Immortal cows of Nouvelle France - Reflections around four variations on modern digital humanities techniques for Zooarcheology

Nicolas Delsol, Éric Drapeau, Samuel Laperle, Josiane Van Dorpe, Grégoire Winterstein

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Immortal cows of Nouvelle France – Reflections around four variations on modern digital humanities techniques for Zooarcheology

Nicolas Delsol[1], Éric Drapeau[2], Samuel Laperle[2],
Josiane Van Dorpe[2], Grégoire Winterstein[2],
[1]Département des Sciences Historiques, Université Laval, Québec (QC), Canada
[2]Département de Linguistique, Université du Québec à Montréal, Montréal (QC), Canada
Correspondence: nicolas.delsol.1@ulaval.ca, winterstein.gregoire@uqam.ca

Abstract

This paper explores the integration of digital humanities techniques into archaeological and historical research, focusing on the historical representations of cattle in New France through archival documents from the 17th and 18th centuries. Our objective is to evaluate the effectiveness of computational methods—such as textometry, word embeddings, topic modeling, and large language model (LLM) representation clustering—in uncovering the semantic and cultural dimensions of bovines in colonial texts. We employ these methods to analyze a corpus of historical documents, aiming to identify recurring themes, associations, and underlying patterns in the portrayal of cattle. The textometric analysis highlighted the frequency and context of bovine-related terms, while word embeddings revealed significant associations, such as the unexpected pairing of *vache* (cow) with *immortelle* (immortal), reflecting legal obligations around perpetual donations of cattle. Topic modeling further illustrated the centrality of cattle in agricultural practices, particularly their wintering and the broader socio-economic implications within the settler communities. Clustering LLM representations allowed us to refine these findings by grouping related terms and exploring their contextual usage across the corpus. The results demonstrate that digital humanities techniques can significantly enhance the study of historical texts, providing deeper insights into the cultural and economic roles of cattle in New France. This interdisciplinary approach not only contributes to our understanding of human-animal relations in colonial settings but also suggests new directions for future research in digital humanities and historical archaeology, particularly in the automated analysis of archival materials.

1 Introduction

This work examines four different digital humanities techniques for the investigation of the semantic space around selected topics within historical corpora. In particular, we evaluate in which measure the clustering of representations provided by recently developed Large Language Models (LLM) for classical French dovetail and correlate with more standard techniques in digital humanities.

The topic of interest is the mention of bovines in writings from and about Nouvelle-France (NF). Nouvelle-France designates the former territories colonized by the French crown in North America, more particularly the settler colonies around the lower Saint-Lawrence valley corresponding to the present day Canadian province of Quebec. The first colonists from Europe began to establish permanent settlements at the beginning of the seventeenth century. Like in the rest of the Americas, many animal species that were central to the European lifestyle and economy did not exist, in particular all the domesticate species such as cattle, pigs, or sheep, providing crucial goods and foodstuffs. Among these animals, cattle held a particularly prominent role by helping the first Euro-Canadian farmers open new crops in the area.

The broader history of their introduction is known through a few historical sources (Trudel, 2016; Desloges, 2009) but many aspects of this process, in particular the origins of the bovine populations and details on their management practices remain unclear. This work is part of a larger project that combines the analysis of archaeological remains of colonial cattle with the automated study of large amounts of historical archival documents. Overall, this project aims at addressing the following research questions: (1) where did the animals come from and what is the overall phylogeographic history of these populations, (2) how did the management practices and uses of cattle evolve over time, and how did the conceptions and representations of cattle change in relation to these changes.

More specifically, this work's purpose is twofold: (a) evaluate the historical representations of cattle,

and their correlates, in a series of archival documents of different type dating from the early seventeenth to the late eighteenth century, and (b) methodologically assess and compare the application of different techniques to the study of historical documents written in Modern French language.

2 Related work

2.1 Zooarchaeology and history of animals in New France

The arrival of cattle in New France was an event that played a major role in shaping the environment, economy and culture of the region during colonial times. Despite its importance there is still much to learn about the pathways of cattle dispersal in eastern Canada and how this introduction affected the daily lives of European settlers and Native communities. To apprehend this phenomenon and its broader anthropological consequences, we have designed a multifaceted research project that aims at integrating archaeological, zooarchaeological, and biomoleculardata (RABBA – "Recherches en archéologie biomoléculaire sur les bovins aux Amériques: origines,mobilité, pratiques") with the automated analysis of large amounts of digitized archival texts (Projet BNF –"Bovins Nouvelle-France"). Our approach involves exploring both material evidence and historical texts to understand how cattle were introduced into New France and how they impacted society and the environment. By combining findings with insights from written records we aim to show how humans interactions with animals like cattle transformed landscapes and livelihoods.

Zooarchaeology, the analysis of archaeological faunal materials (bones), aims at shedding light on the cross cultural and historical trajectory of human-animal relations. While traditionnally focused on ancient diets, this subfield of archaeology has gradually broadened its focus to embrace a wider palette of social and cultural issues (political economy, agrosystems, environment, symbolic use of animals). One way to address such a range of issue has recently found its expression in works focusing on a single species and their historical and cultural itinerary (Sykes et al., 2020; Thornton, 2016). Regarding the question of the introduction of Eurasian domesticates in the Americas on the heels of European colonists, few large regional syntheses have been produced so far (Delsol, 2024).

The zooarchaeology of periods with a written historical record often use these archival resources as a tool to inform its research questions or illutrate its findings. Despite the pervasiveness of the use of historical documents, no attempt at automatically analyzing the written record with approaches based on LLMs has ever been used in such a research program. The approach offered in this work introduces the results of the BNF part of the project, offering new venues of research for historical and zooarchaeological studies.

2.2 Digital humanities

Typically, what is referred to as digital humanities concerns a set of computational methods used to answer classic humanities questions (Vanhoutte, 2016). The fields in which this type of approach can be found generally revolve around history, literature, media studies, etc. (Watrall et al., 2016). Archaeology being categorized as a social science is not directly part of what encompasses digital humanities. This positioning is reflected in the field's use of computational methods. From the 1980s onwards, new technologies enabled the creation of useful tools for the visualization and modeling of excavated sites, and this is pretty much all that can be found as computational use in the literature (Watrall et al., 2016). Recently, however, the massive digitization of archival data may enable the use of automatic language processing to facilitate the extraction of data of interest to researchers (Manjavacas Arevalo and Fonteyn, 2021). Archaeology's primary sources consist in the material traces and vestiges of past societies. As a result, most of the computational methods applied in that field do not relate to language, aiming instead at providing tools to sort and create typologies of elements of past material cultures (Plutniak, 2022). However, historical archaeology (i.e. the archaeology of periods with a written record) quite often relies on the combined analyses of both material and archival sources (Hicks and Beaudry, 2006). Textual sources offer priceless clues to the archaeologist to enrich their perspective of material processes in the past. They can inform and guide research questions as well as offer some context to better understand these phenomena. We aim to harvest methods from digital humanities and natural language processing to explore archeologic data. Specifically, we will use textometric measures, word embeddings, topic modeling and clustering of representations encoded in language models to characterize our data. Textometry focuses

on surface descriptions using statistical methods concerning the frequency and co-occurrences of certain words or expressions (Pincemin, 2011). By identifying these elements, we can note their relative importance within a document and their tendency to co-occur with other terms. Vector models such as `word2vec` (Mikolov et al., 2013) enhance these methods, and offer numerical representations of the terms that appear in a corpus. These vector representations are known to capture complex morphosyntactic and semantic properties of the terms they represent, in particular information about the latent associations of a term with other terms. By training this type of model on our data, we can extract synonyms and words that occur in similar contexts from our keywords. Topic modeling is a technique that creates classes of documents on the basis of the terms that appear in it. Here, document is to be understood as a general term: in practice, a sentence within a text can be treated as a document for the purposes of the method. For our data, we used BERTopic (Grootendorst, 2022), a modular architecture that enables us to use the dynamic embeddings offered by large language models such as those of the BERT family (Devlin et al. 2019, cf. next section). Topics are identified by the terms that served to define them. One can then explore the topics to see if some are defined by certain concepts of interest.

2.3 Large language models for French

Language models are tools that manipulate representations that encapsulate information about linguistic expressions. Following the development of Transformer models (Vaswani et al., 2017), powerful models have been developed, which offer dynamic representations for linguistic expressions, dependent of the context in which those expressions appear: the representation of a given expression will vary according to the linguistic context in which it appears. These models are pretrained on vast quantities of text, giving their representations a general character which can, in principle, be leveraged for a variety of downstream applications on which the model can be fine-tuned. In the context of this research, we focus on bidirectional models of the BERT family (Devlin et al., 2019), whose representations integrate information from both the left and right context of an expression (see Sec. 4.4 for further discussion).

The language found in the texts of interest to the project differs from contemporary French. As discussed in Sec. 3, the historical period covered by our data goes from the mid sixteenth to the late eighteen century, which corresponds to what is referred to as pre-classical and classical French. These varieties of French were used on both sides of the Atlantic, so the language used in NF corresponds to the one used in France that time, specifically to that used in the region around Paris (Gendron, 2013). This allows us to use language models trained on texts from that period, without having to focus on a particular geographical area. To our knowledge, the only model of that sort available to this day is d'AlemBERT (Gabay et al., 2022), trained on the FreEM corpus which covers a historical period ranging from the 16th century to the end of the 19th, and thus matches our period of interest. Of particular interest to us is the robustness of the representations offered by d'AlemBERT across dialectal and diachronic variations. As highlighted by Gabay et al. (2022), d'AlemBERT representations fare well for various linguistic tasks, even when dealing with data that was absent or underrepresented in the training data of the model.

3 Data

We used two main sub-corpora for our analyses.

The first one was manually compiled on the basis of literary works written about NF. The list of works in the corpus can be found in appendix A.

4 Analyses

In this section, we explain the method behind each type of analysis and showcase some of the most relevant outputs of these analyses. Mostly for reasons of space, we do not present the whole set of results, though those were taken into consideration in the discussion in section 5.

4.1 Basic textometry

We began by establishing a list of lexical terms that correspond to the theme of bovines in NF. Those terms are shown in table 1 along with their raw frequencies and frequency of occurrence per million words in the complete corpora.[1]

[1] A list of all keywords and their translations is given in Appendix B.

data	keyword	per million	total
NFN	bestiaux	87.78	746
	vache	57.77	491
	boeuf	54.95	467
	veau	15.3	130
	taureau	5.06	43
Published	bestiaux	40.6	205
	vache	11.29	57
	boeuf	1.58	8
	veau	4.36	22
	taureau	0.99	5

Table 1: Frequencies of our keywords

We then proceeded to look at bigrams, and focus on the ones that involve our target terms and display a strong association via their Pointwise Mutual Information score. Table 2 summarizes some interesting bigrams for the keyword *vache* ('cow') in the NFN corpus and in the published corpus.

Corpus	bigram	N. occ.	PMI
NFN	('vache', 'immortelle')	12	13.67
	('vache', 'prisée')	9	8.88
Published	('boeufs', 'vaches')	9	15.08
	('vaches', 'moutons')	5	14.17

Table 2: Most relevant bigrams for the word *vache*

4.2 Static word embeddings

We trained three `word2vec` models (Mikolov et al., 2013), one per sub-corpus and one for the entire corpus, to obtain static word embeddings. The training configurations were the same for each model with a vector length of 300, a window of 3, and a set number of 5 epochs (using the `gensim` library Rehurek and Sojka 2011). Words with less than 5 occurrences were ignored in the process. We used the embeddings to calculate the cosine similarities of the terms in our list of key words. Table 3 shows the five most similar term for each corpus, for the target term *vache*.

Similar term	Similarity
NFN	
immortelle	0.64
genisse	0.61
cariolle	0.61
jument	0.59
taure	0.59
Published	
barique	0.72
pistolle	0.67
ferrure	0.66
corne	0.64
fermage	0.64
Combined	
pouliche	0.58
camisolle	0.58
cariolle	0.58
truye	0.57
jument	0.57

Table 3: Top five most similar terms for *vache*, per corpus

We can already observe an overlap between the results in Table 3 and the bigrams in Table 2 with the (admittedly unexpected) adjective *immortelle* ('immortal'). We discuss the significance of this association in section 5.1.

4.3 Topic modelling

To automatically extract topics from our data, we used BERTopic on each of our sub-corpora separately and on the combined corpus, using d'AlemBERT embeddings. The number of topics found in each case are shown in Table 4.

File	Number of topic
NFN	11 006
Published	2185
Combined	9745

Table 4: Topics found per corpus

In Table 5 we show the number of topics that involve our set of target terms.

Target term	NFN	Published	Combined
vache	13	0	8
taureau	1	0	0
bestiaux	4	2	3
boeuf	8	0	4

Table 5: Topics associated to target terms, per corpus

As can be seen, the two sub-corpora differ in whether they seem to be organized around topics that involve bovines. While several topics involve our target terms in the NFN sub-corpus, they are much more rare in the published corpus, in which only *bestiaux* ('beasts/livestock') seem to be relevant. This is expected in a way: the NFN sub-corpus contains many notary records listing heritages, the buying and selling of livestock etc., and the target terms are more frequent in the NFN sub-corpus than in the Published one (cf. Table1), though they are not absent either from the Published corpus. Rather, their presence is not tied to an identifiable topic.

As mentioned in section 2.2, though the number of topics identified by the algorithm is large, we are only interested in those that are specific to our terms of interest (given in Table 1). Table 6 shows the most representative topics associated with the keywords *vache, boeuf* and *bestiaux* from the NFN part of our corpus. The count associated to each topic corresponds to the number of documents (i.e. spans of texts) associated to the topic, and the representation is the set of terms that define the topic.

Beyond suggesting other keywords for further and expanded analyses, those topics already suggest the outlines of the place of bovines in NF, and how they were conceptualized. We discuss those findings in section 5.1.

As for the lack of conclusive results from the use of BERTopic on the Published corpus, this confronts us with certain limitations of this kind of approach (see Egger and Yu (2022) for more general limitations). It is possible that the low frequency of the terms we are interested in prevents us from seeing them emerge in interesting clusters. Nevertheless, we do expect the above-mentioned target terms usages to have a particular meaning, and to refer to different facets of bovines, even in the Published sub-corpus. To capture such nuances of meaning requires a complementary approach to topic modelling, to which we now turn to.

Id.	Target	Count	Representation
2614	*vache*	60	'hyverné', 'hyvernée', 'hyverner', 'hyvernés', 'moutons', 'taure', 'hyverne', 'nouriturraux', 'hyvernes', 'pacagée'
9819	*vache*	13	'moutons', 'vaches', 'genisse', 'ccechons', 'chicvat', 'giarre', 'grangé', 'troisueaux', 'vache', 'lochon'
1119	*boeuf*	143	'boeuf', 'boeufs', 'vaches', 'caribous', 'bola', 'besson', 'apellent', 'peaux', 'appelés', 'boeuf'
2792	*boeuf*	57	'labours', 'labour', 'laboureurs', 'labourer', 'laboureur', 'laboratoire', 'laboure', 'labouré', 'labourage', 'boeufs'
3927	*bestiaux*	40	'attaque', 'attaquer', 'attaquée', 'atoucher', 'atriotume', 'attablissement', 'attabues', 'camiral', 'attassine', 'attavoix'

Table 6: Examples of relevant topics associated to target terms about bovines in the NFN corpus

4.4 LLM representation clustering

Analyses based on topic modelling approaches are not particularly suited to the investigation of particular set of terms. In the context of DH, they serve to identify and characterize the topics approached in collection of documents, but there is no guarantee that certain target terms will indeed be part of such topics. In the previous subsection this is what happened with the Published sub-corpus, in which no topic relevant to cows was identified, even though the term does appear a significant number of times in the corpus. To circumvent such obstacles, we leverage techniques that were used by Erk and Chronis (2023) to study the properties of the representations of specific lexical items by large language models. In essence, the technique involves obtaining the LLM representations of all the tokens that correspond to target lexical items in a given corpus. After using dimension reduction techniques, those representations are automatically

clustered together, and the clusters are manually qualified on the basis of the sentences that correspond to the tokens belonging to the cluster. In the rest of this section, we give details on all the steps we used to use this technique on our data.

Given that the data is not clean, in particular in terms of spelling of our target terms, we took precautions in the preprocessing of the data. We used d'AlemBERT (Gabay et al., 2022) to extract dynamic embeddings of terms within our list of target terms. All the texts in the corpus were first tokenized using d'AlemBERT's tokenizer, and split into batches of 512 tokens, the maximal size for a sentence in d'AlemBERT. We found that target terms were tokenized differently by d'AlemBERT if they were preceded directly by a space. As an example, the tokenized term *vache*, was either split by the tokenizer in '*v*' and '*ache*', or kept entirely as *Ġvache* where 'Ġ' represents the space character. We found that terms not preceded by a space were either at the very start of the text or they were fused with another term, possibly due to an OCR error. For example the sequence '*unevache*' appears in the corpus and is tokenized as *'une', 'v', 'ache'*, though the sequence is most likely the result of faulty OCR.

Being aware of this possible issue, we were able to identify the position of each token in every batch. We then extracted the tokens' embeddings from d'AlemBERT's last hidden layer. Whenever the term was split into two or more tokens, we calculated an average embedding to represent the term. The decision to extract only the last layer of the d'AlemBERT is based on the work Erk and Chronis (2023) who found that the latter layers of BERT models encode semantic and pragmatic information in a consistent way, which echoes similar findings in the literature (see a.o. Tenney et al. 2019).

For each term, we reduced the number of dimensions of the gathered embeddings from 718 to 3 and 2 using the Principal Component Analysis (PCA) method. The embeddings with 3 dimensions were used for clustering and creating 3D visualizations while the embeddings with 2 dimensions were only used for creating 2D visualizations. We used k-means clustering to group the embeddings of each term by considering the embeddings of every occurrence of that term. To determine the optimal number of clusters for k-means, we initially explored a wide range of values for *k*, from 1 to 60. We then plotted the inertia - which measures the within-cluster sum of squared distances - and identified a range of values for *k* where an elbow was evident (e.g., from *k*=4 to *k*=10). From there, we calculated the average silhouette score - which measures how similar an instance is to its own cluster compared to other clusters - and selected the *k* value with the highest score. To visualize the clusters, we created both 3D and 2D plots of the embeddings. For each occurrence of a term, we included 12 tokens preceding and 13 tokens following the term to capture its context and see the differences between each occurrence. Within each cluster, we identified the most representative occurrence by finding the one closest to the center of the cluster. Figure 1, Figure 2 and Figure 3 illustrate the results of the method and show the clusters for *vache* in 2D, for each of the two sub-corpus and the total corpus.

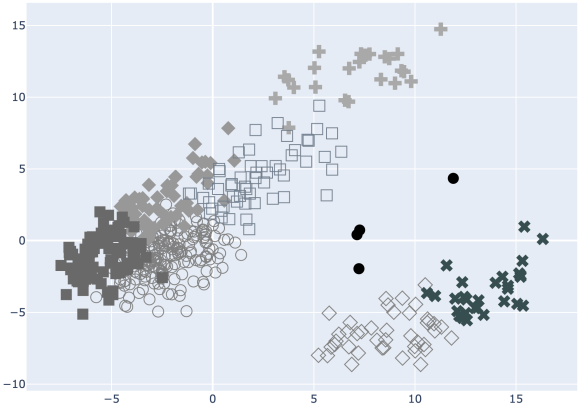

Figure 1: 2D clustering of the occurrences of *vache* in the NFN corpus

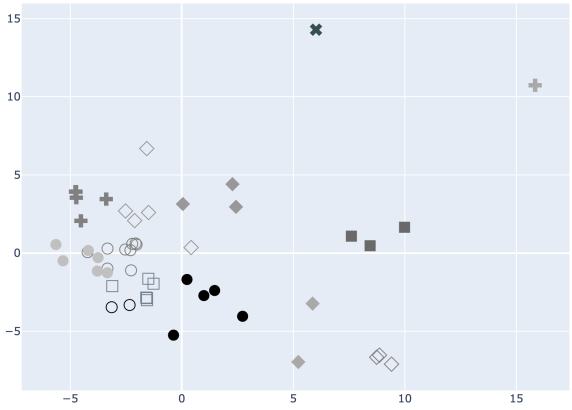

Figure 2: 2D clustering of the occurrences of *vache* in the Published corpus

The method yields particularly legible results for the entire corpus: this is where the clusters appear to be the most interpretable and well separated. Roughly we find a cluster that is related to inventories, such as those made by notaries when dealing

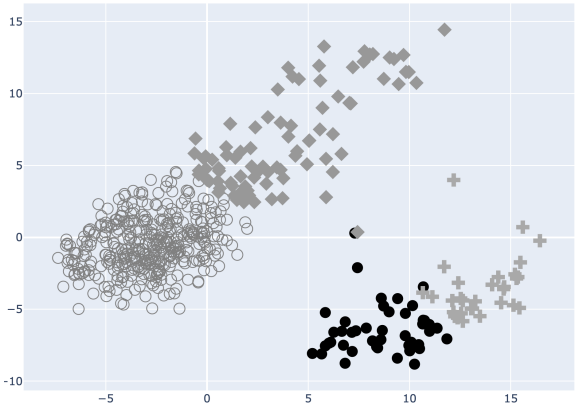

Figure 3: 2D clustering of the occurrences of *vache* in the complete corpus

with inheritance questions, one which is related to conflicts that involve cows, typically of judiciary nature, for which the cow is at the source of the conflict, e.g. because its ownership is being disputed or because the cow caused some damage to one of the parties. Another cluster involves cows in judiciary matters, but in those examples the cow is part of the compensation offered to one of the parties. To a degree, these clusters reflect the general topics that we expect to find in the documents of the NFN corpus which mostly deal with legal matters (see Annnex A.2). It is nevertheless worth noting that the method seems to correlate with different types of judiciary acts.

Figure 4 shows the result for the plural *vaches* ('cows').

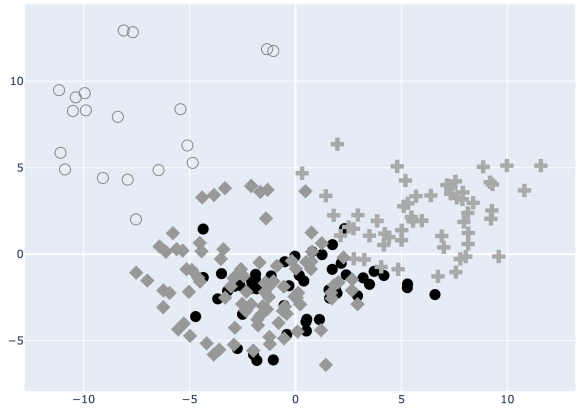

Figure 4: 2D clustering of the occurrences of *vaches* in the complete corpus

We find the same number of clusters as for singular *vache*, though their nature differs. In particular, we find one cluster that seems to revolve around description of locations. There, cows are either mentioned as resources available in a location or as means of transportation. Another cluster involves cows as a comparison point in the description of other animals, suggesting yet another perspective and experiencing of cows.

5 Discussion and openings

The computational methods introduced in this paper and their application to the analysis of New France archival documents offer a critical insight into the mentalities, practices and uses of cattle in the first centuries of European presence in the lower Saint-Lawrence basin. The innovative use of such an approach in the study of historical archaeology and zooarchaeology provides exciting new venues of research in these fields.

5.1 The role of cows in New France through the lens of digital humanities

The application of digital humanities methods to the study of archival documents related to cows in New France reveals a nuanced perspective on the role of bovine in the region between the sixteenth and the eighteenth century. By analyzing historical texts using techniques such as textometry, word embeddings, and topic modeling, we observe that cows were not merely agricultural assets but also symbolic figures deeply embedded in the cultural and economic landscape of the Euro-Quebecois colonial society.

One of the main takeaways of the word similarity analysis is the highly frequent, though unexpected, association of the words *vache* and *immortelle*. The association of these two terms does not refer to any supernatural property of undying animals. It revolves instead around the legal obligation for a donor to replace a cow after its passing. This practice of perpetual donation, very little studied elsewhere but apparently mostly attested in New France legal documents, underlines the crucial importance of cows as providers of food, labor, and manure in the colonial households.

The topics defined by the topic modelling approach illustrate the centrality of cattle to the settlers' survival and the transformation of the landscape. One of the main topics relates to the wintering of the cows (e.g. '*hyverné, hyvernée, hyverner, hyvernés*') and the need to have the animals prepared to survive the long and harsh Canadian winters. Given the major role played by cattle, especially in the opening of new agricultural land (also found through the topic modelling approach:

'*labour, laboureurs, labourer, laboureur*'), the preparation for winter to ensure the survival of the animals was critical, as already implied in the historical scholarship (D'Amour and Cossette, 2002). Other contexts in which cows are mentioned, relating mostly to other agricultural practices, highlight their diversified relevance to the settlers' everyday lives. These other topics include for example the realm of all the farm animals and the topic of reproduction and mating of cattle. Critically, another theme found through the topic modelling approach potentially refers to the conflictual nature of Europeans and the indigenous communities and the role of cattle in these relations. The topic relating to the term '*bestiaux*' revolves around notions of aggression. While the earlier historical scholarship had mentioned that cows were often the target of relaliatory actions by the Indigenous communities (Séguin, 1954), the high frequency of this topic highlights the relevance of this concern to the European colonists and suggests the relative regularity of such attacks.

The clustering of terms related to cows shows that they were often discussed in conjunction with other livestock and agricultural practices, paralleling the results of topic modelling and refining our interpretation of the historical documentation. One theme that stands out is the use of cows as a comparison to describe animals prior unknown to the European settlers. From the onset of the European presence in the Western Hemisphere, natural histories and chronicles describing the natural oddities found in the new continent were a common literary production designed to European audiences (Gerbi, 2010).

Interestingly, other topics appear to be completely absent from the ones identified through these methods. In particular, the question of the origins, the introduction, and the adaptation of bovine populations from Europe to the Americas is not addressed in any of the documents. Over the past decades, the historical scolarship have repeatedly asserted an inferrential narrative stating that cows from New France were imported by French settlers from Nortwestern France (Brittany, Normandy) (Séguin, 1954; Trudel, 2016). Such a narrative seems relatively unsubtantiated by the archival documents, as confirmed by our study, relying instead on circumstancial evidence such as the point of origin of these early settlers. The critical lack of such data underlines the relevance of the other component of our project (RABBA) that aims at investigating these aspects through the biomolecular analysis of archaeological cattle specimens.

5.2 Methodological implications

We find that the four approaches we used to approach our corpus data yield results that overlap to some extent, but remain complementary. Textometry and static word embeddings both pointed to the unexpected concept of 'immortal cow', though the word embeddings certainly provide a more flexible tool to find other latent associations, and especially analogies. Future work will focus on characterizing analogies, for example in the treatment of cows as opposed to other forms of livestock.

The most relevant methodological finding is in the approach discussed in section 4.4 that relies on clustering LLM representations. First, the method provides immediate access to the topics formed by the occurrences of target terms in the corpus, as opposed to topic modelling algorithms which might simply not identify such topics, as was the case for the Published sub-corpus. For that corpus, the clustering approach did provide a relevant cluster. Second, the method also confirms the robustness of the LLM representations offered by models such as d'AlemBERT. Part of the corpus is imperfect, due to OCR errors, and because French spelling proved highly variable in our time period of interest. Yet, the clustering approach was able to provide meaningful and interpretable results. This suggests that the method can reliably be used to investigate other topics in historical documents.

Acknowledgments

This work has been supported by a SSHRC-CRSH Banting Fellowship (PI: Nicolas Delsol), and an NSERC-Discovery Grant (RGPIN-2024-06718, PI: Grégoire Winterstein).

We are particularly thankful to the team of the *Nouvelle France Numérique* project for their help in collecting part of our corpus data, in particular to Maxime Gohier, Dominique Deslandres, Léon Robichaud and Kim Petit.

References

Nicolas Delsol. 2024. *Cattle in the Postcolumbian Americas: a zooarchaeological historical study*, 1st edition. University Press of Florida. OCLC: 1396141429.

Yvon Desloges. 2009. *À table en Nouvelle-France: alimentation populaire, gastronomie et traditions*

alimentaires dans la vallée laurentienne avant l'avènement des restaurants. Septentrion.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Valérie D'Amour and Évelyne Cossette. 2002. Le bétail et l'activité économique en Nouvelle-France: la vente et la location. *Revue d'histoire de l'Amérique française*, 56(2):217–233. Publisher: Institut d'histoire de l'Amérique française.

Roman Egger and Joanne Yu. 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7:886498.

Katrin Erk and Gabriella Chronis. 2023. Word embeddings are word story embeddings (and that's fine). In Shalom Lappin and Jean-Philippe Bernardy, editors, *Algebraic Structures in Natural Language*. Taylor and Francis, Boca-Raton and Oxford.

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. From FreEM to d'AlemBERT: a large corpus and a language model for early Modern French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3367–3374, Marseille, France. European Language Resources Association.

Jean-Denis Gendron. 2013. *D'où vient l'accent des Québécois? Et celui des Parisiens ? Essai sur l'origine des accents. Contribution à l'histoire de la prononciation du français moderne*. Presses de l'Université Laval, Québec.

Antonello Gerbi. 2010. *Nature in the New World: from Christopher Columbus to Gonzalo Fernandez de Oviedo*, 1. paperback ed edition. Univ of Pittsburgh Pr, Pittsburgh, Pa.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Dan Hicks and Mary C. Beaudry. 2006. Introduction: the place of historical archaeology. In Dan Hicks and Mary C. Beaudry, editors, *The Cambridge Companion to Historical Archaeology*, pages 1–10. Cambridge University Press, Cambridge.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2021. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36, NIT Silchar, India. NLP Association of India (NLPAI).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Bénédicte Pincemin. 2011. Sémantique interprétative et textométrie. *Corpus*, 10:259—-269.

Sébastien Plutniak. 2022. What makes the identity of a scientific method? a history of the "structural and analytical typology" in the growth of evolutionary and digital archaeology in southwestern europe (1950s–2000s). *Journal of Paleolithic Archaeology*, 5(1):10.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Robert-Lionel Séguin. 1954. Étude d'histoire économique: les bêtes à cornes et leurs implications historiques en amérique française. *Revue d'histoire de l'Amérique française*, 7(4):538–557.

Naomi Sykes, Piers Beirne, Alexandra Horowitz, Ione Jones, Linda Kalof, Elinor Karlsson, Tammie King, Howard Litwak, Robbie A. McDonald, Luke John Murphy, Neil Pemberton, Daniel Promislow, Andrew Rowan, Peter W. Stahl, Jamshid Tehrani, Eric Tourigny, Clive D. L. Wynne, Eric Strauss, and Greger Larson. 2020. Humanity's best friend: A dog-centric approach to addressing global challenges. *Animals*, 10(3):502.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593—-4601. Association for Computational Linguistics.

Erin Kennedy Thornton. 2016. Introduction to the special issue - turkey husbandry and domestication: Recent scientific advances. *Journal of Archaeological Science: Reports*, 10:514–519.

Marcel Trudel. 2016. *The Beginnings of New France 1524-1663*, volume 2. McClelland & Stewart.

Edward Vanhoutte. 2016. The gates of hell: History and definition of digital| humanities| computing 1. In *Defining Digital Humanities*, pages 119–156. Routledge.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ethan Watrall, Matthew K Gold, and Lauren F Klein. 2016. *Archaeology, the Digital Humanities, and the 'Big Tent.'*, volume 8. JSTOR.

A Corpora

The following sections indicate all the document that we included in our two sub-corpora, along with their initial dates of publication and size in number of tokens.

A.1 Published sub-corpus

Title	Period	Size
Histoire véritable et naturelle de la Nouvelle-France	1664	291805
Jugements et délibérations du Conseil souverain de la Nouvelle-France [microforme] v5v6	1710-1716	549884
Jugements et délibérations du Conseil souverain de la Nouvelle-France v1	1663	458012
Jugements et délibérations du Conseil souverain de la Nouvelle-France v2	1676	490399
Jugements et délibérations du Conseil souverain de la Nouvelle-France v3	1686	495702
Jugements et délibérations du Conseil souverain de la Nouvelle-France v4	1696	504627
Jugements et délibérations du Conseil souverain de la Nouvelle-France; publiés sous les auspices de la Législature de Québec .. V1	1663	459329
Jugements et délibérations du Conseil souverain de la Nouvelle-France; publiés sous les auspices de la Législature de Québec .. V2	1676	496407
Relations des Jésuites contenant ce qui s'est passé de plus remarquable dans les missions des Pères de la Compagnie de Jésus dans la Nouvelle-France	1611, 1626, 1632–1641	762561
Relations des Jésuites contenant ce qui s'est passé de plus remarquable dans les missions des Pères de la Compagnie de Jésus dans la Nouvelle-France v2	1642-1655	460472
Relations des Jésuites contenant ce qui s'est passé de plus remarquable dans les missions des Pères de la Compagnie de Jésus dans la Nouvelle-France v3	1656-1672	430978
Voyage de Kalm en Amérique	1753-1761	66409
Histoire de la Nouvelle-France by Marc Lescarbot	1617	291805
Le grand voyage du pays des Hurons by Gabriel Sagard	1632	90830
Voyage de J. Cartier au Canada Oeuvres de Champlain 1599-1632	1544	45177

A.2 Nouvelle France Numérique sub-corpus

The documents in the Nouvelle France Numérique sub-corpus were graciously shared by the project *Nouvelle France Numérique* (https://nouvellefrancenumerique.info/). All the documents come from archives that were scanned and passed through OCR. After the title of the document we indicate the label used to identify the document in the relevant archive.

Title	Period	Size
41765 - Baillage de Montréal - Registres 1 à 35 BAnQ-MTL | TL2, S11	1665-1693	1407683
55686 - Correspondance générale, Louisiane ANOM | C13A	1694	5108311
77146 - Fonds Viger-Verreau et Fonds Casgrain MCQ | ASQ,O | P32,O94D et MCQ | ASQ,O | P32,O94b	1754-1755	53800
78302 - Ordonnances d'intendant BAnQ-Qc | E1,S1	1705-1707	2190025
129439 - Michel Saindon BAnQ-Rim | CN104,S50	1768 à 1780	276296
178772 - Nicolas-Jean Olide Kervezo BAnQ-Rim | CN104, S44	1742-1755	31347

B Keywords lists

We list the keyterms used as initial search terms as well as those that came out of our various analyses. We provide English equivalents for all the terms when they are unambiguously interpretable.

B.1 Search keywords

Keyword	Translation
bestiaux	*cattle*
boeuf	*beef*
génisse	*heifer*
taureau	*bull*
taurillon	*young bull*
vache	*cow*
veau	*calf/veal*

B.2 Analyses keywords

Keyword	Translation
apellent/appelés	*call(ed)*
atoucher	*touch*
atriotume	
attablissement	
attabues	
attaque (and its derivatives)	*attack*
attasine	
attavoix	
barique	*barrel*
besson	*twin, typ. for sheep*
bola	
camiral	
camisolle	*shirt*
caribous	*caribou*
cariolle	*cart*
ccechons/cochons/lochon	*pig(s)*
chicvat/cheval	*horse*
corne	*horn*
fermage	*farm rent*
ferrure	*metal hardware*
génisse	*heifer*
giarre	
grangé	*barn (and its content)*
hyverné (and its derivatives)	*to winter*
immortelle	*immortal*
jument	*mare*
labour (and its derivatives)	*plow*
moutons	*sheep*
nourituraux	*food*
pacagée	*grazed*
peaux	*skins*
pistolle	(the currency at the time of NF)
pouliche	*filly*
taure	*bull*
troisueaux (poss. trousseau?)	*dot*
truye	*sow*