

The Algorithmic Challenges and Opportunities in Tracing Meaning at Scale in Large Historical Document Collections

Filip Ginter
0000-0002-5484-6103

TurkuNLP
Department of Computing
University of Turku
Email: fgint@utu.fi

Mikko Tolonen
0000-0003-2892-8911

COMHIS
Department of Digital Humanities
University of Helsinki
Email: mikko.tolonen@helsinki.fi

Anna Plassart
0000-0002-3441-2332

History department
The Open University
Email: anna.plassart@open.ac.uk

Jenna Kanerva
0000-0003-4580-5366

TurkuNLP
Department of Computing
University of Turku
Email: jmnbyl@utu.fi

Kira Hinderks
0009-0004-9233-9315

COMHIS
Department of Digital Humanities
University of Helsinki
Email: kira.hinderks@helsinki.fi

Yu Wu
0009-0006-1634-5566

COMHIS
Department of Digital Humanities
University of Helsinki
Email: yu.wu@helsinki.fi

Abstract—The capabilities of modern language models are expanding the questions that computational history can pose. In this position paper we argue for a research program organized as a progression of increasingly semantic forms of text reuse, moving from surface-level lexical matching, through cross-lingual translation mining, toward the indirect forms of engagement through which ideas are actually transformed. We briefly review the state of the art at each step of this progression, and identify the open frontier, the least direct kinds of engagement. We argue that recognizing such indirect text reuse requires encoder models that move beyond paraphrase, and that historians should contribute to the development of such models. The research program is a case study in fruitful two-directional collaboration: AI enables new history research, whose demands in turn can shape new AI.

I. INTRODUCTION

AS the capabilities of large language models increase, so does the range of research questions that computational methods can address. Tasks that were until recently impossible have been addressed with the combination of state of the art AI methodology and high performance computing (HPC). Taking this shift as our premise, in this position paper we outline a research program for computational intellectual history that the current generation of language models makes feasible.

The program is organized around *text reuse*, treated not as a single recognition task but as a graded phenomenon. At one extreme is surface reuse: a passage copied, reprinted, or quoted nearly verbatim. At the other is the form of engagement that is central to intellectual history, in which an author engages with an argument, restates it in a different vocabulary or another language, or argues against it, often without explicitly naming its source. These extremes are connected by a continuum of

types of text reuse, each requiring a distinct computational treatment. We structure the program along this continuum, in order of increasing semantic distance: reuse, translation, meaning, reception, and interpretation.

II. FORMS OF TEXT REUSE IN HISTORICAL LITERATURE

We arrange the forms of text reuse along a single axis ranging from lexical overlap to semantic similarity (Figure 1). At the lexical end, two passages are related if they share surface form, that is, the same words. At the semantic end, two passages are related if they express or engage with the same historically situated idea, possibly with no shared vocabulary. By “idea” we mean not an abstract philosophical concept, but a claim or argument articulated in a particular textual form.

The first category concerns artefacts of the text itself, such as OCR errors and imprint information, which characterize how a book was digitized rather than its content. It is followed by reprints, ranging from partial reprints to new editions, and by quotations and extensive reuse, which span near-verbatim borrowing, exact quotation, and second-hand quotation drawn from an intermediary. Computationally, these categories remain a matter of identifying shared character sequences. The right half of the axis is where reuse cannot be reduced to character matching. Allusion references ideas, themes, or arguments without directly reproducing them. Paraphrasing restates a claim in new words, from light paraphrase to thorough reformulation. Translation mining covers cross-lingual reuse, including full and partial translation as well as cross-lingual paraphrase. Meaning match denotes shared ideas with no textual overlap, and commentary adds discussion, evaluation, or critique to a borrowed idea.

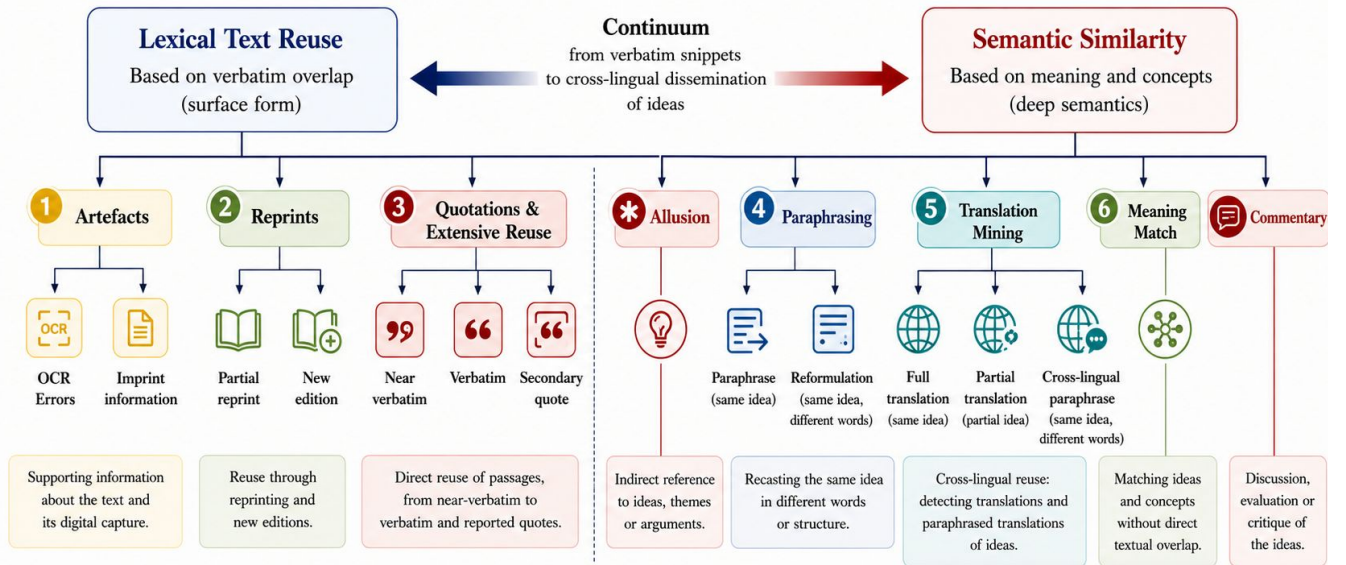


Fig. 1. Conceptualizing text reuse and semantic similarity.

The consequence is methodological. The further on the spectrum a phenomenon lies, the more the analysis must operate on models of meaning. The remainder of this paper proceeds along this axis and reports the state of the methods at each step.

III. OPERATIONALIZING THE PROGRAM

This section summarizes both our own work and that of others at the successive steps of the program, and indicates where the current frontier lies.

A. Lexical reuse

At the lexical end the problems are well understood, and dominated by three challenges. The first is OCR noise: large historical collections are produced by scanning and automatic transcription, and the resulting text contains levels of noise that defeat naive string matching. The second is scale, as the collections of interest contain hundreds of thousands of works and billions of words. The third is that the unit of reuse is not known in advance. A borrowing may be a sentence, a paragraph, or a loosely bounded passage, and a method must recover both the existence of a match and its extent.

Two families of methods address these problems. The first adapts sequence-search techniques from bioinformatics, which were developed specifically for fuzzy matching in very large character-sequence databases. Because even distant overlap is meaningful in bioinformatics applications, these tools can operate even under severe OCR noise. Our work builds on the BLAST tool [1], [2], which can extract matches such as the pair illustrated in Figure 2. The second family relies on n-gram overlap and shingling, either character- or word-based. Less resilient to extreme OCR noise but computationally lighter than BLAST, such methods have featured in recent work on

text reuse [3] and are exemplified by the passim system [4] and the TextPAIR tool [5]. Both families produce the same output: a network of passages that recur across a collection.

cru- i.,i,lil and malfilly to ligiht mndcr C;, balncr ;as;liit't
fin, the worHd, :,i(tli (tcí; ;il a tio contliiie CihrlC'
tiilif'ul ifolli't-'aind flrvant iite(/,ii lilc's encl. Anlleil.
T7jeJZa/nl/ t/h P/ri/?/IJ9', iC J-eing now, dearly hbel ed
bhrc- i. i hrci, that this (Chi/i! by Balp- ;1li re.negncirate,
;land grafted into hile l:;,dy of Chritlt's Chillreh, let iis ix
e thanks unto Alnight y (,odl :or thcle bellneits

crucified, and mlanfully to fight undler hi; banne! aigm nit
fin, the worldJ, and the devil ; an-d to continuoe Chrift's
fa ithfal foldier and ferviant unto his life's end., Amlen.
Theni jh all the Pries3 fay, S E EiIN G now. dearly
belovedi breihren, ththis hi is by Baptifmrigneaeen rftdit
the bodly of Chrifi s Chuirch, let us give tha nks unto
Almhighly God fat theiie benefits

Fig. 2. Text reuse identified among 33M pages of text by the BLAST approach despite severe OCR noise.

A mature, historian-facing tool arising from the work on large-scale text reuse is the Reception Reader [7]. It is built on the text reuse network computed across Early English Books Online and Eighteenth Century Collections Online, which covers on the order of 200K early modern works comprising 33M pages, and it supports three functions: a visual overview of the reception of a chosen work over time through its incoming connections, an interactive survey of the connected documents, and a close reading view that presents each reused passage in its original context. The tool has been applied in substantive historical studies, including

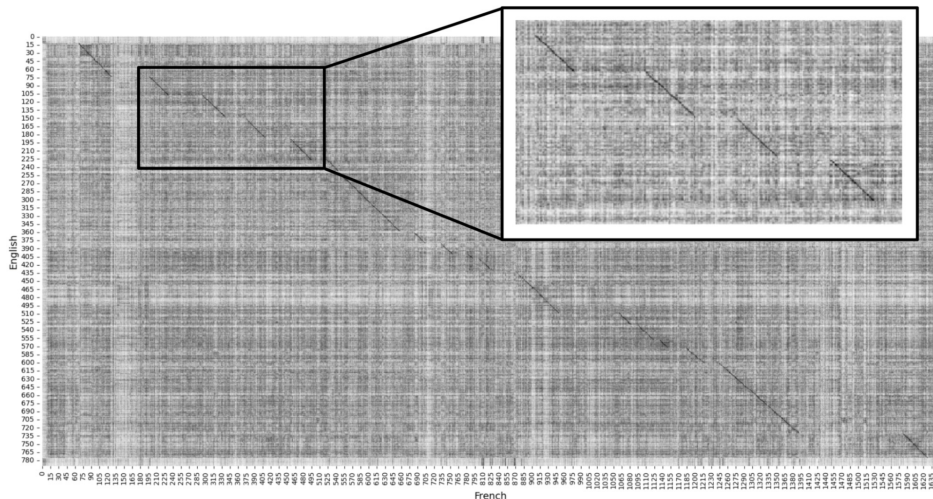


Fig. 3. Segment-by-segment similarity matrix for a pair of works in a translation relation. The diagonal trace marks segments that are mutually more similar than average and occur in the same order, the signal the translation mining pipeline detects. Figure: [6]

a reassessment of the reception of Hume’s *Essays* and a comparative analysis of Mandeville’s works bearing on long-standing questions of authorship attribution [8], [9]. For direct reuse, the computational problem is therefore well addressed, and the open questions concern interpretation and interface.

OCR quality remains a constraint on any work with digitized historical texts, and warrants focused treatment. Given the field’s rapid methodological progress, the obvious solution would be to re-OCR with the latest tools. But because the collections comprise tens of millions of pages, and high-resolution scans are often not openly available, this is not always possible. In recent work we applied open-weight large language models to post-correct OCR output on historical English and Finnish material [10]. The corrected English text improved on every measure considered except the preservation of period spelling, whereas for Finnish no practically useful level of correction was achieved. These results suggest that data quality can be improved, especially for search and retrieval, without re-OCRing.

B. Modeling meaning

Lexical methods are valuable but impose a strict limit on what can be traced in intellectual history, because engagement with an idea is only rarely expressed as direct quotation. More often, the idea is paraphrased, summarised, simplified, embedded in a wider argument, or rebutted by an opponent, none of which provides enough signal for a string-matching method to register the connection.

Two challenges arise on this side of the axis. The first is paraphrase, in which a claim is preserved under a complete change of wording. The second, which breaks lexical methods most decisively, is language. When an idea crosses a linguistic boundary it is translated, and a translation shares essentially no characters with its source, so methods based on shared strings do not apply. This motivates the use of cross-lingual

encoder models, such as the Sentence Transformer [11]. An encoder maps a passage to a vector such that passages of similar meaning are close in the vector space, independently of language, so that similarity becomes a geometric rather than a lexical quantity.

Our translation mining work demonstrates how far this approach can be taken with appropriate engineering [6]. Comparing the English works in ECCO against the French works in Gallica, roughly 206,000 and 287,000 works respectively, amounts to hundreds of millions of text segments per side and on the order of 10^{17} segment-level comparisons. We avoid this exhaustive computation with the FAISS vector search library [12], which prunes most of the search space. What distinguishes translation from chance is geometric: for two works in a translation relation, the matrix of segment-by-segment similarities contains a diagonal trace, the same content expressed in the same order (Figure 3). We detect this trace using the progressive Hough transform [13]. The arrangement of the diagonals is itself informative, as different types of translation, from faithful to partial and rearranged, produce distinct patterns. By plotting the detected diagonals, encoding them with a vision encoder, and clustering the embeddings, the question of what constitutes a translation can be addressed empirically rather than by stipulation, deriving the definition from the data and relating it to historical context.

A second application of semantic modeling, this time mostly within a single language, traces the reception of John Locke through semantic search over the ECCO corpus [14]. Here paraphrase and meaning match exceed verbatim lexical reuse by more than an order of magnitude. The method captures reception inaccessible to lexical search, such as Locke’s philosophy of mind simplified for an educational manual, and follows an idea as it moves between genres and registers of difficulty. Whereas lexical search around a given passage recovered mainly Locke quoting himself, semantic search

revealed a dense network of engagement involving figures such as Berkeley, Boyle, Hume, and Voltaire, reception present in the corpus but invisible to lexical methods.

The remaining steps of the program, allusion and commentary, are the least direct forms of text reuse and remain open. Our working hypothesis is that current state-of-the-art generative LLMs can interpret these relations reliably for a single pair of passages but, as in the translation mining case, cannot do so across a full corpus. Closing this gap requires encoders that model relations well beyond translation and paraphrase, so as to identify suitable candidate pairs for an LLM to interpret. Present-day encoders are trained on a mixture of tasks and respond to an averaged notion of similarity; a prompt appended to the query is meant to indicate the relevant relation, but this conditioning is not yet reliable. New training methods will be needed for encoders that can assign high similarity to, for instance, a passage and one that argues against or otherwise comments on it, enabling detection of qualitatively new forms of engagement at scale. This is an algorithmic problem on the model side and a data problem as well, one for which historians are well positioned to supply the examples that define these harder relations.

IV. CONCLUSION

Advanced methods have already opened research directions that were previously inaccessible. The Locke study, the OCR post-correction experiments, the Reception Reader, and the translation mining pipeline each extend what can be observed in a large historical corpus, and together they cover the program from surface reuse through cross-lingual translation to the first forms of modeled meaning. The remaining steps, the indirect and contested forms of engagement through which ideas are transformed, are not yet tractable at scale, and reaching them will require meaning models and interpretation tools substantially beyond the current state of the art. These are significant algorithmic problems rather than incremental ones, and the program as a whole illustrates the value of sustained collaboration between computer science and the humanities.

ACKNOWLEDGMENT

We are grateful to Cassandra Ledins, Iiro Rastas, Ananth Mahadevan, Rohit Babbar, David Rosson, Michael Mathioudakis, and Eetu Mäkelä for a productive collaboration, and to the Research Council of Finland for its support. Computational resources were provided by CSC – IT Center for Science.

REFERENCES

- [1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990. doi: 10.1016/S0022-2836(05)80360-2
- [2] A. Vesanto, A. Nivala, H. Rantala, T. Salakoski, H. Salmi, and F. Ginter, "Applying BLAST to text reuse detection in Finnish newspapers and journals, 1771–1910," in *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*. Gothenburg: Linköping University Electronic Press, 2017, pp. 54–58.
- [3] G. Roe, "Text reuse as cultural practice: Intertextuality in the 18th-century digital archive," *Digital Enlightenment Studies*, vol. 2, no. 1, pp. 1–30, 2024. doi: 10.61147/des.23
- [4] D. A. Smith, R. Cordell, and A. Mullen, "Computational methods for uncovering reprinted texts in antebellum newspapers," *American Literary History*, vol. 27, no. 3, pp. E1–E15, 2015. doi: 10.1093/alh/ajv029
- [5] C. Gladstone, R. Horton, and M. Olsen, "TextPAIR (pairwise alignment for intertextual relations)," Software, ARTFL Project, University of Chicago, 2021, developed 2008–2021. <https://github.com/ARTFL-Project/text-pair>.
- [6] K. Hinderks, C. Ledins, F. Ginter, and M. Tolonen, "Translation mining: An AI-driven taxonomy of eighteenth-century Anglo-French translation practices," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, pp. 1–25, 2026. doi: 10.1080/01615440.2026.2675558 Advance online publication.
- [7] D. Rosson, E. Mäkelä, V. Vaara, A. Mahadevan, Y. Ryan, and M. Tolonen, "Reception reader: Exploring text reuse in early modern British publications," *Journal of Open Humanities Data*, vol. 9, no. 5, pp. 1–11, 2023. doi: 10.5334/johd.101
- [8] Y. C. Ryan, A. Mahadevan, and M. Tolonen, "A comparative text similarity analysis of the works of Bernard Mandeville," *Digital Enlightenment Studies*, vol. 1, no. 1, pp. 28–58, 2023. doi: 10.61147/des.6
- [9] R. Péter, "Uncovering hidden influences: The reception reader as a tool for intellectual historians," *Global Intellectual History*, pp. 1–13, 2025. doi: 10.1080/23801883.2025.2474486
- [10] J. Kanerva, C. Ledins, S. Käpyaho, and F. Ginter, "OCR error post-correction with LLMs in historical documents: No free lunches," in *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*. Tallinn, Estonia: University of Tartu Library, Estonia, 2025, pp. 38–47.
- [11] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020. doi: 10.18653/v1/2020.emnlp-main.365 pp. 4512–4525.
- [12] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The Faiss library," 2024.
- [13] J. Matas, C. Galambos, and J. Kittler, "Robust detection of lines using the progressive probabilistic Hough transform," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 119–137, 2000. doi: 10.1006/cviu.1999.0831
- [14] Y. Wu, A. Mahadevan, F. Ginter, M. Mathioudakis, and M. Tolonen, "Matching meaning at scale: Evaluating semantic search for 18th-century intellectual history through the case of Locke," 2026, to appear in *Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities (NLP4DH 2026)*, San Diego, CA.