

Matching Meaning at Scale: Evaluating Semantic Search for 18th-Century Intellectual History through the Case of Locke

Yu Wu¹, Ananth Mahadevan^{1*}, Filip Ginter^{2,3}, Michael Mathioudakis¹, Mikko Tolonen¹

¹University of Helsinki ²TurkuNLP, University of Turku ³ELLIS Institute Finland
{yu.wu, ananth.mahadevan, michael.mathioudakis, mikko.tolonen}@helsinki.fi
figint@utu.fi

Abstract

While digitized corpora have transformed the study of intellectual transmission, current methods rely heavily on lexical text reuse detection, capturing verbatim quotations but fundamentally missing paraphrases and complex implicit engagement. This paper evaluates semantic search in 18th-century intellectual history through the reception of John Locke’s foundational work. Using expert annotation grounded in a semantic taxonomy, we examine whether an off-the-shelf semantic search pipeline can surface meaning-level correspondences overlooked by lexical methods. Our results demonstrate that semantic search retrieves substantially more implicit receptions than lexical baselines. However, linguistic diagnostics also reveal a “lexical gatekeeping” effect, where retrieval remains partially constrained by surface vocabulary overlap. These findings highlight both the potential and the limitations of semantic retrieval for analyzing the circulation of ideas in large historical corpora. The data is available at <https://github.com/COMHIS/locke-sim-data>.

1 Introduction

The digitization of vast historical archives, such as Eighteenth Century Collections Online (ECCO) (Tolonen et al., 2022) and Early English Books Online (EEBO) (Mäkelä et al., 2025), together with recent advances in Natural Language Processing (NLP), has created unprecedented opportunities for intellectual history (Tolonen and Ryan, 2026). Digital methodologies now allow scholars to move beyond the study of isolated authorial intent and begin to “recreate the cultural terrain” of historical debate by tracing how ideas circulated across large textual corpora (Edelstein, 2016). At the same time, this shift from close reading to large-scale computational analysis exposes

a central methodological bottleneck: the semantic gap between lexical similarity and meaning-level relationships preserved in textual passages.

Traditional computational approaches to reception history rely heavily on lexical text reuse detection (Rosson et al., 2023); for further discussion in the context of intellectual history, see Tolonen and Ryan (2026). While effective for identifying verbatim quotations, such approaches struggle with implicit reference, paraphrase, Optical Character Recognition (OCR) errors, diachronic lexical change, and the conceptual complexity of philosophical language. As a result, many forms of intellectual engagement between texts remain difficult to detect at scale.

One context in which this semantic gap becomes particularly visible is 18th-century British intellectual history. The period is widely recognised as a locus of Enlightenment contestation and a rapidly expanding print public, where arguments about reason, religion, science, and politics circulated across genres and decades (Porter, 2001). Tracing how such arguments were reformulated and redeployed is essential for understanding the transmission of ideas, yet doing so across millions of publications is beyond the scope of manual scholarship. While recent advances in semantic search based on dense vector retrieval (Reimers and Gurevych, 2019; Douze et al., 2025) make it theoretically plausible to detect semantic similarity beyond exact lexical overlap in the time frame and language domain, their reliability on uncurated historical corpora remains empirically untested.

To operationalise this inquiry, we focus on John Locke (1632–1704)’s *An Essay Concerning Human Understanding* (1689). Because the *Essay* was published at the end of the seventeenth century, its reception spans the entire 18th-century corpus analysed here, allowing its influence to be traced across the full period. This is not optimal for authors such as David Hume, whose major Enlightenment pub-

*Contribution made during doctoral studies.

lications appear only from the late 1730s onwards. Locke’s work is also particularly suitable because his intellectual legacy is more contested and complex than the traditional narrative of the “father of liberalism” suggests (Stanton, 2018). Recent scholarship has highlighted the fragmented nature of his reception, which developed into distinct republican and libertarian traditions (Layman, 2021). At the level of textual practice, Locke’s arguments often appear in structured formulations that are subsequently reformulated or paraphrased in later texts. In such cases, lexical searches fundamentally fail to capture the semantic footprint of the idea.

This paper therefore presents a proof-of-concept evaluation of semantic search for 18th-century historical corpora through the case of Locke’s quotations. We examine whether a minimal, off-the-shelf semantic search setup can identify meaning-level correspondences in the reception of Locke that lexical text reuse detection methods miss. We first operationalize “reception” into a hierarchical framework that distinguishes genuine semantic engagement, such as paraphrases and implicit meaning matches, from both verbatim text reuse and mere topical matches. Guided by this framework, we implement a multi-stage evaluation workflow, annotating and analyzing iteratively in non-lexical hits of our semantic search pipeline. Based on the annotated subset, our mixed-methods analysis demonstrates that semantic search uncovers an order of magnitude more implicit reception than traditional string matching across core historical queries. Finally, through a linguistic diagnostic encompassing vocabulary overlap, OCR noise, and syntactic divergence, we expose a profound “lexical gatekeeping” effect: latent lexical similarities continue to shape retrieval outcomes, even within dense semantic spaces.

To summarize, our main contributions are:

- A proof-of-concept evaluation using Locke as a case study, demonstrating that semantic search retrieves substantially more implicit receptions than lexical text-reuse baselines.
- A conceptual framework for modelling reception that distinguishes between different contextual matches at the level of semantic relationships between passages.
- A reusable iterative evaluation workflow that applies the annotation framework to identify meaning-level receptions and their historical significance.

- A linguistic diagnostic analysis revealing a *lexical gatekeeping* effect despite robustness to OCR noise and syntactic variation.

2 Related Work

Digitized Infrastructure and Lexical Methods

The mass digitization of historical archives (Armitage and Guldi, 2014; Bode, 2018; Langlais, 2021), coupled with expansive structured metadata and relational networks (Hill et al., 2019; Bermès, 2017), has established the infrastructure for computational intellectual history. Retrieving specific conceptual evidence from these corpora has traditionally relied on lexical methodologies. Techniques ranging from Keyword in Context (KWIC) (Luhn, 1960), word-level distributional semantics (Fitzmaurice et al., 2017), and probabilistic topic modeling (Blei, 2012; Goldstone and Underwood, 2014), to advanced hashing and sequence-alignment algorithms for OCR-resilient text reuse (Mahadevan et al., 2025; Düring et al., 2023; Vesanto et al., 2017; Smith et al., 2014), mark the current limits of surface-level matching. Bridging this persistent “semantic gap” requires moving beyond static methods to neural, passage-level models capable of tracking complex conceptual shifts (McGillivray et al., 2024; Giulianelli et al., 2020).

Semantic Search in Historical Corpora To advance beyond lexical navigation, digital humanists increasingly use dense Information Retrieval (IR) models, such as Dense Passage Retrieval (Karpukhin et al., 2020) and Sentence Transformer (Reimers and Gurevych, 2019), to capture compositional semantics, bypassing explicit string matching. Recent applications demonstrate the capacity of bi-encoders to capture complex intertextualities (Schelb et al., 2026) and semantic parallels (Kanerva et al., 2025; Franklin et al., 2024; Alrahabi and Wainstain, 2025) that remain invisible to lexical searches. However, deploying contemporary neural networks on uncurated historical paragraphs exposes severe algorithmic vulnerabilities. Alongside OCR-induced tokenization artifacts (Zhang et al., 2025; Michail et al., 2025), temporal semantic degradation (Lazaridou et al., 2021) and entity-centric popularity biases (Sciavolino et al., 2021), dense models regress to shallow vocabulary matching when confronted with out-of-domain historical syntax (Rizwan et al., 2025; MacAvaney et al., 2022). This limitation is further entrenched by standard IR benchmarks, which

structurally penalize true semantic retrieval through lexically pre-filtered ground truth (Thakur et al., 2021). Therefore, rigorously evaluating semantic architectures within real-world historical corpora is critical to distinguish genuine meaning retrieval from lingering algorithmic biases.

Computational Early Modern Reception Early computational methods successfully mapped early modern intellectual networks using structured metadata from social and epistolary graphs (Warren et al., 2016; Edelstein et al., 2017; Burrows and Curran, 2012), while lexical pipelines tracked verbatim text reuse (Roe, 2024; Tolonen and Spencer, 2025) and extracted explicit quotations (Muzny et al., 2017; Brunner et al., 2020). However, the highly contested intellectual legacy of foundational figures like John Locke (Collins, 2020; Hanck, 2019) exemplifies the requirement of moving to passage-level semantic mappings: his foundational concepts were widely appropriated without explicit citation, relying on implicit references and shifting vocabularies that completely evade traditional lexical searches (Harris, 2023; Carey, 2006). Using Locke’s legacy as an empirical stress test, our study evaluates whether contemporary semantic retrieval can bypass surface-level lexical biases to uncover hidden conceptual reuse within the noisy 18th-century archive.

3 Evaluation Methodology

Our evaluation framework is designed to isolate significant meaning matches from mere lexical reuse. By conceptualizing the retrieval hits as potential reception signals, we aim to uncover their underlying matching distribution and the analytical conditions under which a semantic match constitutes robust historical evidence.

3.1 Heuristic Annotation Taxonomy

To ground the iterative workflow in a stable interpretive framework, we introduce an annotation taxonomy for modelling reception that distinguishes between different types of textual correspondence. The framework is designed as an operational heuristic to facilitate rapid triage rather than imposing a rigid typology. It is a hierarchy of heuristic matches consisting of *Lexical Matches*, *Paraphrase Matches* (*Paraphrases* for short), and *Meaning Matches*, alongside a separate category of *Topical Matches*. *Lexical Matches* capture verbatim reuse, while *Paraphrase Matches* identify cases in which

wording has changed but a clear quotation relationship remains. *Meaning Matches*, by contrast, refer to passages that display semantic engagement suggesting that one text meaningfully interacts with another even without explicit lexical overlap. *Topical Matches*, in turn, refer to passages that address the same subject without semantic evidence of such engagement. For cases of OCR failure or total semantic divergence, we supplement this hierarchy with a *No Match* category, and *Don’t Know* for the cases where the semantic connection is too ambiguous for rapid judgment, requiring extensive close reading to verify. In our evaluation, *Paraphrases* and *Meaning Matches* are aggregated as *significant hits*, representing genuine intellectual reception.

3.2 Iterative Evaluation Strategy

To extract high-signal reception from a noisy retrieval pool, we formalize a multi-stage, human-in-the-loop evaluation framework, illustrated as Fig. 1. We begin by sampling highly reused quotes from Locke’s texts to serve as our search queries. We then deploy a pilot sampling strategy across the broad retrieval pool to understand the basic semantic distribution without presupposing algorithmic success. Finally, we conditionally target our deepening annotation based on these initial results. This progressive annotation process optimizes expert bandwidth, prioritizing genuine reception clusters over algorithmic noise.

3.3 Analytical Framework

To evaluate the pipeline’s utility and its underlying constraints, our framework covers content-level statistical analysis and linguistic-level diagnosis. Firstly, we evaluate the retrieval performance by analyzing how the annotated semantic labels and metadata are distributed across the ranking. Secondly, to interpret the linguistic patterns identified in these results, we first map the annotated English hits into a standard confusion matrix (TP, TN, FP, and FN subsets) based on their ranking intervals and semantic validity. Next, within each quadrant of the matrix, we compute three targeted linguistic features to decode what triggers these specific errors or successes: **Vocab Sim.** (Jaccard similarity of word lemmas) quantifies the model’s dependency on surface-level lexical anchoring, while **OOV** (Out-of-Vocabulary rate on query and target separately) serves as a proxy for historical OCR noise. Additionally, **POS Div.** utilizes Jensen-Shannon divergence over Part-of-Speech distribu-

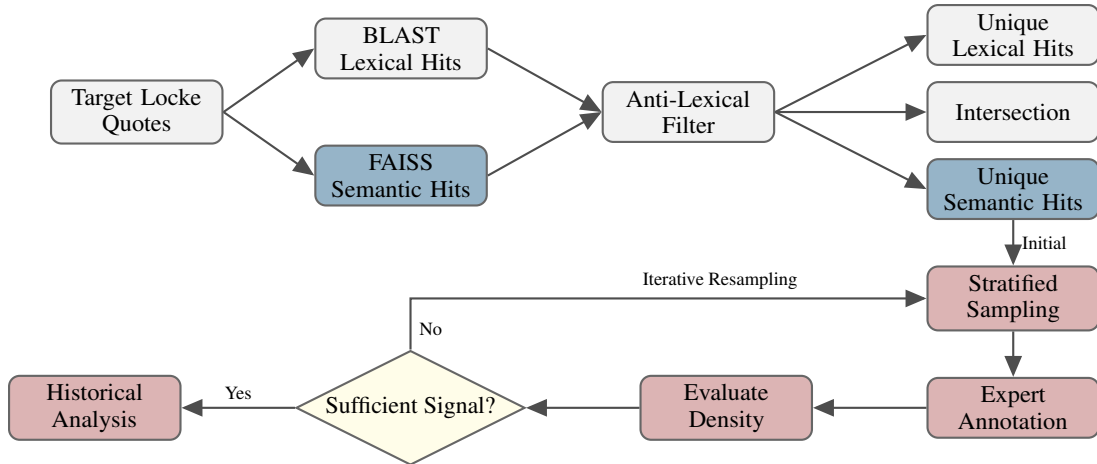


Figure 1: The overall evaluation workflow. The automated search pipeline (top) extracts and filters semantic candidates (blue blocks) against the lexical baseline. The human-in-the-loop workflow (bottom, red blocks) systematically annotates the resulting *Unique Semantic Hits* to evaluate signal density and historical significance.

tions to measure linguistic structural rephrasing independent of vocabulary.

4 Empirical Implementation

This section details the experimental setup and evaluation workflow operated based on our evaluation methodology described in Sec. 3.

4.1 Corpus

Our corpus is the official OCR-extracted text provided by the Eighteenth Century Collections Online (ECCO) (Tolonen et al., 2022), which contains over 32 million book pages from the 18th century. This dataset serves as an ideal testbed for tracing John Locke’s reception, as his influence permeates the period’s diverse genres, from political tracts to educational manuals. The comprehensive corpus undergoes chunking (as detailed in Sec. 4.4) to build the core index for the vectorized search. To manage duplicate editions and ensure bibliographic sufficiency, we integrated metadata from the English Short Title Catalogue (ESTC), specifically the work IDs, authors, publication years of different versions, and genres. The genre labels are assigned based on the Genre Classification Workflow developed by Tiihonen and Hinderks (2025).

4.2 Text Reuse Baseline

As our primary baseline for capturing explicit intellectual reception, we utilize the comprehensive text reuse network generated across the ECCO corpus (Rosson et al., 2023). This infrastructure relies on an adaptation of BLAST (Basic Local Alignment Search Tool). Its fuzzy-matching capabilities inherently tolerate 18th-century orthographic

variations and persistent OCR noise, effectively mapping any text input to verbatim quotations, exact reprints, and lightly edited excerpts across millions of pages. Despite the limitation on semantic-level matching, we repurpose the lexical reuse for three foundational tasks: establishing a comparative baseline, sourcing evaluation queries (Sec. 4.3), and purifying our unique semantic hits (Sec. 4.4).

4.3 Query Source and Construction

In alignment with grounding computational evaluation in 18th-century British intellectual history, we selected John Locke’s *An Abridgment of Mr Locke’s Essay Concerning Human Understanding* (ESTC R22993) as a highly representative, scalable case study. Published prior to the 18th century, this foundational work allows us to track intellectual reception across the entirety of the ECCO timeline. Because its historical significance is universally recognized and it possesses a dense, well-documented network of existing lexical reuses, it provides an ideal, undisputed baseline without the need to overstate the novelty of the source itself.

To capture genuine external lexical reuse rather than isolated fragments, we queried the baseline for reuse text clusters originating from this work, as introduced by Rosson et al. (2023). In this pipeline, we formally define a *quote* not as a punctuated citation, but as a contiguous, verifiable fragment of reused text. We isolated these quotes by enforcing a length constraint of 150–300 characters, and a minimum external reuse frequency ($n \geq 3$)¹.

¹These thresholds were empirically determined to eliminate uninterpretable short snippets and overly long, multi-page sequences that obscure targeted semantic analysis.

4.4 Minimal Semantic Search Pipeline

We adopt a deliberately minimal dense retrieval-based search pipeline to evaluate the conceptual semantic capabilities of semantic search. The implementation of core infrastructure is publicly accessible via the semantic-sim repository². The underlying vector index was built upon a broader Early Modern English corpus (encompassing both 18th-century ECCO and pre-18th-century EEBO texts). We can easily isolate the target ECCO corpus by filtering the retrieved outputs via document IDs. The texts were segmented into 100-token chunks and encoded using the paraphrase-multilingual-mpnet-base-v2 model (Reimers and Gurevych, 2020; Song et al., 2020). This encoder natively supports over 50 languages, including English, French, Italian, and Spanish, while allowing us to evaluate its zero-shot capability on unsupported historical languages such as Latin. Efficient vector indexing and retrieval were executed via FAISS (Facebook AI Similarity Search) (Douze et al., 2025). For each query, this index outputs a ranked list of retrieved segments with their semantic similarity scores to the input. We referenced these ranked outputs against our BLAST lexical baseline. By computing the set difference between the two systems, we filtered out known verbatim overlap.

4.5 Annotation Setup

We deployed a project in the locally hosted instance of Label Studio (Tkachenko et al., 2020-2025) for the manual annotation. As illustrated in Fig. 2, the interface was designed to prevent annotators from evaluating fragments in a vacuum. Alongside the retrieved chunk and Locke’s original quote, the UI simultaneously displayed the corresponding primary author, work title, and publication year, and provided direct access to the broader document context. The annotation was conducted by five domain experts from the Computational History Group (COMHIS) at University of Helsinki. Each candidate was assigned to a single annotator for rapid triage. Ambiguous cases were marked as *Don’t Know* and processed independently.

4.6 Annotation Process

We implemented a workflow following our methodology in Sec. 3 and all the setup above to deploy our iterative evaluation strategy. The empirical out-

²<https://github.com/TurkuNLP/semantic-sim>

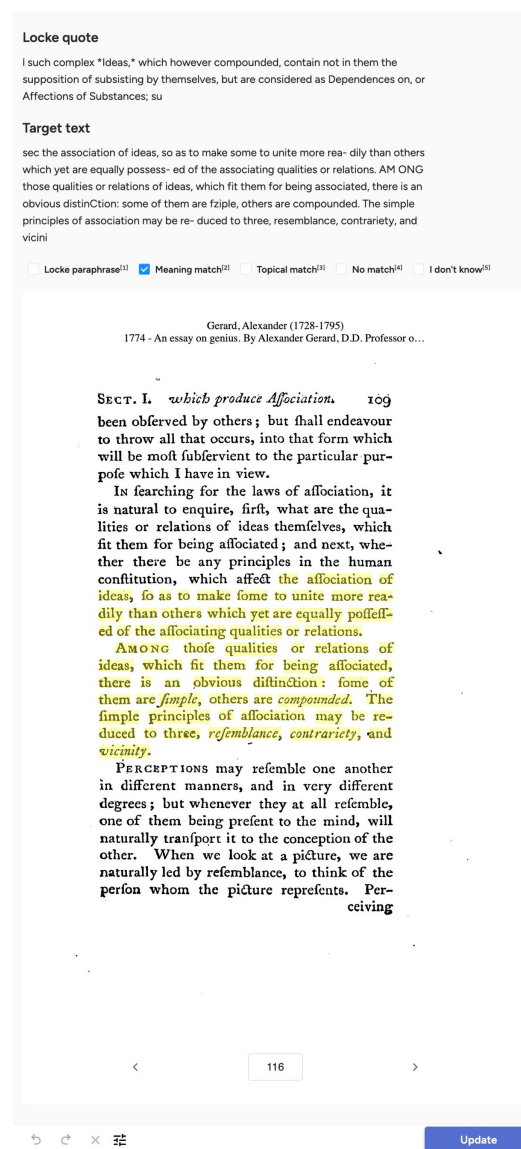


Figure 2: The annotation interface screenshot.

comes and the evidence for iterative deepening are detailed in Sec. 5.

Source Selection From the top 1,000 most frequently reused segments extracted in Sec. 4.3, we selected the top-5 highest-frequency quotes, supplemented by 5 randomly sampled from each of three subsequent frequency tiers (ranks 6-50, 51-150, and 151-1,000), yielding 20 search queries.

Search and Filtering For these 20 queries, we retrieved the top 200,000 semantic hits using our pipeline established in Sec. 4.4. Filtering by document ID retained an average of 173,266 ECCO hits per query (86.6%). To prevent the search space from being dominated by intra-document repetition or frequent reprints, we deduplicated the results by ESTC work ID. Retaining only the highest-scoring hit per distinct work yielded an average of 13,391

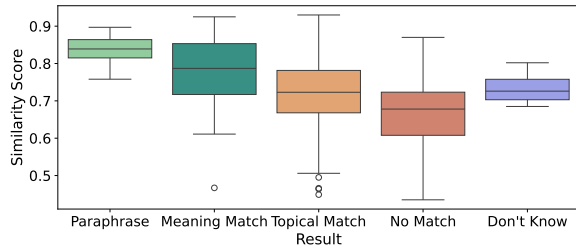


Figure 3: Similarity scores across categories, indicating a lack of global calibration.

unique candidates per query. We then filtered out known verbatim overlap.

Stratified Pilot Sampling We conducted a probing manual annotation of 50 hits per quote, sampling the top 5 results and uniformly distributed intervals across the top 90% of the filtered hits. They were distributed among five domain experts and annotated according to the taxonomy in Sec. 3.1.

Iterative Deepening We annotated more top hits for specific quotes according to a dual criterion: a high density of semantic matches and substantive historical significance. We set the top 200 as an empirical annotation upper bound. Before the exhaustive annotation, we annotated 50 samples within the top-200 hits for a quick triage to verify that the promising yields from our initial broad but sparse sampling were not merely algorithmic outliers. They were the top 20 hits, plus 30 hits sampled iteratively (one every six) from the remainder. By confirming a consistently dense semantic signal first, we mitigated the risk of exhaustively annotating a noise-heavy set that would force us to restart the deepening process. Ultimately, candidates maintaining robust yields triggered the exhaustive annotation of their complete top-200 results for deep-dive historical analysis.

5 Results and Analysis

Our analysis assesses the reliability and technical limitations of the semantic search pipeline.

5.1 Overall Results

Before detailing the annotation outcomes, we quantify the pipeline’s lexical filtering efficacy. Semantic search achieved a 95.5% recall of BLAST exact matches (averaging 103.9 intersection vs. 4.9 unique BLAST hits), confirming it effectively matched OCR-resilient lexical baselines even before capturing semantic variance.

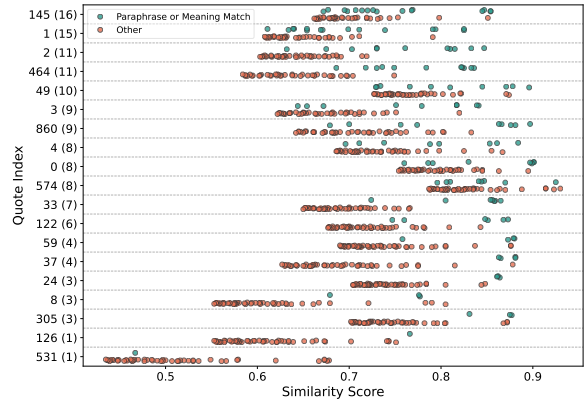


Figure 4: Hit distribution per quote, highlighting query-specific heterogeneity. Queries are sorted by the absolute count of significant hits, indicated in parentheses.

On this highly refined dataset, our heuristic taxonomy facilitated an efficient rapid-triage evaluation, logging a median annotation time of approximately 19.3 seconds per candidate. The resulting annotations demonstrate that the model possesses a robust ordinal capability. As detailed in Tab. 1, the ranking validity, measured by Spearman’s rank correlation (ρ), exhibits a highly significant monotonic relationship with human judgments. The most informative semantic categories (*Paraphrase* and *Meaning Match*) show strong negative correlations with the hit ranks, confirming that the system surfaces substantive semantically matched fragments to the top of the ranking. This efficacy is most pronounced at the very top: the top-5 ranked hits are predominantly valid semantic matches, with zero of *No Match* across 15 of the 20 quotes.

While simply extracting the top- K results suffices for standard ad-hoc retrieval, corpus-level intellectual history requires the complete boundaries of a complete semantic match cluster to quantify macro-level reception patterns. Our evaluation reveals that automating this boundary-drawing is fundamentally bottlenecked by score calibration. As shown in Fig. 3, while relative score ranking is correlated to semantic categories, absolute similarity scores overlap significantly across them, preventing the use of a universal global threshold.

This uncalibrated nature extends deeply to the query level. Fig. 4 demonstrates pronounced query-specific heterogeneity: the score thresholds for significant hits shift drastically depending on the specific query content. More critically, even for individual quotes, these local boundaries remain highly fuzzy, preventing the extraction of a cleanly separated match pool for each quote. While this noise partially reflects the fundamental ambiguity

Category	Overall (%)	In Top-5 Ranked (%)	In Top 20% (%)	In Top 50% (%)	Spearman’s ρ
Paraphrase	2.1	18.0	7.0	3.5	-0.224***
Meaning Match	11.6	44.0	26.3	16.3	-0.294***
Topical Match	31.1	24.0	34.0	32.7	-0.054
No Match	54.8	14.0	32.3	47.0	0.307***
Don’t Know	0.4	0	0.3	0.5	-0.026

Table 1: Distribution and ranking validity of semantic categories. Overall, Top-5 Ranked, Top 20%, and Top 50% yields are reported as mean percentages across the sampled hits, calculated based on original retrieval rankings. For each category, Spearman’s rank correlation (ρ) is calculated between its binary indicator and the local retrieval rank within the sampled subset of each query. Significance levels: *** $p < 0.001$.

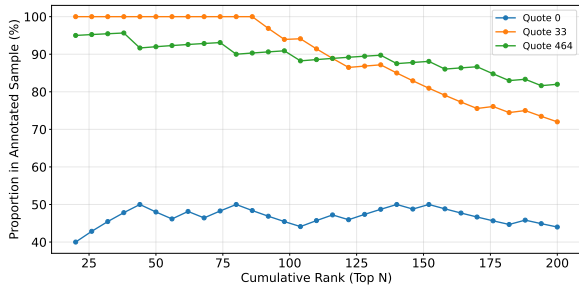


Figure 5: Proportion of significant hits within the 50 validation samples across the top 200 hits for each quote.

of reception research, where the line between a *Meaning Match* and a mere *Topical Match* is often epistemologically contested, it makes individual thresholds hard to set. This failure of both global and local calibration necessitates a methodological pivot. To determine whether semantic search can still provide inspiring insights for intellectual history, we shift from thresholding strategies to targeted case studies. We prioritize specific quotes with clearer empirical score boundaries to assess whether their high-ranking hits yield substantive meaning matches with historical significance.

5.2 Individual Results

To operationalize our iterative deepening (Section 3), we finalized three core queries from an initial mixed pool by balancing algorithmic clarity with historical significance. Quotes 33 and 464 were selected as primary anchors because of their clear semantic boundaries at high similarity scores (> 0.8) alongside profound historical relevance. Conversely, Quote 0, supplemented by domain experts due to its massive lexical footprint, was retained as a comparative baseline. Contrasting with the clear signals of the primary anchors, Quote 0 serves as a rigorous sanity check to verify whether its top-ranked hits reproduce the boundary fuzziness observed during the broad initial sampling.

Before committing to the labor-intensive exhaustive annotation of their top hits, we utilized our

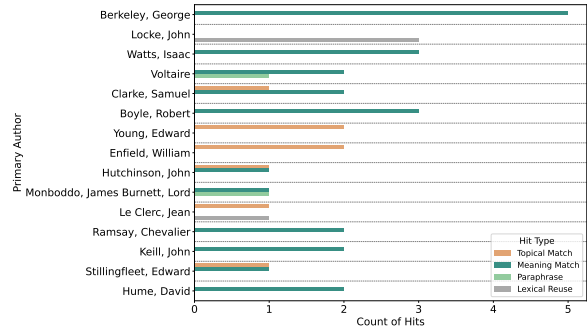


Figure 6: Top 15 authors of the top hits for Quote 464.

50-sample validation set in top-200 hits as a rapid validation mechanism. Fig. 5 tracks the proportion of significant hits across this sampling continuum. While the proportion for Quote 0 keeps low in top hits, confirming its boundary fuzziness, Quotes 33 and 464 maintain robust yields above 70% and 80%, respectively. This sustained density verifies that these top hits represent deeply cohesive semantic clusters rather than anomalous spikes.

Following the curation of the expert-validated dataset with 400 hits from the two core quotes, we conducted a structural analysis of their historical distribution. A fundamental finding is the sheer volume disparity: significant semantic reuse (*Paraphrase* and *Meaning Match*) outnumbers verbatim lexical reuse by over an order of magnitude (313 vs. 22 hits at the work level), effectively broadening the observable horizon of reception studies. This expansion is starkly visible at the authorial level. Taking Quote 464 as a representative example (Fig. 6), traditional lexical matching predominantly isolates citations by Locke; in contrast, semantic search uncovers a dense, previously invisible network of engagement from central intellectual figures, including Berkeley, Boyle, Hume, and Voltaire.

Projecting these results onto temporal and disciplinary axes (Fig. 7) reveals that the distribution of reuse types varies heavily by genre. For instance, semantic reuse is significantly more prevalent in *Philosophy* and *Religion* than verbatim quotation.

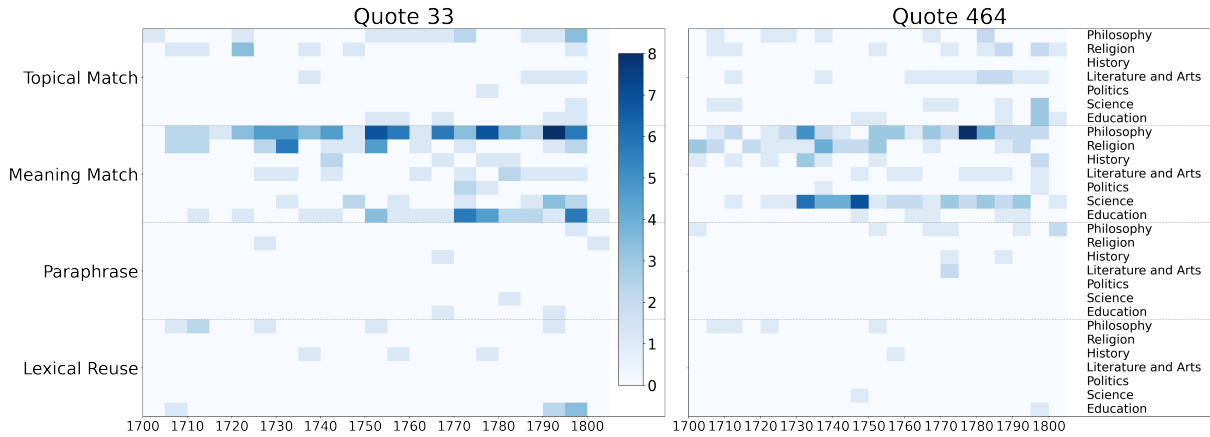


Figure 7: Temporal and disciplinary count distribution of retrieved hits for Quotes 33 (left) and 464 (right). These heatmaps visualize the hits in semantic categories across the seven most dominant 18th-century book genres.

ID	Original Quote Fragment for Query
33	I such complex <i>*Ideas,*</i> which however compounded, contain not in them the supposition of subsisting by themselves, but are considered as Dependences on, or Affections of Substances; su
464	ing their Causes, or any otherwise. The <i>*Atomists,*</i> who define Motion to be <i>*a passage from one place to another,*</i> What do they more than put one synonymous Word for another? For what is <i>*Passage*</i> other than <i>*Motion</i>

Table 2: Representative quotes for the qualitative study.

Temporal analysis further reveals distinct trajectories, with different hit types peaking at different decades, suggesting evolving modes of engagement over time. Furthermore, the two analyzed quotes exhibit highly divergent reception profiles: one forms massive, tightly-knit clusters within philosophical and educational discourse, while the other displays a pronounced volume of reuses across the sciences. This macro-level analysis demonstrates the search system’s utility in capturing structural intellectual shifts invisible to traditional lexical detection. How these divergent profiles manifest in actual rhetorical repurposing will be the focus of our qualitative case studies.

5.3 Case Study

To interpret these macro-patterns, we examined the intersection of genre and publication year through targeted close reading. Tab. 2 presents the original texts for these cases. Focusing on Quote 464, from the right sub-figure of Fig. 7, we observe a distinct peak in reuses within the science category during the 1740s. Qualitative analysis reveals that these hits largely originate from scientific textbooks and manuals. This phenomenon appears closely linked to the consolidation of Newtonianism and the con-

current rebuttal of Cartesian physics, where Lockean epistemological framework concerning motion was repurposed to ground scientific empiricism.

In contrast to this polemical weaponization of Quote 464, Quote 33 illustrates an entirely different trajectory of conceptual drift: pedagogical formalization. As indicated by its dense clustering within the philosophy and education genres, Locke’s complex epistemological framework was systematically adapted into instructional manuals. Rather than engaging with the philosophical nuances, these texts typically simplified and repackaged his core logic to serve didactic purposes.

Crucially, across both divergent trajectories, the 18th-century authors rarely quoted the original sources verbatim. Whether paraphrasing physics for scientific pedagogy or conceptually flattening Locke for educational manuals, this “implicit reception” is successfully captured by semantic search. By retrieving these highly reformulated iterations, the semantic search demonstrates its applicability in tracking ideas as they cross disciplinary boundaries and registers of complexity, uncovering context-dependent intellectual shifts that traditional lexical searches would inevitably overlook.

5.4 Linguistic Diagnostics

While individual analyses and case studies validate the pipeline’s potential, our linguistics-driven analysis reveals the constraints and properties of this success, based on the method proposed in Sec 3.3.

Starting at the macro scale, language identification via Lingua (Stahl, 2021) reveals that the semantic search demonstrates a pronounced capacity for cross-lingual alignment, as detailed in Tab. 3. When compared against the token-weighted corpus

Category (%)	English	French	Latin	Italian	Spanish
ECCO Tokens	92.2	3.3	3.4	0.4	0.1
Paraphrase	100.0	0.0	0.0	0.0	0.0
Meaning Match	90.5	9.5	0.0	0.0	0.0
Topical Match	91.0	7.4	0.0	1.3	0.3
No Match	92.7	6.2	0.9	0.2	0.0
Don't Know	100.0	0.0	0.0	0.0	0.0

Table 3: Language distribution across semantic categories in the 1000 pilot samples. The ECCO token language distribution baseline is calculated via token-level aggregation from the ECCO metadata.

baseline, the highly relevant “Meaning Match” category exhibits a nearly three-fold enrichment of French texts (9.5%). This non-random distribution across semantic tiers confirms that the multilingual embedding space effectively maps English philosophical propositions onto Continental discourse, even under noisy OCR’ed text, bypassing the need for explicit literal translation. Conversely, the model fails to achieve zero-shot semantic transfer for unsupported historical languages like Latin (3.4% of the corpus), which appears exclusively as noise within the *No Match* category (0.9%).

However, this cross-lingual reach introduces a specific structural side effect. A close inspection of the retrieval outcomes reveals that the top-ranked FP contains a noticeably higher proportion of non-English hits (14.0%) compared to the TP (10.0%). This inflation indicates that the model occasionally exhibits a cross-lingual over-eagerness, elevating non-English texts based on coarse-grained topical proximity rather than strict conceptual equivalence, thereby introducing systemic noise at the head of the search results.

To isolate the underlying model mechanics from this cross-lingual noise, evaluating the purely English subsets exposes rigid monolingual bias, processed using spaCy’s large English model (Honni-bal et al., 2020). As illustrated by the linguistic diagnostics in Tab. 4, TPs inherently rely on a higher degree of lexical overlap with the original query, even though the proportion is very small, with a maximum of 0.030. This is further evidenced by other subsets, which crowd the results due to their higher lexical similarity compared to the marginalized matches. In contrast, FNs display near-zero lexical overlap, revealing the existing limitation on retrieving highly non-lexical semantic matches. Because the FNs actually exhibit the lowest Hit OOV and the most similar syntactic structure despite

Subset	N	Vocab Sim.	Quote OOV	Hit OOV	POS Div.
Top P	40	0.030 ± 0.026	4.11 ± 3.31	19.25 ± 7.88	0.348 ± 0.048
Top N	121	0.017 ± 0.020	1.73 ± 2.45	20.62 ± 7.62	0.362 ± 0.061
Tail N	197	0.012 ± 0.017	2.17 ± 2.57	21.23 ± 7.23	0.359 ± 0.064
Tail P	10	0.009 ± 0.019	3.44 ± 3.88	17.46 ± 2.62	0.331 ± 0.045

Table 4: Linguistic feature distributions across English hit subsets. The rows correspond to defined TP: highly ranked (in top 30%) significant hits (*Paraphrase* or *Meaning Match*), FP: highly ranked (in top 30%) non-matches, TN: no-matches in the tail (top 60%–90%), and FN accordingly. N represents the size of the subset.

consistently low query-side noise, this disparity confirms a critical paradox: while the model is remarkably robust against OCR degradation and syntactic drift, its ranking remains anchored to surface-level lexical similarity, systematically marginalizing structurally aligned historical concepts, regardless of their superior textual integrity.

6 Conclusion

This study demonstrates that a human-in-the-loop semantic search workflow can surpass the limitations of traditional lexical matching, uncovering a significant volume of implicit reception within noisy historical corpora. By offering both a methodological proof-of-concept and an expert-curated reference dataset, our purpose is to provide feasibility evidence and a reusable evaluation framework. We plan to scale it into a systematic benchmark to evaluate dense retrieval models across broader historical and low-resource domains.

Acknowledgments

This research was funded by the European Union under the Horizon Europe Marie Skłodowska-Curie Actions (Grant Agreement No. 101119511) and the Academy of Finland (Grant Nos. 347706 and 347709). Computational resources were provided by CSC – IT Center for Science. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or Horizon Europe MSCA Actions. Neither the European Union nor the granting authority can be held responsible for them.

We extend our gratitude to Kira Hinderks, Iiro Tiihonen, Yann Ryan, and David Rosson for their extensive support and feedback during the data annotation phase, alongside the broader Computational History Group (COMHIS) network for their insightful discussions.

References

- Motasem Alrahabi and Tom Wainstain. 2025. Versus: An automatic text comparison tool for the digital humanities. In *Proceedings of the First on Natural Language Processing and Language Models for Digital Humanities*, pages 32–37, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- David Armitage and Jo Guldi. 2014. *The History Manifesto*. Cambridge University Press.
- Emmanuelle Bermès. 2017. Text, data and link-mining in digital libraries: Looking for the heritage gold. In *IFLA Satellite Meeting - Digital Humanities – Opportunities and Risks: Connecting Libraries and Research*.
- David M. Blei. 2012. Probabilistic topic models. *Commun. ACM*, 55(4):77–84.
- Katherine Bode. 2018. *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press, Ann Arbor.
- Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. 2020. Corpus REDEWIEDERGABE. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 803–812, Marseille, France. European Language Resources Association.
- Simon Burrows and Mark Curran. 2012. The French Book Trade in Enlightenment Europe Project and the STN Database. *Journal of Digital Humanities*, 1(3).
- Daniel Carey. 2006. *Locke, Shaftesbury, and Hutcheson: Contesting Diversity in the Enlightenment and Beyond*. Ideas in Context. Cambridge University Press, Cambridge.
- Jeffrey R. Collins. 2020. *In the Shadow of Leviathan: John Locke and the Politics of Conscience*. Cambridge University Press. Google-Books-ID: 1tLKD-wAAQBAJ.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The FAISS Library. *IEEE Transactions on Big Data*, pages 1–17.
- Marten Düring, Matteo Romanello, Maud Ehrmann, Kaspar Beelen, Daniele Guido, Brecht Deseure, Estelle Bunout, Jana Keck, and Petros Apostolopoulos. 2023. Impressed Text Reuse at Scale. An interface for the exploration of text reuse data in semantically enriched historical newspapers. *Frontiers in Big Data*, 6.
- Dan Edelstein. 2016. Intellectual History and Digital Humanities. *Modern Intellectual History*, 13(1):237–246.
- Dan Edelstein, Paula Findlen, Giovanna Ceserani, Caroline Winterer, and Nicole Coleman. 2017. Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project. *The American Historical Review*, 122(2):400–424.
- Susan Fitzmaurice, Justyna A. Robinson, Marc Alexander, Iona C. Hine, Seth Mehl, and Fraser Dallachy. 2017. Linguistic DNA: Investigating Conceptual Change in Early Modern English Discourse. *Studia Neophilologica*, 89(sup1):21–38.
- Brevin Franklin, Emily Silcock, Abhishek Arora, Tom Bryan, and Melissa Dell. 2024. News Deja Vu: Connecting Past and Present with Semantic Search. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 99–112, Mexico City, Mexico. Association for Computational Linguistics.
- Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.
- Andrew Goldstone and Ted Underwood. 2014. The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us. *New Literary History*, 45(3):359–384.
- Tyler Hanck. 2019. *Locke’s Confusion About the Confused Idea of Substance*. Thesis, University of Illinois Chicago.
- James A. Harris. 2023. Of the origin of government: The afterlives of Locke and Filmer in an eighteenth-century British debate. *Intellectual History Review*, 33(1):33–55.
- 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 *Proceedings of the Digital Humanities in the Nordic Countries*, Copenhagen, Denmark.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.
- Jenna Kanerva, Hanna Kitti, Li-Hsin Chang, Teemu Vahtola, Mathias Creutz, and Filip Ginter. 2025. Semantic search as extractive paraphrase span detection. *Language Resources and Evaluation*, 59(1):257–276.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Pierre-Carl Langlais. 2021. Fictions littéraires de Gallica / Literary fictions of Gallica.

- Daniel Layman. 2021. Locke’s Republican and Liberal Legacy. In *The Lockean Mind*. Routledge. Num Pages: 10.
- Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d’Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the Gap: Assessing Temporal Generalization in Neural Language Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 29348–29363. Curran Associates, Inc.
- H. P. Luhn. 1960. [Key word-in-context index for technical literature \(kwic index\)](#). *American Documentation*, 11(4):288–295.
- Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. 2022. [ABNIRML: Analyzing the Behavior of Neural IR Models](#). *Transactions of the Association for Computational Linguistics*, 10:224–239.
- Ananth Mahadevan, Michael Mathioudakis, Eetu Mäkelä, and Mikko Tolonen. 2025. [Text reuse in large historical corpora: Insights from the optimization of a data science system](#). *International Journal of Data Science and Analytics*, 20(5):4631–4643.
- Barbara McGillivray, Federico Nanni, and Kaspar Beelen. 2024. 10. Why Does Digital History Need Diachronic Semantic Search? In *Computational Humanities*. University of Minnesota Press.
- Andrianos Michail, Juri Opitz, Yining Wang, Robin Meister, Rico Sennrich, and Simon Clematide. 2025. [Cheap Character Noise for OCR-Robust Multilingual Embeddings](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11705–11716, Vienna, Austria. Association for Computational Linguistics.
- 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.
- Eetu Mäkelä, James Misson, Devani Singh, and Mikko Tolonen. 2025. [Opening the Black Box of EEBO](#). *Digital Scholarship in the Humanities*, page fqaf086.
- R. Porter. 2001. *Enlightenment: Britain and the Creation of the Modern World*. Penguin Books Limited.
- 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. 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 (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.
- Hammad Rizwan, Domenic Rosati, Ga Wu, and Hassan Sajjad. 2025. Resolving Lexical Bias in Model Editing. In *Forty-Second International Conference on Machine Learning*.
- Glenn Roe. 2024. [Text reuse as cultural practice: Intertextuality in the 18th-century digital archive](#). *Digital Enlightenment Studies*, 2(1).
- David Rosson, Eetu Mäkelä, Ville Vaara, Ananth Mahadevan, Yann Ryan, and Mikko Tolonen. 2023. [Reception Reader: Exploring Text Reuse in Early Modern British Publications](#). *Journal of Open Humanities Data*, 9:5. ArXiv:2302.04084 [cs].
- Julian Schelb, Michael Wittweiler, Marie Revellio, Barbara Feichtinger, and Andreas Spitz. 2026. [Loci Similes: A Benchmark for Extracting Intertextualities in Latin Literature](#). *Preprint*, arXiv:2601.07533.
- Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. [Simple Entity-Centric Questions Challenge Dense Retrievers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- David A. Smith, Ryan Cordel, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson. 2014. [Detecting and modeling local text reuse](#). In *IEEE/ACM Joint Conference on Digital Libraries*, pages 183–192.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, pages 16857–16867, Red Hook, NY, USA. Curran Associates Inc.
- Peter M. Stahl. 2021. Lingua: The most accurate natural language detection library for Python. <https://github.com/pemistahl/lingua-py>. Python bindings for the Lingua language detection library.
- Timothy Stanton. 2018. [John Locke and the Fable of Liberalism](#). *The Historical Journal*, 61(3):597–622.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Iiro Tiihonen and Kira Hinderks. 2025. [Genre Classification Workflow for the English Short Title Catalogue \(ESTC\)](#). *Transformations: A DARIAH Journal, Workflows(Metadata-based workflows)*:14754.

- Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020–2025. [Label Studio: Data labeling software](https://github.com/HumanSignal/label-studio). Open source software available from <https://github.com/HumanSignal/label-studio>.
- Mikko Tolonen, Eetu Mäkelä, and Leo Lahti. 2022. [The Anatomy of Eighteenth Century Collections Online \(ECCO\)](#). *Eighteenth-Century Studies*, 56(1):95–123. Publisher: Johns Hopkins University Press.
- Mikko Tolonen and Yann Ciarán Ryan. 2026. *Computational Methods in Intellectual History*, pages 239–260. Proceedings of the British Academy. Liverpool University Press, United Kingdom.
- Mikko Tolonen and Mark G. Spencer. 2025. [The Reception of David Hume’s Essays in Eighteenth-Century Britain](#). In Max Skjönsberg and Felix Waldmann, editors, *Hume’s Essays*, Cambridge Critical Guides, pages 15–35. Cambridge University Press, Cambridge.
- Aleksi Vesanto, Filip Ginter, Hannu Salmi, Asko Niivala, and Tapio Salakoski. 2017. A System for Identifying and Exploring Text Repetition in Large Historical Document Corpora. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 330–333, Gothenburg, Sweden. Association for Computational Linguistics.
- Christopher N. Warren, Daniel Shore, Jessica Otis, Lawrence Wang, Mike Finegold, and Cosma Shalizi. 2016. Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks. *DHQ: Digital Humanities Quarterly*, 10(3):1.
- Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. 2025. OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17443–17453.