Routledge
Taylor & Francis Group

Historical Methods: A Journal of Quantitative and Interdisciplinary History

ISSN: 0161-5440 (Print) 1940-1906 (Online) Journal homepage: www.tandfonline.com/journals/vhim20

Translation mining: An AI-driven taxonomy of eighteenth-century Anglo-French translation practices

Kira Hinderks, Cassandra Ledins, Filip Ginter & Mikko Tolonen

To cite this article: Kira Hinderks, Cassandra Ledins, Filip Ginter & Mikko Tolonen (23 Jun 2026): Translation mining: An AI-driven taxonomy of eighteenth-century Anglo-French translation practices, Historical Methods: A Journal of Quantitative and Interdisciplinary History, DOI: 10.1080/01615440.2026.2675558

To link to this article: https://doi.org/10.1080/01615440.2026.2675558

HISTORICAL METHODS: A JOURNAL OF QUANTITATIVE AND INTERDISCIPLINARY HISTORY
https://doi.org/10.1080/01615440.2026.2675558

Translation mining: An AI-driven taxonomy of eighteenth-century Anglo-French translation practices

Kira Hinderks[a*], Cassandra Ledins[b*], Filip Ginter[b#] and Mikko Tolonen[a]

[a]Helsinki Computational History Group, Department of Digital Humanities, University of Helsinki, Helsinki, Finland; [b]TurkuNLP, Department of Computing, University of Turku, Turku, Finland

ABSTRACT

This paper introduces the concept of "translation mining", a cross-lingual embedding approach that bridges close reading and large-scale quantitative analysis of historical translations. Focusing on eighteenth-century British (ECCO) and French (Gallica) corpora, we systematically identify semantically aligned passages across hundreds of thousands of documents. The approach provides a comprehensive AI-driven methodology for locating translations of various types. This enables us to build a multi-scalar, data-driven taxonomy of translation practices that offers finer granularity than prior dichotomies of "full" versus "partial" translation. By embedding all texts once using high performance computing, we can iteratively detect subtle textual connections across languages without re-running the entire process. This paper demonstrates how embedding models derived from large language models can capture cross-lingual translation pairs, challenge rigid classifications, and illuminate multi-level cultural transfer, offering an adaptable framework for historical research.

KEYWORDS

Natural language processing (NLP); computational history; reception studies; digital humanities; artificial intelligence (AI)

1. Introduction

Scholars studying historical translations have traditionally faced a dilemma. One option has been to rely on close reading and archival work on a small set of texts (McMurran, 2000; Dow 2014). This produces rich insight into how translators adapted works to new linguistic and cultural contexts, but it does not scale beyond a limited corpus. The other option has been to use quantitative methods built on bibliographic datasets (Raven 2002; Zhou and Sun 2017). Such an approach can trace broad patterns of translation activity across languages and periods, but it seldom engages with the textual substance of translations themselves. While excellent work has been done on large-scale text reuse using primarily lexical methods, such approaches are strongest where reuse preserves shared wording and are less effective when engagement occurs through translation, paraphrase, abridgement, or semantic adaptation (Gladstone and Cooney 2020; Kulkarni et al. 2015; Ryan et al. 2023; Salmi et al. 2021; Smith et al. 2015). Recent infrastructure work has substantially advanced large-scale historical reuse detection, including the Reception Reader and optimized reuse pipelines developed by the Helsinki Computational History group, as well as related European platforms such as impresso (Düring et al. 2023; Mahadevan et al. 2025; Rosson et al. 2023). However, achieving comparable results in the cross-lingual domain requires semantic methods capable of identifying related meaning beyond lexical overlap. Such methods have not previously been implemented at comparable scale in historical multilingual corpora, which is the gap addressed here. Recent advances in AI and natural language processing (NLP) offer a third way—what we call *translation mining*—an approach that enables the systematic detection of translations at scale while preserving textual complexity.

To operationalize the translation mining approach, we draw on text embedding models. These models are closely related to large language models (LLMs) but differ fundamentally from the generative systems typically associated with the term. A primary application of embedding models is to evaluate the semantic similarity of two text segments on a continuous numerical scale, including across languages. Because these models (and the similarity measures they enable) are a core component of modern search engines, the methodology

CONTACT Mikko Tolonen mikko.tolonen@helsinki.fi Department of Digital Humanities, University of Helsinki, PO Box 24, Unioninkatu 40, 00014 Helsinki, Finland.
*Agreed to share first authorship
#Secondary Affiliation: ELLIS Institute Finland, Espoo, Finland.

is well understood, and computationally efficient methods have been developed to make similarity computation feasible even at extremely large scales. In this work, we build on these best practices to compute textual similarity across full-text databases, and apply it to the task of identifying translated text segments. In the present study, we identified almost 3 million matching passages (or "matched pairs") between English-language texts in Eighteenth Century Collections Online (ECCO) and French-language texts from Gallica (Bibliothèque nationale de France). All matched pairs are made available for inspection on a case-by-case basis.

When we think about Early Modern translation activity, we soon understand that these alignments illuminate not just complete translations, but also shorter English passages woven into French-language works and vice versa, typically without any acknowledgement. Because these matches range from book-length translations to fragmentary, reorganized, and embedded forms of reuse, a taxonomy is needed to render this evidence comparable and historically interpretable at scale. Applying translation mining to eighteenth-century Anglo-French materials, we propose six recurrent translation patterns grounded in bottom-up regularities visible in similarity matrices. The result is a heuristic map for moving from detection to explanation—supporting questions about how translation practices varied across genres, formats, and communities, and how texts moved and changed within eighteenth-century print culture. Our taxonomy enables scholars to enquire how a single act of translation was embedded in shared translation practices, which carried normative assumptions about authorship, readership and the cultural role of books. What the taxonomy helps to do is to turn the study of translation into a measurable structure of reception. It offers an analytical dimension to cultural transfer and a systematic path to research how texts are reproduced, fragmented, adapted, and recontextualized across languages, genres, and media.

Our taxonomy is inspired by the empirical tradition of systematization that characterized eighteenth-century scholarship, from encyclopedic ordering projects to scientific classification systems (Withers 1995). A particularly useful analogy is Carl Linnaeus' *Systema Naturae* (1735), not as a model of fixed or essentialist categories, but as an inductive framework developed through the observation of recurring forms. Recent historians of science have shown that Linnaeus did not seek to impose abstract ideals on nature but instead worked bottom-up, refining his categories in response to empirical variation (Müller-Wille 2007; Winsor 2006; Witteveen 2020). Our taxonomy follows the same logic: it groups recurrent structural patterns in translation practices without presupposing stable types or intentional design. The aim is not to define what translations *are*, but to provide a functional classification that allows scholars to compare how texts were transformed, reorganized, and redeployed across eighteenth-century print culture.

In what follows, we treat translation mining as a multi-scalar approach. At one end of the spectrum, the method surfaces micro-level translation pairs that can be inspected case by case—allowing, for example, a historian of Montesquieu or Hume to trace concrete interventions sentence by sentence. At the other end, the same pipeline underpins macro-level analyses of cultural transfer, such as quantifying how translation practices between France and Britain shift across the long eighteenth century. Crucially, the approach also opens up a meso-level of analysis: coherent ensembles such as an author's translation profile, a controversy-bound subset of political pamphlets, a genre in a given period, or the corpus of a single translator or publisher. It is at this intermediate level that translation mining becomes historically particularly productive, because it links interpretable units (authors, genres, controversies, careers) to the broader distributions of translation practices we describe in this paper.

Although developed for eighteenth-century French- and English-language sources in this initial case study, translation mining can be applied to a much broader range of historical materials. Our computational pipeline can be extended beyond its initial source base and can be adapted to work with corpora in other languages and time periods. For large-scale historical corpora, where additions or corrections to digitized texts can occasionally occur, the ability to efficiently re-run partial analyses rather than a full re-embedding is especially valuable. Methodologically, the same pipeline demonstrates how cross-lingual embeddings and similarity matrices can be turned into reusable data for other scholars working on cultural transfer, reception history, or the sociology of translation.

This paper is structured as follows: Section 2 introduces the ECCO and Gallica datasets, Section 3 guides the reader through each development step of our translation mining method, while Section 4 shows how our taxonomy can be applied to study Early Modern translation practices, including the transformation of ideas and the decisions taken by historical actors involved in the translation process. The final section, Section 5, provides a summary of our findings and outlines future research perspectives.

2. Data

Our data come chiefly from ECCO for English texts (206,613 texts) and Gallica for French texts (287,873 texts). ECCO, covering roughly 54% of extant

eighteenth-century British imprints (Tolonen, Mäkelä, and Lahti 2022), provides a broad textual base. Coverage is assessed against the English Short Title Catalog (ESTC), used as the best available bibliographical baseline for what survives from the period. ECCO documents were digitized from microfilms created during "The Eighteenth Century" project (1981), primarily sourced from the British Library, Oxford, and Cambridge. They were later processed with OCR for searchable text (Tolonen, Mäkelä, and Lahti 2022). Our ESTC data is based on an offline version updated in 2020 and curated by COMHIS (Tolonen et al. 2021). For the French side, we rely on Gallica's French-Language Monographs opened for the public in 2023, representing about 300,000 works spanning multiple centuries (https://api.bnf.fr/fr/node/222). While coverage of eighteenth-century French texts is less comprehensive than ECCO, this limitation still allows us to identify meaningful translation instances and patterns.

Both datasets come with familiar challenges: OCR noise, incomplete or inconsistent metadata, and uneven distributions of genres or time periods. OCR errors, in particular, are known to impede many text-mining methods—especially lexical approaches where character-level inaccuracies directly distort token counts, keyword searches, and similarity measures. Uneven genre coverage poses a different difficulty: if certain domains or periods are sparsely represented, historians have limited basis for drawing conclusions about longer-term change or comparing communities on equal terms. NLP techniques, however, help mitigate these limitations. By shifting attention from surface-level tokens to patterned relations—repeated passages, recurrent phrasings, or shared conceptual vocabularies—we can derive meaningful signals even from imperfect data. Our approach builds on this insight by using cross-lingual embedding–based semantic similarity to detect translations and transformations across a spectrum of reuse, reducing dependence on clean OCR and enabling comparisons that would otherwise remain inaccessible.

The amount of data processed throughout the different filtering steps of the study is detailed in Table 1. The initial number of pairs of documents is the maximum number of matching pairs, if all the documents were to match each other. In the next steps, this column represents how many pairs of documents have been retained in the filtering steps of the process. The massive initial data volume is further illustrated by the number of embedded text segments (numerical vectors output by the embedding model) as detailed in Section 3. These embedded segments total approximately 3TiB of numerical data. The second line of the table presents the data resulting from our translation-mining pipeline, comprising 3 million matching passages in 1 million document pairs. Two ranges of the signal-to-noise ratio and semantic similarity scores are suggested to help separate direct translations supported by strong scores, and close topical matches supported by somewhat lower but still notable score values. The matching pair counts at these two score levels are listed in the table as well. Naturally other score thresholds can be set based on the needs of individual use cases.

All 3 million detected translation passages are available for download in tabular format at https://doi.org/10.5281/zenodo.14514300. The file contains metadata for each detected matching passage pair, including authors, titles, original IDs from the Gallica and ECCO datasets, character level offsets of matching passages, and a number of pre-computed scores. It is accompanied by another tabular file containing translation taxonomy classification of each pair, according to a classifier trained on our translation taxonomy suggestion, as explained later in Section 3.2 and Appendix. The matching passages dataset is available via the Helsinki Computational History Group website (https://www.helsinki.fi/en/researchgroups/computational-history/translation-mining), where both the interactive interface—featuring metadata,

Table 1. Number of document pairs and matching passages at different phases of the process and score thresholds.

Phase/Data	Pairs of documents	Matching passages	Documents	Segments
Initial	59,478,304,149	–	206,613 (en) 287,873 (fr)	266,727,600 (en) 393,986,205 (fr) (~3TiB)
After translation-mining	1,013,665	2,992,241	72,546 (en) 45,941(fr)	–
"Topic match" filter (1.5 > signal ratio > 1.25, similarity_score > 0.4)	404,510	723,046	47,185 (en) 30,512 (fr)	–
"Translation" filter (signal ratio > 1.5, similarity_score > 0.4)	129,424	198,567	23,480 (en) 11,526 (fr)	–

Table 2. Quick reference glossary of the most important technical terms in Section 3.

Byte Pair Encoding (BPE): A method used by language models to break text into smaller units called tokens, which correspond to whole words or parts of words. It is used to control the size of the vocabulary used by the model and is critical for computation efficiency.
Cross-lingual model: A language model trained to represent texts from different languages in the same shared semantic space.
Edge detector: An image-processing algorithm that detects boundaries or sharp changes in an image. Here it helps prepare similarity matrices for line detection.
Embedding: A numerical representation of a text segment that captures aspects of its meaning. Embeddings allow passages to be compared on the basis of semantic similarity rather than exact textual overlap.
FAISS: A software library designed for finding nearest neighbors in very large collections of vectors quickly and efficiently.
Full similarity matrix: A grid containing similarity scores for every possible pair of segments between two documents. Unlike the sparse hit matrix, it is exhaustive and therefore more informative, though also more costly to compute.
Hit: A segment from one document appearing among the top most similar segments to a segment from another document. This alone does not yet prove a translation relationship, but indicates its possibility.
Sparse hit matrix: A grid that records the locations of hits between the segments of two documents. When these hits form a diagonal or near-diagonal line in the matrix, the document pair warrants the calculation of the full similarity matrix and a more detailed analysis. It is called "sparse" because most possible comparisons are left empty, with only a limited set of promising matches present.
Hough transform: A standard image-processing technique for efficiently detecting straight lines in noisy visual data.
Language model: A computational model trained on large amounts of text in order to detect statistical patterns in language. In this study, the language model is used to compute embeddings of text segments, rather than generate text.
Nearest neighbors: The items in a dataset that are most similar to a given query item according to a model's numerical representation.
Progressive probabilistic Hough transform: A faster and more scalable variant of the basic Hough transform.
Semantic similarity: Similarity in meaning rather than in exact wording. Two passages may be semantically similar even when they use different vocabulary, or are in a different language.
Sentence Transformer: A type of language model specifically trained to convert text segments into embeddings that can be compared efficiently.
Signal-to-noise ratio: A measure of how strongly a meaningful pattern stands out from surrounding background variation. In our case, it indicates whether a detected translation line is clearly stronger than the average similarity values around it.
Vector: A point in N-dimensional space, used here to encode the meaning of a text segment mathematically.
Vector distance: A numerical measure of how far apart two vectors are in the N-dimensional space.

document excerpts, and similarity matrix visualizations—and the underlying data can be accessed.

3. Methods

The following section details the computational pipeline used to integrate NLP-based semantic similarity with image-based pattern recognition for large-scale analysis of ECCO and Gallica. The methodology allows us to compare hundreds of thousands of works to identify, characterize, and trace different types of translations, importantly including also partial translations and fragmentary borrowing. We will first discuss earlier, lexical-based methods for studying historical translations and outline the advantages of our semantic-level approach. Next, we describe the different steps of capturing translations using this new approach. Finally, we explain the development process of our data-driven taxonomy of translations, which will be applied to historical examples in Section 4. The most important technical terms are summarized in the glossary in Table 2.

Computational history and digital humanities research have traditionally been leaning on lexical-level techniques, which treat words as fixed symbols and compare texts based on exact matches. Lexical-level techniques form the basis of methods like topic modeling, keyword-driven search, and text reuse detection. The most advanced implementations of text reuse detection can deal with documents that have very poor OCR quality (e.g. Vesanto et al. 2017). Historians frequently use lexical-level methods to study the topical composition of a specific source base like the *Encyclopédie* (Roe, Gladstone, and Morrissey 2016), as well as tracking the spread of ideas and concepts such as liberty (de Bolla et al. 2020) and modernity (Eijnatten and Ros 2019) within a single language. However, these lexical methods are fundamentally limited to capturing surface-level similarity of texts as sequences of characters and therefore struggle to assess similarity across languages, including translations, because texts in different languages naturally do not match at the surface level.

Instead of comparing words or sentences directly, our approach focuses on capturing meaning with the use of "embeddings", a well-studied methodology in NLP. An embedding is a numerical vector (typically containing thousands of dimensions) that represents the meaning of a word or a text segment as a point in a high-dimensional space. Crucially for the purpose of studying translations, these embeddings can be compared in terms of their distance; the closer two vectors are, the more similar the meanings they encode are. This property holds also across languages, i.e., the embeddings of a text and its translation will be numerically close, because the meanings of these two texts are naturally very close as well. This is not an accidental property, rather, it is the primary objective followed during the training of the language models which calculate these embeddings, i.e., the models are specifically trained to produce embeddings that have these properties using text similarity and translation datasets (cf. Muennighoff et al. 2023; Tiedemann 2009). Specifically, we employ the *paraphrase-xlm*

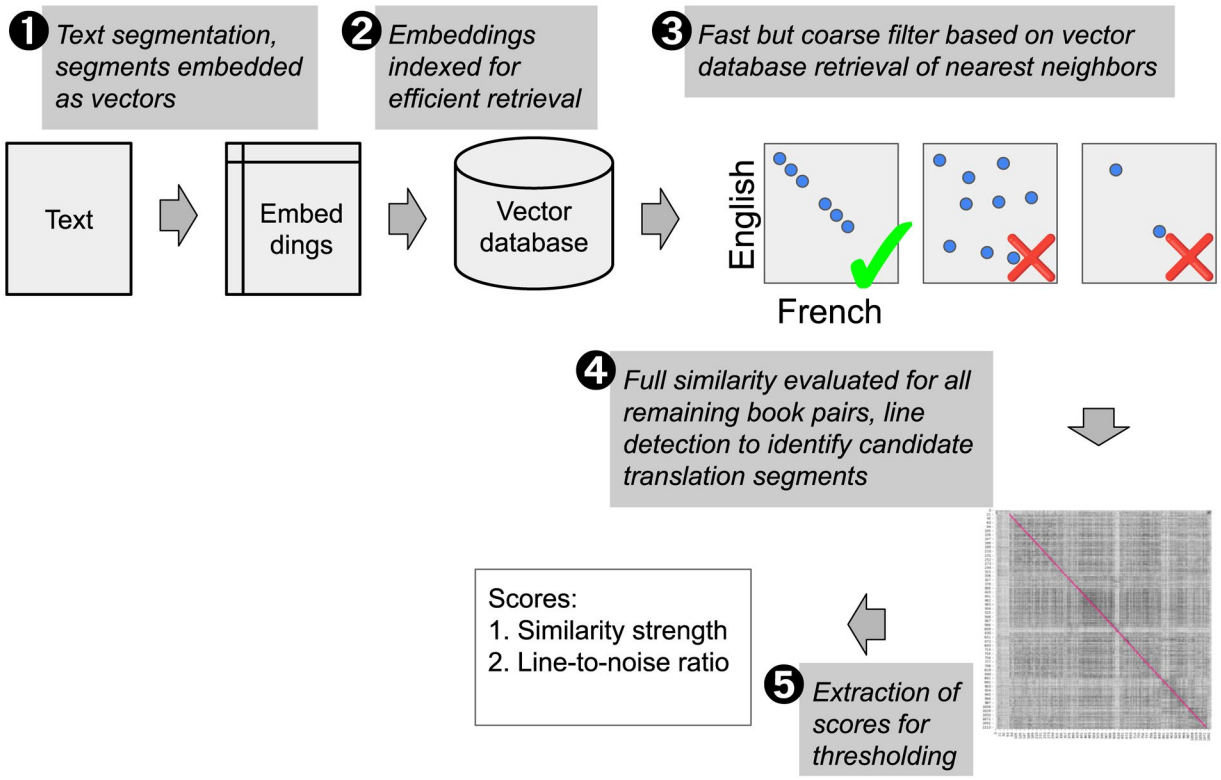

Figure 1. The steps of processing used to detect translation pairs. The texts are divided into overlapping text segments, which are embedded ①, indexed in a vector database ②, and initial comparison is made based on sparse hit matrices ③, promising pairs that contain at least one detectable diagonal are preserved and their full similarity matrix is computed, followed by detection of diagonals characteristic to translations ④, which are subsequently scored by similarity and signal-to-noise ratio ⑤.

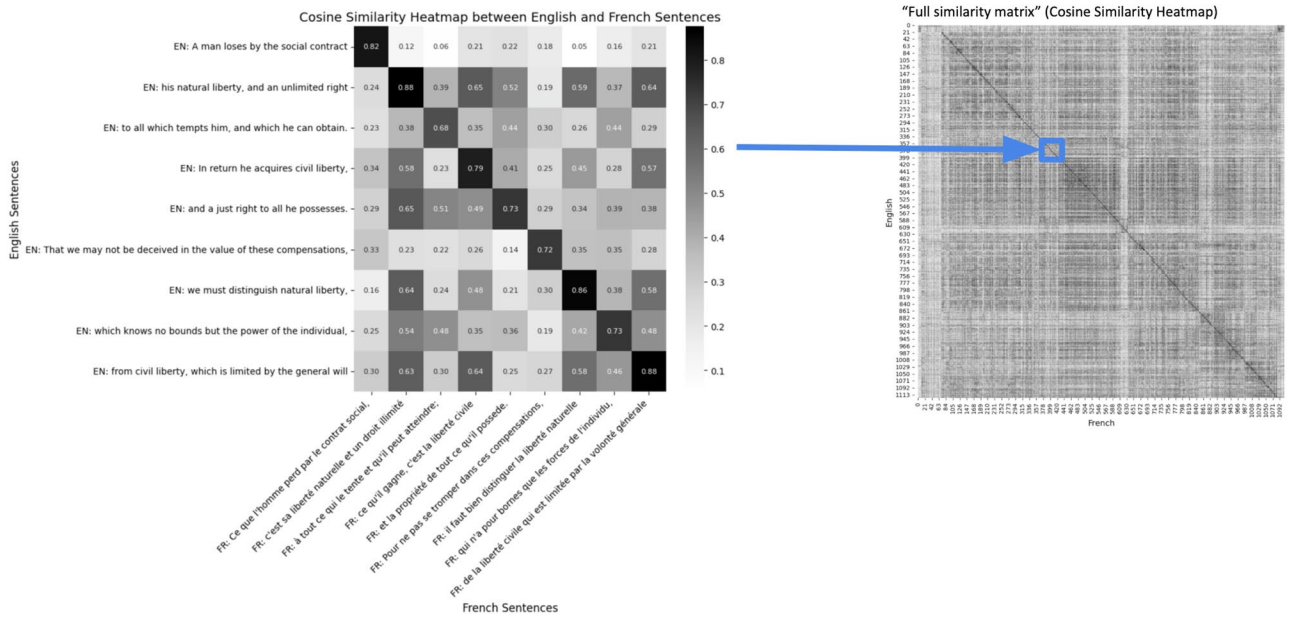

Figure 2. From two texts to one image. This visualization demonstrates the cross-lingual embedding model's ability to calculate text similarity and match translated text. On the left, an extract of Rousseau's *Du Contrat Social,* found in Gallica, is segmented and compared to the matching extract of an eighteenth-century translation found in ECCO. Text similarity, as detected by the model, is shown in levels of gray, the darker shades representing stronger similarity. Pairs of translated sentences match closely in meaning and therefore produce higher similarity scores along the diagonal of the comparison matrix, in contrast to other parts of the matrix where non-corresponding sentence pairs (those that are not mutual translations) are compared. On the right, this same principle is applied to the entire book and its translation, showing a dark diagonal emerging in a heatmap that visualizes the alignment between the entire book and its translation. This diagonal is a central feature of our computational methods and serves as key evidence of translation.

-r-multilingual-v1 Sentence Transformer model, first introduced by Reimers and Gurevych (2020), which achieves a balance of embedding quality as established on public benchmarks (Muennighoff et al. 2023) and sufficiently low computational cost to allow the large-scale application necessary in our study.

3.1 Translation mining: a step-by-step guide

In this section, we outline our process for mining translations, comprising five main steps (also visualized in Figure 1). We begin by splitting each document into comparatively short, overlapping text segments (Step ①). For each segment, we use the cross-lingual model to calculate its vector embedding, therefore representing an entire document as a sequence of these vector embeddings (Step ②). Subsequently, we use a series of filtering steps to discover significant matches between document pairs, and therefore potential translations (Steps ③④). These candidate matches are then scored numerically to determine the degree of similarity (Step ⑤). Finally, a taxonomy of translation practices is used to classify these matches, highlighting common historical translation practices such as inserting brief translated passages into otherwise original texts, or making minor changes to source works.

Our study builds on the ability of cross-lingual embedding models to match translated text. This ability is demonstrated in an example taken from our data, Rousseau's *Du Contrat Social*, in Figure 2. When matching the first part of the extract, "Ce que l'homme perd par le contrat social", to one of its translations found in ECCO, "A man loses by the social contract", the model finds a high similarity of 0.82 (the maximum being 1). On the contrary, the model finds a significantly lower similarity with other parts of this paragraph, scoring for example 0.16 (the minimum being 0) when compared to "we must distinguish natural liberty". Even with a degree of noise, a clear visual signal stands out, when every next sentence of a book matches the next sentence of another book, a dark diagonal of similarity stands out to indicate a longer systematic match and therefore a possible translation. Such a diagonal line necessarily emerges whenever two documents systematically match in meaning on a segment-for-segment basis, which corresponds to a diagonal direction when the similarity of all segment pairs is plotted as in Figure 2.

Applying the full comparison illustrated in Figure 2 on all pairs of books in our data would be computationally infeasible, needing nearly 60 billion comparisons (Table 1). Instead, we employ computational techniques that approximate full comparison at a fraction of the computational cost. Crucially, the methods we employ ensure that any given document is compared only to a fixed, relatively small subset of documents that are most likely to contain a translation. This reduction in the comparison space—rather than exhaustive document-to-document alignment—guides the design of the computational pipeline shown in Figure 1.

Step ①: The first step is to divide each document of both ECCO and Gallica's French-Language Monographs into overlapping segments, ensuring they are short enough for precise comparison but long enough to retain contextual meaning. These text segments are defined in terms of consecutive "tokens". Tokens are a basic unit of text in NLP techniques and correspond to words or parts of words, i.e., there is not a 1:1 correspondence between the number of tokens inputted to the language model and the number of words in the text. This is due to the language model requirement of a fixed-sized vocabulary, typically less than 100K unique items in size, implemented using a technique called Byte Pair Encoding (Gage 1994). This token encoding is established as a part of the training procedure of language models, and from our perspective is therefore "fixed", given by the model we use to compute the embeddings. In our case, one word of text corresponds on average to about 1.3 tokens. In preliminary experiments based on known translation pairs, we found that segments of approximately 75 words in length (100 tokens), with an overlap of approximately 25 words (30 tokens) balances favorably between too short segments inducing noise and lacking in context, and too long segments lacking in resolution. Each text segment is subsequently embedded using the Sentence Transformer model and therefore represented as a single, dense vector. The numerical similarity of two embeddings correlates with the degree of semantic similarity of their corresponding text segments in ECCO and Gallica: higher numerical similarity (low vector distance) corresponds to a closer match of meaning.

Step ②: Since comparing every possible pair of text segments in the whole data is not feasible computationally, we must use an optimized retrieval method to narrow down the search. Instead of exhaustively comparing all embeddings, we rely on extremely efficient database techniques that are designed to retrieve a fixed number of the most relevant (i.e., numerically closest) candidates from among potentially billions of vectors, i.e., easily scalable to our data and potentially beyond, since the number of candidate segment embeddings is in the hundreds of millions (Table 1). Specifically, these vector databases

allow us to efficiently answer the question: given a query vector q, return the list of k nearest neighbors of q in the index where k typically ranges in tens to hundreds. Here it is important to note that, to achieve the necessary computational efficiency, these techniques are approximating the full search, i.e., the retrieved k nearest neighbors is not guaranteed to be exactly corresponding to the actual k nearest neighbors in the data. In this work, we use FAISS (Douze et al. 2024), the de facto standard, highly efficient vector database implementation capable of scaling up to our needs. We index the embedding of every segment in every Gallica document, in total almost 400 million individual vectors.

Step ③: Using this index, we are able to find the list of a small number, 100 in our case, of the most similar segments from Gallica for every segment from ECCO. At this point, it is important to keep in mind that the steps above serve as a first-pass filter; finding similar segments does not mean we have found translation pairs. We still need to address two crucial aspects: (1) the text segments are considered in isolation, any Gallica segment is in principle a candidate for any ECCO segment, there is no document-level information at this stage; and (2) the 100 most similar segments do not imply that translation pairs were found. The vector database will always return the k most similar vectors and does no other filtering; for example, it can still include paraphrases and segments discussing related, but not exactly equivalent topics. Most of the segment pairs identified at this stage, in fact, are not translations. Rather, the purpose of this step is to reduce the search space. Using statistics of known translation pairs in the data, we set $k = 100$ as a suitable cutoff for the vector database retrieval, striking a balance between computational cost and sufficiently reliable retrieval.

Subsequently, we shift from individual text segments to a document-level perspective, aiming to identify texts that are likely to be translations. We aggregate the vector database matches (which we will refer to as *hits*) on the level of documents (ECCO and Gallica titles) and identify pairs of ECCO-Gallica documents which have a comparatively high number of *hits*, i.e., segments from the given Gallica document appearing among the top-100 most similar segments in the ECCO document. This allows us to identify potential translation pairs on document level based on two observations: Firstly, a pair of documents which forms a translation pair will have a high number of hits. As an initial filter, we only retain such ECCO-Gallica pairs where at least 2% of the ECCO document length is covered by segments that have a top-100 hit in the corresponding Gallica document. The value of 2% is a conservative filter, retaining all book pairs with even a very slight chance of forming a translation pair, and has been selected based on statistics of known translation pairs in our data. Secondly, in translation pairs, the hits would be expected to proceed sequentially in both documents. To identify such sequences, we accumulate the hits into *hit matrices*: for any ECCO-Gallica document pair, the hit matrix will have as many rows as the ECCO document has segments, and as many columns as there are segments in the Gallica document. For any pair of documents retained after the abovementioned filter, we can produce the hit matrix (as shown in Figure 1 as the result of step ③) and assess whether it indicates a potential translation using a simple observation: if the two documents form (or contain) a translation spanning a number of segments long sequence, the translation will exhibit itself as a diagonal, or near-diagonal line of hits in the hit matrix (see example in Figure 3 (a)). The absence of such a line in a matrix with numerous hits indicates a document pair with high content similarity, but which does not have the sequential hit arrangement indicative of a translation.

To automate the identification of these diagonal lines, we apply a textbook image-processing technique, known as the *Hough transform*, which is commonly used to detect lines in images. This algorithm first applies an edge detector filter to the matrix and subsequently converts the result to a representation where straight lines are easily detectable by thresholding. We use the *progressive probabilistic Hough transform* (Galambos, Matas, and Kittler 1999), an efficient probabilistic variant of the base Hough transform algorithm, resilient to noise and gaps and, unlike the base algorithm, scalable to our data in terms of computational requirements. We set a very permissive filter at this stage, preserving for further processing hit matrices which contain a line of at least 30 segments in length at an angle of 45°±25°. Even at this permissive setting, once again established on known translation pairs, the number of candidate document pairs proceeding to the next, computationally heavy analysis step is significantly reduced.

Step ④: Once the number of matching document pairs is greatly reduced by these initial filtering steps, it is possible to apply a more detailed computation to each remaining pair. The most important aspect that needs to be addressed in this step is that the hit matrices used in the previous steps are sparse (see Figure 3 (a) for illustration), owing to the inherent limitations brought about by the nearest neighbor

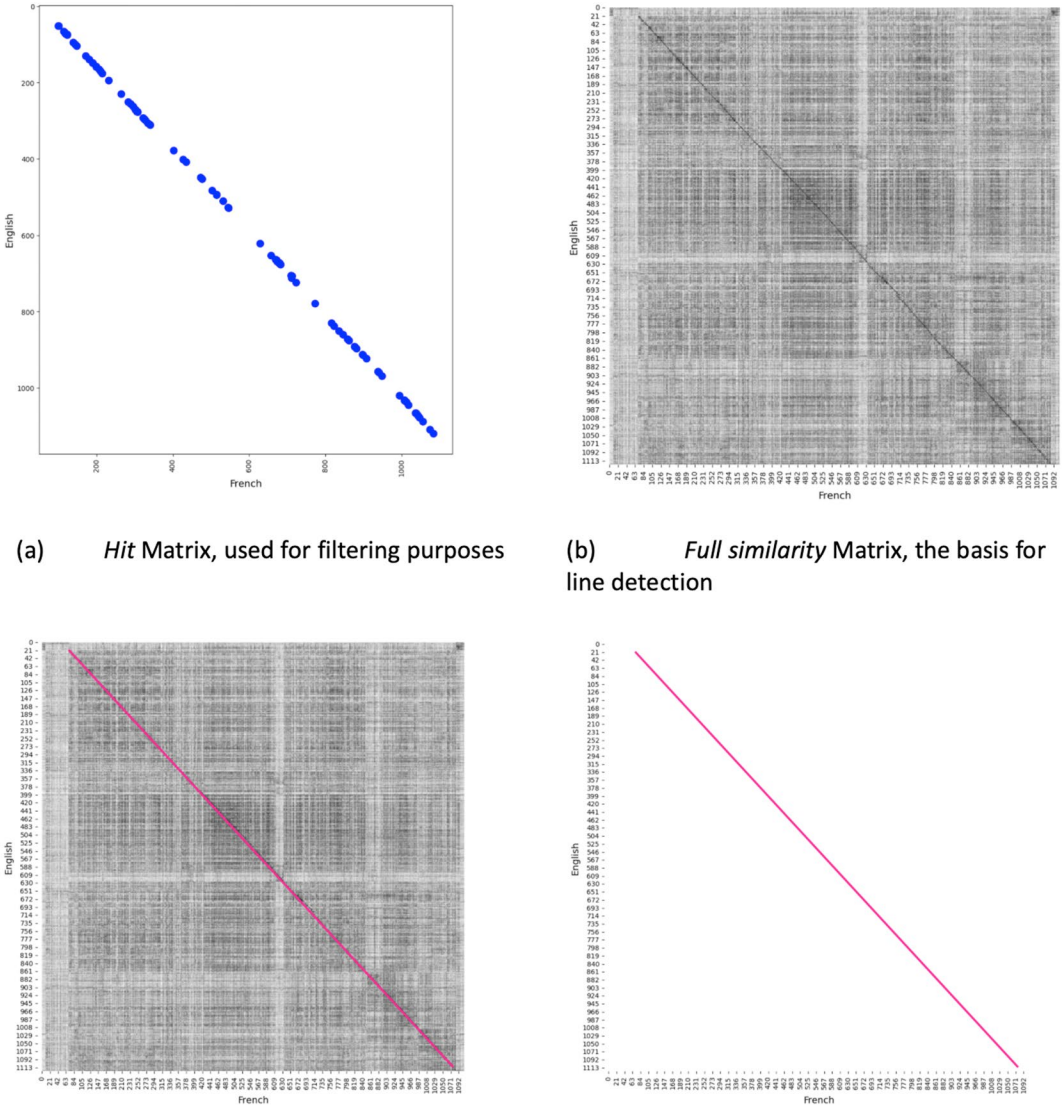

Figure 3. Rousseau's *Du Contrat Social* and one of its eighteenth-century translations as they progress through different steps of the computational process. (a) As a hit matrix, with each dot being a hit of high similarity found through the FAISS index. (b) As a full similarity matrix, a heatmap of similarity between each segment of both texts. c) As a full similarity matrix with a line detected with the *Hough transform*. (d) As only a detected line, as shown to a classifier for simplification purposes.

search among hundreds of millions of candidates. As the initial filtering reduces the number of candidate document pairs from about 59 billion down to 1 million, it is now computationally feasible to calculate full similarity matrices, i.e., for each candidate ECCO-Gallica document pair, calculate the numerical similarity of every segment in the ECCO document with every segment in the Gallica document. Note that the main difference between these full similarity matrices and the hit matrices used in the first filtering stages is that the similarity matrices store the exact similarity value for every segment pair and are therefore not sparse (see Figure 3 (a) versus Figure 3 (b) for visual comparison). We compute the full similarity matrix of each document pair passing the previous filtering steps, thus moving to a considerably more informative, exhaustive comparison, not limited by the sparseness of the hit matrices. In the full similarity matrices, translated segments will also manifest themselves as characteristic diagonals of high similarity values, which can be detected algorithmically. This can be illustrated by plotting the similarity matrix in a black and white scale, with black as the highest possible similarity value (1) and white as the

lowest possible similarity value (0). We therefore re-apply the Hough transform line recognition algorithm to the full similarity matrices, expecting a more reliable line detection with the full matrices, compared to the sparse hit matrices. It is not uncommon for the line detection algorithm to recognize a single long line as a series of potentially partially overlapping lines. We address this issue using a simple post-processing step which, proceeding from left to right, merges lines which are closer than a small threshold. Visually, when plotted in the manner described above, we would expect the translation lines of high similarity to stand out from the (noisy) background of "average" similarity of weakly related or unrelated segments.

Step ⑤: To distinguish true translation patterns from false positives, we use a combination of two scores. The first score is the average similarity value along the detected line, a real value ranging from 0 to 1. A similarity score closer to 0 means that the two segments have no relationship, while a score closer to 1 means that the two segments are nearly identical in their meaning. One of the limitations of the embeddings produced by the presently available language models is that the numerical similarities of the embeddings do not conform to some absolute similarity scale comparable across different texts. Consequently, it is not possible to set a single absolute threshold that would reliably differentiate translations in all texts.

To account for this limitation, we introduce a second score which considers the different levels of "average similarity" in different text pairs. The score is defined in terms of a signal-to-noise ratio, i.e., the similarity value of a point on the line, divided by the average of the similarity values in a small surrounding window. This signal-to-noise ratio is averaged across all points on the line and can be intuitively understood as the degree to which the line visually "stands out" from its surroundings. This ratio score theoretically ranges from 0 to infinity, with any value above 1 signaling that the similarity scores along the detected line are above the surrounding average similarity and the likelihood of spurious detection decreases as the signal to noise ratio increases.

To validate whether these two scores capture genuine translations, we sampled 44 document pairs. To ensure that the evaluation candidates cover the full variety of score combinations, the candidates were sampled evenly across the distribution of the two scores, rather than at random, which would result in a sample concentrated mostly in the low score dense areas (blue dots in Figure 4). A domain expert manually evaluated whether these pairs correspond to a genuine translation or only resemble one. The evaluator first compared the aligned snippets, followed by consulting the original editions in Gallica and ECCO. When we here speak of "genuine" translations, we by no means claim that we have found definitive criteria for establishing not-translationness, since definitions of translations and adaptations are contested, and the notion of intellectual influence is not uncontroversial. As translations, we understand documents where ideas are arranged in the same sequence using words that relate to the original, while potentially modifying, reframing, or contesting its meaning. In contrast, we evaluated texts as "not translations" if they merely exhibit what we describe as "topic similarity". In concrete terms: two authors talking about a topic in a way that is similar enough to be recognizable by both algorithm and a human reader as being about the same topic but not similar enough to indicate definitive influence of one text on the other, such as two histories narrating the same battle but with different foci. For each pair, the evaluator provided detailed feedback on their decision. For instance, the evaluator noted instances when OCR noise was matched to other OCR noise, such as tables or mathematical formulae. Similarly, Bible commentaries in French and English often received a high similarity score, but were not regarded as translations of each other, since both texts share the same Hebrew and Greek source base. Although the number of samples was relatively small, these comments thus helped to discover results that were correctly matched from a technical perspective but undesired regarding our research objectives. The resulting judgements are plotted in Figure 4, showing

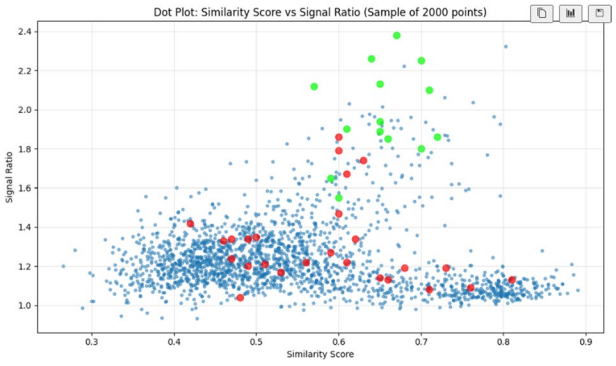

Figure 4. Distribution of identified translation candidates (blue dots, random sample) with respect to the two scores: average similarity on the horizontal axis, and signal-to-noise ratio on the vertical axis. A sample of candidates was manually inspected and judged with respect to being true translation (green dots) or not (red dots).

a clear correlation between the scores and the judgements. It also becomes evident that the similarity score computed from the embedding is not in isolation capable of sufficiently differentiating the translations, and the signal-to-noise ratio plays an important role in ranking the data from most- to least-likely translation pair.

An important aspect driving the design of the processing pipeline is scalability. While the embedding (step ①) and diagonal detection and scoring (steps ④⑤) implement the main analysis method, the indexing and pre-filtering (steps ②③) are only needed to make the analysis computationally feasible at the scale of our data. The main limiting factor is the quadratic nature of the problem, i.e., all documents in the first language must be compared to all documents in the second language. The indexing and pre-filtering steps reduce the number of candidates for each document by 4 orders of magnitude (Table 1), controlled by the k parameter, i.e., the number of nearest neighbors for each segment retrieved in step ③ of the process, which subsequently controls the number of candidate documents for which the full similarity matrix needs to be evaluated. The vector database (FAISS) retrieval time for a single query vector is nearly constant irrespective of the number of indexed vectors, since it uses advanced heuristics to prune the search space. Consequently, the overall pipeline can scale nearly linearly with the number of works, rather than quadratically as a naive implementation of the method would if the indexing and pre-filtering steps were to be skipped. Linear scaling means that the computational demand of the study is proportional to the number of works in the study, while quadratic means that it would be proportional to the square of the number of works, obviously a massive difference for a study involving hundreds of thousands of works. The linearity of the approach potentially allows studies on an even larger scale than we were able to carry out.

Having identified recurring semantic and structural patterns through the computational workflow, the next step is to make these findings historically interpretable. Developing a taxonomy of translation types becomes essential at this stage. It provides a framework for organizing the different forms of translation practices revealed by the models, allowing historians to move from individual detection of pairs toward a broader understanding of how translations operated across genres, periods, and intellectual communities. A clear typology thus links the computational identification of patterns to the substantive historical questions of translation practice, influence, and transmission.

3.2 Developing a data-driven taxonomy

While digital methods continue to advance rapidly, scholars in the humanities have grown increasingly attentive to the limits of different classification schemes, recognizing that categories applied to cultural materials are historically contingent, interpretive constructs rather than fixed ontological kinds (Bode 2020). At the same time, both traditional and computational scholarship inevitably rely on some form of organization when working with large and heterogeneous corpora. Our aim is not to treat data as a direct mirror of historical reality, but to identify recurring patterns that emerge from translation practices across many texts. We think that an analogy with Linnaean taxonomy is useful in this sense: just as Linnaeus imposed a provisional, descriptive order on botanical diversity without claiming to uncover nature's immutable structure, we seek a functional taxonomy that groups observable translation behaviors without presupposing intentional design or stable "species" of translation. The point is simply to provide a systematic framework that helps historians discern recurring cultural practices: patterns that become visible only when viewed at scale and that remain open to revision as new evidence accumulates.

To apply this approach, we first examined how different types of textual alignment appear in similarity matrices. Preliminary exploration of the identified lines confirmed a wide spectrum of ways in which the lines can be arranged. This variation corresponds intuitively to a broad spectrum of ways in which a work or its part can be translated and included in another work. As we are in no way limited to only identifying full document translations (which would correspond to a line spanning from the top-left to bottom-right corner in the similarity matrices, discussed as category 1 below), we established a taxonomy to categorize and interpret the different configurations that were observed. Throughout, we use "translation" in an operational sense to include translation-mediated reuse, ranging from full-volume renderings to fragmentary, reorganized, or embedded passages, while recognizing that historical actors and modern scholars draw these boundaries differently.

We developed the taxonomy inductively by organizing visual representations of detected alignments using an image embedding model and then defining classes through iterative human annotation of recurrent line configurations. A random sample of 225 pairs was processed and displayed in a grid according to

their visual similarity, once again from a model point of view (see Figure 5). These visual representations correspond to plotting the identified lines of alignment (much like the image resulting from step ④ in Figure 1). Six different classes emerged through observation. For this random sample of document pairs, we manually assigned each to one of these six classes based on the arrangements of lines. The properties of each class and the historical translation practices which these classes represent are described in Section 4

This annotated set was then used to train a classifier, i.e., a model which would be able to recognize and categorize translation pairs based on their structural properties. A VGG16 vision model of Simonyan and Zisserman (2015) is shown images from the set to learn what the different categories are supposed to look like. Once it is trained, we obtain a customized classifier, a tool capable of classifying previously unseen images. We applied it to categorize all document pairs.

Our approach, which clusters visual representations of translation pairs and refines them through manual annotation, therefore provides a structured yet adaptable framework for studying textual transmission at scale. While this taxonomy allows us to systematically identify different translation patterns, its true value lies in the historical insights it enables. In the next section, we apply our taxonomy to specific historical research questions.

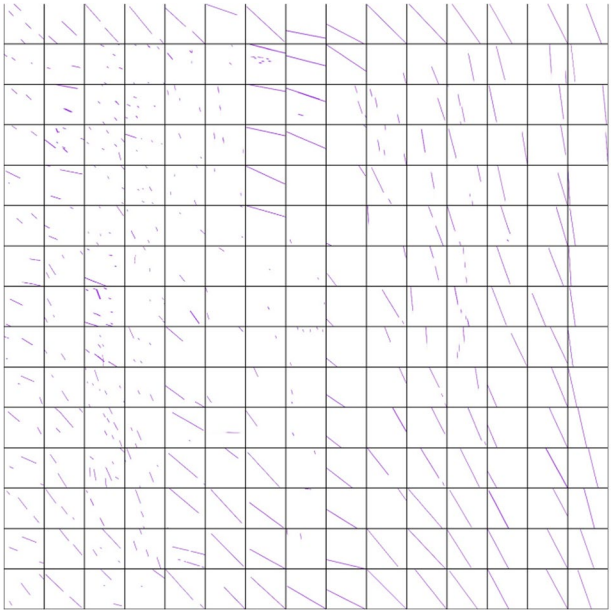

Figure 5. Random sample of document pairs represented as lines. Visually organized in a grid, these pairs were later manually annotated and provided the starting point for our taxonomy development.

4. A taxonomy of eighteenth-century translations

The systematic study of Early Modern translation practices has been limited by difficulties of scale. The manual comparison of two texts has allowed scholars to recover the richness of Early Modern translation within the context of a single act of translation. Quantitative analyses of bibliographic data have added valuable information about the translation market. However, an important part of Early Modern translation, namely highly modified translations and translations that consist of shorter fragments, have remained difficult to identify at scale. Our translation mining dataset thus provides access to a much larger quantity, and types, of translations, including acts of translation which are not recorded in library catalogues because they consist of paraphrases, adaptation and conceptual borrowing. Since the data is organized as document pairs, the taxonomy we have developed provides a structure beyond the comparison of two texts. Our taxonomy therefore helps the user to regard the agents involved in the translation process not merely as participants in a single act of translation, but as part of a culture of translation, which has recognizable shared practices and normative assumptions. With the help of the taxonomy, we can investigate these assumptions, for example by studying in which contexts translated snippets were inserted into otherwise original texts, how the oeuvre of a specific author was translated and modified, whether particular translation practices were especially common among different genres, as well as how Early Modern translation practices relate to other elements of print culture (e.g. ideas of authorship and readership). Different genres and publication formats in the eighteenth century tended to impose distinct constraints on how translation was carried out—ranging from complete book-length translations to selective extraction, adaptation, or reuse within miscellanies. While we do not use genre as a classificatory criterion in this paper, the observed structural patterns are nevertheless likely to intersect with genre conventions and scale as historically contingent outcomes of these practices.

Following a Linnaean logic of classification, this taxonomy comprises six structurally defined classes, which in turn correspond to historically recognizable translation practices (Figure 6). Our taxonomy classifies translations into two broad categories: "conservative" and "modifying" translations. Early Modern translators frequently made modifications to the source text, to the point that highly divergent translations became known as "les belles infidèles" (e.g.

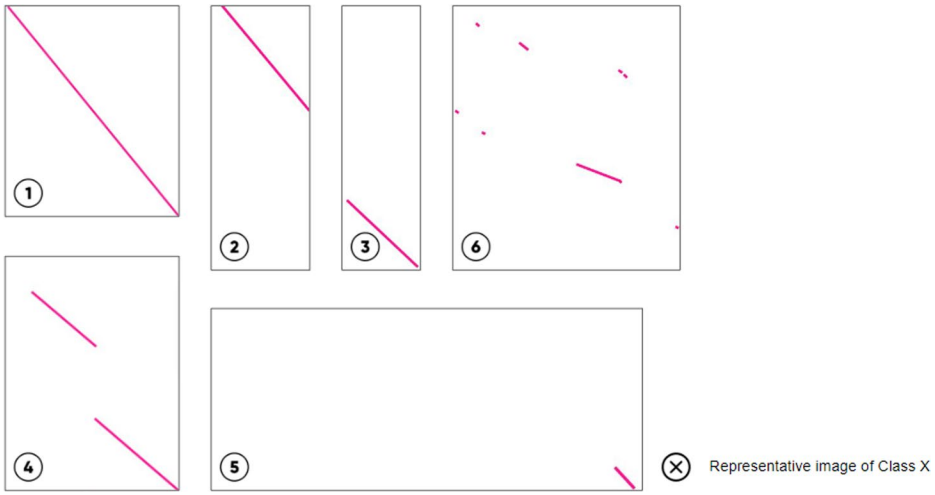

Figure 6. The six classes of the taxonomy: Single Volume (Class 1), Volume Reconfiguration (Class 2), Collections and Compilations (Class 3), Omission/Insertion (Class 4), Fragment (Class 5), Multi-Fragment (Class 6). The diagonal represents the translated part in an ECCO-Gallica document pair. The development process of the taxonomy is described in Section 3.2.

Ferrand 2014). The existing scholarship has tended to focus on modifications made at the sentence-level, such as paraphrases (e.g. McMurran 2000). Our taxonomy operates on a different scale, namely structural modifications. Here, we do not make judgments about whether a certain kind of translation is more "faithful" to the original text but instead view the different translation practices from the perspective of the translator. Hence, when we use the term "conservative", we mean it in a descriptive sense only. Classes 1–3 can be characterized as "conservative translations" since these translations do not alter the overarching structure of the original text. The three conservative classes exist on a spectrum where the main variation consists in the length of the source text compared to the translated text: i.e., whether the source text was translated as a single volume (Class 1), as part of a longer collection, such as the collected works of a single author (Class 3), or is stretched to more volumes or compressed to fewer volumes than the original (Class 2). Classes 4–6, by contrast, are called "modifying translations", since these translations exhibit a much greater degree of structural modification by the translator: e.g. omitting or inserting significant parts (Class 4), only translating a short fragment (Class 5) and performing other radical fragmentations of the original text (Class 6).

Our taxonomy also sheds light on the materiality of the translation process and the transformation of ideas, providing fruitful new research opportunities in fields such as book history and intellectual or cultural history. For book historians we provide a methodology to study the addition and removal of different materials such as prefaces and epilogues at scale to interpret and understand the role of material changes in translation practices. In intellectual or cultural history, translation mining offers a new way to investigate a longstanding historiographical debate—namely, the relationship between the "high" Enlightenment of philosophers and the "low" Enlightenment of pamphlets, popular literature, and ephemeral print culture. By systematically identifying translation practices across a wide textual spectrum, translation mining enables scholars to trace how philosophical ideas circulated beyond elite intellectual circles, including how ideas were adapted and reframed. The following part will introduce each class one-by-one, using a salient example to illustrate its characteristics.

Class 1: Single Volume

Class 1 covers full-volume translations, where an entire text is translated within a single volume with only very minor structural modifications. These structural modifications should not occur in the main text but concern other paratextual elements. By far the most common practice is the insertion or removal of a preface or an epilogue. This class therefore includes both (a) texts where the preface and/or appendix from the original work has been translated, and (b) texts where a newly written preface and/or appendix has been added instead of translating an existing one. Importantly, we can distinguish between the two sub-types based on their different visual representations. A text that does not contain new paratext will be represented as an unbroken diagonal running from the top left to the bottom right. In contrast, a text with a newly added preface or epilogue will have a

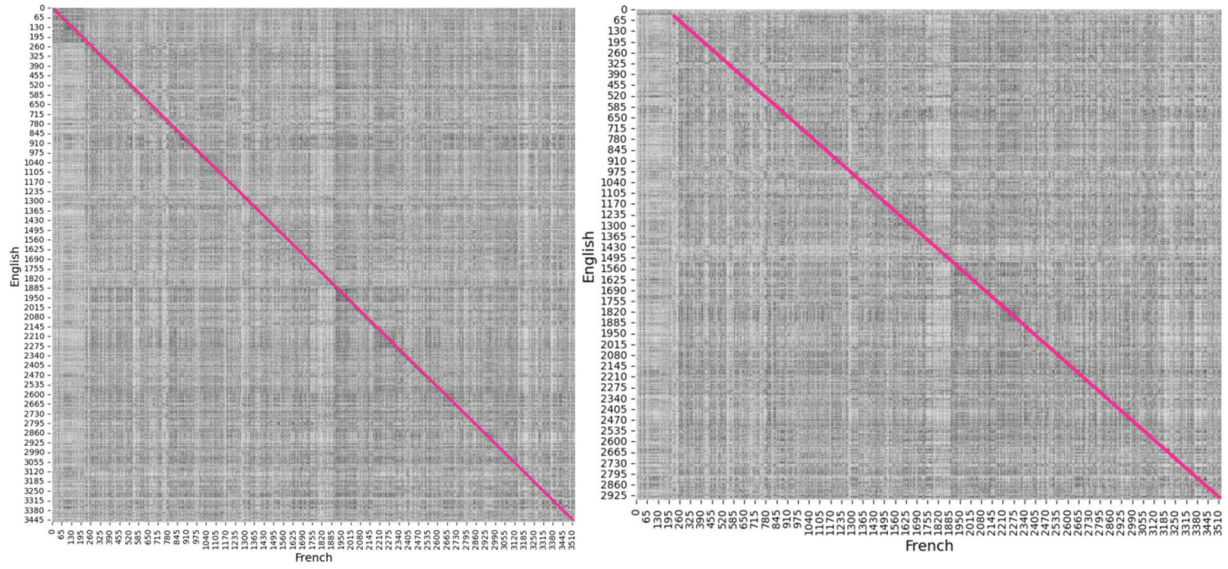

Figure 7. Visual representation of two single-volume translations of Fénelon's *Aventures de Télémaque*. Left: translation that preserves the work's original structure and paratext. Right: offset in the diagonal indicating the removal of the French-language preface.

gap at the beginning or the end and the diagonal will be offset by the length of the inserted paratext.

Being able to track whether these paratexts were translated or written by the translator or editor can give valuable insights into how a work fitted—or could be made to fit—into a different linguistic and cultural context. Prefaces were often a means for Early Modern translators to justify their decisions and to comment on the relevance of the text for its new audience (Coldiron 2019). While the importance of paratext in Early Modern literary culture has been recognized by the recent scholarship, many such studies operate in a monolingual context (see Readioff 2021) and studies focusing on the relationship between paratext and translation tend to analyze individual works or editions (e.g. Belle 2018). Computational methods can capture modifications of paratexts on a large scale, enabling a large-scale study of translation as a site of political negotiation and revealing evolving trends in literary and ideological framing. Our approach can provide a more scalable means for paratext analysis, thus contributing to a more nuanced understanding of the role and agency of translators and editors.

The case of François Fénelon's *Aventures de Télémaque* (1696) illustrates how translators engaged with shifting political contexts through their editorial decisions. Figure 7 shows two single-volume translations of *Télémaque*. The diagonal on the left indicates that Fénelon's work was translated without any structural changes to the original text or paratext, while the right has an offset that indicates two simultaneous changes: the removal of the French preface and the insertion of a new, shorter English-language preface. Having identified these translation practices through their differing visual representations, the researcher can subsequently consult the primary sources to study the context for each translator's decision. The English translation on the left was done by Pierre Des Maizeaux, a Huguenot who took refuge in London in the first half of the eighteenth century and became an important cultural mediator between France and Britain (Thomson 2024). The translator on the right is Francis Fitzgerald, who was a contributor to the *Artist's Repository*, the first British periodical dedicated to fine arts (Roberts 1970). Des Maizeaux' translation, first published in 1742, continued to be reprinted throughout the eighteenth century. Des Maizeaux's English-language *Adventures of Telemachus* also includes a translation of the *Discours sur la poésie épique*, a defense of *Télémaque* that was originally written in French by Fénelon's Scottish-born disciple Andrew Michael Ramsay in 1716. Some scholars read Ramsay's *Discours* primarily as a political piece, indicating his support for the restoration of the Catholic Stuart dynasty to the British throne (García-Minguillán 2022). The translator's decision to retain or replace this paratextual element could therefore be seen as an equally political act. Fitzgerald, in his 1792 translation, omits Ramsay's piece in favor of a new preface praising the principles articulated in *Télémaque*. This new preface is also an adaptation to a changed political context; instead of Jacobitism, which had lost its fervor by the second half of the century, it contains a reference to the French Revolution.

These two ways of translating *Télémaque* are a microcosm of how prefaces and other paratexts were

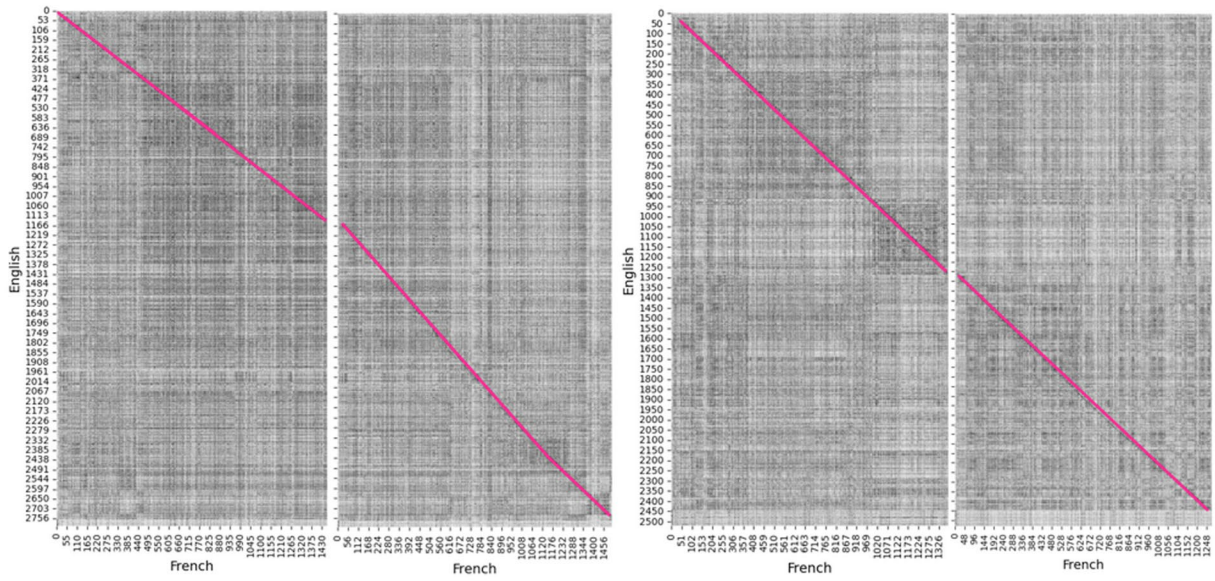

Figure 8. Side-by-side alignment of the 2-volume 1719 English edition of *Robinson Crusoe* and the 4-volume 1817 French translation. The diagonals indicate that one French volume only contains half of the text in the corresponding English volume, and that all four volumes translate the English texts without structural modifications.

used as a soapbox by Early Modern translators: this domain knowledge allows for a broader investigation into whether shifts in translation practices correspond to political changes, such as declining Jacobite influence or growing Revolutionary sentiment, showing how translations reflect the moment in which they produced. Here, our translation mining method and the researcher's domain knowledge work in harmony to capture how individual examples of paratextual modifications can thus signify meso-level or macro-level cultural or political shifts.

Class 2: Volume Reconfiguration

Class 2 largely consists of translations where the number of volumes differs between the original and translation. As is the case with Class 1, individual volumes can contain a newly written preface or epilogue, while the overall structure of the original text is retained. Compared to modifications made to paratexts, which are often interventions based on the content of the text, a change in the number of volumes tends to be motivated by economic concerns. These shifts in book structure reflected broader trends in the book market and printing industry; for example, translators and publishers could modify a text's material form to align with audience preferences or production constraints. Therefore, a single-volume original text could be split into multiple volumes by the translator, or a multi-volume original translated as a single volume.

One example of a differing number of volumes is the 1817 French translation of Daniel Defoe's *Robinson Crusoe*. The 1719 English edition consists of two volumes, while the French translation is stretched to four volumes. When putting all four visual representations of the 1817 translation side-by-side, as done in Figure 8, one can see how both English volumes were translated to French without any large-scale, structural changes.

Having discovered a change in volume numbers, a researcher could return to the entire dataset and look at all the French translations that were matched to this particular edition of *Robinson Crusoe*. Although not all of these matches will signify intellectual descent, it is still possible to draw inferences into how *Robinson Crusoe* was translated into French across the eighteenth century and in later periods, perhaps in response to changing demands of the reading public. This is not to suggest that there is a linear relationship between the number of volumes and the price of a work; on the British book market, for instance, cheap multi-volume editions of Shakespeare plays existed alongside far less affordable multi-volume editions of David Hume's *History of England* (see e.g. Hume 2014; Tiihonen, Lahti, and Tolonen 2024).

The modifications highlighted by Class 2, therefore, offer insight into broader book market trends. Changes in volume numbers may reflect commercial strategies, shifting reading habits, or production constraints. In addition, combining the results of Class 2 with an analysis of the other classes can enable a systematic study of how the materiality of a work was transformed during the translation process. Such an analysis could lead to a deeper understanding of publishing

and readership practices on the Early Modern book market.

Class 3: Collections and Compilations

Many examples in this class consist of the collected works of a single, usually well-known, author. The prevalence of collected works in translation could be used as a proxy for the popularity or canonicity of an author, since this kind of translation was most likely shaped by economic concerns and responsive to market trends: e.g. whether a lesser-known author might prove enough of a draw for readers to buy a collected edition of their works. And yet, despite their important status in the Early Modern book market, collected editions of an author's oeuvre tend to be studied much less than editions of a single text (Cahn 2004). This disparity is without a doubt even more stark for translations of collected works. Our taxonomy (and Class 3 in particular) can provide a valuable framework for investigating how economic and literary factors shaped the reception of authors across linguistic and cultural boundaries. For instance, certain authors might have their whole oeuvre translated as part of a collected edition, whereas others, though perhaps no less productive in terms of output, might only have one or two works that were translated. Individual collected editions can thus speak to broader patterns of literary appreciation or national taste.

In addition to collected works, Class 3 also contains a genre which largely sat at the opposite end in terms of price and prestige: prose miscellanies. Heterogeneous in nature, these miscellanies could consist of any writings that were popular at a certain moment or be organized around a specific theme or purpose: e.g. travel literature, entertaining stories, and tales for teaching children moral behavior (e.g. Benedict 2003). Although there exists a longstanding scholarly tradition concerned with recovering Early Modern reading habits beyond the intellectual elite (e.g. Darnton 1971; Jackson 2004), prose miscellanies—and especially their transfer across linguistic boundaries—are comparatively understudied. One reason for this lacuna could be that, thus far, it has been difficult to determine the origin of a text in a miscellany, as the author is often not indicated. The editor in question might have been seeking to bolster their own originality rather than admitting to having pilfered from elsewhere. Our translation mining approach can help to recover the source texts of such collections across languages, enabling a large-scale investigation of how authors' ideas spread and were transformed in response to the changing tastes and demands of a burgeoning readership.

These compilations can also give insights into how genres were understood differently in Britain and France. One such example is the Early Modern fairy tale: the French *contes de fées* could be highly subversive tales which criticized *Ancien Régime* politics and society (Defrance 2006). Its British equivalent, by contrast, is much tamer: concerned only with the moral education of children rather than social commentary (Grenby 2006). One of the most popular French fairy tale authors was Marie-Catharine Le Jumel de Barneville (1650/51-1705), commonly known as Mme d'Aulnoy. Her works were translated into English at the turn of the eighteenth century and remained very popular: in their bibliographic analysis, Palmer and Palmer (1974) list two collections that were reprinted multiple times. The results from our translation mining project suggest that Mme d'Aulnoy's fairy tales might have been even more widely appreciated in eighteenth-century Britain than previously thought, appearing in collections whose titles give no indication that her work is included. For example, a collection called *The Mirror, or Fairy World Displayed* (1785), contains a translation of d'Aulnoy's *Rameau d'Or*, but nowhere is it indicated that this text is in fact a translation or that d'Aulnoy is the real author (Figure 9). The editor's preface likewise avoids any mention that the collection includes texts that were originally published in French. At an author-level, a scholar could now focus on whether d'Aulnoy's works were plundered systematically or whether specific fairy tales found favor with British editors, which could give insights into her relevance as an author, as well as the judgements of editors about their audience.

Such a finding also opens the door for a broader study of attribution practices in translated collections. It could, for example, be used as a jumping-off point for a more wide-ranging analysis, e.g. whether (not) mentioning the original author of a translated piece was a more common practice for particular genres or types of collections, and how this correlated with perceived literary prestige and notions of originality and authorship.

Class 4: Omission/Insertion

Class 4 consists of translations where there is a significant modification of the original textual structure, e.g. the omission of a section or the insertion of different material between translated parts. These practices could signify editorial interference in response to socio-political pressures (e.g. the removal of sections

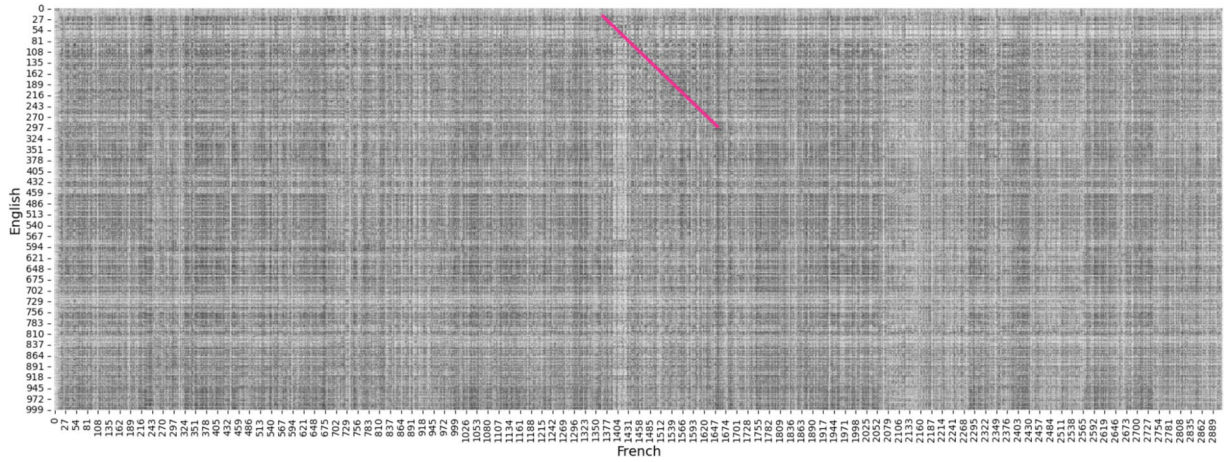

Figure 9. Translated segment of d'Aulnoy's *Rameau d'Or* in *The Mirror, or Fairy World Displayed* (1785). This unattributed translation is captured by our translation mining approach.

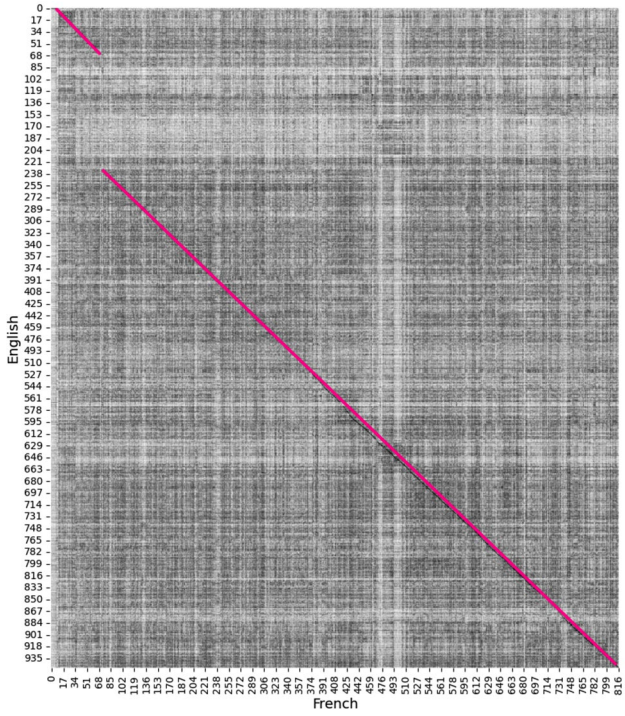

Figure 10. *A Speculative Sketch of Europe* (1798), an anonymously published translation of Charles-François Dumouriez' *Tableau spéculatif de l'Europe* (February 1798). The break in the diagonal after the beginning indicates a lengthy insertion of original English-language material.

deemed to be offensive or subject to censorship) or an adaptation to the new cultural context and local tastes, e.g. removing parts thought to be uninteresting to local readers or inserting a commentary that explains aspects specific to the culture of the original work. Such "transformissions" (Coldiron 2019) show that Early Modern translators did not merely regard themselves as transmitters of the thoughts of others. Instead, translators saw themselves as active participants in shaping texts, and as authors and thinkers in their own right. While the prevalence of adaptation as a translation strategy in the eighteenth century is well-known, the previous scholarship has analyzed the phenomenon largely in connection with the development of the novel, focusing on how translators and editors modified the works of prominent authors (e.g. McMurran 2000). Our translation mining approach offers a new way to approach this phenomenon by enabling scholars to move beyond individual, well-known texts. By combining an analysis of Class 4 with the other classes of our taxonomy, scholars can gain a broader and deeper understanding of how

historical agents decided to translate texts and how the socio-political context affects the translation process. For instance, Charles-François Dumouriez' *Tableau spéculatif de l'Europe*, published in February 1798, was rapidly translated into English in two different fashions. Silas James' translation, entitled *The Political State of Europe Speculatively Delineated* (1798), preserves the structure of the original text in its entirety and belongs to Class 1 of our taxonomy. In contrast, another translation, *A Speculative Sketch of Europe* (1798), published anonymously in Dublin, is a clear example of a Class 4 translation. As can be seen in Figure 10 there is a lengthy insertion after the beginning.

The insertion, following directly after the translated preface, is an English-language original text likely written by the translator himself. It is a critique of Dumouriez' chapter on Britain, and especially his description of the Directory's failed expedition to Ireland in an attempt to stir up a rebellion against the Crown. The translator justifies this intervention by arguing that it is his duty as an Englishman to "rescue the National Character from degradation, and to endeavor to obviate the effect of willful misrepresentations" (p. 36). This insertion is a clear example of a translator who did not merely seek to convey the ideas of the original text to a new audience, but to engage in a dialogue and to challenge cultural assumptions present in the original text.

The second French edition, the *Nouveau tableau spéculatif*—likely published concurrently with the English translations of the first edition in late 1798—was considerably expanded and modified by Dumouriez. These changes are also made visible at scale by our translation mining approach, namely by comparing the second French edition to the English translation of the first edition (Figure 11). The visual representation of this book-pair alignment clearly shows Dumouriez' new insertions, as the diagonal is more broken up compared to that of the first French edition.

Dumouriez likely made all the changes in response to the rapidly shifting political situation of the Revolutionary Wars. For example, the first break in the diagonal represents a newly inserted discussion of the French Directory's invasion of Egypt, which commenced in the summer of 1798 and was led by Napoléon Bonaparte. A scholar can thus easily see how quickly works might be considered "obsolete" and in need of an update, giving insights into Early Modern authors and translators as commercial agents.

Class 5: Fragment

Class 5 consists of translated fragments (e.g. individual sentences, short paragraphs) that are inserted into a text which primarily consists of non-translated material. Acknowledgement of the act varies: some borrowers name the original source; others pretend that the entire text flowed out of their own pen. The origin

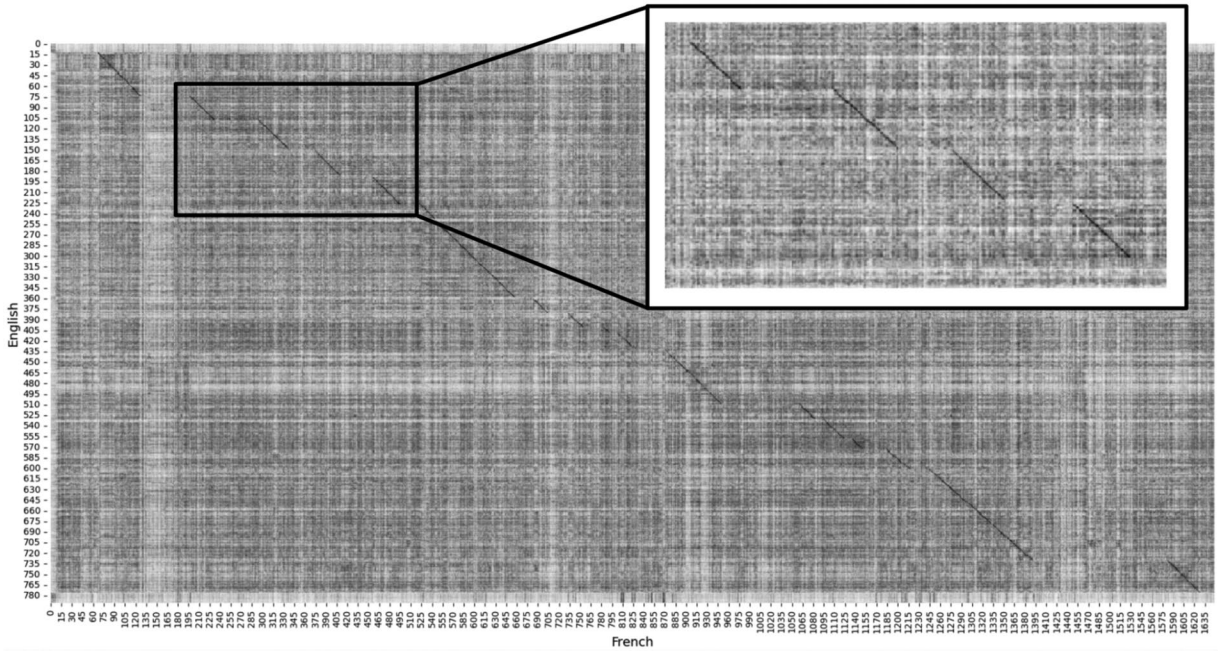

Figure 11. Silas James' English translation of the first edition of Dumouriez' *Tableau spéculatif* (Feb 1798) aligned with the second French edition *Nouveau tableau* (Late 1798). The breaks in the diagonal indicate new material in the second French edition.

of borrowed snippets and their prevalence compared to the original material can give insights into Early Modern quotation practices, as well as the intended audience of a text. For instance, quoting a classical author in Latin (and especially Greek) rather than using a vernacular translation likely signifies that the text is aimed at a more scholarly audience. In contrast, a preference for translation, especially of authors that educated eighteenth-century gentlemen would have read in the original language, could indicate that the text was meant to be accessible to an audience beyond the intellectual elite. Individual instances of translating a classical author reflect meso-level cultural assumptions about learnedness. Therefore, this translation practice reveals important macro-level historical dynamics in knowledge dissemination, particularly in terms of how ideas were selectively made accessible to different social groups.

These snippets often—but by no means always—originate from the works of authors that were well-known and appreciated in the eighteenth century. When the source is openly acknowledged, the borrower tends to deploy the translated fragment for a specific rhetorical purpose, e.g. citing an authority to support their own position. Citation is never a neutral phenomenon but always grounded in a rhetorical intent. Our method helps with the recovery of such moves, since it displays the two texts side-by-side, allowing the researcher to see the original context as well as how the borrower integrates the fragment into a different context. At the meso-level, these practices can give insights into which authors and works were considered authoritative in the eighteenth century.

Unacknowledged borrowing might be more common in certain genres, e.g. histories and encyclopedias, where authors would want to emphasize their originality. Existing computational analyses of eighteenth-century encyclopedias have largely concentrated on well-known projects like the *Encyclopédie* and have tended to identify French-language borrowings only (Allen and Cooney 2010; Edelstein, Morrissey, and Roe 2013). Studies of transnational knowledge transfers have likewise tended to use a specific dictionary as a unit of analysis (Donato and Lüsebrink 2021). In contrast, our translation mining approach operates on the level of two full-text collections, which enables studying these borrowing practices across text genres and enables the discovery of previously unknown and unindicated acts of appropriation across disciplines and contexts.

Even when the original author is indicated, our translation mining approach can help to discover new connections between authors and illustrate the spread of ideas to unexpected places. For instance, inserting a translated sentence from the *Esprit des Lois* was not uncommon in eighteenth-century British political thought, where Montesquieu remained one of the key figures (Israel 2023). However, this method also reveals that his work appeared in less predictable contexts. It is an unexpected but welcome surprise to find Montesquieu not only in sober, abstract, discussions of political systems, but in the depths of local party politics: *Euphrasy*, an anonymous Tory comment on the loss in Ipswich in the general election of 1768 (Figure 12). The visual representation of the texts' alignment shows the length discrepancy between the *Esprit des Lois* and the much shorter *Euphrasy.* This act of translation is an example of how Montesquieu's ideas of government could be weaponized for concrete political purposes, such as delegitimizing election results and attacking opposing political factions.

The main thrust of *Euphrasy* is an attack on Dissenters (i.e., Protestants who were not members of the Church of England), whom the author holds responsible for all societal ills. To save King, Church, and County, the reader should heartily embrace the author's "True Blue" Tory principles. This party-political

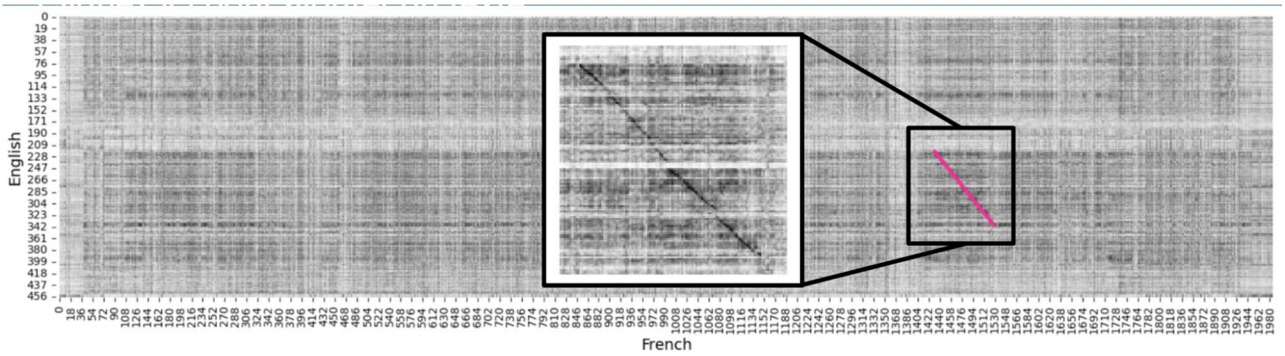

Figure 12. Translated fragment of Montesquieu's *Esprit des Lois* in the Tory pamphlet *Euphrasy* (1768). The diagonal, compared to its surroundings, indicates notable difference in text length and how Montesquieu was embedded in the English text.

harangue is buttressed by a translation of Montesquieu's description of the English constitution as a mixed government of aristocracy, monarchy, and democracy. The author's decision to employ a translation rather than the original French is, at first, a strange one, since the chapter epigraphs are untranslated Latin quotations drawn from the works of prominent ancient authors like Cicero and Horace. However, the author explains that they have provided their own translation of Montesquieu to allow "the more unlettered part of my countrymen" to weigh in on important political topics such as the English constitution (p. 26). These unlettered countrymen would have to have sufficient means to buy a pamphlet such as *Euphrasy* but could not read French or afford a complete English translation of a lengthy work like the *Esprit des Lois*. Shorter translated fragments thus formed an important means of distributing ideas to a wider audience than those who could read the original or buy a complete translation. In theory, it would have been possible to identify the origin of this translated fragment without recourse to computational methods, as the author of *Euphrasy* gives a precise reference to the specific chapter of the *Esprit des Lois*. However, in practice, the chance of a researcher being aware of and consulting a publication as niche as *Euphrasy* is vanishingly slim, let alone to suspect that the work employed a translation of Montesquieu to support its party-political argument in an English local election.

Being able to identify this type of translation at scale allows scholars to, for example, gain a more extensive understanding of eighteenth-century rhetorical practices, e.g. who was regarded as a work's ideal audience, which authors and works were commonly used for appeals to authority, as well as why and in which contexts an author might have preferred a translation over the original.

Class 6: Multi-Fragment

The sixth class contains perhaps the most heterogeneous set of results, covering instances where multiple fragments of a text have been translated. The order in which these fragments have been translated also differs significantly from the structure of the original text. The heterogeneity of this class reflects the intellectual flexibility and agency of Early Modern translators. It shows how translators freely rearranged, omitted, and adapted fragments ranging from sentences to entire essays - practices that reveal a permissive attitude toward textual borrowing and a dynamic engagement with source materials. The heterogeneity of this class reflects the Early Modern period's emphasis on the agency of the translator and its permissive attitudes toward borrowing fragments from other texts, ranging from sentences to entire essays.

Essays were a popular genre of Early Modern writing, their structure often closely influenced by pioneering authors like Michel de Montaigne and Francis Bacon, but also developing into new forms, such as becoming a regular feature in journals and newspapers in the eighteenth century (e.g. Black 2006). As essays could be both highly topical and concerned with timeless aspects of the human condition, they were often considered fertile grounds for borrowing and repurposing. The wholesale reprinting of individual essays and monolingual reuse can be identified with existing computational methods for text reuse detection (Kivistö 2022; Spencer and Tolonen 2025), whereas cross-lingual transfers have thus far been far more difficult to detect. It is therefore hardly surprising that multi-fragment translations of essays and other works have received considerably less scholarly attention than translations that closely preserve the structural integrity of the original text. This holds true even for multi-fragment translations whose nature is clearly indicated in the title of the work, as well as leading eighteenth-century thinkers whose oeuvre is otherwise well studied. Yet, these multi-fragment translations played a crucial role in shaping intellectual discourse, as they allowed translators to tailor Enlightenment ideas to specific national and political contexts.

For instance, research on how Benjamin Franklin's ideas were received and transformed in France at the turn of the eighteenth century tends to utilize more structurally conservative translations of his works. The early translation history of Franklin's autobiography is particularly checkered, hampered by the incomplete original manuscript (e.g. Arch 2009). The first complete edition of the autobiography was not published until 1828, not in English but in French: Augustin-Charles Renouard's *Mémoires sur la vie de Benjamin Franklin* (Hunter 2007). In addition to the *Mémoires*, Renouard also published a more eclectic French-language collection of Franklin's writings, the *Mélanges de morale, d'économie et de politique* (1824). As indicated by the title, the *Mélanges* contain extracts taken from a number of Franklin's writings, including his many essays (Figure 13).

The multi-fragment nature of the translation perhaps helps to explain why the work has received little attention in recent scholarship compared to the *Mémoires*. Hunter, in his careful analysis of the

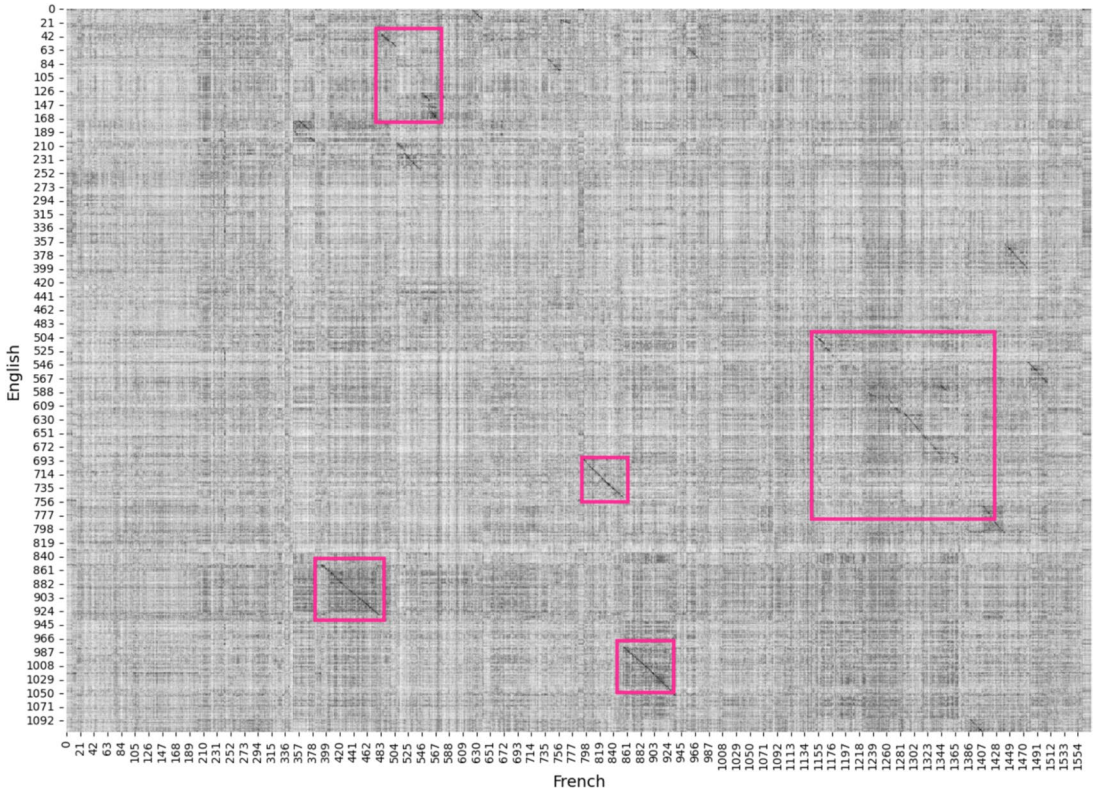

Figure 13. Comparison of the 1799 edition of the *Works of the Late Doctor Benjamin Franklin* and the 1824 French compilation *Mélanges de morale, d'économie et de politique*. The fragments translated into French are clearly identifiable as short diagonals.

Mémoires, merely mentions Renouard's authorship of the "relatively popular" *Mélanges* (Hunter 2007, 491). The older Franklin scholarship offers more insights into its contents: according to John McBride (1956, 118), Renouard's "forty-three selections cover all aspects of Franklin's writings except the scientific". And yet, McBride stops short of conducting a more in-depth analysis of the origin of these extracts.

Our translation mining results can help to provide new insights into the dissemination and reception of essays, e.g. which topics were considered to be of interest to readers and why certain essays might have been omitted or rearranged by the translator for the sake of the local readership. For instance, the French chapter title "La Science du bonhomme Richard, ou le Chemin de la fortune" indicates that the essay was taken from Franklin's *Poor Richard's Almanack*. Undoubtedly the most popular and influential almanack to come out of Colonial America, *Poor Richard* was published yearly between 1733 and 1758 (Miller 1961). Without further information, such as the publication year, it is difficult to identify the edition where the essay originated, unless one is already intimately familiar with the translation history of *Poor Richard*. Our translation mining results show that the "Science du bonhomme Richard" is the preface to the 1758 edition of the almanack, later known in English as "The Way to Wealth". While it is widely interpreted as a foundational text of American capitalism (Reinert 2015), its strategic placement within the Mélanges suggests that Renouard instead sought to highlight the moral dimension of Franklin's economic philosophy.

Our results show that Renouard arranged the translated essays by a different topic structure compared to the late eighteenth-century English editions of Franklin's works. Our algorithm matched Renouard's *Mélanges* to the 1799 edition of the *Works of the Late Doctor Benjamin Franklin,* published in two volumes. In the Millar edition, the preface to the 1758 *Poor Richard* is preceded by a high-level summary of eighteenth-century economic theory and followed by an essay describing a nascent American identity. Renouard does not translate either of these two essays, perhaps due to a perceived lack of relevance, and instead places the "Science du bonhomme Richard" between an essay giving advice on personal saving and a letter discussing how the wrong motives such as greed and ambition lead to unhappiness. Embedded in this different context, the "Way to Wealth" is transformed from a theory of capitalism to a piece emphasizing how an individual's relationship with money relates to morality.

The example of Renouard's multi-fragment translation of Benjamin Franklin's work has demonstrated

how Early Modern translators not only mediated knowledge but actively reshaped political and economic ideas; they did so by recontextualizing fragments and radically changing the structure of the original work to alter the message conveyed by the text. These individual acts of recontextualization thus become fertile ground for meso- and macro-level studies of political thought, political economy, morality, and broader cultural and intellectual currents of the eighteenth century.

5. Conclusion

This paper has introduced translation mining as a new approach that offers access to eighteenth-century French and English language translations at unprecedented scale and bridges the reach of quantitative methods with the contextualization of qualitative methods.

We have shown that deep learning–based AI methods emerging from NLP research in meaning representations can be applied effectively to large historical corpora across different languages. Our results indicate that such approaches can tolerate both OCR errors and the linguistic distance between modern embeddings and historical texts, enabling cross-lingual text correlation more robustly than earlier, primarily lexical models. After embedding each document into a semantic space, further analyses (like identifying cross-lingual matches, applying a translation taxonomy, or refining sub-segment matches) can be performed repeatedly without re-running the entire process. However, implementing these methods at scale remains computationally intensive and generally requires substantial technical resources. At present, this makes them more suitable for large, one-time projects generating re-usable data rather than for quick, ad hoc inquiries by individual scholars (for an earlier similar project on text reuse, see Rosson et al. 2023).

In addition to widening access to the material, our approach offers scholars a structured way of analyzing it through a taxonomy of Early Modern translation practices. This taxonomy is developed inductively from the systematic detection of textual similarities and modifications—a bottom-up, Linnaean-style model of classification grounded in observable patterns rather than predetermined categories. Its strength lies in allowing patterns in eighteenth-century translation to guide the formation of classes that remain provisional, flexible, and open to revision as new evidence emerges. In this way, the taxonomy complements existing humanistic methods by showing how empirically derived classifications can support historical interpretation and provide a reusable framework for future work on cultural transfer and translation practices.

Taken together, these results show that translation mining supports historical inquiry at three interlocking levels. At the micro-level, it allows historians to recover individual acts of translation, omission, and insertion that would otherwise remain hidden—such as the insertion of Montesquieu into local electoral pamphlets or the recontextualization of Franklin's economic thought. At the meso-level, it becomes possible to reconstruct translation regimes around specific authors, genres, controversies, or translators' careers, and to examine how conservative and modifying practices cluster within these more manageable units. At the macro-level, the same data reveal broader shifts in eighteenth-century cultural transfer between France and Britain, such as the relative prevalence of fragmentary versus full-volume translations over time. In this sense, translation mining acts as an infrastructure across humanities research: it provides the connective tissue between close reading and system-level claims.

By making this intermediate, meso-level systematically visible, our approach helps to bridge the gap between studies of canonical figures and work on popular, ephemeral print. Translation practices such as unattributed snippets in miscellanies or rearranged essay fragments cease to be anecdotal curiosities and become part of a structured landscape of cultural transfer, one that can be explored both quantitatively and through targeted case studies. In the future, integrating questions of genre and book formats are natural next steps to carry the work further (Lahti et al. 2019; Tiihonen and Hinderks 2025; Zhang et al. 2022).

Finally, it is important to recognize that the rhythms of humanities research and methodological innovation in NLP are not aligned. While computational models evolve rapidly, historical inquiry typically proceeds through slower cycles of interpretation, validation, contextualization, and publishing. The methods employed here are therefore chosen not because they represent the very latest stage of NLP development, but because they are sufficiently robust and transparent to support sustained historical analysis. This asymmetry places a premium on approaches that remain usable as tools change, but it also underscores a more basic point: methodological choices must be judged by their capacity to generate meaningful historical insight in the present. Advances in technique matter insofar as they enable us to answer substantive questions about the past, rather than

deferring significance to future computational improvements.

Another crucial aspect of this work is the necessity to rely on supercomputing resources. The computationally most intensive step is calculating the embeddings, which amounted to less than 10,000 GPU hours, which even at peak consumption corresponds to 5,000 kWh of energy, less than one family's yearly consumption in Finland. In this work, we were able to use the Finnish national supercomputing resources, which are hosted in a datacenter using 100% renewable energy sources and waste heat is reused in district heating. (CSC - IT Center for Science, 2025) Within the EU, researchers have access to similar supercomputing resources through the EuroHPC resource pool. The primary dataset comprising the extracted translation pairs is made available, removing the need to replicate our runs by subsequent users. It is nevertheless clear that the application of present-day NLP methodology to any field with large textual collections, history included, requires both supercomputer-level computational resources and the technical skill to use them.

When it comes to future work, given the recent developments in NLP, it can be reasonably expected that repeating our study with the latest available embedding models would sharpen the distinction between matches and near-matches, allowing the retrieval of even shorter passages and translation pairs that have possibly been missed. Continual advances in NLP mean that the latest cross-lingual embedding models already surpass those used in this study (see, e.g., Lee et al. 2024). What matters for the humanities is that methodological development and historical research are mutually formative: humanities questions guide computational work, while new data enable concrete analyses that reimagine the cross-lingual cultural landscape and trace the forms of influence that moved across it. In most cases, the ultimate bottleneck remains access to data—but that is a different story altogether.

Acknowledgements

We gratefully acknowledge the members of the HPC-HD project, COMHIS, and TurkuNLP for their support. We also extend our thanks to GALE for providing access to the ECCO data, and to the National Library of France (BNF)—with special appreciation to Jean-Philippe Moreux—for making the Gallica data available. We would also like to thank Anna Plassart for comments on an earlier draft of this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Finnish Research Council under grant numbers 1333716, 1347706, 347708. Computational resources were provided by CSC - IT Center for Science, Finland.

Data availability statement

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.14514300

References

Allen, T., and C. M. Cooney. 2010. Plundering philosophers: Identifying sources of the encyclopédie. *Journal of the Association for History and Computing* 13 (1). http://hdl.handle.net/2027/spo.3310410.0013.107.

Arch, S. C. 2009. Benjamin Franklin's autobiography, then and now. In *The Cambridge Companion to Benjamin Franklin*, ed. Carla Mulford, 159–71. Cambridge Companions to American Studies. Cambridge: Cambridge University Press. doi:10.1017/CCOL9780521871341.013.

Belle, M.-A. 2018. *Thresholds of translation: Paratexts, print, and cultural exchange in early modern Britain (1473-1660).* New York, NY: Springer Berlin Heidelberg.

Benedict, B. M. 2003. The paradox of the anthology: Collecting and Différence in eighteenth-century Britain. *New Literary History* 34 (2):231–56. doi:10.1353/nlh.2003.0013.

Black, S. 2006. *Of essays and reading in early modern Britain*. Palgrave Studies in the Enlightenment, Romanticism and Cultures of Print. Basingstoke (GB) New York: Palgrave Macmillan.

Bode, K. 2020. Why you can't model away bias. *Modern Language Quarterly* 8 (1):95–124. doi:10.1215/00267929-7933102.

Cahn, M. 2004. Opera omnia: The production of cultural authority. In *History of science, history of text*, ed. Karine Chemla, 81–94. Boston Studies in Philosophy of Science. Boston: Kluwer Academic.

CSC - IT Center for Science. 2025. Sustainability report for 2024. https://csc.fi/app/uploads/2025/04/CSC_Sustainability_report_2024.pdf

Coldiron, A. E. B. 2019. Translation and transformission; or, early modernity in motion. *Canadian Review of Comparative Literature / Revue Canadienne de Littérature Comparée* 46 (2):205–16. doi:10.1353/crc.2019.0018.

Darnton, R. 1971. Reading, writing, and publishing in eighteenth-century France: A case study in the sociology of literature. *Daedalus* 100 (1):214–56.

de Bolla, P., E. Jones, P. Nulty, G. Recchia, and J. Regan. 2020. The Idea of Liberty, 1600–1800: A Distributional

Concept Analysis. *Journal of the History of Ideas* 81 (3):381–406. doi:10.1353/jhi.2020.0023.

Defrance, A. 2006. La politique du conte aux XVIIe et XVIIIe siècles. *Féeries* 3 (February):13–41. doi:10.4000/feeries.137.

Donato, Clorinda, and Hans-Jürgen Lüsebrink, eds. 2021. *Translation and Transfer of Knowledge in Encyclopedic Compilations, 1680-1830*. The UCLA Clark Memorial Library Series. Toronto; Buffalo; London: University of Toronto Press.

Douze, M., A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou. 2024. "The Faiss Library". arXiv. https://arxiv.org/abs/2401.08281

Dow, G. 2014. Translation, cross-channel exchanges and the novel in the long eighteenth century. *Literature Compass* 11 (11):691–702. doi:10.1111/lic3.12183.

Düring, M., M. Romanello, M. Ehrmann, K. Beelen, D. Guido, B. Deseure, E. Bunout, J. Keck, and P. Apostolopoulos. 2023. impresso text reuse at scale. An interface for the exploration of text reuse data in semantically enriched historical newspapers. *Frontiers in Big Data* 6:1249469. doi:10.3389/fdata.2023.1249469.

Edelstein, D., R. Morrissey, and G. Roe. 2013. To quote or not to quote: Citation Strategies in the 'Encyclopédie. *Journal of the History of Ideas* 74 (2):213–36. doi:10.1353/jhi.2013.0012.

Eijnatten, J. v, and R. Ros. 2019. The eurocentric fallacy. A digital-historical approach to the concepts of 'Modernity', 'Civilization' and 'Europe' (1840–1990). *International Journal for History, Culture and Modernity* 7 (1):686–736. doi:10.18352/hcm.580.

Ferrand, N. 2014. Translating and illustrating the eighteenth-century novel. *Word & Image* 30 (3):181–93. doi:10.1080/02666286.2014.938522.

Gage, P. 1994. A new algorithm for data compression. *The C Users Journal Archive* 12:23–38.

Galambos, C., J. Matas, and J. Kittler. 1999. Progressive probabilistic hough transform for line detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), 1:554-560 Vol. 1 doi:10.1109/CVPR.1999.786993.

García-Minguillán, C. 2022. Andrew Michael Ramsay's defence of Fénelon: Classical poetics or political discourse? *Romance Studies* 40 (2):61–75. doi:10.1080/02639904.2022.2114627.

Gladstone, C., and C. M. Cooney. 2020. Opening new paths for scholarship: algorithms to track text reuse in ECCO. In *Digitizing enlightenment: Digital humanities and the transformation of eighteenth-century studies*, ed. Simon Burrows and Glenn Roe, 353–74. Oxford: Voltaire Foundation.

Grenby, M. O. 2006. Tame Fairies make good teachers: The popularity of early British Fairy Tales. *The Lion and the Unicorn* 30 (1):1–24. doi:10.1353/uni.2006.0006.

Hume, R. D. 2014. The value of money in eighteenth-century England: Incomes, prices, buying power—and some problems in cultural economics. *Huntington Library Quarterly* 77 (4):373–416. doi:10.1525/hlq.2014.77.4.373.

Hunter, C. 2007. From print to print: The first complete edition of Benjamin Franklin's 'Autobiography. *The Papers of the Bibliographical Society of America* 101 (4):481–505. doi:10.1086/pbsa.101.4.24293662.

Israel, J. 2023. Montesquieu and the enlightenment. In *The Cambridge companion to Montesquieu*, ed. Keegan Francis Callanan and Sharon Ruth Krause, 266–88. Cambridge Companions to Philosophy. Cambridge: Cambridge University Press. doi:10.1017/9781108778923.016.

Jackson, I. 2004. Approaches to the history of readers and reading in eighteenth-century Britain. *The Historical Journal* 47 (4):1041–54. doi:10.1017/S0018246X04004091.

Kivistö, M. 2022. Retaining popularity: studying eighteenth-century prominence of the spectator with computational methods. Master's thesis., Helsinki, Finland: Helsingin yliopisto. http://hdl.handle.net/10138/343515.

Kulkarni, V., R. Al-Rfou, B. Perozzi, and S. Skiena. 2015. Statistically significant detection of linguistic change. *Proceedings of the*. 24th International Conference on World Wide Web: 625–635. doi:10.1145/2736277.2741627.

Lahti, L., J. Marjanen, H. Roivainen, and M. Tolonen. 2019. Bibliographic data science and the history of the book (c. 1500–1800). *Cataloging & Classification Quarterly* 57 (1):5–23. doi:10.1080/01639374.2018.1543747.

Lee, C., R. Roy, M. Xu, J. Raiman, M. Shoeybi, B. Catanzaro, and W. Ping. 2024. NV-embed: improved techniques for training LLMs as generalist embedding models. arXiv. https://arxiv.org/abs/2405.17428.

Mahadevan, A., M. Mathioudakis, E. Mäkelä, and M. 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):4631–43. doi:10.1007/s41060-025-00742-x.

McBride, J. 1956. Benjamin Franklin as Viewed in France during the Bourbon Restoration (1814-1830). *Proceedings of the American Philosophical Society* 100 (2):114–28.

McMurran, M. H. 2000. Taking Liberties: Translation and the Development of the Eighteenth-Century Novel. *The Translator* 6 (1):87–108. doi:10.1080/13556509.2000.10799057.

Miller, C. W. 1961. Franklin's 'Poor Richard Almanacs': Their printing and publication. *Studies in Bibliography* 14:97–115.

Muennighoff, N., N. Tazi, L. Magne, and N. Reimers. 2023. MTEB: Massive Text Embedding Benchmark. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2014–2037. Dubrovnik, Croatia: Association for Computational Linguistics. doi:10.18653/v1/2023.eacl-main.148.

Müller-Wille, S. 2007. Collection and collation: Theory and practice of linnaean botany. *Studies in History and Philosophy of Biological and Biomedical Sciences* 38 (3):541–62. doi:10.1016/j.shpsc.2007.06.010.

Palmer, N., and M. Palmer. 1974. English editions of French 'Contes de Fees' attributed to Mme D'Aulnoy. *Studies in Bibliography* 27:227–32.

Raven, J. 2002. An antidote to the French?: English novels in German translation and German novels in English Translation 1770-99. *Eighteenth-Century Fiction* 14 (3-4):715–34. doi:10.1353/ecf.2002.0032.

Readioff, C. 2021. Recent approaches to paratext studies in eighteenth-century literature. *Literature Compass* 18 (12):e12650. doi:10.1111/lic3.12650.

Reimers, N., and I. 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. Association for Computational Linguistics. doi:10.18653/v1/D19-1410.

Reimers, N., and I. Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. Proceedings of the 2020 conference on empirical methods in natural language processing. pages 4512–4525. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.365.

Reinert, S. A. 2015. The way to wealth around the world: Benjamin Franklin and the globalization of American capitalism. *The American Historical Review* 120 (1):61–97. doi:10.1093/ahr/120.1.61.

Roberts, H. E. 1970. British art periodicals of the eighteenth and nineteenth centuries. *Victorian Periodicals Newsletter* 9:1–183.

Roe, G., C. Gladstone, and R. Morrissey. 2016. Discourses and disciplines in the enlightenment: Topic modeling the French encyclopédie. *Frontiers in Digital Humanities* 2 . doi:10.3389/fdigh.2015.00008.

Rosson, D., E. Mäkelä, V. Vaara, A. Mahadevan, Y. Ryan, and M. Tolonen. 2023. Reception Reader: Exploring text reuse in early modern british publications. *Journal of Open Humanities Data* 9 (5):1–11. doi:10.5334/johd.101.

Salmi, H., P. Paju, H. Rantala, A. Nivala, A. Vesanto, and F. Ginter. 2021. The reuse of texts in Finnish newspapers and journals, 1771–1920: A digital humanities perspective. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 54 (1):14–28. doi:10.1080/01615440.2020.1803166.

Simonyan, K., and A. Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In 3rd International Conference on Learning Representations (ICLR 2015), 1–14. Computational and Biological Learning Society.

Smith, D. A., R. Cordell, and A. Mullen. 2015. Computational methods for uncovering reprinted texts in antebellum newspapers. *American Literary History* 27 (3):E1–E15. doi:10.1093/alh/ajv029.

Spencer, M. G., and M. Tolonen. 2025. The reception of Hume's essays in eighteenth-century Britain. In *Hume's Essays: A Critical Guide*, ed. Felix Waldmann and Max Skjönsberg, 15–35. Cambridge: Cambridge University Press. doi:10.1017/9781009047227.003.

Thomson, A. 2024. Pierre Des Maizeaux and the (Huguenot) business of translation in the early eighteenth century. In *Ideas across Borders: Translating visions of authority and civil society in Europe c.1600-1840*, ed. Gaby Mahlberg and Thomas Munck, 83–98. New York: Routledge.

Tiedemann, J. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. *Recent Advances in Natural Language Processing* V:237–48.

Tiihonen, I., and K. Hinderks. 2025. Genre classification workflow for the English short title catalogue (ESTC). *Transformations: A DARIAH Journal* 1:1–33. doi:10.46298/transformations.14754.

Tiihonen, I., L. Lahti, and M. Tolonen. 2024. Print culture and economic constraints: A quantitative analysis of book prices in eighteenth-century Britain. *Explorations in Economic History* 94 (October):101614. doi:10.1016/j.eeh.2024.101614.

Tolonen, M., E. Mäkelä, and L. Lahti. 2022. The anatomy of eighteenth century collections online (ECCO). *Eighteenth-Century Studies* 56 (1):95–123. doi:10.1353/ecs.2022.0060.

Tolonen, M., M. J. Hill, A. Ijaz, V. Vaara, and L. Lahti. 2021. Examining the early modern canon: The English short title catalogue and large-scale patterns of cultural production. In *Data visualization in enlightenment literature and culture*, ed. I. Baird, 63–119. Cham, Switzerland: Palgrave Macmillan. doi:10.1007/978-3-030-54913-8_3.

Vesanto, A., A. Nivala, H. Rantala, T. Salakoski, H. Salmi, and F. Ginter. 2017. Applying BLAST to text reuse detection in finnish newspapers and journals, 1771–1910. Proceedings of the 21st Nordic Conference of Computational Linguistics, Gothenburg, Sweden, May 23–24, 2017. Linköping: Linköping University. http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf.

Winsor, M. P. 2006. Linnaeus's biology was not essentialist. *Annals of the Missouri Botanical Garden* 93 (1):2–7.

Withers, C. W. J. 1995. Geography, natural history and the eighteenth-century enlightenment: Putting the world in place. *History Workshop Journal* 39 (1):137–64. doi:10.1093/hwj/39.1.137.

Witteveen, J. 2020. Linnaeus, the essentialism story, and the question of types. *Taxon* 69 (6):1141–9. doi:10.1002/tax.12346.

Zhang, J., Y. C. Ryan, I. Rastas, F. Ginter, M. Tolonen, and R. Babbar. 2022. Detecting sequential genre change in eighteenth-century texts. In *Proceedings of the computational humanities research conference 2022*, ed. Folgert Karsdorp, Alie Lassche, and Kristoffer Nielbo, vol. 3290. Aachen: CEUR. https://ceur-ws.org/Vol-3290/#short_paper2630.

Zhou, X., and S. Sun. 2017. Bibliography-based quantitative translation history. *Perspectives* 25 (1):98–119. doi:10.1080/0907676X.2016.1177100.

Appendix

In this appendix, we summarize technical details of our implementation which are not relevant to all readers but provide necessary detail for possible result replication efforts.

Embeddings and similarity search

To calculate the embeddings of text, we used the model paraphrase-xlm-r-multilingual-v1 from the SentenceTransformers model collection. The similarity k nearest neighbors search was carried out using the FAISS (Facebook AI Similarity Search) index, set up with the python faiss library (Douze et al. 2024). Before the Hough transform, the similarity matrices were thresholded to increase contrast.

Translation taxonomy classifier training

Translation type images, i.e., images with the drawn lines, were embedded using the VGG16 model of Simonyan and Zisserman (2015). This model embeds each image into a dense vector representation which allows image clustering by visual similarity and serves as the basis of training image classifiers.

The classifier for recognition of translation types was trained in several stages of development:

1. 586 examples were randomly selected, manually annotated for the taxonomy classes, and used to train an initial image classifier based on the VGG16 model.
2. We observed suboptimal performance on the less naturally represented classes 1–4. Further 500 images were randomly selected such that the lines span at least 70% of one of the documents and there are fewer than 5 lines in total and annotated, thus enriching the data with examples of classes 1–3. A similar process was applied to enrich the data by examples of class 4: images with lines spanning at least 50% and line count between 2 and 5 were randomly sampled and annotated. These two steps increased the size of the manually labeled dataset to 1164 examples.
3. 200 examples underwent double-annotation by two independent annotators; initial agreement rate was 89.5% and the remaining 10.5% of examples with a disagreement were jointly resolved to arrive at a single label.
4. Hyperparameters were set on a held-out portion of the training data, and the image classifier was evaluated on the test data from step (3)

Overall, the classification accuracy was 84.5% when considering the final unique ground truth after the annotation conflicts have been resolved (step 3). The accuracy increases to 88.5% if the label given by any one of the two annotators can be considered correct.