BRILL

JOURNAL OF DIGITAL ISLAMICATE RESEARCH
2 (2024) 1–29

Journal of Digital Islamicate Research
brill.com/jdir

Exploring Gulf Manumission Documents with Word Vectors

Suphan Kirmizialtin | ORCID: 0000-0001-5020-0578
Division of Arts and Humanities, New York University Abu Dhabi, United Arab Emirates
suphan@nyu.edu

David Joseph Wrisley | ORCID: 0000-0002-0355-1487
Division of Arts and Humanities, New York University Abu Dhabi, United Arab Emirates
Corresponding author
djw12@nyu.edu

Received 7 October 2024 | Accepted 26 November 2024 | Published online 27 December 2024

Abstract

In this article we analyze a corpus related to manumission and slavery in the Arabian Gulf in the late nineteenth- and early twentieth-century that we created using Handwritten Text Recognition (HTR). The corpus comes from India Office Records (IOR) *R/15/1/199 File 5*. Spanning the period from the 1890s to the early 1940s and composed of 977K words, it contains a variety of perspectives on manumission and slavery in the region from manumission requests to administrative documents relevant to colonial approaches to the institution of slavery. We use word2Vec with the WordVectors package in R to highlight how the method can uncover semantic relationships within historical texts, demonstrating some exploratory semantic queries, investigation of word analogies, and vector operations using the corpus content. We argue that advances in applied computer vision such as HTR are promising for historians working in colonial archives and that while our method is reproducible, there are still issues related to language representation and limitations of scale within smaller datasets. Even though HTR corpus creation is labor intensive, word vector analysis remains a powerful tool of computational analysis for corpora where HTR error is present.

Published with license by Koninklijke Brill BV | DOI:10.1163/27732363-BJA00005

Keywords

Handwritten Text Recognition (HTR) – word vector models (WVM) – India Office Records (IOR) – manumission – Gulf Studies – colonial archives – slavery

1 Introduction

Persian Gulf archival records, shaped by colonial processes of collection and preservation, are scattered across global institutions, reflecting historical asymmetries in power that continue to determine which narratives remain accessible.[1] Amid this fragmented and politically charged archival landscape, the Qatar Digital Library (QDL) stands as a critical access point for Gulf histories and offers an unprecedented resource for computational analysis. Within this expansive digital repository, our study focuses on a specific subset – File 5 Slave Trade (IOR/R/15/1/199-234) from the India Office Records (IOR) – comprising approximately 14,000 pages related to manumission and slavery in the Gulf during the late nineteenth and early twentieth centuries (hereafter, IOR File 5). These documents, including manumission requests recorded by British officials on behalf of enslaved and indentured individuals, offer a unique window into the lived experiences of servitude as well as the colonial discourse surrounding liberation.

Responding to scholarly calls, such as those by Zdanowski (2011), for critical engagement with manumission documents to uncover embedded forms of violence and power, we have created a machine-readable corpus of IOR File 5 using Handwritten Text Recognition (HTR) technology. Employing both word vector analysis to map thematic structures across the corpus and close reading to explore contextual layers within individual narratives, our approach bridges computational and traditional textual analysis. In balancing these computational and interpretive methods, we follow an approach informed by existing scholarship on close and distant reading, where iterative engagement with the text and the computational model helps illuminate both large-scale patterns and specific, nuanced narratives (Underwood 2019; Kirilloff 2022; Gavin 2019). This dual approach allows us to examine the language of manumission with a balance of computational scale and historical specificity, advancing the study of Gulf histories within this complex, distributed archival framework.

1 For a detailed discussion of the fractured nature and the political dynamics of archives in the Gulf region, see Bsheer 2020.

2 Word Vectors for the Study of Historical Corpora

In this article, we employ word vector analysis to examine our collection of historical documents. Word vector analysis is a computational technique that creates mathematical representations of individual words as vectors in a continuous vector space derived from the corpus. This approach approximates semantic affinities of words by mapping them into a "spatial analogy to the relationship between words" (Schmidt 2015). The underlying principle is that words appearing in similar contexts tend to have similar or related meanings (Verheul et al. 2022). These numerical representations capture relationships between words without incorporating external knowledge about their semantics. It is important to note that while similar vector contexts can indicate similar meanings, proximity in a word vector model does not always imply synonymity; it might also indicate antonyms, abbreviations, or other instances of words that appear in similar contexts.

In our study, we utilize a modified version of the word2Vec algorithm, specifically the wordVectors package (Schmidt and Li 2017), adapted by the Women's Writers Project at Northeastern University.[2] Although pre-trained vector models are available for analyzing text collections in various languages (Pennington et al. 2014; fastText 2022), we believe that the unique characteristics of our historical corpus lend themselves to custom-trained models. The wordVectors package is particularly effective for training our custom vector space, validating it, making queries based on semantic relationships, investigating word analogies, and performing vector operations using our colonial corpus content. We will elaborate on these operations later; however, it is important to note that, like many other digital historians, we use this method for exploratory data analysis (EDA) – both to uncover insights from the manumission corpus and to reflect on the method's limitations for similar corpora. Such exploration requires familiarity with both the topical content of the corpus and the computational parameters of its creation, enabling a deeper investigation into the interrelated concepts it contains.

3 Source Material

The archival collections at the heart of our research are sourced from the extensive India Office Records, held in London by the British Library, now digitized

2 Our thanks go to Julia Flanders, Sarah Connell and the others in digital scholarship at the Women Writers Project at Northeastern University for making this method accessible within the context of their Advanced Topics in the Digital Humanities seminar.

and made publicly available by the QDL. More specifically, our focus lies on the records of the British Residency in the Persian Gulf that pertain to slavery and manumission in the region.[3] Spanning from the 1890s to the early 1940s, these documents provide insights into manumission as well as the experiences of enslaved or indentured individuals in the Gulf region during a time of significant social and economic change. The collection includes manumission requests as well as a variety of administrative records related to the slave trade, including official publications, regulations, procedures concerning manumission, and correspondence detailing the fates of freed individuals. They also contain a vast array of additional documentation, contextualizing the complex institution of slavery and its integral role within the socio-economic fabric of the Indian Ocean world in general and the Arabian Peninsula in particular.

The manumission statements, a key source in this collection, were typically created through a process shaped by the constraints and practices of colonial administration. Most applicants were illiterate and communicated their situations orally, typically in Arabic, to assistants at British agencies or consulates across the region. The assistants would then translate the statements into English for official records. Each statement typically began by noting the applicant's place of birth and origin, followed by a brief account of their enslavement and life experiences and a justification for fleeing their masters. These statements concluded with a formal request for a certificate of manumission, typically validated by the applicant's thumbprint as a mark of authenticity (Zdanowski 2011).

It is crucial to acknowledge that these documents are heavily mediated through the colonial lens. Even at moments when the voices of the enslaved people come through, the narratives are shaped by British officials, who recorded and translated accounts using formulaic language and selecting information deemed relevant to colonial administration. The power dynamics embedded in the creation and preservation of these documents often obscure the full experiences of enslaved individuals, imposing silences and reinforcing the priorities of colonial authorities.

These records exhibit a range of material conditions – from pages marred by bleeding ink, especially evident in typewritten texts, to others left intentionally blank, likely to prevent ink transfer between pages. The documents' complex layout, featuring handwritten and typewritten pages in English and Arabic, further complicates the efforts to render them computable as textual data.

3 A list of the documents used in this study can be found in Appendix 1. For more on the records of the office of the British Residence in the Gulf, see "The Political Residency, Bushire."

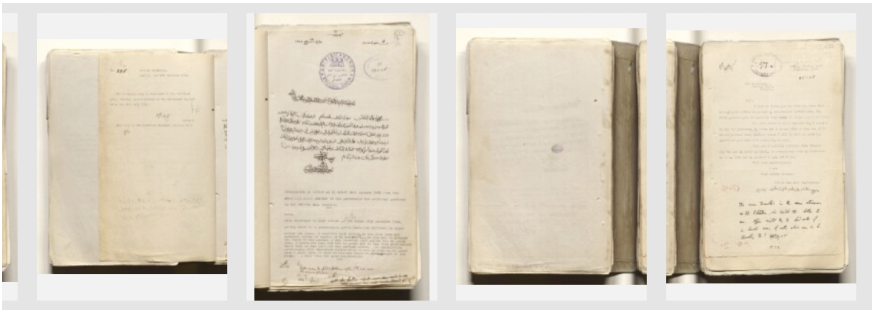

FIGURE 1 Sample pages from the volume File 5/190 II Manumission of slaves at Muscat: individual cases showing examples of handwritten English and Arabic together with typewritten English and blank pages. Depicted here are 149r–151r (images 306–310). Digitized by the Qatar Digital Library with an Open Government License.

4 Historical Background: Slavery in the Gulf

In the history of the Indian Ocean world, the phenomenon of slavery stands out for its complexity and persistence, particularly in the age of abolition. Despite the British Parliament's Act for the Abolition of the Slave Trade in 1807, and subsequent naval patrols around the African coast to curb the shipments of enslaved people, the trade in the Indian Ocean world not only persisted but, in some areas, intensified (Campbell 2013, 23).[4] This surge was partly due to European slave traders circumventing British naval patrols in the Atlantic by rerouting their operations through the Mozambique Channel into the Indian Ocean zone, leading to a significant increase in slave shipments (Allen 2013, 187).[5] It was also due to the rise of plantation economies in Zanzibar and

4 On the complex and sometimes conflicting goals of British imperial policy in the region regarding slavery, Campbell writes: "During the late nineteenth-century imperialist surge in the Indian Ocean World, abolition formed a central justification for the imposition of European colonial rule. Moreover, 'liberated' slaves were a potentially vital source of both taxation and manpower under colonial regimes governed by precepts of self-financing. A colonial priority was thus to transform the local working population into a free wage-labor force. Yet it was also vital to retain the goodwill of local slave-owning elites, whose assistance was required to administer the colony." (Campbell 2013, 34).

5 Allen highlights the significant role Europeans played in the Indian Ocean slave trade: "These figures, when added to those on British, Dutch, French, and Portuguese slave trading within the Indian Ocean noted earlier, suggest that Europeans were involved in the purchase and

Pemba, operated predominantly by Omani Arabs with slaves brought from the East African coast (Harms 2013).

The Gulf region, interconnected with the broader Indian Ocean world, experienced its particular dynamics of slavery and manumission against the backdrop of economic and colonial pressures of the nineteenth and early twentieth centuries (Hopper 2015). Slavery in the Gulf was deeply entwined with the economic and social fabric of the region, particularly the pearl diving industry, which operated as a significant employer of slave and indentured labor (Bishara 2017). In addition to pearling, enslaved men were utilized in diverse capacities as mercenaries, agricultural hands, and other forms of manual labor. Enslaved women, on the other hand, primarily served as domestic servants and concubines, highlighting the deep-seated dependence on enslaved labor within both the economic and domestic spheres of the Gulf societies where general emancipation did not take place until the mid-twentieth century (Zdanowski 2011, 864).[6]

As evident from the information above, the questions of slavery and manumission in the Gulf region are complex and multi-layered. Thomas McDow points out that the status of slaves in this particular context cannot be fully understood through a simple "slave versus free dichotomy" as the enslaved people "existed in hierarchies of dependency." The rights and circumstances of enslaved people varied significantly depending on the type of servile labor they performed. Household or personal slaves were more likely to gain significant independence or full freedom due to the obligations their masters had towards them, shaped by Islamic and local customs, whereas plantation slaves were rarely emancipated. Many enslaved individuals sought to improve their social standing by securing influential patrons and gaining greater autonomy. For some, remaining a slave client under a powerful patron was economically and socially more advantageous than achieving complete freedom (McDow, 2013).

The definition and practices of manumission in this context are just as complex and multi-layered as the institution of slavery itself (Hopper 2015; Bishara 2017). The idea of freeing slaves was not introduced to the region by the British. Islamic teachings emphasize the moral and religious virtues of freeing slaves, and the practice of manumission was deeply embedded in the region's cultural and social fabric long before the British influence (Zdanowski 2011, 866). Interestingly, Zdanowski argues that British emancipation efforts in these

transportation of at least 826,000 and perhaps as many as 1,089,700 slaves within and beyond the confines of the Indian Ocean between 1500 and 1850." (Allen 2013, 188).

6 For the early 1900s, Zdanowski estimates that enslaved individuals made up approximately 14.5% of the Gulf's population, with 36,880 out of a total population of 253,000.

Islamic societies led to significant ambiguity. Since individuals freed by British authorities were not emancipated according to Islamic law and prevailing social conventions, they were often perceived as merely being purchased by colonial authorities. Consequently, these freed individuals were sometimes referred to as "slaves of the government" or "slaves of the Consul" (Zdanowski 2011, 866).

Another issue complicating manumission practices was the diverse origins of enslaved people in the Gulf. In manumission statements, individuals were required to identify their ethnic or geographic origins, often revealing three main categories: those kidnapped from British colonies, those originating from lands identified as British protectorates, and those born into slavery to parents who might have been born as free people but who had been kidnapped or brought to the Gulf from the former two categories.[7] This distinction was crucial, as the British abolition of slavery in 1833 ensured immediate freedom for those who could prove their origins in British colonies. However, for individuals from protectorates and those born into slavery, the situation was more complex, as their status did not automatically guarantee manumission.

Throughout the period under study here, British policies fluctuated between efforts to suppress the slave trade, striking agreements with local authorities to curtail slave imports, and advocating for individual cases of emancipation. However, they refrained from exerting significant pressure on local leaders to implement full-scale abolition, recognizing the importance of these leaders as allies. The British were concerned that pushing too hard for abolition could incite political unrest and rebellion against these local authorities, thereby destabilizing the region (Freamon 2013).

In the Trucial Coast, the economic downturns of the 1930s – marked by the decline of the pearl industry due to overfishing, the advent of cultured pearl production in Japan, and the global financial crisis – significantly impacted the lives of slaves, prompting many to seek manumission. There was a sudden increase in manumission requests around 1935, with the overwhelming majority of applications coming from men employed on pearling boats (Zdanowski 2011, 877).[8] Their primary complaints included inadequate food and clothing provided by their masters, being forced to dive, and not receiving payment for their work.

It is important to reiterate that although British manumission certificates provided emancipated individuals with some protection and a means to

7 In addition to the East African coast, IOR records make mention of enslaved people from Yemen, Baluchistan and India.

8 Zdanowski identified 956 manumission statements in the records of the British Residency in the Gulf for the years between 1921 and 1946, with 545 of those submitted after 1934.

seek assistance against re-enslavement or mistreatment, British policy was not aimed at dismantling the institution of slavery in the region as a whole. Instead, the gradual abolition of slavery in the Gulf unfolded over the early to mid-twentieth century (Zdanowski 2011, 880).

5 Research Methodology and Workflow

Our workflow with IOR File 5 involved transforming the digitized documents available at the QDL into computable plain text and then analyzing them with the computational technique of word vectors. Three critical aspects of our documentary base are essential to highlight in this context: the discursive diversity of documents, their scale, and their multilingual nature. These factors are crucial as they inform our approach at each stage of the workflow, from selecting the appropriate HTR models and determining the parameters for training the word vector models to interpreting our results. The pipeline developed for this project is tailored to our specific study; yet it is presented here in detail to facilitate adaptation to analogous scenarios, highlighting the versatility of HTR and computational analysis frameworks in historical research.

5.1 *HTR for the Automated Recognition of Documents*

In digital humanities, the transition from traditional archival research to computational analysis of digitized archival documentation presents numerous opportunities and challenges. The potential of algorithmic reading and summarization of many thousands of pages of documentation – while theoretically imaginable – quickly confronts obstacles in its implementation, given the complexities of the documents at hand and the relative immaturity of digital archival systems. Our project to machine-read IOR File 5 occupies a critical position in this transition: while significant resources were allocated to digitizing the archives of the QDL, comparatively little attention was given to preparing them as fully machine-readable documents.[9]

The twenty-nine volumes of manumission material in our corpus, like many historical archives, are an assemblage of various types of documentation. Browsing them is akin to reading a scrapbook: a heterogeneous collection of materials loosely connected by subject, yet presented in idiosyncratic layouts, in both handwritten and typewritten formats, and across multiple languages.

9 It is worth noting that in the QDL, typewritten documents have been processed using Optical Character Recognition (OCR), and transcriptions of acceptable quality are available at the page level for browsing. This is not the case for the handwritten pages.

To transform these historical documents into computationally tractable text, the most practical technology that we had at our disposal for archival documents was HTR.[10] For this project, we employed state-of-the-art HTR technology available via the Transkribus platform.

The choice of Transkribus was a pragmatic one. Recent advancements in Transkribus HTR have significantly enhanced its ability to process documents containing both typewritten and handwritten text simultaneously, a capacity that was, until recently, out of reach. Transkribus has introduced 'super models' powered by transformer AI architecture, enabling the HTRing of documents in multiple European languages with greater accuracy and versatility. As the Transkribus user community has historically been focused on archives located in European countries, these super models have been trained on vast amounts of data in multiple European languages and in Latin scripts. This allowed us to process IOR File 5 in the most efficient manner, the lion's share of which is in handwritten and typewritten English, both in the original and in translation.

Prior to the advent of these advanced models, transcribing multilingual archives would have necessitated the creation of language-specific bespoke HTR models, customized to accommodate the diverse handwriting styles and the mixture of handwritten and printed documents in our collection. Language detection would have had to be integrated into layout analysis in a computationally expensive process. The introduction of transformer-based, general models has streamlined such research, enabling us to transcribe all of the English-language documents in our collection simultaneously and rather quickly, though not without challenges related to transcription accuracy and error management (ReadCoop 2024).[11]

Our choice to work with HTR super models marks a significant departure from other HTR projects that focus on creating very clean text with bespoke training. While we recognize that bespoke HTR models potentially yield higher accuracy rates in the transcription output, we find considerable practical and scholarly value in using general models, especially when working with large,

10 The distinction between digitized and computer-readable documents is crucial for anyone interested in digital humanities methods, particularly for scholars working with Arabic or other non-Latin script writing systems. Achieving computer readability means that each word and character string within the documents can be processed computationally, as opposed to merely being displayed as static digital images.

11 In this study, we utilize the general model named Text Titan available to us at the time of writing; this model was trained on historical documents in German, French, Dutch, Finnish, Swedish, and English, spanning from the 16th to the 21st century. Text Titan boasts a Character Error Rate of 2.95%.

multi-language corpora. Our choice reflects a broader trend in computational text analysis, which prioritizes the exploration of large-scale patterns and themes within the corpus over perfect accuracy in text capture, with analytical methods designed to minimize and mitigate the effects of transcription errors (Cordell and Smith, 2022).[12]

At the same time, we recognize the potential downstream challenges of using general HTR models. The transcription output generated by off-the-shelf transformer models presents specific limitations that influence how the corpus can be employed for distant reading and conceptual analysis. Two key issues in this regard warrant mention at the outset as they shape our approach to the corpus; additional considerations will emerge in the analysis below.

First, the process of corpus creation using HTR involves semi-supervised or unsupervised steps that predict the words on the page, sometimes resulting in misspellings, particularly of proper names (Kapan et al. 2023; Zhou et al. 2024). Additionally, orthography in IOR File 5 is inherently unstable. To analyze such materials effectively, it is essential to employ text analysis methods that are not reliant on strict string matching, accounting for both historically non-standardized spelling and transcription artifacts introduced by HTR.

An additional challenge lies in the page layout, abbreviations, and hyphenation present in the manumission documentation. Our corpus includes various types of documents, many of which are handwritten in an informal style. Common words, titles, and metrological terms are frequently abbreviated, often inconsistently. Moreover, clerks copying handwritten documents broke words at syllable boundaries, sometimes inserting hyphens and other times omitting punctuation altogether to fit text blocks within the documents. As a result, the corpus contains numerous orthographic variances, errors, and fragmented words. The implications of these issues are addressed in the analysis section below.

Second, as of the time of writing, the Transkribus user community includes very few groups working with Arabic text, with only one Arabic public model currently available.[13] While efforts are underway to address this gap, the off-the-shelf transformer models used to create our corpus of IOR File 5 were not

12 Corpus-based approaches are sometimes aligned with rule-based NLP techniques such as part-of-speech tagging and lemmatization. However, the impact of applying these rule-based techniques to text generated by HTR, which can be messy and inconsistent, is not yet fully understood. Further research is needed to evaluate the effectiveness of these methods on HTR-generated text.

13 At NYU Abu Dhabi we have accumulated ground truth for a public handwritten Arabic model. We have been combining crowdsourcing approaches with synthetic data for its creation. See John 2023 and NYU Abu Dhabi HTR Working Group 2024.

trained on Arabic-language materials. Consequently, pages written in Arabic or containing Arabic segments remain untranscribed. Although the absence of Arabic-language transcription might seem problematic for an archive containing Arabic materials, the distant reading approach adopted for IOR File 5 mitigates this issue to some extent. Many of the Arabic texts in our corpus appear alongside equivalent or roughly translated English texts, allowing the HTR transcription to capture key concepts from the English segments even if they are missed in Arabic. However, even if the Arabic texts were transcribed, additional challenges would arise when working with multilingual corpora, including the training of word vector models. These issues, which extend beyond Arabic texts, are discussed further in Section 5.2.

These challenges highlight the broader implications of working with general HTR models for multilingual corpora. The model used to transcribe IOR File 5, while effective for English text, often overfits its output based on the training set, resulting in inconsistent transcriptions for other languages present in the collection. For example, while the model captures some French and German words found in the British materials, it frequently misinterprets Arabic handwritten text as French. While setting a frequency threshold during word vector training can filter out infrequent and potentially erroneous words, such inconsistencies in transcription still pose challenges for querying the corpus effectively.

Had we opted for bespoke HTR models, some of the issues highlighted above, such as handling abbreviations and hyphenated words, might have been mitigated through approaches like the "smart model" strategy (Rabus 2022) or post-processing techniques. However, the size of the corpus made such post-transcription corrections impractical, both in terms of time and resources. Given the sheer volume of material and our decision not to pursue extensive editing, conducting word vector analysis directly on the HTR output required a degree of interpretive flexibility during the analysis. To extract meaningful insights and use the imperfect HTR outputs for historical interpretation, it remains essential to understand both the complexities of the archival materials and the inherent imperfections of the trained word vector models. These issues will be explored further in Section 6.

5.2 *Training Word Vector Models with the HTR Output*

Our workflow converts the digital scans of the IOR documents into computer-readable plain text files, processed and exported by the HTR system volume by volume, as detailed in Appendix I. Using this output, we custom-trained Word Vector Models (WVM) specifically for our historical corpus. This approach diverges from many NLP projects that rely on pre-trained embeddings like

BERT, which are optimized for modern language tasks. Scholars have highlighted the enduring relevance of static embedding models such as word2Vec for digital humanities research, particularly in tracing the evolution of ideas and language over time (Ehrmanntraut et al. 2021). This relevance aligns with the particular needs of our project, where understanding the historical specificity of the corpus requires a bespoke approach.

Several factors informed our decision to train a custom WVM. First, the historical context of these documents required an approach tailored to our corpus, as pre-trained embeddings based on modern language were less suitable for analyzing semantic relationships within this specific historical framework. Second, the corpus contains numerous proper names of persons and places from the Gulf region, which are transliterations from Arabic script and exhibit orthographic instability. Capturing these was essential to the topical analysis of slavery and manumission. Finally, a custom WVM was also necessary to address abbreviations and truncated words, as discussed in Section 5.1.

With more than 10,000 pages (excluding blank ones), IOR File 5 represents a substantial corpus for traditional historical reading. The HTR output for the English-dominant pages alone resulted in approximately 977,000 words – a significant volume for human reading.[14] However, as an experiment in WVM training and analysis, this corpus falls at the lower end of what is typically recommended for generating stable embeddings and conducting longitudinal studies, such as tracing conceptual history or semantic change (Wevers and Koolwen 2020).

Moreover, creating a WVM for a corpus is not a one-size-fits-all process; instead, different models can be trained depending on the choice of various parameters.[15] Our research on training WVMs for HTR output from colonial documents revealed that varying these parameters can highlight different kinds of evidence within the data. Put another way, we paid particular attention to whether different training parameters created viable WVMs that responded to typical queries, and also to the varied layers of evidence that differently trained models surfaced. Whereas there are a fair number of possible parameters in the wordVectors package, some of our most interesting results were found by

14 Blank pages were not processed with Transkribus, nor were Arabic-only pages. Some pages included both English and Arabic.

15 Detailed information on the parameters used for model training can be found in the documentation at https://rdrr.io/github/bmschmidt/wordVectors/man/train_word2vec.html. A lesson for an analogous word2Vec method in Python has been published by the Programming Historian, including a discussion of parameters (Blankenship et al. 2024).

varying the training using n-grams, which, in the end, allowed us to access different sorts of discourse. For example, we found that training the WVM for multi-word n-grams was particularly useful for word groupings common in the narrative and formulaic manumission requests or for Arabic *idafa* constructions. By contrast, single-word n-grams provided access to a more conceptual, abstract vocabulary found in administrative and political documents of the collection.

For this study, we have trained multiple models using the same 977K-word corpus, each with different parameters. We have listed those models and their parameters in Appendix II. These variations allowed us to examine how different configurations affect the outcomes and interpretations of the models. In the following section, we discuss querying these models to illustrate salient observations.

6 Observations and Results: Querying the Models

The wordVectors package enables querying the WVMS trained on one's own corpora and facilitates the exploration of the corpus and the variety of common contexts for words found therein. Interrogating the corpus is a significantly faster and less resource-intensive step than the previous two steps of generating textual transcriptions with HTR and training the WVMS. In other words, once the initial steps of transcription and training are complete, iterative inquiries into the corpus become rather straightforward and efficient.

It is worth highlighting again, however, that the speed and efficacy with which one can query the WVMS depend on the historian's pre-existing knowledge of the corpus and the complexity of the discourses it contains. Querying a model, in other words, is ideally aligned with informed historical inquiry. Moreover, experimenting with training parameters can also sometimes produce models that are difficult to query, yielding results that are not (easily) interpretable.

WVM analysis has some notable limitations. First, an instance of any given n-gram in a WVM is a highly abstract concept, one that can conflate homonyms or fail to account for nuanced distinctions present in the original text. Additionally, once a WVM is created, the ability to contextualize terms through direct reading, as provided by a keyword-in-context (KWIC) approach, is lost. For researchers accustomed to wildcard searching in databases, querying a WVM may initially seem like a rigid and less flexible environment. However, when combined with string searching and other natural language processing

techniques, querying a WVM can become a powerful method for conducting distant conceptual reading, offering new perspectives on a given corpus.

We emphasize the importance of "learning to query" a WVM because the semantic relationships it reveals often require interpretive strategies that differ significantly from traditional close reading. These models encourage historians to integrate multiple modes of analysis – close and distant reading – while critically engaging with forms of visualization to uncover insights and identify new terms for exploration. In the sections that follow, we discuss three key analytic frameworks for understanding and working with WVMs: cosine similarity, vector arithmetic, and multi-keyword analysis.

6.1 *Cosine Similarity of Query Terms*

One of the simplest queries to perform with a WVM involves measuring the cosine similarity of words or n-grams. Cosine similarity, represented as a value between 0 and 1 (as shown to the right of the two n-gram columns in Table 1), indicates the closeness of context between a query term and another set of terms. High cosine similarity in a corpus does not necessarily imply a direct relation of synonymity but rather signifies strong contextual proximity between strings.

In the query demonstrated in Table 1, we look at a trope of food and resource deprivation found in the numerous requests for manumission made to British officials by individuals from the Gulf, a query that relates to Zdanowski's observation about an increase in manumission requests during the collapse of the pearling industry in the 1930s. The results of the query for expressions similar to the 2-word gram "sufficient_food" reveal a strong contextual relationship with ill-treatment, physical violence, deprivation of food and clothing, as well as non-payment or seizure of earnings. These common motifs recur throughout the many hundred manumission requests documented in our HTR-created corpus, highlighting the conditions under which such documents presented to British officials formed a compelling situation for action.[16]

Additionally, positions 45 and 48 in the list ("each_season" and "diving_every_year") suggest a connection to the specific situation of pearl divers. This particular query also reveals two artifacts of the HTR (position 13: "my carnings"

16 Despite the closeness of this query to Zdanowski's observation, we did not develop our queries directly from the critical literature on enslavement, but rather from our selective prior reading of the corpus combined with a list of results from a number of exploratory cosine similarity queries. It is important to underscore that this kind of querying process in a corpus will not reveal much of the critical vocabulary used by modern historians, but returns a surface-level reading of the terms used in the corpus.

TABLE 1 A cosine similarity list for 50 similar 1-, 2-, 3- or 4-word grams for the expression "sufficient_food" based on a word vector (WVM) calculated using our QDL Manumission corpus. The model has been trained for 1-grams: 150 dimensions, 20 iterations, 6 word window with negative sampling of 5.

Ranking	n-gram	Cosine similarity	Ranking	n-gram	Cosine similarity
1	sufficient_food	1	26	giving_me	0.6509297126
2	never_gave_me	0.8524095979	27	was_always_illtreating_me	0.64557276
3	sufficient_food_and_clothing	0.8311532476	28	giving_me_much_trouble	0.6412911897
4	i_therefore_managed	0.7989247005	29	wear	0.6395323222
5	enough_food	0.7910946824	30	me_every_year	0.6327960503
6	was_not_supplying	0.7485174453	31	beating_me	0.6295357085
7	given_sufficient_food	0.7470261857	32	illtreat	0.6215113354
8	all_my_earnings	0.7180173073	33	without_any_reason	0.6166959792
9	and_clothing	0.7088835207	34	was_taking_all	0.6142218414
10	illtreating_me	0.7060400566	35	take_my_earnings	0.6130203257
11	with_sufficient_food	0.7039543774	36	taking_all_my_earnings	0.6111727031
12	was_illtreating	0.6964416133	37	taking_my_earnings	0.6109077116
13	my_carnings	0.68670742	38	food_and_clothing	0.6105478785
14	not_giving_me	0.6751625405	39	treat_me	0.6057736642
15	was_illtreating_me	0.6739397096	40	clothes	0.6022455236
16	he_was_always	0.6732428887	41	ford	0.6008772035
17	cruelly	0.6714686123	42	me	0.6004679594
18	very_cruel	0.671422939	43	treating_me	0.5973677692
19	supplying_me	0.6697214	44	illtreating	0.5941988309
20	me_anything	0.6654017771	45	each_season	0.5928885704
21	beat_me	0.6591838264	46	was_always	0.5910927736
22	escape_from_him	0.6587396284	47	therefore_ran_away_from	0.5905988372
23	my_earnings	0.6581637793	48	diving_every_year	0.5836055206
24	give_me_anything	0.6564914341	49	clothing	0.5821238728
25	did_not_give	0.6537808057	50	ghaus	0.580989743

(earnings, sic), position 41: "ford" (food, sic)), illustrating that even though the corpus contains spelling errors that can impede human understanding of the expressions, generally speaking, the WVM approach detects them as belonging to similar semantic contexts. We have found that HTR artifacts in the corpus are not as distracting as we previously imagined.

In this example, there is a relatively close semantic relationship between the n-grams, but this is not always necessarily the case. Although the corpus contains extensive information compiled by British colonial officials regarding manumission, it is not composed entirely of the manumission statements mentioned by Zdanowksi, as this query demonstrates. The many thousands of other documents articulated other discourses about the institution of slavery. The fact that this specific query returns so many highly similar expressions underscores how repetitive, even formulaic, the nature of these requests was. Even though we are working with a corpus of nearly one million words, by choosing such a specific query term close to one of the discourses found therein, we have, in effect, found a particular "corner" of the WVM. Had we chosen a less specific term – or a one-gram such as "food," "clothing," or "diving," the query would have returned a wider variety, and perhaps not as obviously related or clear, set of related terms. While useful for building semantic clusters in a corpus based on iterative, individual query terms, cosine similarity's measure of the position of various words in a WVM can also be seen as somewhat single-faceted, with limited discovery potential.

6.2 *Vector Arithmetic with Multiple Query Terms*

Since cosine similarity is calculated based on the value of a given string within vector space, more complex and combined queries are also possible. Vector arithmetic, for example, allows us to perform mathematical operations on WVMs to explore relationships between a number of given query terms. And importantly, as the term "arithmetic" suggests, the exploration is not always an additive one, but works like boolean logic. By adding and subtracting multiple vectors, semantic relationships and patterns can emerge at the intersection of concepts of interest to the historian. Such relationships might be thought of as combined or contrasting similarities that can yield more complex relationships akin to analogy. Such an arithmetic approach could be very useful in exploring concepts that have received limited attention by historians in the critical literature, say, the gendered aspects of manumission in the Gulf.

For example, in Table 2, the operation combining both subtraction and addition ("boat" – "house" + "master") helps identify words that are contextually similar to the idea of the maritime context of sailing ("boat") combined with the authority figure ("master"), but without the influence of the domestic

TABLE 2 Top results for the vector operation "boat" – "house" + "master," highlighting terms associated with maritime labor and authority in the context of enslavement. The model has been trained for 1-grams: 150 dimensions, 20 iterations, 6 word window with negative sampling of 5.

Word <chr>	**Similarity to "boat" – "house" + "master" <dbl>**
boat	0.6147923
master	0.5823759
beat	0.5259384
nakhuda	0.4721254
dhahi	0.4298578
sailed	0.4278552
jolly	0.4163437
lauuch	0.4066766
bankes	0.4001707
diving	0.3976721

context ("house"). Similarly, in Table 3, the query ("house" – "boat" + "master") identifies words that are contextually similar to "house" combined with "master," but without the influence of "boat." Here, the operation of subtraction denoted by the minus sign effectively removes the contextual influence of the subtracted term, isolating the relationships influenced by the remaining terms. In Tables 2 and 3, we see that the order of the query terms in vector arithmetic is significant, creating a list of words and cosine similarities that illustrate the gendered inflections of enslavement in the Gulf.

These findings can be easily contextualized within the historical backdrop of slavery in the Gulf, as detailed above in Section 4. The experiences of enslaved individuals varied significantly depending on their roles. Women were predominantly employed in domestic service, while men were more commonly engaged in labor outside the household, particularly in pearl diving. This gendered division of labor is also reflected in the manumission requests in our corpus, most of which were submitted by pearl divers.

In Tables 2 and Table 3, any number of query words could have been chosen, but we use "house" and "boat," suggesting the gendered quality of spaces, allowing us to investigate how these contexts intersect with the concept of a figure of oppressive authority, represented by the term "master" – a term frequently used by enslaved persons requesting manumission to refer to slave owners. The comparison indicates the diverging experiences of male and

TABLE 3 Top results for the vector operation "house" – "boat" + "master," highlighting terms associated with domestic labor and authority in the context of enslavement. The model has been trained for 1-grams: 150 dimensions, 20 iterations, 6 word window with negative sampling of 5.

Word <chr>	**Similarity to "house" – "boat" + "master" <dbl>**
house	0.6893260
master	0.6671767
husband	0.5175228
naster	0.5140424
mistress	0.5006217
died	0.4928051
death	0.4899177
serri	0.4820387
hasir	0.4780189
grandmother	0.4758750

female enslaved persons. The term "beating" appearing in the "boat" context suggests that "master" was frequently mentioned alongside physical violence in relation to men working on boats. In contrast, in the "house" context, the term "husband" emerges, suggesting a different form of authority and sexualized power dynamics for women.

It is worth mentioning that for the two arithmetic operations, there is a slightly higher overall cosine similarity score for "house" – "boat" + "master," suggesting a richer relationship between these terms than for the second set. The choice of query terms is crucial in such an inquiry since it situates us in the zones of the WVM where related terms lie. The choice of "house," "boat," and "master" would almost certainly not lead us directly to legal, administrative, or diplomatic discussions of manumission, but as we saw in Section 6.1, they situate us squarely in the part of the corpus with the manumission requests.

In a corpus that captures a range of voices and discourses, vector arithmetic is invaluable for identifying localized semantic "neighborhoods" – clusters of terms sharing contextual similarities within a specific conceptual field, such as we have seen here with domestic or maritime settings. However, understanding the broader relationships between these clusters requires an approach that can map complex, multi-dimensional associations. By utilizing principal component analysis (PCA) in the next section, we can extend this exploration to visualize how multiple terms and their contextual neighborhoods interact

within a larger semantic space. This method allows us to observe not only how individual "neighborhoods" are organized but also how they overlap and contrast, shedding light on the underlying discursive fields in the manumission corpus and highlighting both expected and novel connections.

6.3 *Terms Associated with Multiple Keywords with PCA*

When analyzing a corpus, certain concepts might be nuanced or multi-faceted, making it difficult to capture their full meaning with a single word or keyword (Schmidt and Connell 2021). They might also be used by different voices or perspectives within a corpus. One last example of how we use the wordVectors package to explore complex word relationships in our custom WVM is by plotting terms associated with multiple keywords using Principal Component Analysis (PCA). This approach has proven useful given the heterogeneity and multi-layered quality of the IOR File 5. By using PCA to plot terms associated with multiple input keywords, we can explore how these concepts interact and overlap within the corpus. By examining these relationships, we gain insights into the semantic connections between terms, enabling a richer understanding of the corpus's multi-layered nature.

Figure 2 represents a query using five key terms: "nakhuda" (a boat captain), "powers," "weapons," "traffic," "clerk," and "mistress." For example, the term "powers," chosen as one of the search terms, is used by colonial forces to refer to themselves, and it can be found in the upper right quadrant of the plot. Nearby, we see related terms like "power," "undertake," "regulations," "signatory powers," and "limits," all of which reflect a colonial discourse on legality and sovereignty. While this terminology has not appeared in earlier analyses in this article, it is prominent in IOR File 5, where the colonial understanding of enslavement intersects with the economic and political concerns of empire. By selecting terms that represent key topics in the corpus, this type of analysis and visualization highlights the nuances and interconnectedness of these concepts.

As noted earlier in the discussion of cosine similarity queries, the keywords for this analysis were not selected based on critical literature about enslavement in the Gulf. Instead, they emerged through iterative cycles of close and distant reading of IOR File 5. For demonstration purposes, we chose terms that produced a clear and evenly distributed arrangement in Figure 2. This distribution highlights distinct semantic fields within the corpus, with one notable exception. In the center-right area of the plot, the terms "traffic" and "weapons" appear much closer to each other. The term "weapons," in fact, is partially obscured by the red vector line and its related terms, emphasizing that these two keywords are semantically the most connected in this particular query.

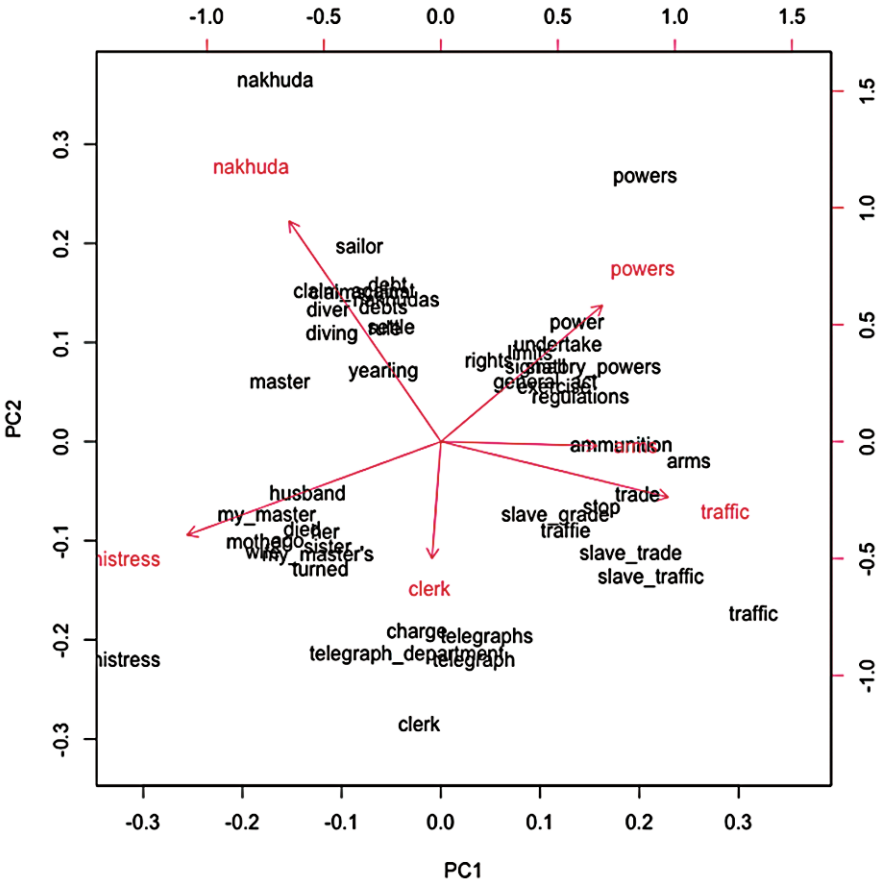

FIGURE 2 Key terms ("nakhuda" [upper left], "powers" [upper right], "weapons" [center right], "traffic" [center right], "clerk" [bottom center], and "mistress" [lower left]) in red and their contextual associations in black plotted in a two-dimensional space, generated using PCA on our custom trained WVM. The figure illustrates how specific words are spatially distributed based on their contextual similarities. The model has been trained for 2-grams: 150 dimensions, 20 iterations, 6 word window with negative sampling of 5.

This proximity suggests a meaningful relationship in the corpus between the trade in weapons and the trafficking of enslaved people – an area worth further research exploration.

Another observation about the PCA plot concerns the overall orientation of its axes. The spatial positioning of terms can sometimes reflect broader patterns in the data. In Figure 2, clusters on the right side of the plot largely correspond to colonial infrastructure, Western sovereignty, and the arms and slave trades. In contrast, the left side contains terms related to the lived experiences

of enslaved individuals in the Gulf. This dichotomy provides insight into the contrasting discursive landscapes within the corpus.

The distinct clusters in Figure 2 were chosen deliberately to illustrate key discursive "neighborhoods" in the data. By using multiterm queries, we can uncover nuanced relationships between terms, offering a more granular understanding of how concepts are interconnected and contributing to broader themes in the debates surrounding slavery. This exploratory approach, like those discussed in Sections 6.1 and 6.2, reveals both familiar and unexpected patterns, deepening our interpretation of IOR File 5. Had we selected rare or anachronistic terms – those reflecting modern critical terminology rather than the language of the documents themselves – the results would likely have been less informative. By carefully choosing keywords, we can identify meaningful semantic clusters, which serve as starting points for further comparative close reading. This iterative process of "reading" and "rereading" enables researchers to navigate higher-level concepts in the corpus, ultimately bridging the diverse discursive fields in a more holistic and comprehensive way.

6.4 *Discussion*

Our research into the manumission and slavery documents from the Gulf region underscores the complexities and limitations inherent in applying computational methods to non-European, multilingual corpora. Such documents, which include a wealth of entities from non-European languages and explore topics both inside and outside Western perspectives, present unique challenges. The models and tools that have been developed and refined, drawing on decades of digital humanities expertise in Western environments, are not always easily or directly transferable to such contexts. This disparity is reflective of a broader issue at hand: the digital infrastructures and resources that support high-level computational analysis at scale (HTR models, NLP, pretrained language models) are more advanced in Western contexts than in the Islamicate world and other non-Western regions.

In this article we have used HTR to create a corpus and then used a single R package written by digital humanists, demonstrating three analytical querying approaches with custom WVMs. The package uses a "bag of words" approach to custom train its model. Our approach should be considered, therefore, an initial and exploratory one for understanding the larger digitization of the IOR by the QDL. Not only are other querying and evaluation methods available within the same package, but there are other methods for working with word vectors including dynamic vectors, for example, if we are interested in semantic change over time, or even others that stretch beyond the word2Vec methodology (Pedrazzini and McGillivray 2022). While such approaches may

seem desirable, our corpus in its current state does not have enough data yet to benefit from such methods.[17]

This disparity becomes particularly evident when considering the scale and scope of our corpus. While computational methods like word vector analysis are powerful when applied to large datasets, our corpus is relatively small compared to those used in Western digital humanities projects. This limitation has implications both for the kinds of analysis we can perform and the conclusions we can draw. Had we been working with a much larger document collection or a more homogeneous corpus, the patterns and associations we could uncover might be more robust and revealing. Scaling up the corpus, however, may not be feasible while maintaining the focus on research questions specifically concerning slavery and manumission. The idea of scale is not only complex for discourses of manumission but might even be impractical for the way we traditionally conceive many projects in the historical humanities.

The creation of a humanities corpus is an expensive process, especially in terms of human labor. In our research, corpus creation not only involved digitizing documents, but also making critical decisions about how to process them and choosing the best methods for their analysis. This challenge is amplified when dealing with non-European sources, where the methods for text creation do not always lend themselves to straightforward querying or analysis.

Another significant challenge is the issue of scaling in humanities research. Unlike the natural sciences, where data can often be collected in large quantities through automated processes, humanities research frequently involves manual data collection and curation from discrete archival collections. The question of scaling humanities data is an important one, but it lies behind the difference in the approach of the two interlinked methods featured in this article. HTR with off-the-shelf transformer models allows us to create much more text for analysis than bespoke models, and yet custom WVMs allow us to focus on the specifics of the language we are studying, including the artifacts of a scaled HTR process. Although the future of computational methods used in archives is uncertain, it is tempting to suggest that a similar combination of methods will become popular in historical research: text creation as a more generic (and perhaps archive-led) process and iterative analysis by researchers using scripting languages such as Python or R.

We discussed the issue of multilingual corpora above, but the process of working with the QDL materials and with these two fundamentally different computational processes has also raised many questions of method for us.

17 Code available here: https://github.com/Living-with-machines/DiachronicEmb-BigHistData.

While it may be possible to bring together very large collections of multiple millions of words for certain moments of time about the Gulf region, there is neither consistency nor representativeness within those archival data. How can we be sure that in a big bag of words that we will be able to balance different kinds of voices? How can we be sure that any chosen model's transcription of typewritten and handwritten materials is of equal quality? And the list of questions goes on. As we keep these questions in mind, we also contend that computational methods in historical research should not be judged solely by whether they produce anticipated results; rather, their value often lies in the insights gained through experimentation, even when unexpected outcomes arise. In our study, each analytical stage has provided new understandings that inform our approach, particularly in refining corpus selection, digitization, transcription, and broader infrastructure design for future research. These iterative adjustments are essential in developing robust digital resources and methodologies that reflect the complexity of historical sources like those in IOR File 5.

7 Future Works and Conclusion

While the documents we have selected provide insight into the history of manumission and slavery in the Gulf region, they represent just one part of a much larger and more complex historical picture captured in the India Office Records (IOR). The results of our initial analysis should be viewed as one perspective among many, rather than a definitive account of the question.

We are fortunate to have the QDL for the study of the historical Gulf from the perspective of the IOR. It contains a wealth of information about how the British were trying to make sense of questions such as manumission and its position within Islamic society. The IOR materials also provide perspectives through which this moment in Gulf history can be studied from the perspective of coloniality.

Some doubt remains about using the method we have discussed here at scale. While working on a robust HTR model for Arabic remains an important objective, we are not sure that it will provide significant insight into the contents of the IOR archives given the method we have laid out in this article. The QDL is indeed one of the largest online repositories of digitized material in the Middle East, but since so much of the Arabic-language material contained in it was removed before the collections were transferred back to London in the twentieth century, it is an open question whether a custom word vector model would be a feasible approach. An alternative approach using HTR-

created text and pre-trained vector models may need to be adopted. It is also unclear whether pre-trained models will capture the historical specificity of the corpus.

Nonetheless, the method we propose of creating a corpus using HTR from the digitized IOR collections is not only transferable to other non-Western or Arab world projects – given the caveats about language and scale mentioned above – but it is also refinable in the multilingual colonial archive. In our case, we believe that there is promise in identifying other IOR subject files of interest that are topically connected, such as those about pearling and piracy. We expect that the topical diversity within them will be similar to that of the results discussed in Section 6.3, but will provide new vantage points from which to explore more deeply the interconnected discourses of coloniality in the Gulf with computational methods.

Bibliography

Allen, Richard B. 2013. "Slave Trading, Abolitionism, and 'New Systems of Slavery' in the Nineteenth-Century Indian Ocean World." In *Indian Ocean Slavery in the Age of Abolition*, edited by Robert W. Harms, Bernard K. Freamon, and David W. Blight, 181–204. New Haven: Yale University Press.

Bishara, Fahad Ahmad. 2017. *A Sea of Debt: Law and Economic Life in the Western Indian Ocean, 1780–1950*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316659083.

Blankenship, Avery, Sara Connell, and Quinn Dombrowski. 2024. "Understanding and Creating Word Embeddings." *Programming Historian*, January 31, 2024. https://doi.org/10.46430/phen0116.

Bsheer, Rosie. 2020. *Archive Wars: The Politics of History in Saudi Arabia*. Stanford: Stanford University Press. https://doi.org/10.1515/9781503612587.

Campbell, Gwyn. 2013. "Servitude and the Changing Face of the Demand for Labor in the Indian Ocean World, c. 1800–1900." In *Indian Ocean Slavery in the Age of Abolition*, edited by Robert W. Harms et al., 181–204. New Haven: Yale University Press.

Cordell, Ryan, and David Smith. 2022. *Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines*. Accessed August 6, 2024. https://viraltexts.org.

Ehrmanntraut, Anton, Thora Hagen, Leonard Konle, and Fotis Jannidis. 2021. "Type- and Token-based Word Embeddings in the Digital Humanities." In *Computational Humanities Research*, edited by Maud Ehrmann, Folgert Karsdorp, Melvin Wevers, Tara Lee Andrews, Manuel Burghardt, Mike Kestemont, Enrique Manjavacas,

Michael Piotrowski, and Joris van Zundert, 16–38. Amsterdam: CEUR-WS.org. https://www.ceur-ws.org/Vol-2989/long_paper35.pdf.

fastText. Accessed August 6, 2024. https://fasttext.cc/.

Freamon, Bernard. 2013. "Straight, No Chaser: Slavery, Abolition, and Modern Islamic Thought." In *Indian Ocean Slavery in the Age of Abolition*, edited by Robert W. Harms, Bernard K. Freamon, and David W. Blight, 181–204. New Haven: Yale University Press.

Gavin, Michael. 2019. "Is There a Text in My Data? (Part 1): On Counting Words." *Journal of Cultural Analytics*, September 17, 2019. https://doi.org/10.22148/001c.11830.

Harms, Robert. 2013. "Introduction." In *Indian Ocean Slavery in the Age of Abolition*, edited by Robert W. Harms, Bernard K. Freamon, and David W. Blight, 1–15. New Haven: Yale University Press.

Hopper, Matthew S. 2015. *Slaves of One Master: Globalization and Slavery in Arabia in the Age of Empire*. New Haven: Yale University Press.

Kirilloff, Gabi. 2022. "Computation as Context: New Approaches to the Close/Distant Reading Debate." *College Literature* 49 (1): 1–25. https://doi.org/10.1353/lit.2022.0000.

John, Fady. 2023. "Building Handwritten Ground Truth for HTR with the Google Vision API in Google Drive." *OpenGulf*. https://opengulf.github.io/htr/building-handwritten-gt/.

Kapan, Almazhan, Suphan Kirmizialtin, Rhythm Kukreja, and David Joseph Wrisley. 2022. "Fine-Tuning NER with spaCy for Transliterated Entities Found in Digital Collections from the Multilingual Persian Gulf." In *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference* (*DHNB 2022*), edited by Karl Berglund, Matti La Mela, and Inge Zwart, 288–296. Uppsala, Sweden, March 15–18, 2022. http://ceur-ws.org/Vol-3232/.

McDow, Thomas F. "Deeds of Freed Slaves: Manumission and Economic and Social Mobility in Pre-Abolition Zanzibar." In *Indian Ocean Slavery in the Age of Abolition*, edited by Robert W. Harms, Bernard K. Freamon, and David W. Blight, 160–181. New Haven: Yale University Press, 2013.

NYU Abu Dhabi HTR Working Group. 2024. "Arabic Handwritten 18th–20th Centuries." *HTR Model*. https://www.transkribus.org/model/arabic-khat-17-20-century-hand written.

Pedrazzini, Nilo, and Barbara McGillivray. 2022. "Diachronic Word Embeddings from 19th-Century British Newspapers [Data set]." Zenodo. https://doi.org/10.5281/zenodo.7181682.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2024. "GloVe: Global Vectors for Word Representation." Stanford University. Accessed August 6, 2024. https://nlp.stanford.edu/projects/glove/.

Rabus, Achim. 2022. "Handwritten Text Recognition for Croatian Glagolitic." *SLOVO* 72: 181–192. https://doi.org/10.31745/s.72.5.

ReadCoop. 2024. "Introducing Transkribus Super Models: Get Access to the 'Text Titan I'." *ReadCoop* [blog]. Accessed September 15, 2024. https://readcoop.eu/introducing-transkribus-super-models-get-access-to-the-text-titan-i/.

Schmidt, Ben. 2015. "Vector Space Models for the Digital Humanities." *Bookworm* [blog]. Accessed July 30, 2024. https://bookworm.benschmidt.org/posts/2015-10-25-Word-Embeddings.html.

Schmidt, Ben, and Sarah Connell. 2021. "Word Vectors Visualization" [notebook]. Accessed July 30, 2024. https://github.com/NEU-DSG/wwp-public-code-share/blob/main/WordVectors/Word-Vectors-Visualization.Rmd.

"The Political Residency, Bushire." 2024. *Qatar Digital Library*. Accessed August 6, 2024. https://www.qdl.qa/en/political-residency-bushire.

Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change*. Chicago: University of Chicago Press.

Verheul, Jaap, Hannu Salmi, Martin Riedl, Asko Nivala, Lorella Viola, Jana Keck, and Emily Bell. 2022. "Using Word Vector Models to Trace Conceptual Change Over Time and Space in Historical Newspapers, 1840–1914." *Digital Humanities Quarterly* 16 (2). Accessed August 6, 2024. http://digitalhumanities.org:8081/dhq/vol/16/2/000550/000550.html.

Wevers, Melvin and Marijn Koolwen. 2020. "Digital Begriffsgeschichte: Tracing Semantic Change Using Word Embeddings." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53 (4): 226–243. https://doi.org/10.1080/01615440.2020.1760157.

Women Writers Project, Northeastern University. Accessed August 6, 2024. https://wwp.northeastern.edu/.

Zdanowski, Jerzy. 2011. "The Manumission Movement in the Gulf in the First Half of the Twentieth Century." *Middle Eastern Studies* 47 (6): 863–883. Accessed August 6, 2024. https://www.tandfonline.com/doi/full/10.1080/00263206.2010.527121.

Zhou, Jolie, Camille Lyans Cole, and Annie T. Chen. 2024. "Basreh or Basra? Geoparsing Historical Locations in the Svoboda Diaries." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (*Volume 4: Student Research Workshop*), 291–304. August 11–16, 2024. https://doi.org/10.18653/v1/2024.acl-srw.33.

Appendix 1

India Office Records File 5 – Slave Trade files used in this study from their digitized copies located in the Qatar Digital Library.

Title of QDL document	Document URL
File 5/104 II,III – Miscellaneous slave trade correspondence	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000af
File 5/168 IV – Manumission of slaves on Arab Coast: individual cases	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000b5
File 5/168 V – Manumission of slaves on Arab Coast: individual cases	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000b6
File 5/168 VII – Manumission of slaves on Arab Coast: individual cases	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000b8
File 5/168 VIII – Manumission of slaves on Arab Coast: individual cases	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000b9
File 5/183 (D 31) – Manumission of slaves at Kuwait	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000ba
File 5/187 I – Proclamation prohibiting slave trade	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000bb
File 5/188 I, 189 I – Expenses incurred as a result of slaves taking refuge in consulates and agencies; manumission of slaves and general treatment of slave trade cases	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000bc
File 5/190 II – Manumission of slaves at Muscat: individual cases	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000bd
File 5/190 III – Manumission of slaves at Muscat: individual cases	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000be
File 5/190 IV – Manumission of slaves at Muscat: individual cases	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000bf
File 5/190 V – Manumission of slaves at Muscat: individual cases	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000c0
File 5/191 – Kidnapping of Baluchis and Indians on the Mekran Coast and exporting them for sale at Oman and Trucial Coast	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000c2

(*cont.*)

Title of QDL document	Document URL
File 5/191 II – Individual slavery cases	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000c3
File 5/191 IV Individual slavery cases	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000c5
File 5/193 II (B 38) – Slavery in the Gulf	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000c7
File 5/193 II (B 38) – Slavery in the Gulf	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000c6
File 5/193 IV (B 55) – Slavery in the Persian Gulf	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000c9
File 5/194 I, 195 I, 179 III, 169 II, 104 IV – Kidnapping of individuals; manumission of slaves at Kuwait and Bushire; miscellaneous slavery cases	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000ca
File 5/194 I, 195 I, 179 III, 169 II, 104 IV – Kidnapping of individuals; manumission of slaves at Kuwait and Bushire; miscellaneous slavery cases	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000b0
File 5/197 I – Absconding of Slaves from Sharja and Henja	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000cc
File 5/198 I – Kidnapping on the Trucial Coast	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000cd
File 5/198 II, 199, 200 – Kidnapping of persons on the Trucial Coast; purchase and export of slaves from the Trucial Coast; Saudi Government's regulations on the slave traffic	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000ce
File 5/201 – Manumission of slaves and rules relating to cases arising out of the pearling industry	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000cf
File 5/6 I – Brussels Conference and general rules and procedure on slave traffic	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000ac

(*cont.*)

Title of QDL document	Document URL
File 5/65 1 – Question of disposal of emancipated slaves and proposal to check traffic between Muscat, Oman ports and Zanzibar	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000ad
File 5/74 – Practice attributed to British authorities of surrendering fugitive slaves	https://www.qdl.qa/en/archive/81055/vdc_100000000193.0x0000ae

Appendix II

We trained four Word Vector Models with the parameters described in the table below. These models are available at https://doi.org/10.5281/zenodo.14192194.

Name of model	n-gram	Dimensions	Iterations	Window	Negative samples	Total training words
HTR997K1g150 d20i6w5ns.bin	1	150	20	6	5	914596
HTR997K2g150 d20i6w5ns.bin	2	150	20	6	5	857994
HTR997K2g150 d20i6w15ns.bin	2	150	20	6	15	820391
HTR997K3g150 d20i6w5ns.bin	3	150	20	6	5	741542