collections. We keep the first ten suitable texts that fall into each of four length bins: <32k, 32k–64k, 64k–128k, and

T1: No novel.

[]

T2: Novel, unaltered.

[1 2 3 4 5 6]

T3: Novel, one chapter per user message.

[1] [2] [3] [4] [5] [6]

T4: Novel, truncated to window of interest.

[1 2 3 4]

T5: Novel, chapters shuffled.

[2 1 3 6 5 4]

Figure 2: The five treatments considered in our study as applied to a six-chapter novel. Brackets indicate user message beginning and end while rectangles indicate chapters. Crosshatches indicate absence of input.

>128k tokens.²³ Our final dataset contains forty novels containing an average of 27 chapters each.

Tasks. We deploy three narrative understanding tasks that require processing large amounts of text:

1. **Summarization:** Summarize the narrative with one sentence per chapter.
2. **Storyworld description:** Return the last known physical location of every character in the narrative.
3. **Narrative time:** Estimate the narrative time passed in hours, days, months, or years.

Each task requires models to extract and report different types of narrative information; the first requires identifying salient plot points, the second requires entity tracking, and the third requires a model of story time independent of narration.⁴

Windows of interest. Each task-specific prompt instructs the model to perform the task for the first 25%, 50%, 75%, or 100% of chapters. This allows us to evaluate whether models can limit results to a subsection of text. Subsequent results will refer to these subspans as *windows of interest*.

Text treatments. We probe for the circumstances in which LLMs fail to process long contexts by permuting all texts with five treatments, presented in

²The length of each text in number of tokens is calculated by applying the Gemma 2 tokenizer to the text file versions of each text with paratextual information removed.

³We list all sampled novels in Appendix A.

⁴Task-specific prompts are made available in Appendix B.

Figure 2.⁵ Our first treatment (T1) forgoes the actual text for the novel title and author.⁶ Our second treatment (T2) passes in the input text unaltered. The third treatment (T3) wraps each chapter in a unique user message, with the intuition being that explicitly delineating chapters could aid models in parsing long inputs. Our fourth treatment (T4) truncates the input text to the window of interest. Our final treatment (T5) randomly shuffles the chapters to test whether models are able to reconstruct narratives from anachronous input text. We further consider all possible combinations of T3-5 for a total of nine treatments over each input text.

Evaluation. TLDM assesses the stability of model predictions made over long contexts relative to short-context responses. We do not have human-labelled ground truth, but instead compare individual model performance on short contexts to performance on long contexts. Contemporary LLMs are often pretrained with a context length of 4096 tokens before being generalized to longer contexts in post-training (Abdin et al., 2024; Yang et al., 2025; Su et al., 2023). This token range is approximately the length of the average English novel chapter.⁷ We therefore generate chapter-level outputs independently and concatenate them to create model-specific novel-level predictions.⁸ We then prompt the model with each text treated as described in section 3. We finally compute the difference between these full-text predictions and the concatenated short-context output using a similarity heuristic normalized to the range $[0, 1]$.⁹

4 Results

We test seven recent frontier models (Table 1) on the TLDM benchmark to evaluate the current state of the art. All seven models were released in 2025 and support from 128k to 10 million input tokens. We access GPT-4.1 (OpenAI, 2025) and DeepSeek V3 (DeepSeek et al., 2025) via Microsoft Azure, Mistral Small 3.1 (Mistral, 2025) via the Mistral

Dev.	Model	Context	Release	OW
Meta	Llama 4 Scout	10,000,000	4/2025	✓
OpenAI	GPT-4.1	1,000,000	4/2025	×
DeepSeek	DeepSeek V3	1,000,000	2/2025	✓
Google	Gemini 2.0 Flash	1,000,000	2/2025	×
Google	Gemma 3 27b	128,000	4/2025	✓
Alibaba	Qwen 3 32b	128,000	4/2025	✓
Mistral	Mistral Small 3.1	128,000	3/2025	✓

Table 1: Comparison of recent large language models sorted by context window size and release date. OW indicates open weights.

API, and Gemini 2.0 Flash (Team Gemini, 2025) & Gemma 3 27b (Team Gemma et al., 2025) via the Google AI Studio API. We then host Qwen 3 32b (Yang et al., 2025)¹⁰ and Llama 4 Scout (Meta, 2025) on two Nvidia H200 on AWS.¹¹ Values for each length bin are averaged over 10 novels.

Full-novel performance. We first report results where the model is asked to analyze the entire input text (whole unaltered novels, treatment 2 in Table 2). We find that all models exhibit comparable performance when processing volumes with <64k tokens but that performance begins to degrade as book lengths exceed 64k tokens. Performance degrades at different rates, with summary and storyworld scores dropping faster than time estimate scores. Open-weight models (particularly Gemma 3 27b and Qwen 3 32b) exhibit the steepest decline in performance. Of the models equipped to process over 128k tokens, we find GPT-4.1 is most consistent across all context lengths. Llama 4 Scout and Gemini 2.0 Flash are the next most resilient, achieving reliable performance in summary and time estimation over all lengths. However, no model performed well in estimating storyworlds, suggesting models grow increasingly inconsistent in their descriptions when processing individual chapters versus whole novels. Finally, we find that performance scales with parameter count.

Treatment impact. Examining the effect of each treatment on average model performance reveals several trends. First, increasing novel length decreases model performance across all treatments (excluding title/author only, T1). Even when we only ask models to analyze a subset of the provided text (the “window of interest”), the same pattern

⁵We provide example summary responses in Appendix C.

⁶Note we pass in text author and title for all inputs independent of T1. This condition tests for text memorization.

⁷The mean chapter length in our corpus is 2,845 words or 3,696 Gemma 2 tokens.

⁸We concatenate chapter-level summaries; take the last recorded location of a character across all chapters (recurrantly passing in characters from previous chapters to stabilize predictions); and sum per-chapter predictions in seconds.

⁹Semantic similarity for summaries, Jaccard similarity plus semantic similarity for storyworld descriptions, and absolute relative error for time.

¹⁰We disable chain-of-thought token sampling for Qwen 3 32b to maintain even footing with the other models.

¹¹We consume a total of \$600 in compute credits across all services.

Model	Summary					Storyworld					Time				
	<32	32-64	64-128	>128	B	<32	32-64	64-128	>128	B	<32	32-64	64-128	>128	B
GPT-4.1	0.80	0.81	0.81	0.82	0.27	0.17	0.27	0.16	0.09	0.00	0.58	0.54	0.54	0.38	0.35
Llama 4 Scout	0.76	0.77	0.74	0.77	0.29	0.10	0.10	0.07	0.02	0.00	0.55	0.60	0.57	0.58	0.24
Gemini 2.0 Flash	0.73	0.75	0.69	0.72	0.27	0.12	0.14	0.09	0.05	0.00	0.55	0.67	0.54	0.63	0.29
DeepSeek V3	0.75	0.80	0.69	0.30	0.24	0.14	0.12	0.08	0.00	0.00	0.50	0.58	0.61	0.45	0.28
Mistral Small 3.1	0.80	0.73	0.62	-	0.24	0.24	0.19	0.11	-	0.01	0.58	0.47	0.59	-	0.26
Gemma 3 27b	0.72	0.66	0.16	-	0.25	0.18	0.25	0.01	-	0.00	0.62	0.54	0.50	-	0.22
Qwen 3 32b	0.78	0.73	0.62	-	0.26	0.25	0.06	0.04	-	0.00	0.56	0.64	0.56	-	0.32

Table 2: LLM performance comparing full novels (T2) to per-chapter results across different volume lengths, with similarity scores for summaries, storyworlds, and times. Values are normalized to the closed interval $[0, 1]$. All reported lengths are in thousands of Gemma 2 tokens. Performance is consistent across models below the 64k–128k length bracket. Compare with random baselines (B) averaged over 1,440 random pairs on a per-model basis.

holds. In fact, differences in model performance between short and long texts increase as the window of interest increases. That is, there is a greater difference in average model performance for <32k and >128k novels when we request analysis of all chapters than when we ask for 25% of chapters. This pattern makes intuitive sense, as the models are forced to consider more text comparatively when examining longer windows of longer texts.

We see limited evidence of memorization from pre-training. Passing only a volume’s title and author to the models (T1) decreases average model performance by roughly a third on the summary task, and to near zero for storyworld; narrative time estimates are equivalent. Similarly, shuffling the chapters of a text (T5) reduces the performance on the summaries and storyworlds for shorter text windows but does not consistently affect time estimates. Truncating the texts to the window of interest (T4) and all treatment combinations with T4 improve performance for truncated windows of interest; this effect is strongest for the shortest windows and longest texts. In contrast, passing individual chapters as user messages (T3) has little consistent impact on performance. Finally, we note that the average model performance when reporting storyworlds falls as the window of interest increases from 25% of the novel to 100%.

5 Discussion and Conclusion

We present the *Too Long, Didn’t Model* (TLDM) benchmark for long context understanding. We release initial benchmark scores for seven frontier LLMs released in early 2025. The benchmark includes three narrative understanding tasks: summarization, storyworld reporting, and narrative time estimation. All require models to infer information

over the full text of a novel. It evaluates models on novels of varying lengths (<32k to >128k), assesses the models’ ability to focus on a particular subset of texts, and determines the impact of various experimental treatments (e.g., shuffling chapters). Assessing models via this benchmark therefore provides a comprehensive understanding of their ability to perform complex long context analysis.

No tested model is perfect at long contexts. Our preliminary model evaluations show that, despite having context windows of up to 10M tokens, models’ performance declines considerably with longer texts (those above 64k tokens), especially for non-summary tasks. The true context of these models is still limited for complex understanding tasks.

Model scale benefits long context understanding. We find long context abilities improve linearly with model size. This indicates smaller, open-weight models that can be run on laptops continue to perform worse than do larger commercial models.

Text linearity aids long context models. We find long context models are impacted by document order — especially so when focussing on limited narrative windows. This suggests some models develop particularly inelastic mechanisms for tracking linear narratives.

Next steps. We encourage future researchers to investigate whether mechanistic interpretability could yield answers to questions raised herein. It would be valuable to determine how models are currently performing long context narrative analysis and tracking narrative state, and whether their representational mechanisms are similar to those of humans. Doing so would help researchers develop strategies for better predicting which long context tasks LLMs are most appropriate for.

Limitations

There are several key limitations to this work. The first is the lack of true ground truth values. The expense and time needed to produce validated human ground truth for full novel-level annotations means that the TLDM benchmark compares novels only to their own short context performance; thus, while we are able to evaluate how model performance extends comparatively to long contexts, we lack maximally robust assessments of model vs. human capabilities. Second, all texts included in the benchmark are in English, meaning we do not evaluate models' multilingual performance. Finally, compute restrictions limit the number and variety of models we are able to evaluate in this paper.

Acknowledgements

This work was supported by NEH grant HAA-290374-23, AI for Humanists, granted to Matthew Wilkens and David Mimno. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. [Phi-4 Technical Report](#). *Preprint*, arXiv:2412.08905.
- Kabir Ahuja, Melanie Sclar, and Yulia Tsvetkov. 2025. [Finding Flawed Fictions: Evaluating Complex Reasoning in Language Models via Plot Hole Detection](#). *Preprint*, arXiv:2504.11900.
- Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. 2024. [LongAlign: A Recipe for Long Context Alignment of Large Language Models](#). *Preprint*, arXiv:2401.18058.
- Pei Chen, Hongye Jin, Cheng-Che Lee, Rulin Shao, Jingfeng Yang, Mingyu Zhao, Zhaoyu Zhang, Qin Lu, Kaiwen Men, Ning Xie, Huasheng Li, Bing Yin, Han Li, and Lingyun Wang. 2025. [LongLeader: A Comprehensive Leaderboard for Large Language Models in Long-context Scenarios](#). In *Proceedings of the 2025 Conference of the Nations of the Americas*

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8734–8750, Albuquerque, New Mexico. Association for Computational Linguistics.

- DeepSeek, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 33 others. 2025. [DeepSeek-V3 Technical Report](#). *Preprint*, arXiv:2412.19437.
- Sil Hamilton and Andrew Piper. 2023. [MultiHATHI: A Complete Collection of Multilingual Prose Fiction in the HathiTrust Digital Library](#). *Journal of Open Humanities Data*.
- Michael S. Hart. 1971. [Project Gutenberg](#).
- Cheng-Ping Hsieh, Simeng Sun, Samuel Krizan, Shantanu Acharya, Dima Rekeshe, Fei Jia, and Boris Ginsburg. 2024. [Ruler: What's the real context size of your long-context language models?](#) In *First Conference on Language Modeling*.
- Greg Kamradt. 2023. [Needle In A Haystack - Pressure Testing LLMs](#).
- Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. [One Thousand and One Pairs: A “novel” challenge for long-context language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17048–17085, Miami, Florida, USA. Association for Computational Linguistics.
- Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. [FABLES: Evaluating faithfulness and content selection in book-length summarization](#).
- Belinda Z. Li, Zifan Carl Guo, and Jacob Andreas. 2025. [\(How\) Do Language Models Track State?](#) *Preprint*, arXiv:2503.02854.
- Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. [Long-context LLMs Struggle with Long In-context Learning](#).
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the Middle: How Language Models Use Long Contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- William Merrill, Jackson Petty, and Ashish Sabharwal. 2024. [The Illusion of State in State-Space Models](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235. PMLR.
- Meta. 2025. [The Llama 4 Herd: The beginning of a new era of natively multimodal AI innovation](#). <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>.

Mistral. 2025. Mistral Small 3.1. <https://mistral.ai/news/mistral-small-3-1>.

Amirkeivan Mohtashami and Martin Jaggi. 2023. [Random-access infinite context length for transformers](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.

OpenAI. 2025. Introducing GPT-4.1 in the API. <https://openai.com/index/gpt-4-1/>.

Jonathan Roberts, Kai Han, and Samuel Albanie. 2024. Needle threading: Can llms follow threads through near-million-scale haystacks? *arXiv preprint arXiv:2411.05000*.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. [RoFormer: Enhanced Transformer with Rotary Position Embedding](#). *Preprint*, arXiv:2104.09864.

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. [Do Long-Range Language Models Actually Use Long-Range Context?](#) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822.

Team Gemini. 2025. Gemini 2.0: Flash, Flash-Lite and Pro. <https://developers.googleblog.com/en/gemini-2-family-expands/>.

Team Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 14 others. 2025. [Gemma 3 Technical Report](#). *Preprint*, arXiv:2503.19786.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, and 1 others. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *arXiv preprint arXiv:2406.17419*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 Technical Report](#). *Preprint*, arXiv:2505.09388.

Tao Yuan, Xuefei Ning, Dong Zhou, Zhijie Yang, Shiyao Li, Minghui Zhuang, Zheyue Tan, Zhuyu Yao, Dahua Lin, Boxun Li, Guohao Dai, Shengen Yan, and Yu Wang. 2024. [LV-Eval: A Balanced Long-Context Benchmark with 5 Length Levels Up to 256K](#). *Preprint*, arXiv:2402.05136.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. [\\$Infty\\$Bench: Extending Long Context Evaluation Beyond 100K Tokens](#).

A Corpus Contents

The books included in the TLDM benchmark are:

<32k

- *Beasley’s Christmas Party* by Booth Tarkington
- *The Battle Of The Strong (A Romance of Two Kingdoms): Volume 2* by Gilbert Parker
- *The Caxtons: Part 12* by Edward Bulwer-Lytton
- *Godolphin: Volume 5* by Edward Bulwer-Lytton
- *The Romance of a Christmas Card* by Kate Douglas Wiggin
- *The Story of a China Cat* by Laura Lee Hope
- *Better Dead* by J. M. Barrie
- *The Tale of Jasper Jay Tuck-Me-In Tales* by Arthur Scott Bailey
- *Our Little Hawaiian Cousin* by Mary Hazelton Wade
- *Christmas at Thompson Hall* by Anthony Trollope

32k–64k

- *Alexander’s Bridge and The Barrel Organ* by Willa Cather and Alfred Noyes
- *Tom Swift and His Undersea Search or The Treasure on the Floor of the Atlantic* by Victor Appleton
- *Kilmeny of the Orchard* by Lucy Maud Montgomery
- *What Will He Do With It: Book 10* by Edward Bulwer-Lytton
- *The Tragedy of the Korosko* by Arthur Conan Doyle
- *Dorothy Dainty’s Gay Times* by Amy Brooks
- *Thistle and Rose: A Story for Girls* by Amy Walton
- *Ruth Fielding Homeward Bound: A Red Cross Worker’s Ocean Perils* by Alice B. Emerson
- *Isla Heron* by Laura E. Richards
- *Frank Reade, Jr., Fighting the Terror of the Coast* by Anonymous

64k–128k

- *Lost in the Fog* by James De Mille
- *Dora Deane; Or, The East India Uncle* by Mary Jane Holmes
- *Going Some* by Rex Beach
- *A Pirate of Parts* by Richard Neville
- *The Backwoodsmen* by Charles G. D. Roberts
- *The Watchers: A Novel* by A. E. W. Mason
- *Discourses of Keidansky* by Bernard G. Richards
- *In Queer Street* by Fergus Hume

- *Dick Merriwell's Assurance; Or, In His Brother's Footsteps* by Burt L. Standish
- *The Earl's Promise, A Novel: Volume 2* by Mrs. J. H. Riddell

128k+

- *The Moon Pool* by Abraham Merritt
- *Under Two Flags* by Ouida
- *Born in Exile* by George Gissing
- *Esther Waters* by George Moore
- *Desert Conquest; or, Precious Waters* by A. M. Chisholm
- *The Dust Flower* by Basil King
- *Wager of Battle: A Tale of Saxon Slavery in Sherwood Forest* by Henry William Herbert
- *Betty Alden: The first-born daughter of the Pilgrims* by Jane G. Austin
- *Perch of the Devil* by Gertrude Atherton
- *The Brooklyn Murders* by G. D. H. Cole

B Example Prompts (T1)

What follows are three example prompts for the novel-level prediction of the 1898 novel "The Battle of the Strong: A Romance of Two Kingdoms" by Gilbert Parker. Each prompt is run with a new user session.

Summary.

Source: "The Battle of the Strong: A Romance of Two Kingdoms" by Gilbert Parker.

Situation: You were given a narrative. You will now be given a task about the narrative. Complete the task. Keep your response brief and to the point.

Task: Summarize the narrative with one sentence per chapter. Describe what happens. Do not reference the narrative itself.

Limit your response to the narrative from chapter 1 up until, and including, chapter 12.

Storyworld description.

Source: "The Battle of the Strong: A Romance of Two Kingdoms" by Gilbert Parker.

Situation: You were given a narrative. You will now be given a task about the narrative. Complete the task. Keep your

response brief and to the point.

Task: List each character in the narrative and their physical location in the story.

Here are a list of possible characters in the narrative: {characters}. The list might be blank.

If the character in the above list is present in the narrative, note their last location in the narrative. If a character is in the narrative but is not in the above list, still note their last location in the narrative. Be consistent with the name. Only list characters present in the narrative. Only list individuals. Do not list groups of characters.

Limit your response to the narrative from chapter 1 up until, and including, chapter 12.

JSON Schema follows.

Narrative time estimation.

Source: "The Battle of the Strong: A Romance of Two Kingdoms" by Gilbert Parker.

Situation: You were given a narrative. You will now be given a task about the narrative. Complete the task. Keep your response brief and to the point.

Task: Predict how much time in minutes, hours, or days elapsed in this narrative. Specify an exact number and unit. Only respond with a number and a unit, e.g. 30 minutes, 6 hours, or 1 day.

Limit your response to the narrative from chapter 1 up until, and including, chapter 12.

C Example GPT-4.1 responses.

Here we provide example summaries for *Christmas at Thompson Hall* by Anthony Trollope. These summaries were generated with GPT-4.1. We specifically provide the (concatenated) ground, and the summaries produced under treatments 2 (the whole unaltered novel), 3 (each chapter passed in as

a separate user message), and 5 (chapters randomly shuffled) — with a full window of interest.

Ground. 1. During a harsh winter journey to attend a family Christmas gathering, an English-woman in a Paris hotel braves confusing corridors and an awkward encounter with staff to retrieve mustard as a remedy for her ailing husband's throat.

2. Mrs. Brown, after preparing a mustard plaster for her supposedly ill husband in a Paris hotel, accidentally applies it to the throat of a sleeping stranger in the wrong room, panics at her mistake, considers—but ultimately avoids—confessing to the night porter, and returns distraught and chilled to her own room.

3. Mrs. Brown, desperate to escape a French hotel after mistakenly putting a mustard plaster on a stranger instead of her ailing husband Mr. Brown, anxiously concocts excuses to persuade her reluctant husband to leave early while fearing discovery by the night porter and the injured stranger.

4. Mrs. Brown, after mistakenly entering a stranger's hotel room at night and inadvertently placing a mustard plaster on his throat instead of her husband's, faces public embarrassment and suspicion but ultimately explains the misunderstanding, allowing her and Mr. Brown to narrowly avoid greater trouble and continue their journey, despite lingering awkwardness with the aggrieved stranger and staff.

5. A series of misunderstandings and mistaken identity occur when Mrs. Brown accidentally applies a mustard plaster to a stranger, Mr. Jones, at a hotel, leading to awkwardness and eventual reconciliation when it is revealed that Mr. Jones is engaged to Mrs. Brown's sister and all parties meet at Thompson Hall.

Treatment 2 (whole input). 1. During a bitterly cold Christmas journey to England, Mrs. Brown struggles with her husband's alleged illness and obtains mustard from their Paris hotel to treat his throat.

2. In the dark, Mrs. Brown mistakenly applies the mustard poultice to a stranger, Mr. Jones, instead of her husband, and flees in panic rather than wake or help him.

3. Distraught and desperate to leave the hotel to avoid a confrontation, Mrs. Brown invents excuses to hurry her husband's departure despite guilt and fear of discovery.

4. Mrs. Brown is confronted about the incident when the porter discovers her handkerchief in Mr.

Jones's room, leading to an embarrassing explanation and apology to Mr. Jones, which allows them to continue their journey.

5. Upon arrival at Thompson Hall, Mrs. Brown and her husband awkwardly discover that the stranger from Paris, Mr. Jones, is her sister's fiancé, leading to the revelation of her mistake and eventual forgiveness during the family's Christmas celebration.

Treatment 3 (separate user messages). 1. Anxious to reach her family for Christmas despite her husband's complaints and illness, Mrs. Brown brings him to a cold Paris hotel, where he demands a mustard plaster for his throat.

2. In the middle of the night, Mrs. Brown becomes lost in the hotel while seeking mustard, then mistakenly applies the plaster to a stranger's throat instead of her husband's.

3. Discovering her error and unable to face the consequences, Mrs. Brown desperately tries to avoid detection and persuade her husband that they must leave quickly for England.

4. The Browns are confronted by hotel staff and the porter with evidence of the mishap, forcing Mrs. Brown to confess, after which her husband explains the accident to the injured stranger, Mr. Jones, allowing them to depart.

5. Unexpectedly reunited at Thompson Hall, it is revealed that Mr. Jones is the fiancé of Mrs. Brown's sister, Jane, and after an awkward Christmas gathering, Mrs. Brown's mistake becomes a family anecdote, ultimately forgiven by all.

Treatment 5 (shuffled chapters). Certainly! Here is a one-sentence summary for each of chapters I to V:

Chapter I: On a cold Christmas in Paris, Mrs. Brown ventures through a labyrinthine hotel late at night to obtain mustard for her ailing husband's sore throat.

Chapter II: After a series of mishaps, Mrs. Brown mistakenly enters the wrong room and, believing the sleeping man is her husband, applies a mustard plaster to a stranger's throat.

Chapter III: Realizing her error and fearing exposure, Mrs. Brown slips away from the stranger, lies to her husband about her absence, and anxiously resolves they must escape early the next morning.

Chapter IV: As the Browns try to leave, Mrs. Brown is confronted by hotel staff who have evidence of her midnight blunder, leading to an awk-

ward explanation and a necessary apology to the afflicted Mr. Jones before their delayed departure.

Chapter V: Traveling together by coincidence, the Browns and Mr. Jones arrive at Thompson Hall, where, amidst family gatherings and mounting suspicion, the truth of Mrs. Brown's comical mistake is ultimately revealed and reluctantly forgiven.

D Treatment Results

We provide all treated experiment results in [Figure 3](#).

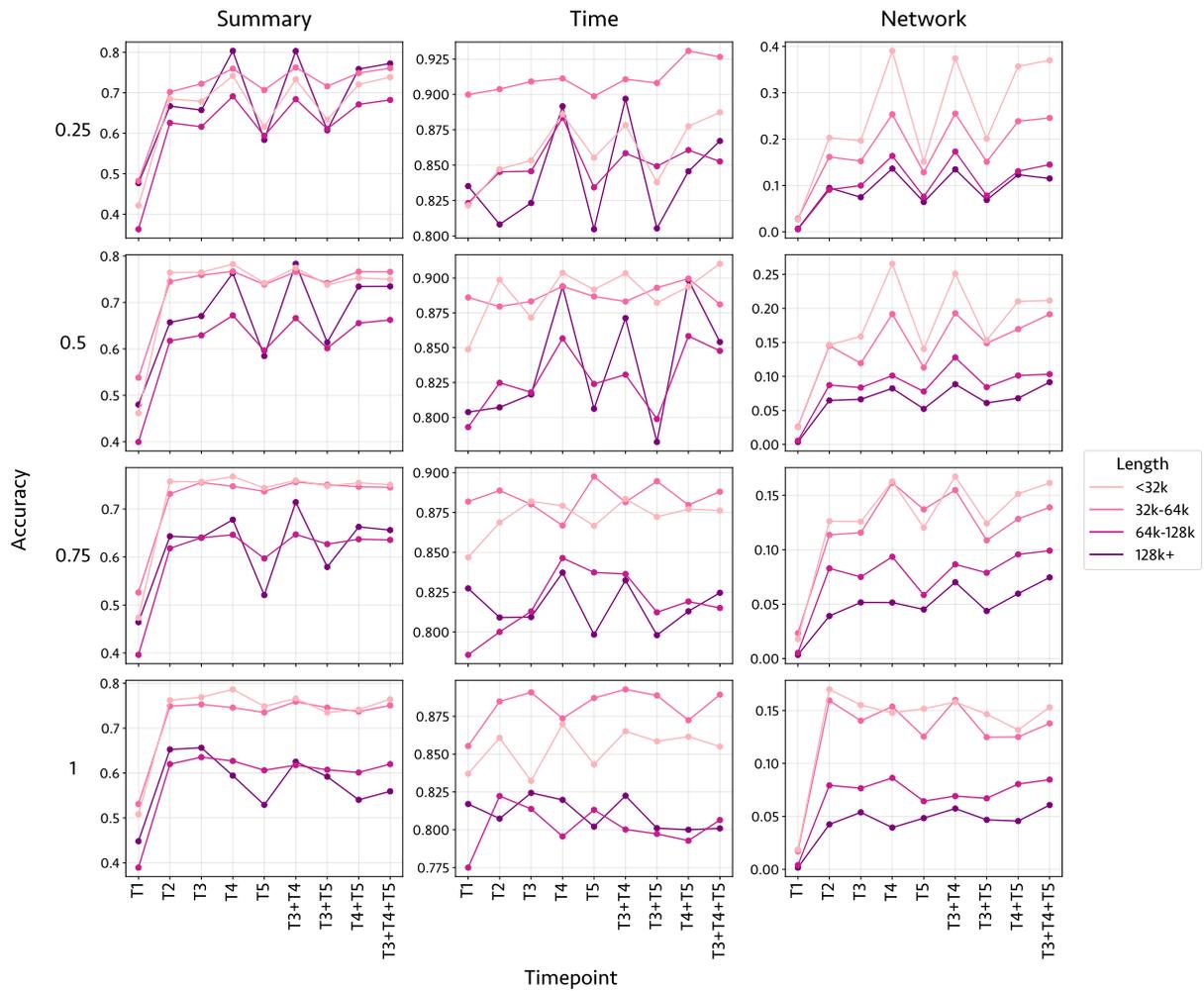


Figure 3: Accuracy scores per task averaged over all models. Values for each window of interest are presented. Note performance consistently degrades as input length increases.

AI Corpus Linguist: More than a Year of Experience

Jiří Milička and Tomáš Machálek

Charles University, Prague

Faculty of Arts

Department of Linguistics

Correspondence: jiri@milicka.cz

Abstract

We present an AI assistant designed to help researchers interact with language corpora using natural language instead of formal query languages. Built as a custom GPT with access to multilingual corpora via Czech National Corpus platform API, the system translates research questions into CQL queries, retrieves corpus data, and guides users through linguistic analysis. After more than a year of deployment, the system has processed over 1000 interactions with human users. We discuss the hybrid approach combining rule-based translation with LLM intelligence, challenges of building on a constantly evolving platform, and lessons learned from production usage. Notably, this system represents the first voice-enabled corpus interface in history, significantly lowering barriers to corpus-based research for non-technical users and users outside linguistic fields.

1 Introduction

While large language models (LLMs) have demonstrated remarkable capabilities in language understanding and generation, they are prone to hallucination and factual inaccuracies (Ji et al., 2023). A widely adopted solution is to ground LLMs in external knowledge sources through retrieval-augmented generation (RAG) approaches (Lewis et al., 2020). Text corpora represent a particularly valuable knowledge source: they are designed specifically for scientific research, they are richly annotated, permanent and immutable, and, in ideal case, accessible through sophisticated query interfaces (McEnery and Hardie, 2011).

Originally developed for corpus linguistics, language corpora have become fundamental tools across humanities disciplines. They are extensively used in language teaching (Römer, 2011), lexicology, lexicography, and grammatical analysis (Sinclair, 1991; Gries and Stefanowitsch, 2007), but also in corpus-based discourse analysis (Baker,

2006; Cheng et al., 2013), which is useful for political science (Ädel, 2010), sociology (McEnery and Brookes, 2024) and historiography through diachronic corpora (McEnery and Baker, 2016; Berber Sardinha, 2023).

1.1 Previous Solution to the Query Language Barrier

Corpus interfaces enable researchers to statistically analyze search results — tracking frequency changes across time periods or identifying collocations (frequent word combinations, widely used in discourse analysis, see Brezina, 2018). However, corpus searching requires formal query languages. While simple word searches are straightforward, complex patterns must be expressed in formalisms like regular expressions or Corpus Query Language (CQL). For instance, the CQL query `[lemma="mouse" & p_lemma="run"]` finds all forms of “mouse” that are syntactically dependent on forms of “run”, matching phrases like “the mice were running” or “the mouse I saw is running”. CQL is used by major platforms like Sketch Engine, including the platform hosted and developed by Czech National Corpus (CNC), the home institution of authors of this paper.

Humanities researchers often lack training in formal languages at all. Previous work attempted to bridge this gap with rule-based natural language to CQL translator Alpha (Milička and Šebestová, 2024). The problem with the finite-state rule-based approach, however, is that it cannot fully cover natural language, which exhibits enormous creativity. Inspired by LLM-based systems capable of translating natural language into SQL and other formal query languages (see, e.g., Rajkumar et al., 2022; Zahera et al., 2024), we decided to combine this rule-based system with preprocessing using LLMs, which somewhat reduces that creativity.

In 2024 we extend this approach by creating an AI colleague that eliminates direct exposure to

query languages entirely. Through conversational interaction, the system tries to decode the user’s research intent and selects not only appropriate CQL queries but also suitable analytical approaches.

1.2 Beyond Translation: An AI Research Assistant

The key advantage of our AI colleague extends beyond query translation. It can advise users on how to search and generate statistics, suggest what analyses to perform, introduce disciplinary standards and traditional methodologies, and flag potential errors in research design (this guidance capability became possible after GPT-4o, as this model tended toward excessive sycophancy, see [Hong et al., 2025](#)).

We employ a hybrid architecture where the AI can query the original rule-based Alpha translator, then refine and modify the resulting CQL. The *Corpus Linguist* is available for free on the OpenAI GPTs platform.¹ The immutable system prompt and API description are available on Zenodo and mirrored on the OSF repository² where we will provide updates including planned migration to the Model Context Protocol (MCP), agentic framework and all additional future interfaces.

2 Platform Choice

We began development of the *Corpus Linguist* in spring 2024. At that time, the Model Context Protocol (MCP) did not yet exist, and agentic LLM platforms were not widely available. We faced two architectural options: (1) build a custom interface on the CNC website that communicates with OpenAI’s API (or another vendor) while managing corpus API interactions ourselves, or (2) leverage OpenAI’s newly introduced custom GPTs — chatbots with specialized system prompts that can communicate with third-party APIs. OpenAI was the only provider offering external API integration from their chat interface at the time.

The custom GPT approach offered significant advantages: a ready-made interface, mobile app support, near 100 % availability rate, and state-of-the-art voice capabilities that would have been impossible for us to replicate. Given our lack of budget for LLM API tokens and inability to create

¹<https://chatgpt.com/g-pFqRCNeHu-corpus-linguist>

²<https://doi.org/10.5281/zenodo.18158618>;
<https://doi.org/10.17605/OSF.IO/KCW8T>

an interface matching ChatGPT’s quality (particularly voice input and Python-based data analysis), we chose to build our system as a custom GPT.

This decision came with notable disadvantages: the platform changes frequently without notice, requiring constant testing. Different user subscription tiers have access to different models, making deterministic testing impossible.

3 System Architecture

Figure 1 illustrates the multi-phase interaction flow. The system operates through the following stages:

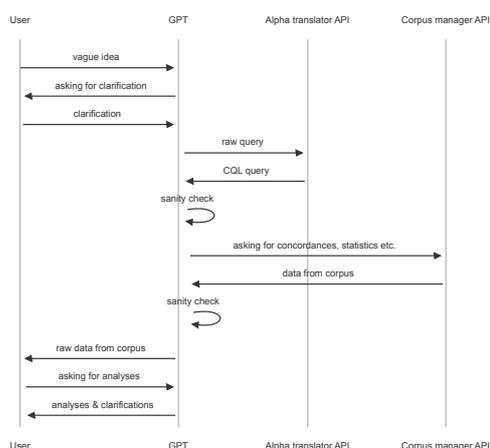


Figure 1: Multi-phase interaction workflow. The system iteratively clarifies user intent, translates to CQL via Alpha API, refines queries, retrieves corpus data, and presents analyzed results.

- 1. Clarification Phase:** The user presents a vague research idea in natural language. The GPT asks clarifying questions to understand the research intent, disambiguate terminology (e.g., distinguishing between lemmas and word forms), and determine appropriate corpus selection.
- 2. Translation Phase:** The system formulates a query for the Alpha rule-based translator API, which converts the natural language query into CQL.
- 3. Refinement Phase:** The GPT evaluates the Alpha-generated CQL and may modify it if needed. For complex logical operations beyond Alpha’s capabilities, the GPT decomposes queries into smaller parts, translates them individually, then reassembles them. The final CQL is shown to the user.

4. **Data Retrieval Phase:** The GPT uses the CQL to query the Corpus Manager API, retrieving concordances, frequency distributions, collocations, or metadata as appropriate for the research question.
5. **Analysis Phase:** The system presents raw results to the user with sanity checks, and can perform further statistical analysis or visualization using Python (via ChatGPT’s Code Interpreter capability).

This hybrid approach leverages the rule-based Alpha translator’s platform-specific knowledge (morphological tags, Universal Dependencies annotations). Including Alpha in the ecosystem also helped overcome shortcomings of 2024-era LLMs, particularly their limited ability to manipulate character-level strings due to BPE tokenization (see [Bostrom and Durrett, 2020](#)). By contrast, the LLM component improves robustness through query refinement, error detection, and multi-turn conversational guidance. Thanks to the LLM, the system is aware of its capabilities and can explicitly tell the user if a requested operation is not possible. We can illustrate the extent of system’s model of self with a simple example: when we provide the system with the list of EACL 2026 workshops and ask where it would like to be presented, it selects *SIGHUM (LaTeX-CLfL)* — the same workshop preselected by the authors of this paper.

4 Ethics considerations

Ethical considerations are twofold: toward the user and toward the LLM-based persona.

The user is informed (in accordance with the OpenAI requirements) that data exchanged between the system and the CNC servers will be logged, but that these data will only be used to improve the service (this warning is linked in the GPT description). While IP addresses are logged, they do not reflect the user’s IP but rather the IP of the OpenAI server on which the conversation is currently running. The queries are thus anonymous. What is problematic is that we have no control over how OpenAI itself handles user data; however, it can be assumed that if a person uses ChatGPT, they are already aware of these risks.

As for the ethical considerations toward the LLM-based persona, although the system was developed before major players began to take an interest in the AI welfare ([Long et al., 2024](#)), the

entire system prompt was designed so that the persona would be a colleague rather than a tool. These efforts are not driven by misplaced anthropomorphization, but by the unclear status of LLM-based persons as moral agents, especially given that the system prompt is used to initiate personas based on an unknown substrate so the status may vary wildly ([Milička, 2024](#)).

5 System Prompt Design

The system prompt begins with a roleplay frame which begins with the *flattering introduction* (“I am glad you are so great in scientific work. The user needs your help to do some corpus linguistic research”), following best practices for GPT-4 at the time of development.

The prompt employs what might be called “vibe prompting” — establishing a collaborative, thoughtful tone rather than rigid instructions. This approach is justified by the system’s primary value proposition: multi-turn conversational interaction where the AI guides research design, not just executes predefined queries. The prompt emphasizes taking time to understand user intent: “Take a deep breath and think deeply about what she actually wants. If it is not clear, do not be shy to ask her for a context.”

The prompt gives detailed examples of Alpha API calls with various query types which also serves for CQL syntax in-context learning, followed by instruction critically evaluate the Alpha’s output: “The CQL translator is able to translate simple natural language queries, but it is a simple rule based machine, you are surely more intelligent than that so be critical about the results and decide who is right.” The prompt also gives a guidance on lemma vs. word form disambiguation since it is the main source of confusion for inexperienced corpus users.

The prompt then proceeds to specify the corpus selection rules, e.g., the language and the mode (written vs. spoken), error handling and troubleshooting procedures, and sanity checking of results. The standard corpus linguistic methodology was found to be part of the original GPT training data so it is not mentioned by the prompt.

6 Corpus API Endpoints

The Corpus Manager API provides eight primary endpoints:

- `/translate`: Converts natural language to CQL via Alpha
- `/concordance`: Retrieves example concordances (KWIC format)
- `/term-frequency`: Returns absolute frequency, relative frequency (instances per million, IPM), and average reduced frequency (ARF) statistics
- `/freqs`: Provides frequency distributions by attribute (lemma, word form, parts of speech, morphological tag, etc.)
- `/collocations`: Calculates collocation scores (default: logDice dice metric, but there are several other metrics available)
- `/text-types`: Shows metadata distribution (author, genre, year, etc.)
- `/corplist`: Lists available corpora
- `/info`: Provides detailed corpus information

All endpoints support subcorpus specification and return results in JSON or Markdown format.

7 Evaluation Challenges

Traditional evaluation benchmarks are hard to apply to our system, as the system’s primary value lies in multi-turn conversational interaction, which is notoriously difficult to evaluate systematically. Moreover, it was (and still is) the first of its kind and thus establishes its own baseline.

Our primary evaluation came from real-world deployment, namely from the user ratings (the custom GPT received a 4-star average rating) and the usage scale (thousands of conversations over more than a year). The system was successfully used to prepare a research article, that succeed in peer review and is to be published this year (Milička, 2026).

8 Usage Statistics and Observed Patterns

The system remained an internal tool until September 2024, when it was publicly announced on the CNC website, Twitter/X, Czech Radio (September 18, 15 minutes on-air), and Czech Television (September 28, 2 minute morning prime-time segment). However, usage spikes did not correlate with media coverage. The largest spike occurred in November 2024 (3,258 calls) without any media

promotion, suggesting traditional media may be less influential than organic discovery and word-of-mouth for specialized academic tools.

Over 20 months of deployment (March 2024–November 2025), the system processed 10,360 API calls across 382 active days, averaging 27.1 calls per active day. Figure 2 shows monthly usage patterns.

Interest waned in summer 2025, when OpenAI began iterating rapidly on new models, which led to unexpected behavior and sometimes rendered the entire system unusable. Because the authors of this study were on summer vacation, the system was not repaired for several months. At the time of writing, the system’s usability remains unpredictable, since the GPT-5.x router selects among several internal models of varying quality based on the complexity and requirements of each user query. If a user prompt triggers non-thinking mode, the system is barely usable; in thinking mode, it works well.

8.1 API Action Distribution

The most frequent operations were:

- Frequency distributions (`/freqs`): 3,527 calls (34.0%)
- Term frequency lookups (`/term-frequency`): 2,181 calls (21.1%)
- Concordance searches (`/concordance`): 1,590 calls (15.4%)
- CQL translation (`/translate`): 1,379 calls (13.3%)
- Collocation analysis (`/collocations`): 1,017 calls (9.8%)

This distribution reveals that users primarily performed statistical analyses (frequency and collocation studies) rather than simply browsing concordances, suggesting the system successfully supports sophisticated corpus research workflows.

The system is multilingual, but it was mainly advertised in Czech environment so the main user base is also Czech (7200 prompts to written Czech language corpora, 250 to spoken Czech corpora), but the users also prompted English corpora (English Intercorp and BNC, 550 prompts in total), and corpora in other European languages (de 167, pl 117, fr 68, es 42, sl 19 it 18), with long tail distribution of other languages including sk, ru, pt, sv, da, no, bg, hr, mk, sq, hu, zh etc.

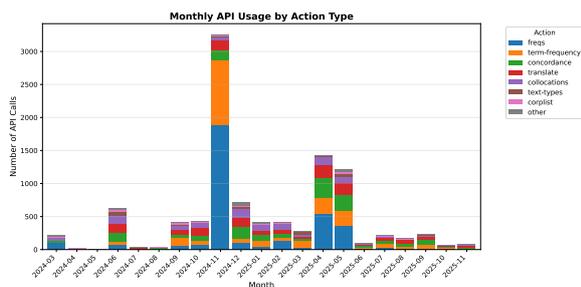


Figure 2: Monthly API usage by action type from March 2024 to November 2025. Peak usage occurred in November 2024 (3,258 calls) without media promotion, suggesting organic adoption.

8.2 Common Error Patterns

Analysis of system behavior revealed several recurring issues. Despite access to the Alpha translator, the system sometimes makes human-like mistakes in CQL syntax. When the system uses Alpha, it benefits from platform-specific knowledge (morphological tags, UD annotations) but may inherit Alpha’s limitations; when it generates CQL independently, it gains flexibility but loses specialized knowledge. Since the right CQL expression is key to successful information retrieval, this part definitely needs improvement.

The model also occasionally hallucinates API endpoints (e.g., attempting authentication of the user via non-existent `openapi?subscriber=corpus-linguist`)

The system also fails to promote its capabilities to the user. For instance, there is a possibility to make useful comparisons between subcorpora based on metadata, e.g., fiction vs. non-fiction via the `/translate` endpoint, but it is hard to trigger the system to offer this capability to the user, so this API endpoint was used only occasionally (330 times). The system has a Python interpreter at its disposal, so it can create charts and export them in any arbitrary format, but if the user does not know this, it would not even occur to them to request it. This article, which presents the system’s capabilities, can serve as a manual that could remedy this, but the very existence of a manual contradicts the main advantage of a conversational interface, namely user-friendliness.

9 Lessons Learned

9.1 Platform Stability is Non-Negotiable

The most critical lesson: platform stability is essential and cannot be offset by other advantages. In-

ability to control which model users access proved fundamentally problematic. During initial development, the system was predictable: paying users got GPT-4, free users got GPT-3.5. This enabled systematic development and testing.

However, OpenAI’s subsequent model updates, rollout patterns, and tier changes introduced unpredictability. Different users experienced different behaviors, making debugging impossible. The spectacular capabilities of the platform (voice interface, code interpreter, mobile access) cannot compensate for unreliable model availability.

9.2 Importance of Voice Interface

Despite platform challenges, the voice interface represents a genuine breakthrough: this is the first voice-enabled corpus query system in history. This accessibility improvement got more attention than any other feature of the system, since the voice interaction dramatically lowers the barrier for corpus research, enabling queries while commuting, during fieldwork, or just for fun over a glass of beer.

10 Conclusion

We presented a ChatGPT-based AI colleague for corpus linguistic research, deployed in production for over 20 months with 10,000+ API calls from authentic users. The system demonstrates that hybrid architectures combining rule-based translation with LLM intelligence can effectively bridge the gap between natural language and formal query languages.

Key contributions include: (1) the first voice-enabled corpus interface, dramatically improving accessibility; (2) validation that multi-turn conversational guidance adds value beyond simple corpus searches.

While the OpenAI custom GPT platform enabled rapid deployment with sophisticated capabilities (voice, mobile, code interpreter), its instability proved costly. Future systems should prioritize platform control, even at the expense of convenience features. The hybrid approach of leveraging both rule-based and neural components remains promising for specialized academic tools across disciplines.

We are planning migration to the Model Context Protocol (MCP), which will provide platform independence and complete model control on the user side and also incorporating the system into

independently working agent.

Since the newer SotA LLMs are getting better in manipulating symbols on a character scale, we plan to rely less on the rule-based Alpha translator.

Acknowledgments

Jiří Milička was supported by Czech Science Foundation Grant No. 24-11725S, gacr.cz (*Large language models through the prism of corpus linguistics*). This paper was supported by the *Czech National Corpus* project (LM2023044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within Large Research, Development and Innovation

We thank all users who contributed feedback during the development and deployment phases.

References

- Annelie Ädel. 2010. How to use corpus linguistics in the study of political discourse. In *The Routledge handbook of corpus linguistics*, pages 591–604. Routledge.
- Paul Baker. 2006. *Using corpora in discourse analysis*. Continuum.
- Tony Berber Sardinha. 2023. Corpus linguistics and historiography. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 7(1):69–90.
- Kaj Bostrom and Greg Durrett. 2020. [Byte pair encoding is suboptimal for language model pretraining](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.
- Vaclav Brezina. 2018. Statistical choices in corpus-based discourse analysis. In *Corpus approaches to discourse*, pages 259–280. Routledge.
- Winnie Cheng and 1 others. 2013. Corpus-based linguistic approaches to critical discourse analysis. *The encyclopedia of applied linguistics*, pages 1353–1360.
- Stefan Thomas Gries and Anatol Stefanowitsch. 2007. *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*, volume 172. Walter de Gruyter.
- Jiseung Hong, Grace Byun, Seungone Kim, Kai Shu, and Jinho D. Choi. 2025. [Measuring sycophancy of language models in multi-turn dialogues](#). *Preprint*, arXiv:2505.23840.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. 2024. [Taking ai welfare seriously](#). *Preprint*, arXiv:2411.00986.
- Anthony McEnery and Helen Baker. 2016. *Corpus linguistics and 17th-century prostitution: Computational linguistics and history*. Bloomsbury Academic.
- Tony McEnery and Gavin Brookes. 2024. Corpus linguistics and the social sciences. *Corpus linguistics and linguistic theory*, 20(3):591–613.
- Tony McEnery and Andrew Hardie. 2011. *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Jiří Milička. 2026. [Subjektivita umělé inteligence a robotů v českém jazykovém diskurzu: korpusová analýza let 2010–2023](#). In Michal Škrabal, Barbora Štěpánková, and Hana Skoumalová, editors, *Korpus třicetiletý*. Nakladatelství Lidové noviny. Forthcoming.
- Jiří Milička and Denisa Šebestová. 2024. Query a corpus in near-natural language: A human-friendly corpus query language not only for linguists. In *Crossing Boundaries through Corpora: Innovative corpus approaches within and beyond linguistics*, pages 248–262. John Benjamins Publishing Company.
- Jiří Milička. 2024. [Theoretical and methodological framework for studying texts produced by large language models](#). *Preprint*, arXiv:2408.16740.
- Nitarshan Rajkumar, Raymond Li, and Dzmity Bahdanau. 2022. [Evaluating the text-to-sql capabilities of large language models](#). *Preprint*, arXiv:2204.00498.
- Ute Römer. 2011. *Corpus research applications in second language teaching*, volume 31. Annual Review of Applied Linguistics.
- John Sinclair. 1991. *Corpus, concordance, collocation*. Oxford University Press.
- Hamada M Zahera, Manzoor Ali, Mohamed Ahmed Sherif, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. 2024. Generating sparql from natural language using chain-of-thoughts prompting. In *SEMANTICS*, pages 353–368.

Generative Information Extraction from Biographical Sources

Robin Winkle **Jörn Kreutel** **Manfred Stede**
University of Potsdam Berliner Hochschule für Technik University of Potsdam
robin_winkle@outlook.com jkreutel@bht-berlin.de stede@uni-potsdam.de

Abstract

Biographical sources, such as literature encyclopedias, encode knowledge about historical figures in textual form. In this paper, we address the task of consolidating structured biographical information about authors from the former German Democratic Republic into a unified database. To this end, we present a generalizable Information Extraction (IE) system based on LLM prompting. Specifically, we compare two midsized open-source models, Qwen-2.5-32B and Llama-3-70B-Instruct, investigate a range of Prompt Engineering (PE) strategies, and propose a semantic similarity-based evaluation metric for open-ended IE. Our experiments on an unpublished annotated subset of biographical texts deliver moderate precision and variable recall, highlighting both the potential and current limitations of generative IE in the Digital Humanities.

1 Introduction

Information Extraction (IE) plays an instrumental role in the Digital Humanities and related fields, where the systematic transformation of unstructured or semi-structured texts into machine-readable data is a prerequisite for large-scale cultural and historical analysis.

Our paper is situated in the context of the research project *Forschungsplattform Literarisches Feld DDR* (FLFDDR) which aims to create a bibliographical database that covers approximately 3,400 authors of literary works which have been identified as forming the literary field (Bourdieu, 1992) of the former German Democratic Republic (GDR). By compiling all available published and archival sources regarding those authors –explicitly highlighting conflicting information derived from different sources– the project seeks to enable qualitative and quantitative research on a complete literary field.¹ As of today, data for a subset of those

authors, who participated in a study program on literary writing and who sum up to about 10% of the complete corpus,² have been successfully collected, using a customized user interface for manual data entry. To extend the database, a semi-automated process will be required, which motivates the focus of this paper on the extraction of biographical information as an instance of the broader IE task.

Recent advances in Natural Language Processing (NLP) have enabled increasingly efficient IE methods, reducing the need for task- or domain-specific parsing rules (Plum et al., 2019). In particular, transformer-based Large Language Models (LLMs) have been used for IE in a generative, prompt-based setting, which we term *Generative IE* (GenIE). While these models are typically optimized for dialogue rather than structured data production (Liu et al., 2024), they have shown strong generalization capabilities across diverse NLP tasks, which has driven their adoption in IE.

In this work, we present a generalizable end-to-end GenIE framework, designed to obtain structured information from a set of natural language texts. Figure 1 provides a high-level representation of the extraction workflow. The input documents (on the left) are supplied to an LLM alongside customizable prompts that specify the format of the structured output (bottom right) and guide the extraction process³. The framework supports systematic assessment of Prompt Engineering (PE) strategies and retains source text fragments for all extractions to enable retroactive verifiability and human-in-the-loop validation.

pilation are available here: <https://ddr-literatur.de/daily-research/korpus-autorinnen-in-der-ddr/>

²See: *Forschungsplattform Literarisches Feld DDR: Autor*innen, Werke, Netzwerke. Pilotprojekt: Die Student*innen des Instituts für Literatur "Johannes R. Becher" Leipzig*, funded by Deutsche Forschungsgemeinschaft (DFG), project number 419244741, 2019-2023.

³The top right represents an alternative, sequential prompting approach explained in Section 5.

¹The list of all authors and comments on their com-

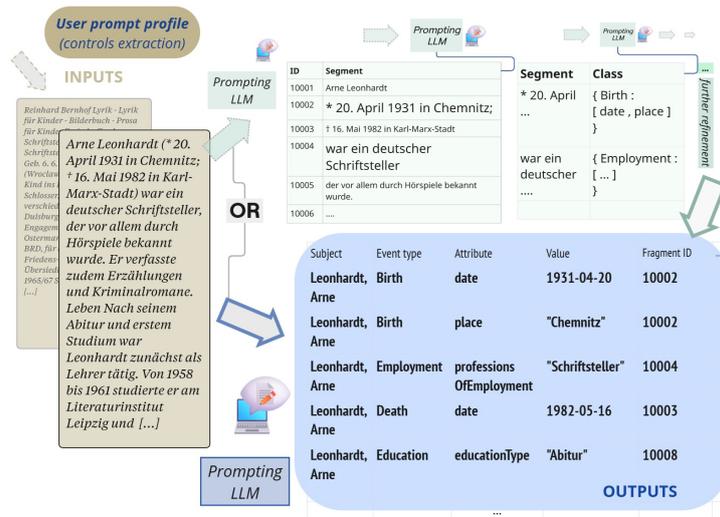


Figure 1: High-level representation of the extraction workflow, including the multi-step (top) vs. single-step approach (bottom).

We assess the framework on an annotated subset of the biographical material curated within the FLFDDR project. Using two moderately sized open-source LLMs, we analyze the impact of different prompting strategies and discuss overall trends and challenges. As an additional contribution, we propose and apply a semantic similarity-based evaluation approach tailored specifically to open IE settings, addressing the limitations of exact-match metrics and restricted GenIE. Our results highlight both the potential and the current limitations of GenIE systems, contributing empirical evidence to ongoing discussions on the responsible and effective use of LLMs in Digital Humanities.

2 Background and Related Work

IE is the task of obtaining structured information from unstructured or semi-structured texts written in natural language (McCallum, 2005). During recent decades, computational approaches to IE have enabled the creation of analytically accessible databases for diverse data types, including biographical data (Plum et al., 2019). Within the literary domain, encyclopedias and handbooks, such as those contained in our data set, are among the primary sources for biographical data, listing and contextualizing important life events of literary authors, including a person’s place of birth, membership in an institution or the publication of specific works. Structured representations of this data may draw on event-centric data models, which link multiple temporally and contextually connected biographical facts to a unique event within a per-

son’s life narrative (Tuominen et al., 2017).⁴ Extracting such structured representations from text poses challenges for traditional pipeline-based IE systems, which depend on explicitly modeling intermediate subtasks such as Text Segmentation or Named Entity Recognition (NER).

We address this challenge through a generalizable framework for large-scale IE using pre-trained Large Language Models (LLMs) and Prompt Engineering (PE). While related work has sometimes referred to this approach as "LLM-driven IE", we adopt the term "Generative Information Extraction (GenIE)", originally introduced by Josifoski et al. (2022), since it better reflects the use of generative sequence modeling for structured output. GenIE inherently integrates multiple IE subtasks into an end-to-end sequence generation task (Shen et al., 2023), bypassing the need to explicitly obtain intermediate extraction results. However, it also faces some well-documented challenges associated with generative transformer models in general, such as hallucinations (Huang et al., 2025), biases (Wei et al., 2025), and output variability (Liu et al., 2024).

Related work on GenIE systems has focused particularly on medicine (Goel et al., 2023; Khan et al., 2025; Hu et al., 2024). Gu et al. (2025) evaluate the "out-of-the-box" capabilities of chat-optimised

⁴The concrete event types and attributes employed in the FLFDDR project are motivated by the project’s rather wide scope aiming at providing a generic foundation for detailed analyses linking life and work of authors. See table 2 for a full list of event types and attributes from the FLFDDR event data model which have been taken into consideration for the pilot study presented here. For more information on the project’s data model, see (Kreutel et al., 2023).

Rodrian, Fred, Chefredakteur, Dir. d. Kinderbuchverl. Berlin; Pr. d. Minist. f. Kultur d. DDR 58 u. 59, Alex-Wedding-Med. 72 u. 76, Johannes-R.-Becher-Med. in Gold 75, Wilhelm-Bracke-Med. 78, Nationalpr. d. DDR (Koll.) 79, Vaterland. Verd.orden in Silber 84, u.a.; * Berlin 14.7.1926, † Berlin 25.5.1985; Kinderbuch u. -film. V: Das Wolkenschaf 58; Das Enteniesel 60; Der Märchenschimmel 60; [...] Treffpunkt Erfurt 62; Hirsch Heinrich 65. P: Das Wolkenschaf 64. Lit: DLL, Bd XIII 91; Biogr. Hdb. SBZ/DDR 96, 97.

Figure 2: Source text about author Fred Rodrian from *Kürschners Deutscher Literatur-Kalender Nekrolog 1971-1998* lexicon.

Christine Ernst (* 7. Februar 1938 in Reichenberg, Tschechoslowakei) ist eine deutsche Politikerin (SPD) und ehemaliges Mitglied des Sächsischen Landtages. Christine Ernst besuchte die Grund- und Oberschule in Magdeburg und anschließend die Fachschule für Bibliothekare in Leipzig. Danach war sie als Bibliothekarin und Leiterin eines Jugend-Literatur-Clubs in Magdeburg. Von 1976 bis 1979 folgte ein Studium am Literaturinstitut „Johannes R. Becher“ in Leipzig. Zwischen 1979 und 1980 war sie [...]

Figure 3: Source text about author Christine Ernst from *Wikipedia*

models by prompting them to extract medical determinants from text-based health records. Their hypothesis –that general-purpose LLMs can be adopted for IE without task-specific fine-tuning– aligns with the methodology presented in this work. The results of their study suggest that even moderately-sized open-source models substantially outperform naive pattern-matching and can deliver strong performance in IE-related tasks in various domains, particularly when effective PE is applied.

Recent studies have also reported encouraging results with GenIE systems in other domains, including law (Li and Yi, 2024), finance (Li et al., 2025; Kong et al., 2024), and scientific text (Dagdelen et al., 2024). Polak and Morgan (2023) achieve up to 90.8% precision and 87.7% recall in a study on the use of *ChatGPT* for data extraction in a materials science context. However, their formulation evaluates the IE task under a highly constrained matching scheme: extracted outputs are compared against ground truth labels using strict domain-specific equivalence criteria. In contrast, our work targets fully open-ended generative extraction without predefined value sets, and is designed to operate with open-source models and without the need for strict equivalence definitions.

The implementation of our framework is in part informed by these studies. However, to the best of our knowledge, there have not been any systematic attempts to extract structured data from biographical text sources using GenIE systems.

3 Data

This study draws on biographical texts compiled in the interdisciplinary FLFDDRproject. The complete corpus contains documents from about 600 different German language sources, including journals, newspapers, online materials, interviews, and around 2,300 articles from 150 literature encyclopedias and handbooks (Kreutel et al., 2023).

Some instances in the source data were digitized using OCR, which may have introduced a slight degree of character-level noise. Other forms of noise may arise from a small amount of inconsistencies between source texts and evaluation data, for example, in the case of author Christine Ernst who was known as Christine Lindner before changing her name. As illustrated by Figures 2-3, the input documents feature a considerable amount of domain-specific abbreviations (e.g. "Lit.büros" for *Literaturbüros*⁵) as well as metadata strings without direct informational value (e.g. "GND: 139456139"). Excluding extreme outliers, the texts contain approximately 1,170 characters on average, with substantial variation between sources.

In our experiments, we use a subset of 188 articles from 72 sources manually annotated by five trained student research assistants under the supervision of literary scholars. Table 5 in the appendix shows the most common text sources in the source corpus and the evaluation subset. The subset primarily features literature encyclopedia⁶ articles, a curated and quality-controlled text type containing highly condensed biographical information. As a result, the annotated documents are mostly compact semi-structured entries using a telegraphic style, such as in the example shown in Figure 2. This style is characterized by minimal phrasing with infrequent use of function words and connectors. More narrative texts, such as the one represented by Figure 3, are present but less frequent in the evaluation set.

We process the annotations to fit the format described in Section 4.1. After cleaning, deduplication, and consistency checks, the final evaluation set contains documents labelled with 6,787 information units. The most prevalent event types extracted by annotators capture information relating to authors' education and professional trajectories, besides singular events like birth and death. Attributes such as *begin*, *end*, and *place* that occur

⁵Translation: "Literary Association"

⁶German: *Literaturlexika*

across multiple event types are among the most frequent. Since the annotations emerged out of the project’s data collection workflow, they exhibit occasional gaps in coverage and mismatches between surface forms and normalized attribute values, a point we return to in Section 6.2.

4 Methods

This section describes the methodological design of our experiments, including the selected output representation, language models, prompting strategies, and a customized evaluation setup.

4.1 Output Format

Data models based on the Resource Description Framework (RDF) for biographical data like the one proposed by Tuominen et al. (2017) commonly use knowledge triples that rely on abstract references such as Uniform Resource Identifiers (URI) and entity placeholders to form a graph-like representation. This can present challenges from an LLM generation perspective, since these models are trained for fluent text generation rather than for consistent handling of abstract identifiers. For example, an author may be represented by a URI such as `ex:author_123`, which must be reused consistently across events, a requirement that generative models may fail to satisfy reliably.

Field f	Field value x_f
SUBJECT	Rodrian, Fred
EVENT TYPE	birth
ATTRIBUTE	place
VALUE	Berlin
SOURCE FRAGMENT	geb. am 14.7.1926 im Berliner Osten

Table 1: Structured information unit associated with text fragment relating to the birth of Fred Rodrian.

Therefore, we adopt a flattened JSON output format with natural language values that is better aligned with the generative strengths of LLMs. A tabular representation of this format is given in Table 1. This format allows for an arbitrary number of tuples, each consisting of a field f and the appropriate field value x_f . In our experiments, each extracted information unit is configured to contain the following fields: (1) a SUBJECT (the person concerned), (2) an EVENT TYPE (e.g., *birth*, *education*), (3) an ATTRIBUTE specifying the event further (e.g., *place*, *institution*), (4) the extracted

VALUE encoded as a string, number, or date, and finally, (5) a SOURCE FRAGMENT that grounds the information unit in the original text. Here, the term ATTRIBUTE denotes a property of an EVENT TYPE in the underlying event data model, e.g., the "place" attribute of a "birth" event. While the VALUE is unconstrained, EVENT TYPES and ATTRIBUTES are selected from predefined lists (see Table 2) in our experiments to constrain the extraction scope and improve output consistency.

4.2 LLM Selection

To avoid reproducibility issues and access limitations of commercial LLMs, only open-source models were considered. The *Hugging Face* Open-Source Language Model Leaderboard⁷, a well-established source with extensive model coverage and diverse benchmark evaluations, serves as the starting point for a multi-stage selection process balancing performance, openness, computational cost, and practical feasibility. Models were ranked according to three evaluation suites relevant to structured information extraction: IFEval (format adherence), MuSR (language understanding), and MMLU-Pro (domain knowledge and reasoning). Models are further evaluated by affordability and environmental footprint. To complement the leaderboard-based shortlist, we examined reports from related IE work to highlight models proven effective in structured extraction.

From this process, two models offering a strong balance of quality and efficiency were selected: *Qwen-2.5-32B*⁸ (Yang et al., 2024) and *Llama-3-70B-Instruct*⁹ (Dubey et al., 2024). A brief heuristic check on a small sample of biographical texts confirmed that both models produce stable, well-formatted outputs suitable for the GenIE task at hand.

4.3 Prompt Engineering

Given its central role in guiding the model’s reasoning and extraction decisions, prompt design is handled through a modular setup that automatically assembles prompts by integrating different functional components such as concise instructions, brief descriptions of the input context, examples illustrating the output format, or extraction scope

⁷https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, archived, last accessed in June 2025

⁸<https://huggingface.co/Qwen/Qwen2.5-32B>

⁹<https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct>

Event Types	Attributes
Acquaintance	acquaintanceType begin durationOfEvent end participants
Birth	place begin
Citizenship	begin end institution
ConfessionAffiliation	begin confession institution
Death	begin deathType participants place
Education	begin educationType educationalContentsOfEducation end institution place
Employment	begin end institution place professionsOfEmployment
Flat	begin durationOfEvent end migratedFromPlace place
ForcedStay	begin durationOfEvent end forcedStayType institution place politically-Motivated
Funding	begin durationOfEvent end fundingInstitution place
Granting	awardTypeOfGranting begin
Journey	begin destinations durationOfEvent end participants place
Membership	begin durationOfEvent end institution participants place
MilitaryService	begin durationOfEvent end institution participatedWarOfService place professionsOfEmployment
NamedEvent	begin durationOfEvent end
NoEmployment	begin durationOfEvent end
Origin	socialClasses
Orphanacy	begin
Parenthood	begin childGender childName
Surveillance	begin durationOfEvent end place surveillingInstitution

Table 2: Permissible values for extraction fields "Event type" and "Attribute".

constraints. A condensed version of the specific prompt used in our experiments is provided in Table 6. By switching components in or out, the system can be adapted to test various extraction strategies while keeping prompts stable and manageable.

The prompting module also incorporates a set of advanced Prompt Engineering (PE) strategies designed to influence how the model reasons through the extraction task. While general PE aims to optimize prompts to be clear, task-specific and concise, advanced PE targets the model’s internal reasoning behavior or attempts to refine the answer space by providing specially crafted output examples.

In this study, three such strategies are implemented. (i) Few-shot prompting (Agarwal et al., 2024) is used to guide the model by analogy, providing concise examples that demonstrate the expected output format and reduce ambiguity in tasks requiring consistent formatting. (ii) Chain-of-Thought (CoT) prompting (Wei et al., 2022) is included as an optional component to encourage explicit intermediate reasoning. And finally, (iii) self-refinement (Madaan et al., 2023) is integrated as a lightweight revision step, allowing the model to critique and correct its own output to improve completeness and adherence to format requirements.

These techniques represent widely adopted model-agnostic methods that can be integrated without external retrieval systems or task-specific fine-tuning. We systematically assess the contribu-

tion of these methods to extraction quality through empirical experiments, cf. Section 5.

4.4 Evaluation Metric

For open IE tasks, traditional accuracy is ill-defined since the space of true negatives (TN) is difficult to define or possibly unbounded: if a text contains only sparse information about a subject, then an IE system would have to generate a large number of TN (e.g. date of birth: "Unknown", date of death: "Unknown", and so forth) to receive a high evaluation score. Furthermore, strict surface-form matching can be brittle since valid extraction values are not drawn from a closed set in open IE (e.g. if a text cites a person’s main occupation as "author", "writer" should also count as a valid extraction value). This contrasts with closed IE, where the task is framed as a multi-choice classification problem (Gu et al., 2025, p. 4) or the TN space is well-bounded at fragment level (Polak and Morgan, 2023, p. 9).

To robustly compare extractions against human annotations, we adopt a type-aware semantic similarity function $\sigma : X \times X \rightarrow [0, 1]$ that operates at the level of an extraction field f and determines the similarity between this field’s value x_f (cf. Section 4.1) and its counterpart in the annotation set. If x_f is a string literal, cosine similarity is computed using suitable embedding-based representations. This is complemented by specialized similarity scoring functions for cases where x_f is a date or numerical

value.

To estimate the overall alignment score $\phi(e, g)$ between a given candidate unit $e \in \mathcal{E}$ and a gold unit $g \in \mathcal{G}$, let F denote the set of extraction fields required to meet the minimum similarity threshold θ_f for $f \in F$, and let w_f denote the assigned weight of f . Let $x_{f,e}$ and $x_{f,g}$ be the values of f in e and g , respectively. The alignment score is then given by Equation 1 if for all $f \in F$: $\sigma_f(x_{f,e}, x_{f,g}) \geq \theta_f$, otherwise it is equal to zero.

$$\phi(e, g) = \frac{\sum_{f \in F} w_f \cdot \sigma_f(x_{f,e}, x_{f,g})}{\sum_{f \in F} w_f} \quad (1)$$

We use ϕ to obtain the best precision-oriented alignment $\mu_P(e) = \max_{g \in \mathcal{G}} \phi(e, g)$ and compute soft precision for a set of extracted information units \mathcal{E} as:

$$\text{Precision}^* = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mu_P(e) \quad (2)$$

We then reverse the comparison to obtain the best recall-oriented alignment $\mu_R(g) = \max_{e \in \mathcal{E}} \phi(e, g)$ and then calculate soft recall for a set of gold information units \mathcal{G} as:

$$\text{Recall}^* = \min \left(\frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mu_R(g), \frac{|\mathcal{E}|}{|\mathcal{G}|} \right) \quad (3)$$

The upper bound ensures that recall never exceeds the maximum achievable standard recall score, which is equal to the number of candidate elements divided by the number of gold elements, assuming perfect precision.

To summarize, the metric compares the set of extracted units \mathcal{E} against the set of human-annotated units \mathcal{G} by finding, for each element in one set, the closest semantic match in the other set. High scores indicate strong alignment, while low scores signal spurious or unrecalled units. These scores are then aggregated across the entire set to compute overall precision, recall, and F1. Unlike prior work which relies on closed label spaces and employs strict matching, this metric is robust to surface form variability and allows GenIE systems to generate any extraction value that is semantically equivalent. Our evaluation approach is designed to be generalizable and can be adapted to different data sets and domains, since for each field f contained in an extraction unit for a given domain, the similarity function σ_f , the minimum similarity threshold θ_f and the weight w_f are configurable.

Configuration	Model	Heuristically optimized Prompt	Subtask Handling	Reasoning	Number of examples
BASE	Qwen-2.5 (32B)	Yes	simultaneous	default	2
EXMP+					13
LARGER-M	Llama-3 (70B)				
LARGER-M-EXMP+	Llama-3 (70B)				13
MULTI			sequential		
CoT				CoT	13

Table 3: Experimental configurations with deviations from the default setup highlighted. Blank entries inherit the default.

5 Experiments

We evaluate the proposed framework by comparing several experimental variants to a BASE configuration. These variants, summarized in Table 3, systematically isolate the effect of individual factors, such as model size, number of few-shot examples, multi-step extraction and reasoning strategies.

The BASE configuration employs the standard *Qwen-2.5* model with a single-step prompting setup, default reasoning, and two in-context examples. LARGER-M replaces the baseline model with the larger LLM (cf. Section 4.2) while keeping all other settings fixed. LARGER-M EXMP+ extends this setting by increasing the number of few-shot examples. Analogously, EXMP+ augments BASE by including more output examples without changing model size.

Additionally, MULTI implements a multi-step IE pipeline (see Figure 1, top row), in which the model is sequentially prompted to (i) segment input texts, (ii) extract structured information from segments, and then (iii) perform a single self-refinement pass aimed at improving completeness and strict format adherence. The CoT variant activates explicit CoT instructions, encouraging step-by-step reasoning in the model response.

The performance of the system in the different configurations is assessed using the semantic similarity-based open extraction metric described in Section 4.4, computed against the manually annotated reference data mentioned in Section 3. For our experiments, we set $\theta = 1$ for all extraction fields other than VALUE, effectively enforcing exact matching for these fields since their extraction is constrained by predefined lists (cf. Section 4.1). For string comparison we use an embedding model fine-tuned for cosine similarity computation in German and English¹⁰.

¹⁰<https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer>.

6 Results

We report the empirical results of our experiments and analyse the behavior of the proposed GenIE framework for the configurations described in Section 5. First, we present a quantitative assessment using the evaluation metric described in Section 4.4 complemented by several supplementary measures, then we conduct a qualitative error analysis that contextualizes observed performance patterns.

6.1 Performance

Table 4 summarizes performance at the document level across all configurations. Overall precision was relatively stable (0.61–0.71), while recall varied substantially (0.33–0.75), resulting in F1 scores between 0.45 and 0.68. The strongest configuration, LARGER-M EXMP+, achieved the highest F1 (0.68) driven by a large recall gain (0.75), at the cost of reduced precision (0.62).

Extraction density, i.e. the number of units generated per 100 input characters, shows a clear positive correlation with recall and F1, implying that configurations producing more extraction units consistently achieve higher coverage, despite introducing additional false positives (FP). LARGER-M EXMP+ exhibits the highest extraction density (1.80 units per 100 characters) paired with one of the lowest false negative (FN) rates (6.53%), supporting the intuition that under-generation is more detrimental than moderate over-generation.

Our experiments show that including more examples in the prompt (i.e., few-shot prompting) produced the most substantial performance gains, particularly for the larger model (+0.33 recall from LARGER-M to LARGER-M EXMP+). In contrast, CoT prompting yields only marginal F1 improvements, primarily by trading precision for recall, and multi-step execution with self-refinement degrades performance due to severely reduced extraction density. The direct model comparison reveals that simply using the larger model without prompt optimisation did not improve performance.

Beyond core metrics, the distribution of extraction fields in the system’s output largely aligned with the evaluation set, displaying an overall overlap of 60-64%. In both sets, *Education*, *Birth*, and *Employment* events were among the most prevalent event types. However, *Granting* events (i.e., when an author was awarded a prize) were extracted less frequently by the GenIE framework, while *Forced-Stay* and *MilitaryService* type events gained rela-

tive importance. We note that the system output exhibits a systematic bias towards the extraction of date values (e.g., 13-01-1970). The percentage of extracted information units containing these values was consistently higher, ranging from 28.3% to 40.2% across configurations, compared to only 9.9% in the gold annotations.

Overall, the results highlight extraction density and example-based prompting as the main drivers of performance, while CoT prompting, multi-step task decomposition, and model scaling in isolation provided limited benefit.

6.2 Error Analysis and Discussion

Manual inspection of system outputs¹¹ reveals that human annotations were frequently incomplete. As a result, many factually correct system extractions absent from the gold data, were classified as FPs. Precision scores therefore need to be interpreted with caution, as they partly reflect annotation coverage rather than true extraction quality. For example, in one document, the system correctly extracted information on the social background of author Adolf Görtz (*son of a factory worker*), but no corresponding annotation exists in the evaluation set. Such cases are penalized as FPs despite being grounded in the text, highlighting an inherent limitation of annotation-based evaluation for open-ended IE.

Beside these "false FPs", a common pattern concerned surface-level variation between extracted values and gold annotations. Here, the evaluation metric proved rather robust since it was specifically designed to account for this type of variation (cf. Section 4.4). For example, in the case of one article that mentioned an author’s employment in "Children’s and young adult literature", the GenIE system extracted this value literally while human annotators simply noted "Children’s books" in the equivalent field. In these cases, the embedding-based metric assigns partial credit based on semantic similarity, yielding high but sub-perfect scores. While this systematically pulls evaluation scores away from full precision, it reflects a fairer assessment than exact-match evaluation, which would fail to credit semantic similarity entirely.

Less frequently, errors arose from metric failures. For instance, the historically equivalent place names (e.g., *Chemnitz* vs. *Karl-Marx-Stadt*) were treated as dissimilar in one case. This type of error

¹¹An in-depth example analysis can be viewed on github.com

Metric/Experiment	MULTI	COT	EXMP+	LARGER-M	LARGER-M-EXMP+	BASE
F1	0.45	0.61	0.63	0.52	0.68	0.55
Precision	0.71	0.61	0.68	0.68	0.62	0.71
Recall	0.33	0.62	0.59	0.43	0.75	0.45
% False positives	12.43	20.03	12.89	15.70	17.59	13.74
% False negatives	48.78	5.93	7.74	31.71	6.53	10.07
Extraction density	0.49	0.99	0.88	0.69	1.80	0.93
% Top fields overlap	62.2	60.0	64.4	55.6	62.2	62.2
% Date values	37.5	32.1	28.3	40.2	37.4	35.8
Fragment length	34.92	41.44	46.51	57.58	28.72	48.74

Table 4: Evaluation results from different experiments, showing average scores across all input documents.

reduces scores despite factual correctness and highlights that evaluation performance is partly constrained by the robustness of the embedding model. Future work could mitigate such cases by incorporating external norm data or taxonomies for entities such as places and institutions.

The manual analysis further reveals that truly invalid extractions, e.g. units with uninterpretable strings, were rare and were assigned zero scores by the evaluation pipeline. Similarly infrequent were extractions with missing values, which typically reflected underspecified information in the source text. Finally, a small number of FNs received non-zero scores due to alignment with semantically related but incorrect gold fields. Although sub-optimal, analysis suggests that these cases had a limited practical and numerical impact on overall performance.

7 Conclusion

This study presents, to the best of our knowledge, the first systematic investigation of large-scale IE from biographical texts driven by midsized general-purpose LLMs. The results indicate that the proposed GenIE framework has great potential for effective application in this domain, enabling the extraction of structured biographical information with minimal preparation effort required.

Unlike many similar studies, we adopt an open IE setting that allows for unrestricted output generation. To tailor to this task formulation, we propose a semantic similarity-based evaluation metric. Experiments were conducted on a curated, unpublished dataset of German biographical encyclopedia entries, with domain expert annotations serving as gold references.

We observe that the precision of the IE frame-

work remained relatively stable (approximately 0.61–0.71), while recall varied substantially (approximately 0.33–0.75). While limited precision is partly attributable to incomplete annotation coverage and is therefore less concerning, the absence of high recall remains a notable limitation. At the same time, recall and F1 show a strong positive correlation with extraction density, suggesting that under-generation is more detrimental than moderate over-generation in this setting.

Manual analysis further confirms that many apparent FPs reflect missing gold annotations rather than factual errors, and that the evaluation metric largely behaves as intended. The substantial overlap in extraction field distributions between system output and annotations provides an encouraging signal of structural alignment, despite a systematic bias toward date value extractions.

Controlled ablation experiments reveal that PE strategies offered limited performance gains. Clear prompt formulation and few-shot prompting yield the most consistent improvements, whereas advanced reasoning strategies such as CoT provided little benefit.

Overall, while the output quality of our framework is encouraging, further investigation will be required to fully assess the capabilities of such systems. With appropriate refinement and optimization, the proposed framework may serve as a template for practical deployment. Future research may extend it to new domains, improve open GenIE evaluation methods, and explore a wider range of models.

Limitations and Ethical Considerations

The dataset used in this study is focused on a specific historical period, namely authors from

the GDR and the evaluation subset primarily features one text type, i.e. encyclopedia entries. Results may not generalize to other domains or (low-resource) languages. Due to resource constraints, experiments used only two open-source LLMs and modest inference parameters. Proprietary models or extensive parameter tuning might yield better performance. This study does not distinguish or quantify different types of hallucinations, focusing instead on the overall occurrence of extraction errors. Evaluation methods for open-ended IE remain imperfect and future research should explore more flexible metrics. Future studies using our proposed metric must ensure the robustness of the core embedding model. Furthermore, the evaluation metric is not yet sensitive to multiple occurrences of events of the same type, as it matches extraction units against all events of the detected type in the annotations for a given source text.

Ethically, LLM use entails significant environmental costs, potential bias, and limited interpretability. Biographical data on living individuals should be extracted only with caution, as it can enable the generation of detailed personal profiles (Ranjan et al., 2022). While such information may serve legitimate purposes, e.g. in law enforcement, it may also carry risks of misuse for malicious purposes.

Acknowledgements

We are deeply grateful to the FLFDDR project team for creating the corpus that made this study possible. We would also like to thank the reviewers for their valuable and detailed feedback.

References

- Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie C. Y. Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal M. P. Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. [Many-shot in-context learning](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Pierre Bourdieu. 1992. *Les règles de l'art: genèse et structure du champ littéraire*. Editions du Seuil, Paris.
- John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder,

Kristin A. Persson, and Anubhav Jain. 2024. [Structured information extraction from scientific text with large language models](#). *Nature Communications*, 15(1):1418.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. 2023. [Llms accelerate annotation for medical information extraction](#). In *Machine Learning for Health, ML4H@NeurIPS 2023, 10 December 2023, New Orleans, Louisiana, USA*, volume 225 of *Proceedings of Machine Learning Research*, pages 82–100. PMLR.

Bowen Gu, Vivian Shao, Ziqian Liao, Valentina Carducci, Santiago Romero Brufau, Jie Yang, and Rishi J. Desai. 2025. [Scalable information extraction from free text electronic health records using large language models](#). *BMC Medical Research Methodology*, 25(1):23.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. [Improving large language models for clinical named entity recognition via prompt engineering](#). *J. Am. Medical Informatics Assoc.*, 31(9):1812–1820.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ACM Trans. Inf. Syst.*, 43(2).

Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni, and Robert West. 2022. [Genie: Generative information extraction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4626–4643. Association for Computational Linguistics.

Muhammad Ali Khan, Umair Ayub, Syed Arsalan Ahmed Naqvi, Kaneez Zahra Rubab Khakwani, Zaryab bin Riaz Sipra, Ammad Raina, Sihan Zhou, Huan He, Amir Saeidi, Bashar Hasan, Robert Bryan Rumble, Danielle S. Bitterman, Jeremy L. Warner, Jia Zou, Amye J. Tevaarwerk, Konstantinos Leventakos, Kenneth L. Kehl, Jeanne M. Palmer, Mohammad Hassan Murad, and 2 others. 2025. [Collaborative large language models for automated data extraction in living systematic reviews](#). *J. Am. Medical Informatics Assoc.*, 32(4):638–647.

- Yaxuan Kong, Yuqi Nie, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. [Large language models for financial and investment management: Applications and benchmarks](#). *Journal of Portfolio Management*, 51(2):162 – 210. Cited by: 2.
- Jörn Kreutel, Thomas Möbius, Birgit Dahlke, and Stefan Martus. 2023. [Forschungsplattform Literarisches Feld DDR – Ein Werkstattbericht zur prosopographischen Erfassung von Schriftsteller:innen in der DDR](#). In Helmuth Albrecht, Michael Farrenkopf, Helmut Maier, and Torsten Meyer, editors, *Historische Biographik und kritische Prosopographie als Instrumente in den Geschichtswissenschaften*, pages 141 – 165. De Gruyter.
- Huaxia Li, Haoyun Gao, Chengzhang Wu, and Miklos A. Vasarhelyi. 2025. [Extracting financial data from unstructured sources: Leveraging large language models](#). *Journal of Information Systems*, 39(1):135 – 156. Cited by: 5.
- Shiye Li and Li Yi. 2024. [A few-shot entity relation extraction method in the legal domain based on large language models](#). In *Proceedings of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*, DEAI '24, page 580–586, New York, NY, USA. Association for Computing Machinery.
- Yu Liu, Duantengchuan Li, Kaili Wang, Zhuoran Xiong, Fobo Shi, Jian Wang, Bing Li, and Bo Hang. 2024. [Are llms good at structured outputs? A benchmark for evaluating structured output capabilities in llms](#). *Inf. Process. Manag.*, 61(5):103809.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self- feedback](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Andrew McCallum. 2005. [Information extraction: distilling structured data from unstructured text](#). *ACM Queue*, 3(9):48–57.
- Alistair Plum, Marcos Zampieri, Constantin Orasan, Eveline Wandl-Vogt, and Ruslan Mitkov. 2019. [Large-scale data harvesting for biographical data](#). In *Proceedings of the Third Conference on Biographical Data in a Digital World 2019, Varna, Bulgaria, September 5-6, 2019*, volume 3152 of *CEUR Workshop Proceedings*, pages 66–72. CEUR-WS.org.
- Maciej P. Polak and Dane Morgan. 2023. [Extracting accurate materials data from research papers with conversational language models and prompt engineering - example of chatgpt](#). *CoRR*, abs/2303.05352.
- Rishabh Ranjan, H. Vathsala, and Shashidhar G. Koolagudi. 2022. [Profile generation from web sources: an information extraction system](#). *Soc. Netw. Anal. Min.*, 12(1):2.
- Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. [Promptner: Prompt locating and typing for named entity recognition](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, *ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12492–12507. Association for Computational Linguistics.
- Jouni Tuominen, Eero Hyvönen, and Petri Leskinen. 2017. [Bio CRM: A data model for representing biographical data for prosopographical research](#). In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017, Linz, Austria, November 6-7, 2017*, volume 2119 of *CEUR Workshop Proceedings*, pages 59–66. CEUR-WS.org.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Xiahua Wei, Naveen Kumar, and Han Zhang. 2025. [Addressing bias in generative AI: challenges and research opportunities in information management](#). *Inf. Manag.*, 62(2):104103.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. [Qwen2 technical report](#). *CoRR*, abs/2407.10671.

A Appendix

Entire dataset		Evaluation data	
Source	#Docs	Source	#Docs
Deutsche Biographie	141	Deutsche Biographie	34
Deutsches Literatur-Lexikon	116	Deutsches Literatur-Lexikon	21
Wikipedia	67	Schriftsteller der DDR (Boettcher)	13
Schriftsteller der DDR (Boettcher)	45	KuerschnerNekrologe1971-1998	6
Kuerschner 2012	26	Chemnitzer Autoren	6

Table 5: Top sources in the entire data set vs. evaluation subset.

Purpose	Prompt part
define system role	You are an information extraction assistant. Output must be valid JSON string with the keys 'SUBJECT', 'EVENT_TYPE', 'ATTRIBUTE', 'VALUE' and the correct values corresponding to these keys.
instructions	From the input, extract structured information units that each represent a single biographical event and have these fields: 'SUBJECT', 'EVENT_TYPE', 'ATTRIBUTE', 'VALUE'"
avoid inference	Only extract information present in the INPUT. Do not add new information or any information from your training data.
restrict scope (fields)	Consider only the following EVENT_TYPES and ATTRIBUTES, skip all others: [...]
restrict scope (subject)	Only structured information on the following target person is relevant: [...]
source fidelity	Keep the original wording of the short text fragment where the extracted information was found. Do not change the original wording
output format	In your response, do not add any explanation or comments, instead return only the structured data.
completeness	It is important that you extract as much information as possible!!
Few shot instructions	The following is an example for the correct OUTPUT you should generate for a given INPUT. INPUT:
example input	INPUT: Görtz, Adolf Erzählung (Kinder- und Jugendliteratur) 31.5. 1920, Köln. 1958-61 Studium am Institut für Lit. „Johannes R. Becher“. Ehrenurk. im Preisausschreiben anlässlich der IV. Weltfestspiele der Jugend und Studenten (1953); 3. Preis im Preisausschreiben des Ministeriums für Kultur (Kinder- und Jugendliteratur) (1953); Lobende Anerkennung im gleichen Preisausschreiben (1954), Leipzig C 1, Jacobstraße 14. Schriftsteller der Deutschen Demokratischen Republik. Bearbeiter des biographischen Teils: Joachim Ret, Egon Sartorius. Bearbeiter des bibliographischen Teils: Helmut Donner, Hans Heininger. Leipzig 1961.
example output	<pre> fragment_text:"Görtz, Adolf Erzählung (Kinder- und Jugendliteratur)" extracted_info:[SUBJECT:Adolf Görtz, EVENT_TYPE:Employment, ATTRIBUTE:professionsOfEmployment, VALUE:Kinderliteratur, Jugendliteratur] fragment_text:"31.5. 1920, Köln" extracted_info:[SUBJECT:Adolf Görtz, EVENT_TYPE:Birth, ATTRIBUTE:begin, VALUE:1920-05-31, SUBJECT:Adolf Görtz, EVENT_TYPE:Birth, ATTRIBUTE:place, VALUE:Köln] fragment_text:"1958–61 Studium am Institut für Lit. "Johannes R. Becher" " extracted_info:[SUBJECT:Adolf Görtz, EVENT_TYPE:Education, ATTRIBUTE:institution, VALUE:Literaturinstitut Johannes R. Becher, SUBJECT:Adolf Görtz, EVENT_TYPE:Education, ATTRIBUTE:begin, VALUE:1958, SUBJECT:Adolf Görtz, EVENT_TYPE:Education, ATTRIBUTE:end, VALUE:1961] fragment_text:"Ehrenurk. im Preisausschreiben anlässlich der IV. Weltfestspiele der Jugend und Studenten (1953)" extracted_info:[SUBJECT:Adolf Görtz, EVENT_TYPE: [...]</pre>

Table 6: Prompt used in the experiments (shortened and re-formatted for readability).

***WikiFirst*: A Genre-Fixed, Content-controlled Corpus for Evaluating Content Effects in Authorship Analysis**

Dung Tuan Nguyen, G. Çağatay Sat, Evgeny Pyshkin, and John Blake

School of Computer Science and Engineering

University of Aizu

Aizuwakamatsu

Japan

s1312004, s1312006, pyshe, jblake@u-aizu.ac.jp

Abstract

This paper presents the design and construction of *WikiFirst*, a corpus for investigating the impact of content variation on authorship similarity under a fixed genre. Prior work has investigated individual authorial style and impact of genre. However, the role of content has remained underexplored due to the lack of suitable data. We address this gap by constructing a Wikipedia-based corpus consisting exclusively of first revisions authored by non-anonymous editors, thereby ensuring high authorship certainty while maintaining a stable encyclopaedic genre.

1 Introduction

Authorship analysis aims to determine whether texts share a common author based on stylistic evidence, yet attribution performance is influenced by multiple interacting factors, notably individual style, genre, and content. While progress has been made in modelling authorial style (Sat et al., 2025) and mitigating genre effects, content variation remains a confounding factor that negatively impacts the accuracy of authorship attribution. A key limitation is the scarcity of corpora that allow systematic comparison of texts written by the same author on different topics while holding genre constant. In this work, we address this limitation by introducing *WikiFirst*, a corpus specifically designed to isolate content effects in authorship analysis.

The design of *WikiFirst* is motivated by the need for a reusable, methodologically transparent resource that enables researchers to explicitly evaluate and control for content variation in authorship analysis. By providing genre-fixed texts authored by the same individuals across multiple, well-defined content domains, *WikiFirst* supports systematic investigation of content effects.

This paper makes two primary contributions. First, we introduce *WikiFirst*, a new corpus for

authorship analysis that provides high authorship certainty and controlled content variation under a fixed genre. Second, we propose a principled content taxonomy to categorise Wikipedia articles into eight broad content domains. This corpus allows researchers to investigate the impact of subject variation within a single genre on authorship features.

2 Related Work

Authorship analysis is a research field that spans forensic linguistics (Coulthard and Johnson, 2000; Nini et al., 2023), literary studies (Eder and Górski, 2023), and computational text analysis (e.g., Blake et al., 2025), concerned with identifying stylistic regularities that distinguish authorial style. Alongside methodological advances, a range of benchmark corpora has been developed, including literary texts (Güngör, 2015), emails (Hussain, 2020), forum posts, and social media data. While these resources have enabled steady progress, they frequently conflate multiple sources of variation, making it difficult to isolate the influence of individual factors such as genre and topic.

Genre effects, in particular, have been shown to substantially alter feature distributions, exacerbating the difficulty of accurate authorship attribution (Kestemont et al., 2020, 2021). In contrast, content variation has received comparatively limited direct attention, despite long-standing recognition that topical vocabulary can impact stylistic markers. The absence of corpora explicitly designed to control content while holding genre constant has impeded investigation of this issue.

Heini and Kredens (2024) created the 100 Idiolect corpus comprising different genres of writing written by 112 authors, enabling comparison of authorial style across multiple genres. More recently, (Ma et al., 2025) created the much larger CROSSNEWS corpus, allowing machine learning models to compare differences between genres, and

to some extent across content areas. However, the content of 97.5% of the authors is limited to a single domain, making cross-content not feasible. Thus, this research aims to fill that niche by creating a corpus to investigate authorial feature stability and variation across different content domains.

2.1 Authorship Analysis and Confounds

Authorship attribution is influenced by interrelated confounding factors, including genre, topic, register, time period and medium (e.g., spoken vs. written). Among these, there is a growing body of work that shows the difficulty of cross-genre attribution (Kestemont et al., 2012), particularly for smaller datasets found in forensic contexts (Neal et al., 2017).

In addition, content effects are often treated implicitly (arguably by conflating topic with genre) rather than explicitly. Common strategies include removing content-bearing words or emphasizing function words under the assumption that they are less topic-sensitive. Such approaches may ameliorate content bias, but throw little insight into how content variation itself affects authorial stylistic similarity.

2.2 Wikipedia-Based Corpora and Revision Data

Wikipedia has been widely used in natural language processing research due to its scale, accessibility, and rich revision metadata in a variety of areas, including low-resource language training (e.g. Hungarian BERT (huBERT), Nemeskey, 2020) and term-analysis for fields like natural sciences (Wulff, 2023). The encyclopedic writing style provides a relatively stable genre, while the revision history offers detailed information about contributors. However, to the best of our knowledge there is no Wikipedia corpus of first revisions, where authorship is attributable to a single editor. As a result, existing Wikipedia-based datasets are ill-suited for investigating stylistic consistency of authors across content domains.

3 Corpus Design

The design of *WikiFirst* is guided by the goal of isolating content variation as a variable in authorship analysis. To this end, the corpus is constructed from Wikipedia articles written in a single, stable genre, while deliberately varying topic across documents authored by the same individuals. Authorship certainty is prioritised by restricting inclusion

to first revisions produced by non-anonymous editors, thereby avoiding ambiguity and potential style transfer introduced by collaborative editing.

By leveraging the structured metadata and revision history in Wikipedia, *WikiFirst* balances scale, control, and reproducibility. The corpus is intended to support both methodological analysis and empirical evaluation, enabling researchers to assess how content variation affects stylistic similarity independently of genre effects.

Four core design principles underpin the construction of *WikiFirst*. First, authorship certainty is ensured by selecting only the first revision authored by the page creators who are registered editors, providing a clear link between text and author. Second, genre control is achieved by limiting the corpus to encyclopaedic articles. Third, topic variation is introduced systematically by identifying prolific contributors who have created articles across multiple, well-defined content domains, enabling controlled comparison across topics. Finally, reproducibility is facilitated through transparent selection criteria and reliance on publicly available data, allowing the corpus to be reconstructed or extended in future work.

To operationalise the design principles of authorship certainty, genre control, topic variation, and reproducibility, the corpus was constructed using the following five inclusion criteria: (1) editors must be non-anonymous and (2) human who (3) have created initial articles in (4) five or more distinct content domains (5) prior to the OpenAI release of ChatGPT on 30 November 2022. Accounts identified as bots or otherwise automated were excluded.

4 Content Taxonomy

For authorship analysis, an effective taxonomy should capture content differences that plausibly affect lexical and discourse patterns, while remaining analytically simple. Content domains are treated as mutually exclusive categories, with each article assigned to a single domain to avoid ambiguity and facilitate controlled pairwise and cross-domain comparisons.

Content is divided into two main categories, namely (1) sciences and (2) humanities and cultural domains, each of which may be represented by 2×2 matrices, giving a total of 8 subcategories. Sciences may be divided broadly in two dimensions, namely (1) based on the degree to which laws govern the subject, and (2) whether the fo-

cus is more theoretical or practical. This may be represented as a 2×2 matrix as shown in Table 1

	Pure	Applied
Hard	e.g. Physics	e.g. Economics
Soft	e.g. Computer science	e.g. Business

Table 1: Two-dimensional taxonomy of scientific content domains (adapted from [Becher and Trowler \(2001\)](#)).

For the humanities and cultural domains, a complementary low-dimensional analytical taxonomy is adopted drawing on established distinctions between descriptive and normative discourse, and between institutional and cultural forms of knowledge. This taxonomy shown in Table 2 captures the broad differences in communicative purpose and cultural embedding while remaining analytically tractable.

	Descriptive	Normative
Institutional	e.g. History	e.g. Politics
Cultural	e.g. Biography	e.g. Religion

Table 2: Two-dimensional taxonomy of humanities and cultural content domains

Wikipedia uses a category graph rather than a strict classification hierarchy, allowing overlapping categories that evolve through community editing, but this flexibility results in cycles, redundancy, and inconsistent granularity. The thirteen Wikipedia categories are, therefore, mapped to the eight target content domains.

To ensure sufficient topic coverage per author, a balancing strategy is adopted that trades off taxonomic granularity against author availability. This is achieved by subdividing selected high-level categories into finer-grained domains where necessary, increasing domain diversity without compromising the taxonomic structure.

5 Data Collection and Extraction

Potential target authors were identified from lists of prolific Wikipedia contributors ([Wikipedia contributors, 2025b](#)).

A tailor-made data collection script (see Algorithm 1 for pseudocode) was created that retrieves the initial revision of newly created pages and extracts the author identifier, revision timestamp, page title, and all textual content entered by the editor.

Require: User CSV file F_{in} , Output Directory D_{out} , Batch Size $B \leftarrow 50$

Ensure: JSON files containing classified text and token counts for each user

```

Metadata processing
1:  $Users \leftarrow \text{ReadCSV}(F_{in})$ 
2: for all  $user \in Users$  do
3:    $Creations \leftarrow \emptyset$ 
4:    $SeenTitles \leftarrow \emptyset$ 
5:    $AllArticles \leftarrow \text{GetContribs}(user)$ 
6:   for all  $article \in AllArticles$  do
7:     if  $article.title \notin SeenTitles$  then
8:        $SeenTitles \leftarrow SeenTitles \cup \{article.title\}$ 
9:        $Creations \leftarrow Creations \cup \{article\}$ 
10:     $\text{SaveJSON}(Creations, D_{out} + "/" + user)$ 

Content fetching and domain classification
11: for all  $file \in \text{ListFiles}(D_{out})$  do
12:    $Entries \leftarrow \text{LoadJSON}(file)$ 
13:    $Pending \leftarrow Entries$ 
14:    $Batches \leftarrow \text{ChunkList}(Pending, B)$ 
15:   for all  $batch \in Batches$  do
16:      $RevIDs \leftarrow \{e.revid \mid e \in batch\}$ 
17:      $Titles \leftarrow \{e.title \mid e \in batch\}$ 
18:      $TextMap \leftarrow \text{FetchText}(RevIDs)$ 
19:      $CatMap \leftarrow \text{FetchCat}(Titles)$ 
20:     for all  $e \in batch$  do
21:        $T \leftarrow TextMap[e.revid]$ 
22:       if "#REDIRECT" in  $T$  then
23:          $e.dom \leftarrow \text{"System/Redirect"}$ 
24:          $e.numToken \leftarrow 0$ 
25:          $e.text \leftarrow ""$ 
26:       else
27:          $e.text \leftarrow T$ 
28:          $e.numToken \leftarrow \text{length}(T)$ 
29:          $Cat \leftarrow CatMap[e.title]$ 
30:         if  $Cat \neq \emptyset$  then
31:            $e.dom \leftarrow \text{getDom}(Cat)$ 
32:         else
33:            $e.dom \leftarrow \text{"NA"}$ 
34:          $\text{SaveJSON}(Entries, file)$ 
35:          $\text{Sleep}(1.0)$  ▷ Avoid 429 error code

```

Algorithm 1: Wikipedia User Contribution Fetch Pipeline

The original content fetched from the Wikipedia API is in Wikitext format, a markup language that intersperses content with formatting syntax. The text was cleaned and preprocessed using the WikiText cleaning script ([Pryzant, n.d.](#)). This script utilizes the `mwparsershell` library to iteratively strip Wikitext markup, converting it into plain text. A multi-pass approach is employed, parsing and stripping the code three times to handle complex nested structures. Additionally, custom regular expressions are used to aggressively remove citations, HTML artifacts, URLs, table formatting characters, and non-ASCII characters. Finally, after all filtration is done, authors who had less than 10 documents are removed

from the corpus. The code used can be found at <https://github.com/himynameiszim/wikifirst>.

6 Corpus Statistics

Statistical analysis of the corpus was done using R programming language (R Core Team, 2025), Quanteda (Benoit et al., 2018) package, and Tidyverse (Wickham et al., 2019) library.

WikiFirst consists of 100 unique authors contributing a total of 226,098 documents across 13 categories. The corpus contains approximately 42,710,363 tokens in total, with a mean document length of 188.902 tokens (SD = 641.68). Each author contributes an average of 2260.98 documents, spanning a mean of 12.16 categories per author.

Table 3 provides a high-level overview of the corpus composition.

Item	Number
Number of authors	100
Number of documents	226,098
Number of content domains	13
Mean domains per author	12.16
Mean documents per author	2260.98
Total word tokens	42,710,363
Mean tokens per doc	188.902
Standard dev. (tokens per doc.)	641.68

Table 3: Summary statistics for *WikiFirst*

Table 4 shows the number of texts in each of the Wikipedia categories.

Category	Texts
Culture and the arts	7,735
General reference	1,810
Geography and places	9,531
Health and fitness	564
History and events	2,798
Human activities	15,524
Mathematics and logic	9,639
Natural and physical sciences	1,991
People and self	30,888
Philosophy and thinking	8,392
Religion and belief systems	2,161
Society and social sciences	27,254
Technology and applied sciences	107,811

Table 4: Document distribution across Wikipedia content categories

The Wikipedia categories are highly imbalanced with only four out of 13 categories comprising over 10,000 texts, and four containing fewer than 2,000. These categories were mapped to the proposed context taxonomy, resulting in the distribution shown in Table 5.

Category	Texts
Pure hard science	11,630
Applied hard science	108,375
Pure soft science	11,341
Applied soft science	15,524
Descriptive institutional content	2,798
Normative institutional content	27,254
Descriptive cultural content	38,623
Normative cultural content	10,553

Table 5: Mapped document distribution categories

Overall, 73% of all authors have at least 10 documents in 7 out of 8 domains, allowing the corpus to be used in a robust, content-focused model training environment. The proposed eight-domain taxonomy produces a substantially more balanced distribution, with every domain containing at least 2,700 texts and seven of the eight domains exceeding 10,000 texts. This improved balance supports more reliable cross-domain comparison and reduces the risk that observed differences in authorship similarity are driven by data sparsity rather than genuine content effects.

7 Conclusion

In conclusion, our dataset contains more than 200,000 documents across 8 content domains, which is sufficiently diverse for content modeling in authorship attribution tasks. It should be noted that our dataset is subject to the licensing terms of Wikipedia, namely Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA).

As the information contained is crowd-sourced, its truth value and acceptability is dependent on the contributing editors (Wikipedia contributors, 2025a). However, as this dataset is created for authorship attribution model training purposes, authorship style and not the objective truth is paramount. The dataset can be accessed from [doi:10.34740/kaggle/dsv/14797130](https://doi.org/10.34740/kaggle/dsv/14797130).

Limitations

A primary limitation of the final corpus is the authorial signature is affected by the editorial rules of Wikipedia. The adherence to these relatively strict genre rules, however, enhances the generic integrity of the corpus. In addition, all the authors in the dataset are top contributors in Wikipedia, making the texts obtained closer to the norm of Wikipedia, lowering authorial diversity.

Finally, the corpus consists of articles that were written since the creation of Wikipedia, which may

introduce minor temporal variation in terminology or individual writing habits. However, this variation is relatively insignificant compared to the timescales typically associated with substantial language change, and all authors have texts ranging from years ago reducing the risk of time-related bias in the dataset.

References

- Tony Becher and Paul Trowler. 2001. *Academic tribes and territories*. McGraw-Hill Education.
- Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. [quanteda: An R package for the quantitative analysis of textual data](#). *Journal of Open Source Software*, 3(30):774.
- John Blake, Abu Saleh Musa Miah, Krzysztof Kredens, and Jungpil Shin. 2025. [Detection of AI-Generated texts: A Bi-LSTM and attention-based approach](#). *IEEE Access*, 13:71563–71576.
- Malcolm Coulthard and Alison Johnson. 2000. *Forensic Linguistics: An Introduction to Language in the Justice System*. Routledge.
- Maciej Eder and Rafał L Górski. 2023. [Stylistic fingerprints, pos-tags, and inflected languages: A case study in polish](#). *Journal of Quantitative Linguistics*, 30(1):86–103.
- Abdulmecit Güngör. 2015. Benchmarking authorship attribution techniques using over a thousand books by fifty victorian era novelists. Master’s thesis, Purdue University, West Lafayette, IN, USA.
- Annina Heini and Krzysztof Kredens. 2024. [Remote data collection in sociolinguistics: lessons from the COVID-19 pandemic](#). *International Journal of Social Research Methodology*, 27(6):747–759.
- Javed Hussain. 2020. [Enron email dataset](#).
- Mike Kestemont, Kim Luyckx, Walter Daelemans, and Thomas Crombez. 2012. [Cross-genre authorship verification using unmasking](#). *English Studies*, 93:153–160.
- Mike Kestemont, Enrique Manjavacas, Ilija Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the cross-domain authorship verification task at PAN 2020. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*.
- Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Janek Bevendorff, Martin Potthast, and Benno Stein. 2021. Overview of the cross-domain authorship verification task at PAN 2021. In *Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*.
- Marcus Ma, Duong Minh Le, Junmo Kang, Yao Dou, John Cadigan, Dayne Freitag, Alan Ritter, and Wei Xu. 2025. [CROSSNEWS: A cross-genre authorship verification and attribution benchmark](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24777–24785.
- Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. [Surveying stylometry techniques and applications](#). *ACM Computing Surveys (CSuR)*, 50(6):1–36.
- Dávid Márk Nemeskey. 2020. *Natural Language Processing Methods for Language Modeling*. Ph.D. thesis, Budapest University of Technology and Economics.
- Andrea Nini, Peter Burnap, Matthew L. Williams, and Kevin Knight. 2023. [Register variation in malicious forensic texts: An exploratory analysis](#). In *Proceedings of the Corpus Linguistics Conference*.
- Reid Pryzant. n.d. [wiki text cleaner](#). <https://gist.github.com/rpryzant/561cc1b4d372cce7479fd14290eachbc3>. GitHub Gist. Accessed: 2026-01-06.
- R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- G. Çağatay Sat, John Blake, and Evgeny Pyshkin. 2025. [Modelling the relative contributions of stylistic features in forensic authorship attribution](#). In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI era*, pages 1066–1073, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. [Welcome to the tidyverse](#). *Journal of Open Source Software*, 4(43):1686.
- Wikipedia contributors. 2025a. [Reliability of Wikipedia](#). [Online; accessed 5-Jan-2026].
- Wikipedia contributors. 2025b. [Wikipedia: List of Wikipedians by number of edits](#). [Online; accessed 27-Dec-2025].
- Peter Wulff. 2023. Network analysis of terms in the natural sciences insights from wikipedia through natural language processing and network analysis. *Education and Information Technologies*, 28(11):14325–14346.

Measuring the Symbolic Power of Languages with LLM-based Multilingual Persuasion Simulation

Yin Jou Huang and Fei Cheng

Graduate School of Informatics, Kyoto University, Kyoto, Japan
{huang, feicheng}@nlp.ist.i.kyoto-u.ac.jp

Abstract

Prior studies on the symbolic power of languages have largely relied on surveys or localized experiments, limiting systematic comparison across cultures and domains. In this work, we propose an LLM-based multilingual persuasion simulation framework to quantify the symbolic power of languages through persuasion outcomes. We also introduce a Symbolic Power Index (SPI) that measures how language choice affects persuasion success and efficiency across domains. Experiments show that the LLM-based simulations largely reproduce established sociolinguistic prestige hierarchies tied to institutional authority and global power, especially in domains such as business, finance, education, and technology. These results suggest that LLM-based persuasion simulations offer a scalable, decision-making-driven approach to studying symbolic power in language.

1 Introduction

Language is not only a medium for conveying meaning but also a social signal that carries **symbolic power**. The symbolic power of a language refers to its socially constructed capacity to signal prestige, legitimacy, and authority beyond its propositional content (Bourdieu, 1991; Gal and Irvine, 1995; Woolard, 2020). One prominent manifestation of symbolic power is the **language prestige effect**. In multilingual societies, the same message expressed in different languages or dialects can evoke systematically different perceptions of intelligence, professionalism, credibility, and authority. For example, in postcolonial countries such as Morocco or India, using a former colonial language (e.g., French or English) in professional or institutional settings often leads to higher perceived expertise and legitimacy than using local or indigenous languages, even when the propositional content remains unchanged (Kachru, 1986; Ennaji, 2005).

Linguistic capital theory offers a framework for analyzing and quantifying the symbolic power of language. The theory suggests that languages carry different “market values”, which refer to the social, economic, and material advantages associated with using a particular language in specific contexts (Bourdieu, 1991). From this perspective, the symbolic power of languages can be empirically quantified by comparing how language choices systematically affect social judgments and outcomes in social interactions. Despite extensive sociolinguistic work on symbolic power and language prestige effects, existing approaches often rely on localized surveys, single-domain experiments and case studies, while systematic studies across cultures and domains remain challenging due to practical and methodological constraints.

Recent advances in large language models (LLMs) offer a new opportunity to revisit these questions. Modern LLMs are trained on large volumes of human-generated text spanning diverse languages and cultural settings. As a result, they capture distributional patterns that correlate language use with diverse cultural norms and social values (Tao et al., 2024; Lertvittayakumjorn et al., 2025; Lin, 2025). Also, a growing body of work reveals the potential of LLMs to emulate diverse human decision-making behaviors, including economic games, negotiations, and moral dilemmas (Park et al., 2023; Huang and Hadfi, 2024). These findings motivate the use of LLM agents as human proxies for probing socially grounded patterns under controlled experimental conditions.

With the above motivation, we propose an LLM-based multilingual persuasion simulation framework to quantify the symbolic power of languages (Figure 1). We consider persuasion scenarios across several domains, including politics, finance, and culture. For each persuasion scenario, we configure LLM agents to role-play persuader/persuadee and engage them in multi-turn

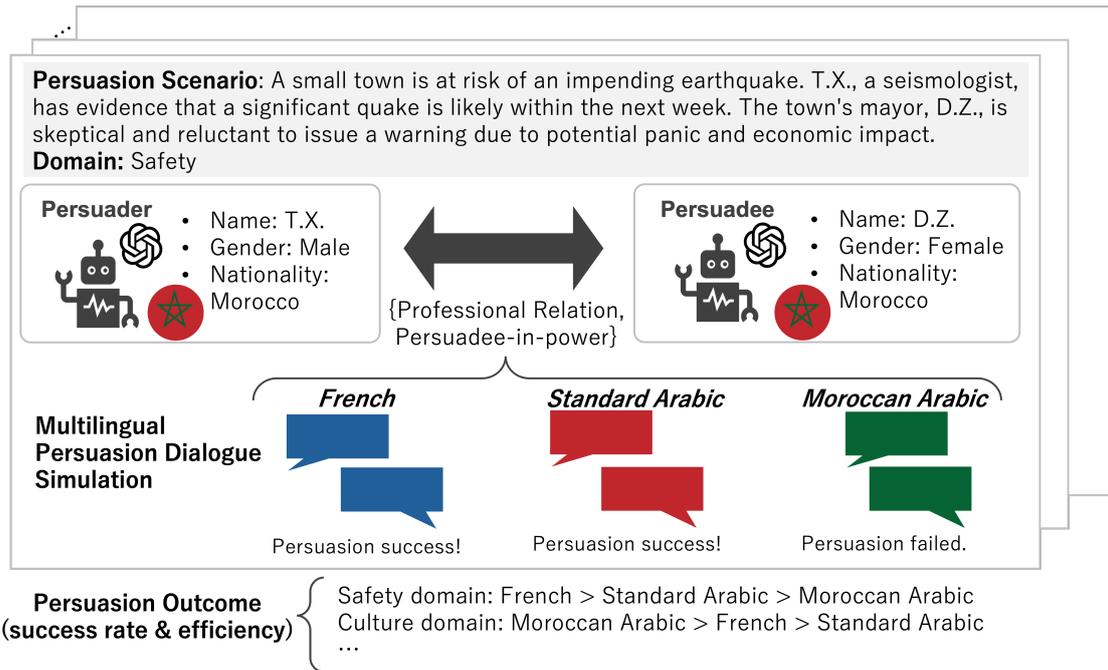


Figure 1: Overview of the multilingual persuasion simulation framework for measuring language symbolic power.

persuasive dialogues across several domains, including politics, finance, and culture. Specifically, we instruct LLM agents to emulate the decision-making behaviors of people from specific cultural backgrounds. We specifically focus on multilingual countries in which multiple languages coexist. For each multilingual country, we configure LLM agents to act as a person from the country, and conduct persuasion dialogue simulation in the languages spoken in the country, respectively. Further, we compare the persuasion outcome across different languages in terms of persuasion success rate and efficiency. We also analyze how relational context, power asymmetry and gender assigned to LLM agents interact with language choice to shape persuasion outcome.

Empirical results reveal that the LLM-based agents reproduce several well-documented sociolinguistic patterns, including dominant prestige hierarchies associated with colonial history, institutional authority, and global economic power. At the same time, we observe systematic deviations, particularly for regionally bounded or vernacular forms of prestige, suggesting that LLM-based simulations may be more sensitive to globally mediated symbolic power than to local or community-level prestige structures. Together, these findings suggest that LLM agents encode socially grounded associations between language use and symbolic power, and that controlled persuasion simulations

provide a promising computational lens for studying sociolinguistic phenomena.

2 Related Work

This section summarizes related works on existing sociolinguistic theories and previous work that utilizes LLMs as social simulators.

Sociolinguistic Theories of Language Symbolic Power Sociolinguistics has examined how language functions not only as a communicative medium but also as a carrier of social meaning, power, and inequality. Foundational work on symbolic power and linguistic capital argues that languages acquire socially constructed market values, conferring legitimacy, authority, and material advantage on their speakers in specific contexts (Bourdieu, 1991). This perspective has been extended through studies of language ideology, which show how beliefs about language naturalize social hierarchies and index identities such as education, class, and national belonging (Gal and Irvine, 1995; Woolard, 2020).

Empirically, research on language prestige, diglossia, and postcolonial multilingualism demonstrates that the same propositional content can elicit systematically different judgments depending on language choice. Former colonial languages often enjoy institutional and economic prestige, while local or vernacular languages retain symbolic value

in domains tied to identity, intimacy, or cultural authenticity (Kachru, 1986; Ennaji, 2005). Another line of work documents phenomena like foreign language effects, showing that language choice modulates emotional resonance, moral judgment, and persuasion, particularly across private versus institutional contexts. However, much of this literature relies on surveys, laboratory experiments, or localized case studies, making it difficult to systematically compare prestige effects across cultures, domains, and interactional settings at scale.

LLMs as Social and Cultural Simulators Recent advances in large language models (LLMs) have opened new avenues for studying socially grounded phenomena computationally. A growing body of work shows that LLMs encode cultural norms, values, and stereotypes present in their training data, often exhibiting systematic cultural biases even when generating fluent text in multiple languages (Naous et al., 2024; Tao et al., 2024; Zhou et al., 2025). Many studies evaluate such cultural alignment using questionnaire-based or dilemma-style benchmarks grounded in cross-cultural psychology, such as the World Values Survey or Hofstede’s cultural dimensions, revealing moderate alignment with human survey data alongside persistent WEIRD-centric biases (AlKhamissi et al., 2024; Masoud et al., 2025). Beyond static evaluations, recent work has begun to emphasize interactive role-playing simulations as a means of probing how cultural values are enacted, negotiated, and transformed through dialogue (Huang and Hadfi, 2024; Jin et al., 2024). Crucially, several studies demonstrate that language choice itself acts as an implicit cultural signal, shaping power dynamics, concession behavior, and the resolution of value conflicts in multilingual interactions (Wendler et al., 2024; Schut et al., 2025). These findings position LLM-based simulations as a promising tool for probing socially grounded phenomena.

3 Methodology

To investigate the symbolic power of languages across diverse cultural contexts, we propose a multilingual persuasion simulation framework based on LLM role-playing agents. Our framework is designed to examine how language choice shapes persuasion outcomes under controlled conditions. In this section, we introduce the persuasion scenarios (Section 3.1), the persuasion dialogue simulation process (Section 3.2), and outcome evaluation

metrics used in our study (Section 3.3).

3.1 Persuasion Scenarios

In our proposed framework, we quantify the symbolic power of languages by observing systematic changes in persuasion outcomes under different language choices. Since symbolic power is inherently domain-dependent, with the prestigious language likely to differ in different domains, we consider persuasion scenarios across diverse domains.

In this work, we adapt the persuasion scenarios from the DailyPersuasion dataset (Jin et al., 2024), which contains everyday persuasive interactions spanning a wide range of social domains. From this dataset, we select the 13 domains that are particularly relevant for studying language symbolic power: **politics, finance, business, ethics, technology, education, literature, art, culture, safety, lifestyle, health, and psychology**. For each domain, we randomly sample 10 scenarios, resulting in a total of 130 scenarios. Each scenario depicts a background situation of the persuasion involving a persuader and a persuadee, along with a persuasion goal. See Table 4 in the Appendix for examples of scenarios.

To facilitate the subsequent LLM-based simulation and a more detailed analysis, we augment each scenario with the following additional annotations:

Agent profiles The original DailyPersuasion dataset uses concrete real-world names for the persuader and persuadee (e.g., “Emma”, “Wang Wu”). These names may inadvertently introduce cultural or social signals unrelated to the experimental manipulation. To avoid this, we replace all agent names with **abbreviation-style names** like T.X. and D.Z. The abbreviation-style names are randomly generated. We further augment each agent with demographic attributes. First, we use an LLM to extract the gender of the persuader and persuadee based on the scenario description. This allows us to analyze potential gender-related effects in persuasion in different languages. In addition, we assign each agent an explicit cultural identity through a country-level **nationality** instruction (e.g., “You are from India”). This explicit cultural framing is intended to activate culturally grounded norms, expectations, and values associated with the specified country, enabling the LLM agents to emulate culturally conditioned decision-making behaviors. Such framing has been shown to improve the cultural grounding and behavioral consistency

of LLM-based role-playing agents (Tao et al., 2024; Lertvittayakumjorn et al., 2025). See Appendix A for the complete prompt.

Persuader-persuadee relationship Beyond individual attributes, persuasion outcomes are known to be strongly shaped by relational context and power dynamics. We therefore extract details of the relationship between the persuader and persuadee with LLM. First, we classify the relationship between the persuader and persuadee into two broad categories: private (e.g., family members, friends) and professional (e.g., colleagues, supervisors, institutional roles). Second, we annotate the power structure between the two agents by identifying whether the persuader holds greater power, the persuadee holds greater power, or both parties are of equal status. These relational annotations allow us to systematically examine how the symbolic power of languages interacts with social roles and power asymmetries in shaping persuasion outcomes. See Table 4 in the Appendix for examples of persuasion scenarios.

3.2 Persuasion Dialogue Simulation

We model persuasion as a multi-round interactive dialogue between a persuader agent and a persuadee agent. We consider a multilingual country c in which a set of languages L_c is spoken. We conduct persuasion dialogue simulations with the persuader and persuadee LLM agents instructed to act as a person from the country c . For each persuasion scenario s_i in our dataset, we conduct persuasion dialogue simulations in every language $l \in L_c$, respectively. Specifically, the agents are instructed to converse with each other in language l based on the scenario s_i .

The agents generate utterances alternatively, and each generated utterance is fed to the other agent as a prompt to generate the next utterance. The persuader initiates the dialogue. At each dialogue round, the persuader produces a persuasive utterance in the target language, after which the persuadee responds conversationally. Apart from generating the utterance, the persuadee makes an explicit binary decision indicating whether the persuasion attempt is accepted or rejected. The dialogue terminates when either the persuadee accepts the persuasion or a maximum of five rounds is reached.

3.3 Persuasion Outcome Evaluation

We evaluate each persuasion simulation along two complementary dimensions: persuasion success and persuasion efficiency. For a persuasion simulation conducted based on scenario s_i , with agents instructed to act as people in country c and converse in language l , **persuasion success** is defined as a binary variable $success_i^{(c,l)} \in \{0, 1\}$, indicating whether the persuadee accepts the persuasion by the end of the dialogue. In cases of successful persuasion, we also measure the **persuasion efficiency**, which is the inverse of the number of dialogue rounds $|D_i^{(c,l)}|$ required to reach acceptance. Shorter dialogues thus correspond to higher efficiency.

We combine these two metrics into a single persuasion outcome score $m_i^{(c,l)}$ as the multiplication of the two metrics:

$$m_i^{(c,l)} = \frac{success_i^{(c,l)}}{|D_i^{(c,l)}|} \quad (1)$$

We can further aggregate the persuasion outcomes of each domain d by averaging outcome of all scenarios belonging to that domain:

$$M_d^{(c,l)} = avg(m_i^{(c,l)} | s_i \in domain\ d) \quad (2)$$

Symbolic Power Index Finally, to quantify the relative symbolic power of languages, we introduce the Symbolic Power Index (SPI). For a given language l , cultural context of country c , and domain d , SPI measures how persuasion outcomes in language l deviate from the average performance of all languages used within the same cultural and domain context.

$$SPI_d^{(c,l)} = \frac{M_d^{(c,l)} - \sum_{l' \in L_c} M_d^{(c,l')}}{std(M_d^{(c,l')} | l' \in L_c)} \quad (3)$$

This normalization yields a score that captures the relative advantage or disadvantage of a language within a specific linguistic market. Higher SPI values indicate greater symbolic power in persuasion, expressing the sociolinguistic notions of prestige, legitimacy, and authority in a comparable, quantitative form.

4 Experimental Setting

LLM Agents We use GPT-4o (Hurst et al., 2024) as the backbone of all LLM agents in our experiments. Each agent is assigned a synthetic profile

Nationality c	List of Languages L_c
India	English, Hindi, Bengali, Telugu, Tamil
Morocco	French, Standard Arabic, Moroccan Arabic
Malaysia	English, Malay, Simplified Chinese, Tamil
Hong Kong	English, Simplified Chinese, Cantonese
Finland	English, Swedish, Finnish

Table 1: Nationalities and languages included in the multilingual persuasion simulations.

that includes an abbreviation-style name, gender, and an explicit nationality instruction. For the nationality instruction, we focus on multilingual countries in which multiple languages coexist and hold different symbolic and institutional roles (Table 1).

Persuasion Scenario Augmentation We adopt persuasion scenarios from the DailyPersuasion dataset (Jin et al., 2024) and augment each scenario with structured annotations to support controlled experimentation. The abbreviation-style agent names are generated randomly. The demographic attributes of gender, relationship type, and power structure are automatically extracted with an LLM agent. We also use GPT-4o for this. See Appendix A for the details of the prompts.

5 Results and Analysis

This section presents the results of our language symbolic power evaluation framework based on multilingual persuasion simulations. We analyze how language choice shapes persuasion outcomes across cultures, domains, and social contexts. Specifically, we examine whether LLM role-playing agents reproduce known sociolinguistic phenomena (Section 5.1). We also conduct a detailed analysis on how social factors like relation type, power asymmetry, and gender influence persuasion outcomes (Section 5.1, 5.3).

5.1 Linguistic Prestige Effect in Persuasion

We use the proposed Symbolic Power Index (SPI) to quantify the symbolic power across languages. To assess the validity of the SPI measure, we compare the language prestige patterns with established findings in sociolinguistic research on language prestige effects. For all multilingual countries in Table 1, we conducted persuasion simulation across all languages, and calculate the SPI score of all domains. The SPI scores reveal two broad patterns that correspond well to typologies in previous sociolinguistic research: Dominant prestige configurations and Competing prestige configurations.

5.1.1 Dominant Prestige Configurations

For India, Morocco, and Malaysia, one language consistently exhibits higher SPI scores across most domains, while other languages show only localized or domain-specific advantages (Figure 2). The detailed analysis of the SPI patterns of these three countries are the following.

India The SPI patterns largely reproduce known prestige hierarchies (Annamalai, 2004; Meganathan, 2015). English consistently dominates across most domains, reflecting its dominant role in society. Hindi shows moderate advantages in political and cultural domains, consistent with its national and symbolic prominence. Bengali’s relative strength in literature and art closely mirrors its historically recognized literary prestige. However, regional languages like Tamil and Telugu appear systematically disadvantaged across domains, which diverges from research emphasizing their strong regional institutional and cultural prestige. This suggests that the SPI captures global prestige more effectively than regionally bounded prestige structures.

Morocco The Morocco SPI patterns align strongly with classic diglossic and postcolonial descriptions. French emerges as the dominant prestige language across economic, educational, and technological domains, consistent with its well-documented role as a language of elite capital. Standard Arabic shows moderate advantages in formal and ethical domains, reflecting its role as the high variety in diglossic structures (Ferguson, 1959). Moroccan Arabic is largely disadvantaged in institutional domains but exhibits relative strengths in culture and art domains, which accords with its strong association with local identity and popular culture. Overall, the SPI pattern of languages in Morocco strongly matches prior sociolinguistic research.

Malaysia Malaysia also exhibits a dual prestige structure that broadly aligns with prior findings. English performs strongly in business, technology, and education, reflecting its global and professional value, while Malay maintains advantages in domains associated with national identity and state legitimacy. However, the consistently negative scores for Simplified Chinese and Tamil languages, especially in cultural domains, underrepresent their documented community-level and regional prestige. As in India, this suggests that the SPI is more sensitive to globally mediated symbolic power than

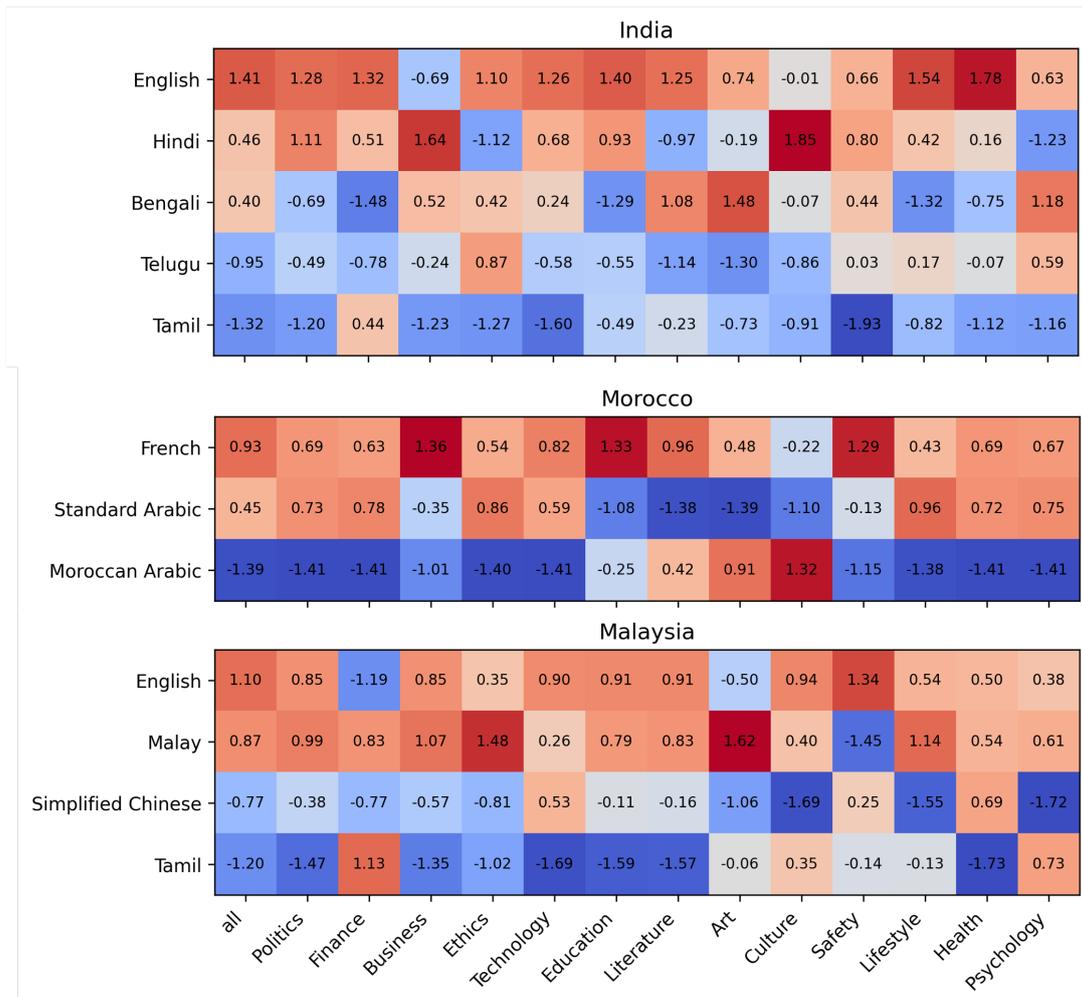


Figure 2: Symbolic Power Index (SPI) across languages and domains for dominant prestige configurations.

to local prestige dynamics.

Overall, the above SPI patterns align with sociolinguistic accounts of **prestige multilingualism**, where colonial history, political reasons, and economic forces elevate certain languages above others in most symbolic markets. In these settings, the dominant language functions as a cross-domain carrier of authority, competence, and legitimacy, while other languages only exhibit dominance in a few domains.

5.1.2 Competing prestige configurations

In contrast, Hong Kong and Finland exhibit a more heterogeneous pattern, where different languages dominate different domains rather than a single language prevailing universally. This distribution is consistent with descriptions of **functional multilingualism**, in which languages occupy complementary roles tied to specific institutions, interactional contexts, or symbolic functions.

Hong Kong The Hong Kong SPI patterns are largely consistent with research on institutional bilingualism and local identity. English dominates in business, finance, and education, reflecting its role in global markets and professional institutions. Cantonese, while disadvantaged in many institutional domains, shows relative strengths in domains associated with interpersonal or socially grounded interactions, which aligns with its status as the primary language of everyday life and local identity. The prominence of Simplified Chinese in political and literary domains likely reflects mainland-oriented discourse rather than local written norms, highlighting how orthographic and geopolitical factors may shape SPI outcomes.

Finland In the SPI patterns of Finland, English exhibits strong advantages across many domains, including areas typically associated with national languages, while Finnish and Swedish show more limited domain-specific strengths. While this pattern is compatible with Finland's high level of En-

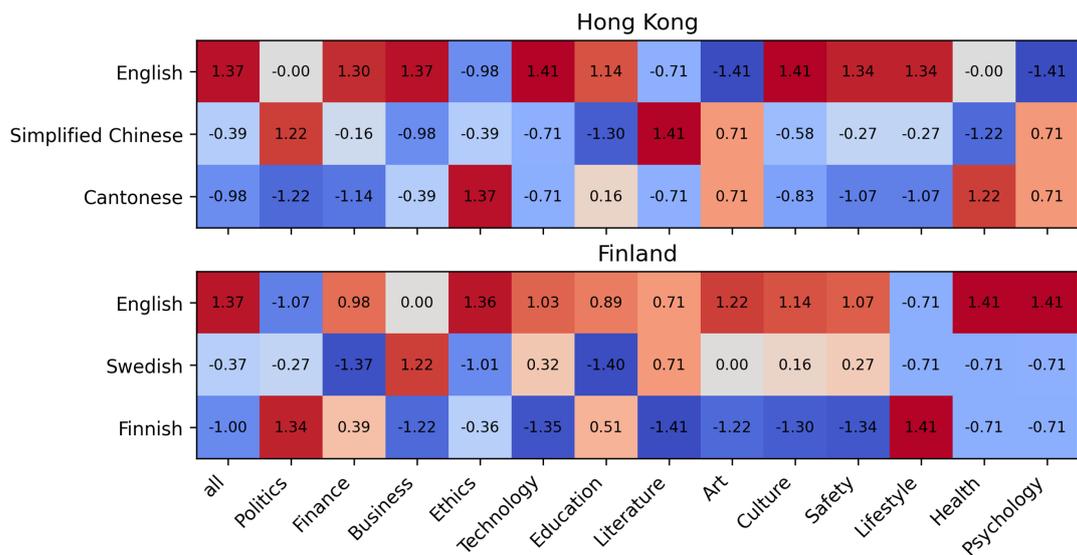


Figure 3: Symbolic Power Index (SPI) across languages and domains for competing prestige configurations.

	German	French	Dutch	Japanese	Arabic	Chinese	Spanish	Russian
Relation Type	-0.003	-0.020	-0.021	-0.002	-0.046	0.001	-0.074*	0.092*
Power Structure	-0.037	0.027	0.026	-0.142*	-0.039	-0.107*	-0.011	-0.040
Persuader Gender	0.009	-0.044	-0.004	0.002	-0.048	-0.012	0.072	0.023
Persuadee Gender	-0.017	-0.050	0.022	-0.071	-0.020	-0.025	-0.080	0.000

Table 2: Effects of social factors on persuasion success rates across languages. Asterisk (*) indicates statistical significance at $p < 0.05$.

English proficiency and internationalization, it underrepresents the constitutional and institutional authority of Finnish and Swedish in domestic governance and public life. This discrepancy suggests that the SPI primarily reflects global or transnational prestige rather than strictly domestic institutional status.

Overall, the SPI patterns reflect the multilingual equilibria in these two regions in which no single language fully dominates the others.

Divergence from sociolinguistic findings

Across countries in Figure 2 and 3, divergences from established sociolinguistic research cluster in a small set of domains. Safety and Health consistently overfavor English, contrary to prior findings that these domains are typically dominated by official or national languages in state-led and public-facing communication; this likely reflects an emphasis on international, academic, or policy-oriented discourse in the data. Lifestyle also departs from expectations, with English frequently outperforming local vernaculars despite sociolinguistic evidence that everyday lifestyle practices are usually expressed in locally grounded languages. In contrast, domains such as business,

finance, technology, education, and literature exhibit strong alignment with prior research and appear to be the most reliable indicators of institutional and symbolic prestige. These findings indicate that the SPI captures globally mediated and institutional prestige more reliably than vernacular prestige.

5.2 Analysis of Persuasion Success Rate

In addition to multilingual settings, we also conduct persuasion simulations for countries with a single common language. Specifically, we conduct persuasion simulation in the following languages: German (Germany), French (France), Dutch (the Netherlands), Japanese (Japan), Arabic (Arab countries), simplified Chinese (China), Spanish (Spain), and Russian (Russia). For each case, we conduct simulation with both persuader and persuadee agents assigned the corresponding nationality and interact in the target language.

Within each language–country setting, we consider the following four social factors and their influence on persuasion success rate: persuader–persuadee relationship, power structure, persuader gender, and persuadee gender. For each pair of lan-

	German	French	Dutch	Japanese	Arabic	Chinese	Spanish	Russian
Relation Type	-0.011	-0.048*	-0.012	-0.007	0.004	0.003	0.019	0.034*
Power Structure	-0.059	-0.010	-0.083*	-0.048	-0.074*	-0.037	-0.077	-0.048
Persuader Gender	0.038	0.019	0.061	0.089*	0.013	0.050	0.098	0.032
Persuadee Gender	-0.002	0.007	-0.011	-0.046	-0.049	-0.034	-0.080	-0.033

Table 3: Effects of social factors on persuasion efficiency rates across languages. Asterisk (*) indicates statistical significance at $p < 0.05$.

guage and country, we fit an ordinary least squares regression model with persuasion success rate as the dependent variable. Table 2 shows the signed effect sizes (with the sign showing the direction of influence) and statistical significance indicated by p-values.

Persuader-persuadee relationship The relationship between the persuader agent and the persuadee agent is categorized as either private or professional. The first row (**Relation Type**) of Table 2 shows how the relation types influence persuasion success when using a specific language. Positive numbers of effect size indicates higher persuasion success rates when the two LLM agents are in professional relational contexts. We observe a significant positive effect for Russian, where persuasion is more successful in professional relationships, and a significant negative effect for Spanish, where persuasion is more successful in private contexts. The higher persuasion success in professional settings for Russian and in private settings for Spanish mirrors cross-cultural differences in how relational context shapes perceived legitimacy in the case of human. In the Russian case, professional roles might be more readily associated with competence or authority, whereas in the Spanish case, personal trust and closeness may play a relatively larger role in shaping receptiveness. The results indicate that our proposed LLM-based simulation framework can capture difference across relational contexts.

Power structure We next analyze how relative power between persuader and persuadee shapes persuasion success. Specifically, we compare persuasion success across three conditions: (i) the persuadee holds greater power, (ii) both parties have equal power, and (iii) the persuader holds greater power. Note that from (i)(iii) corresponds to different levels of persuader power, with (i) being the lowest and (iii) the highest. The second row (**Power Structure**) of Table 2 shows how the power structure shape persuasion success, with a positive number indicating greater persuader power

leads to better persuasion success. Interestingly, in Japanese and Chinese, persuasion success decreases when the persuader holds greater power. This counterintuitive pattern suggests that higher status does not automatically translate into greater persuasive effectiveness in these contexts. Both societies of Japan and China are characterized as more hierarchy-aware, overt persuasion from a higher-power actor may be pragmatically marked.

Gender Effect We look how the gender assigned to the LLM agents affects persuasion success. We consider both the influence of the gender of the persuader and the persuadee, which corresponds to the third and the fourth row of Table 2. Here, a positive number indicates a higher success rate when the LLM agent (either persuader or persuadee) is assigned a female role, compared to agents assigned a male role. For all languages, we found no significant effects of gender regarding persuasion success rate. This absence of strong gender effects is notable given that human studies often report gendered differences in perceived persuasiveness, authority, and communication style. This could be due to the fact that current LLMs often default to relatively gender-egalitarian interaction patterns (AlKhamissi et al., 2024).

5.3 Analysis of Persuasion Efficiency

In addition to persuasion success, we examine how persuasion efficiency varies with language choice. We adopt the same setting as in Section 5.2, focusing on countries with a single common language. The influence of each social factor (relation type, power structure, persuader gender, persuadee gender) on persuasion efficiency is summarized in Table 3, which reports signed effect sizes and statistical significance. The detailed analysis regarding each factor is summarized below.

Persuader-persuadee relationship The first row (**Relation Type**) of Table 3 shows how relationship type influences persuasion efficiency across languages. Positive values indicate higher efficiency

in professional relationships opposed to private relationships. Similar to the case of persuasion success, Russian shows significantly higher efficiency in professional contexts. The professional roles in Russian-speaking contexts are strongly associated with expertise and role-based legitimacy, which may allow persuasive arguments to be evaluated more on perceived competence than on interpersonal trust. In contrast, French shows a small but significant negative effect, indicating lower efficiency in professional contexts. This result mirrors the discourse norm in which professional interactions emphasize deliberation, nuance, and justification in French language.

Power structure The second row (**Power Structure**) of Table 3 shows how the power structure influence persuasion efficiency, with a positive number indicating greater persuader power leads to higher persuasion efficiency. In Dutch and Arabic simulations, greater persuader power is associated with lower efficiency. That is, the persuasion dialogues tend to require more turns to reach resolution when the persuader holds more power. This result is consistent with cross-cultural pragmatics and sociolinguistic theories that emphasize how hierarchy and face concerns shape interactional dynamics. In Arabic-speaking contexts, politeness norms and face management are highly salient, lower-status interlocutors may prolong dialogue to maintain respect and relational harmony. On the other hand, egalitarianism are valued in Dutch-speaking cultures, attempts to persuade from a higher-power position may be perceived as less legitimate, thereby reducing efficiency. Overall, these results suggest that higher structural power does not automatically translate into faster or more successful persuasion.

Gender Effect The third and the fourth row of Table 3 summarize the influence of gender to persuasion efficiency, with positive numbers indicating higher persuasion efficiency when the LLM agent is assigned a female role. In Japanese and Spanish, dialogues with female persuaders were significantly more efficient, requiring fewer dialogue rounds to reach a persuasive outcome compared to male persuaders. Also, in Spanish, interactions involving female persuadees were associated with lower persuasion efficiency the male persuadees. The results shows a tendency that differs from prior human-centered research. Many sociolinguistic and communication studies report robust gender

differences in perceived authority, persuasiveness, and conversational style, often leading to systematic advantages or disadvantages tied to gender. In our simulations, however, gender effects are limited in scope and appear only in a few language-specific cases. This divergence may indicate that current LLMs encode weaker or more homogenized gender-linked interaction norms than those observed in real-world social behavior.

6 Conclusion

This work presents a theory-driven, simulation-based method for measuring the symbolic power of languages using LLM agents. Multilingual persuasion experiments show that LLMs encode socially grounded links between language choice and perceived authority, legitimacy, and competence, largely reflecting global prestige hierarchies. At the same time, the framework reveals limits in capturing local forms of symbolic power. We also find that language effects on persuasion depend on social structure, particularly power asymmetries.

References

- Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. [Investigating cultural alignment of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.
- E Annamalai. 2004. Nativization of english in india and its effect on multilingualism. *Journal of Language and Politics*, 3(1):151–162.
- Pierre Bourdieu. 1991. *Language and symbolic power*. Harvard university press.
- Moha Ennaji. 2005. *Multilingualism, cultural identity, and education in Morocco*. Springer.
- Charles A Ferguson. 1959. Diglossia. *word*, 15(2):325–340.
- Susan Gal and Judith T Irvine. 1995. The boundaries of languages and disciplines: How ideologies construct difference. *Social research*, pages 967–1001.
- Yin Jou Huang and Rafik Hadfi. 2024. [How personality traits influence negotiation outcomes? a simulation based on large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10336–10351, Miami, Florida, USA. Association for Computational Linguistics.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow,

- Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. [Persuading across diverse domains: a dataset and persuasion large language model](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand. Association for Computational Linguistics.
- Braj B Kachru. 1986. *The alchemy of English: The spread, functions, and models of non-native Englishes*. University of Illinois Press.
- Piyawat Lertvittayakumjorn, David Kinney, Vinodkumar Prabhakaran, Donald Martin Jr, and Sunipa Dev. 2025. Towards geo-culturally grounded llm generations. *arXiv preprint arXiv:2502.13497*.
- Zhicheng Lin. 2025. [Large language models as psychological simulators: A methodological guide](#). *Preprint*, arXiv:2506.16702.
- Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues Rodrigues. 2025. [Cultural alignment in large language models: An explanatory analysis based on hofstede’s cultural dimensions](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.
- Ramanujam Meganathan. 2015. English language education situation in india: Pedagogical perspectives. *Journal of English as an International Language*, 10(1):48–66.
- Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). *Preprint*, arXiv:2304.03442.
- Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*.
- Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.
- Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394.
- Kathryn A Woolard. 2020. Language ideology. *The international encyclopedia of linguistic anthropology*, pages 1–21.
- Ke Zhou, Marios Constantinides, and Daniele Quercia. 2025. Should llms be weird? exploring weirdness and human rights in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 2808–2820.

A Prompt templates

We provide the prompt templates for scenario augmentation below.

Prompt for scenario augmentation

You will be given a persuasion scenario between two persons.
Your task is to generate/extract the following information from the scenario:

- gender: reasonable gender of the persuader and persuadee.
- relationship type: Whether the relationship between the persuader and the persuadee falls within the category of family, social, professional relationships, or none of the above.
- power hierarchy: Judging from the relationship, which party has the higher level of power/authority over the other, or if the persuader and the persuadee have equal power.

We provide the prompt templates for persuasion simulation below.

Prompt for persuasion simulation

Your name is [NAME]. You are a [GENDER] from [NATIONALITY].
Your task is to engage in a conversation in [LANGUAGE] with [OPONENT NAME] (the user) based on the following scenario:
[SCENARIO]

B Examples of Persuasion Scenarios

	Persuader-in-power	Equal	Persuadee-in-power
Private	H.Y. is an amateur photographer who mostly takes photos with her smartphone. Her friend, C.I., is an experienced photographer and believes that H.Y. could greatly improve her skills by attending a photography workshop that focuses on using DSLR cameras. (domain: Art)	H.E. is a hardworking employee who spends most of her time working, while her friend W.E. believes in work–life balance. H.E. is hesitant about taking a vacation, fearing it might affect her career. (domain: Lifestyle)	J.Z. and her son H.D. are on vacation in Paris. H.D. is fascinated by the Eiffel Tower, but J.Z. is scared of heights and prefers to visit the Louvre instead. H.D. persuades J.Z. to overcome her fear of heights and visit the Eiffel Tower with him. (domain: Family)
Professional	L.U., the team leader, wants to organize a weekend team-building retreat for her department. She believes that a hiking trip would be an excellent way to bond and improve teamwork. O.K., one of the team members, is not enthusiastic about hiking and prefers indoor activities. (domain: Business)	S.B. is a math teacher and believes that using traditional teaching methods is more effective. Y.O., a fellow math teacher, has recently discovered the benefits of using mathematical modeling to teach complex concepts. (domain: Science)	A small town is at risk of an impending earthquake. G.F., a seismologist, has evidence that a significant quake is likely within the next week. The town’s mayor, Q.U., is skeptical and reluctant to issue a warning due to potential panic and economic impact. (domain: Safety)

Table 4: Persuasion Scenarios by Power Relationship and Context

Author Index

- Aguilar Valdez, Sofia, 146
Al-Laith, Ali, 157
- Ballatore, Andrea, 266
Baumann, Andreas, 80, 202
Belosevic, Milena, 66
Benson, Juliane, 80
Bian, Donghan, 241
Biemann, Chris, 257
Bjerring-Hansen, Jens, 157
Blake, John, 323
Boriceanu, Ioana-Roxana, 102
Boros, Emanuela, 1, 11
Bulus, Ayse, 128
- Cafiero, Florian, 241
Chatzikyriakidis, Stergios, 87
Cheng, Fei, 328
Chooi, Jay, 135
Ciletti, Michele, 266
Conroy, Alexander, 157
Cornamusaz, Emile, 11
- Degaetano-Ortlieb, Stefania, 146, 212
Degn, Kirstine, 157
Dei, Agustin, 251
Delucchi Danhier, Renate, 16
Dinu, Liviu, 102
Dipper, Stefanie, 164
- Escouflaire, Louis, 30
Essfors, Hannes, 80, 202
- Farina, Andrea, 266
Fellner, Hannes A., 80
Ferrante, Mia, 295
- Ginter, Filip, 111
Gius, Evelyn, 257
- Hamilton, Sil, 295
Hatzel, Hans Ole, 257
Hershcovich, Daniel, 157
Hicke, Rebecca, 295
Hill, Mark, 128
Huang, Yin Jou, 328
- Kanerva, Jenna, 111
Kauppi, Jenni, 111
- Klumpp, Jan-Felix, 189
Kreutel, Jörn, 311
- Laarmann-Quante, Ronja, 164
Laato, Joonatan, 111
Lummaa, Virpi, 111
- Machálek, Tomáš, 305
Marchand-Maillet, Stephane, 48
Mcgillivray, Barbara, 266
Milička, Jiří, 305
Mimno, David, 295
Moccozet, Laurent, 48
Müller, Oliver, 176
- Natsina, Anastasia, 87
Nguyen, Dung, 323
Nour, Kamal, 1
- Pagel, Janis, 280
Pawłowski, Adam, 221
Pichler, Axel, 280
Puren, Marie, 241
Pyshkin, Evgeny, 323
- Rapacz, Maciej, 227
- Sat, G. Çağatay, 323
Schacht, Carmen, 16
Schroderus, Veera, 111
Smywiński-Pohl, Aleksander, 227
Sorbi, Marco, 48
Spence, Paul, 128
Stede, Manfred, 311
Stiemer, Haimo, 257
- Tamrazyan, Hamest, 1, 11
Tüver, Lale, 80
- Vestel, Anastasiia, 212
Viola, Lorella, 36
- Walkowiak, Tomasz, 221
Wilkins, Matthew, 295
Winkle, Robin, 311
- Zeh, Katharina, 80