Disentangling language change: sparse autoencoders quantify the semantic evolution of indigeneity in French

Jacob A. Matthews[*], **Laurent Dubreuil**[*], **Imane Terhmina**[*], **Yunci Sun**,
Matthew Wilkens, **Marten van Schijndel**
Cornell University
{jam963, ld79, it228, ys463, wilkens, mv443}@cornell.edu

Abstract

This study presents a novel approach to analyzing historical language change, focusing on the evolving semantics of the French term "*indigène(s)*" ("indigenous") between 1825 and 1950. While existing approaches to measuring semantic change with contextual word embeddings (CWE) rely primarily on similarity measures or clustering, these methods may not be suitable for highly imbalanced datasets, and pose challenges for interpretation. For this reason, we propose an interpretable, feature-level approach to analyzing language change, which we use to trace the semantic evolution of "*indigène(s)*" over a 125-year period. Following recent work on sequence embeddings (O'Neill et al., 2024), we use k-sparse autoencoders (k-SAE) (Makhzani and Frey, 2013) to interpret over 210,000 CWEs generated using sentences sourced from the French National Library. We demonstrate that k-SAEs can learn interpretable features from CWEs, as well as how differences in feature activations across time periods reveal highly specific aspects of language change. In addition, we show that diachronic change in feature activation frequency reflects the evolution of French colonial legal structures during the 19th and 20th centuries.

1 Introduction

The concept of indigeneity, particularly its relationship to language and culture, is the subject of longstanding, interdisciplinary academic study. Postcolonial researchers and writers, for example, have repeatedly emphasized the role of language as central to understanding colonial power (Said, 1977; Bhabha, 2004; Dubreuil, 2013, inter alia). The effects of historical and contemporary structural inequalities have also been addressed in computational linguistics, as low-resource indigenous languages pose technical (Mager et al., 2018; Stap and Araabi, 2023) and ethical (Wiechetek et al., 2024) challenges related to their use as training material for contemporary natural language processing (NLP) systems. More broadly, the United States has only recently recognized Indigenous People's Day as a national holiday, reflecting increasing popular and political attention towards the history of indigenous communities in North America (The White House, 2021).

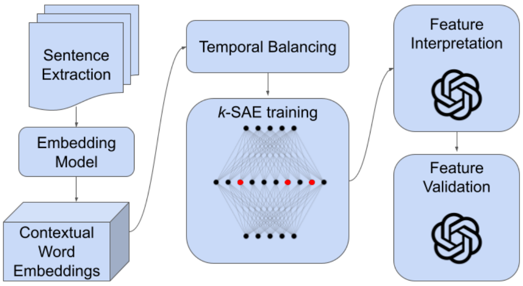


Figure 1: We use k-SAEs to identify interpretable features in over 200,000 contextual word embeddings corresponding to the word "*indigène(s)*". Features identified by a k-SAE are subsequently given natural language descriptions by GPT-4o and validated by GPT-4o-mini.

Just as indigenous peoples and languages are themselves not interchangeable and must be understood within their appropriate historical and linguistic contexts, the category of "indigenous" is likewise contingent on the specific context in which it is deployed. While the French cognate "*indigène*" has a similar contemporary definition to "indigenous", it also has its own historical, cultural, and linguistic connotations related to French colonial rule, which was often overtly violent and oppressive towards subjects deemed "indigène(s)" (Mann, 2009). To understand the historical and linguistic impact of French colonial policies related to indigeneity, we examine here how "*indigène*" changes in meaning over the 19th and early-20th centuries. Due to the complexities and sensitive nature inherent to the study French colonialism,

[*] Equal contribution.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11208–11222
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

we strove to maximize the breadth, specificity, and interpretability of our analyses, while at the same time leveraging the power of contemporary methods and technologies.

With this in mind, we present an interpretable method for analyzing historical language change through a new large-scale French-language diachronic dataset, which we collected. Whereas contemporary approaches to measuring historical language change often compute word similarities or clusters directly from dense contextual word embeddings (CWE) extracted from large language models (LLMs; Periti and Montanelli, 2024), we instead use sparse autoencoders (SAE) to identify interpretable features from dense CWEs corresponding to our target word, "*indigène(s)*". This method, directly adapted from recent work in interpretability research on LLMs (Gao et al., 2024) and sequence embeddings (O'Neill et al., 2024), has two key advantages. First, it enables us to find and validate natural language descriptions of features learned from CWEs without supervision. Second, it allows for feature-level analysis of diachronic change, as opposed to discrete sense-level or word-level approaches. Through our approach, we find that temporal changes in specific feature activations reflect known events and periods of French colonial history.

Our method and results represent relevant contributions to several domains. Notably, we demonstrate how SAEs can be used for analyzing historical language change when similarity- or clustering-based approaches are intractable or undesirable. In addition, our method provides additional insight into precisely what CWEs can capture about word-level meaning, as well as the current capabilities of automated feature labeling techniques.

2 Motivation and Background

The word "*indigène(s)*" differs from its English cognate "indigenous" both grammatically and semantically. It can be used as either an adjective (as in "*plantes indigènes*", "indigenous plants") or as a substantive (as in "*les indigènes*", "the indigenous [people]", or "the natives") (Trésor de la Langue Française informatisé, 2024). Specific to "*indigène(s)*" are its highly charged historical connotations tied to French colonial rule, including its deployment as a legal category. Early precedents include the *Sénatus-consulte* of April 22, 1863, which legally defined indigenous landholdings (Archives Nationales d'Outre-Mer, 1863), and the Crémieux Decree in 1870 (France, 1870), which selectively granted French citizenship to Algerian colonial subjects on the basis of religion and ethnicity. Those deemed "*indigène(s)*" were subsequently subjected to social segregation policies analogous to Jim Crow in the United States, under a set of laws referred to as the *Code de l'indigénat* ("Indigenous code"; Jakus, 2017; Hayes, 2021), which took form between 1881 and 1946 (Mann, 2009, inter alia). By the 1960s, it had been noted in Algeria that "*indigène(s)*" had largely taken on a pejorative connotation and had therefore begun to disappear from common use (Bousquet, 1961).

Given these factors, we wanted to determine whether historical trends in the use of "*indigène(s)*" were visible in large-scale textual data and would reflect the evolution of these legal and social frameworks, such as ethnic and racial segregation. More specifically, we hypothesized that the botanical or zoological senses of "*indigène(s)*" would be more prominent in the early 19th century, whereas the uniquely human, specifically ethnic, legal, or military senses would be more visible from the late 19th century onward.

While there has been longstanding interest in the relationship between semantics, history, and society in the humanities (Benveniste, 2017, inter alia), there are also a variety of approaches to semantic shift detection (SSD) in the computational linguistics literature. Prior work typically relies on contextual word embeddings (CWE), word-level representations extracted from pretrained masked language models like BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019), which are subsequently averaged or clustered and then compared across time steps (Periti et al., 2022; Periti and Montanelli, 2024). Other supervised or more specialized approaches also exist, such as in Hoffman et al. (2020), which uses graph representations of social networks in addition to temporal labels to enrich language model representations with additional sociotemporal context. More generally, there exist diverse computational historical analysis techniques that have been developed for other modalities, such as for historical change in melody (Hamilton and Pearce, 2024) and tracing the circulation of archival images (Du et al., 2024).

At the same time, there is a growing body of work showing that using CWEs generated by LLMs may come with significant drawbacks. These include anisotropy (Gao et al., 2019; Ethayarajh,

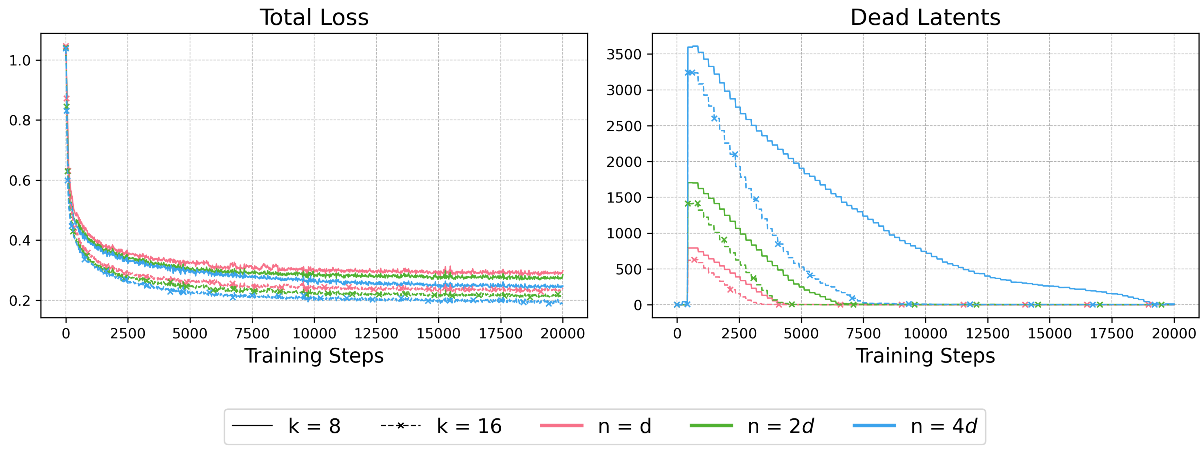


Figure 2: Total loss and number of dead latents for all k-SAE configurations during training.

2019; Hämmerl et al., 2023), "rogue dimensions" (Timkey and van Schijndel, 2021), lack of noise robustness (Matthews et al., 2024), and social bias (Guo and Caliskan, 2020). These studies suggest that even though CWEs perform reasonably well on downstream tasks, comparing CWEs directly (particularly with measures such as cosine similarity) may not be a reliable proxy for semantic similarity.

Within the mechanistic interpretability literature, however, there have been significant recent advancements using sparse autoencoders (SAE) to interpret LLM representations (Cunningham et al., 2023; Bricken et al., 2023; Gao et al., 2024). Individual neurons in an LLM (or, in our case, dimensions in a CWE) may correspond to multiple human-interpretable features, a phenomenon referred to as "superposition" (Elhage et al., 2022). SAEs, by learning to represent dense model representations as higher-dimensional sparse representations, encourage greater "monosemanticity", where individual neurons in the SAE hidden state correspond to a single, human-interpretable feature (Bricken et al., 2023). More recently, this technique has been applied to sequence embeddings, further demonstrating that SAEs can identify highly interpretable, monosemantic features even from pooled or sequence-level representations (O'Neill et al., 2024). An additional benefit of these approaches is the relative simplicity of finding and validating natural language descriptions for each feature using another language model (Bills et al., 2023), as well as the ability to analyze the SAE's learned feature representations (that is, columns in its decoder weight matrix; O'Neill et al., 2024).

We offer several original contributions that draw from each of these research areas. Building on prior work by O'Neill et al. (2024) and Gao et al. (2024), we demonstrate how SAEs can be used to identify interpretable features from CWEs, even when these CWEs correspond to only two distinct lexical items ("*indigène*" and "*indigènes*"). We go on to show that features with high levels of diachronic variation have activation frequencies that correspond to known events in French history which influenced the signification of the term. Our method and findings highlight how interpretability methods can be applied productively to difficult, nuanced questions in historical language change. Finally, we find support for our hypotheses regarding the semantic shift of "*indigène(s)*" in our dataset.

3 Dataset Construction

To our knowledge, there is no existing, historical French-language dataset suitable for diachronic analysis. For this reason, we cooperated with the French National Library (BNF) to source monographs preprocessed with optical character recognition (OCR) along with metadata, including date of original publication. All of the monographs we requested were free of copyright restriction as of 2023. While our original request encompassed nearly all monographs in the BNF's digital collection, the present study is limited to the first batch retrieved from their servers, which includes approximately 71k volumes. Additional information regarding this data is available in Appendix A.

From these complete texts, we extracted all sentences that included the word *indigène* or *indigènes* using a regular expression. We retain sentences longer than ten whitespace-delimited words in length in order to ensure that our CWEs are gen-

n	k	Dead Latents	Mean Acc.	Mean F_1	Mean Corr.	Highly Interp.	% Highly Interp.
d	8	**0**	0.6050	0.4442	0.2885	63	6.15
	16	**0**	0.5757	0.4082	0.2138	33	3.22
$2d$	8	3	0.6069	0.4634	0.2858	108	5.27
	16	**0**	0.5863	0.4249	0.2340	76	3.71
$4d$	8	5	**0.6265**	**0.4959**	**0.3225**	**281**	**6.86**
	16	1	0.5973	0.4456	0.2564	200	4.88

Table 1: Feature interpretability results for trained k-SAEs with different values of k and n. We report the total number of dead latents, mean Predictor accuracy (Mean Acc.), mean Predictor F_1 (Mean F_1), mean Predictor correlation (Mean Corr.), total number of highly interpretable features (Highly Interp.), and highly interpretable features as a percentage of n (% Highly Interp.). Highly interpretable features have Predictor F_1 and correlation scores > 0.8.

erated with sufficient context. To avoid ambiguity in our diachronic analysis, we excluded sentences sourced from texts that lacked a precise publication date (either due to missing metadata or where publication date was recorded as a range of years). Due to extreme data sparsity in very early or late time periods, we limited our analysis to sentences between 1825 and 1950. In total, we include 210,305 sentences in our sentence dataset.[1] In order to facilitate diachronic analyses over 125 years (though not necessarily to align with typical periodizations of French history), we assigned each of these sentences to one of five 25-year time periods based on date of publication. We note that these time periods are severely imbalanced: our smallest time period (1851-1875) comprises 4,942 sentences, while our largest (1901-1925) includes 92,767 sentences. We address this particular challenge further in Section 4.2.

To generate CWEs from these extracted sentences, we use a masked language model, CamemBERT (Martin et al., 2019), a variant of RoBERTa (Liu et al., 2019) optimized for French-language applications. We extract the last hidden state activations of the pretrained `camembert-large` model that correspond to the tokens in "*indigène(s)*", then mean pool token-level representations to construct a single CWE for each sentence.

4 Interpreting Contextual Word Embeddings with Sparse Autoencoders

4.1 k-Sparse Autoencoders

SAEs are simple MLPs comprising two layers: an *encoder* that maps an input vector $\mathbf{x} \in \mathbb{R}^d$ to a sparse hidden representation $\mathbf{h} \in \mathbb{R}^n$, and a *decoder* layer, which reconstructs the input vector $\mathbf{x}$ as $\hat{\mathbf{x}}$ from the sparse hidden representation $\mathbf{h}$. Following prior work (Gao et al., 2024; O'Neill et al., 2024), we use k-sparse autoencoders (k-SAEs; Makhzani and Frey, 2013) to identify features in CWEs. k-SAEs enforce sparsity by setting all but the top k values in $\mathbf{h}$ to zero. The encoder for our model is therefore given by

$$\mathbf{h} = TopK(W_e\mathbf{x} + \mathbf{b}_e) \tag{1}$$

where W_e is the encoder weight matrix, $\mathbf{b}_e$ is the encoder bias vector, and $TopK$ is an activation function that sets all but the k-largest values to zero, ensuring that $\mathbf{h}$ is sparse. Likewise, our decoder is given by

$$\hat{\mathbf{x}} = W_d\mathbf{h} + \mathbf{b}_d \tag{2}$$

where W_d is the decoder weight matrix and $\mathbf{b}_d$ is the decoder bias vector. Our k-SAE has learnable parameters $\theta_e = \{W_e, \mathbf{b}_e\}, \theta_d = \{W_d, \mathbf{b}_d\}$.

Gao et al. (2024) and O'Neill et al. (2024) describe techniques for minimizing "dead latents" during training. In this context, a dead latent is simply a feature that is never activated by $\mathbf{h}$. More precisely, we say that a latent i is "dead" when $\mathbf{h}_i = 0\, \forall \mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ is the set of input vectors in training data. Dead latents can be minimized or eliminated during training using the following auxiliary loss function:

$$\mathcal{L}_{aux}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{d}\|(\mathbf{x} - \hat{\mathbf{x}}) - (W_d\mathbf{h}_{aux} + \mathbf{b}_d)\|_2^2 \tag{3}$$

where

$$\mathbf{h}_{aux} = TopK_{aux}(W_e\mathbf{x} + \mathbf{b}_e) \tag{4}$$

and $TopK_{aux}$ sets all but the top k_{aux} dead latents to zero, which we identify after a predetermined

[1] Our sentence dataset and code are available at `https://github.com/jam963/disentangling-language-change`.

number of samples. Put simply, this auxiliary loss function is just the mean squared error of the difference between the model's overall reconstruction error and a reconstruction using only the top k_{aux} dead latents. Our composite loss function is then

$$\mathcal{L}(\theta_e, \theta_d) = \frac{1}{d}\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \alpha\mathcal{L}_{aux}(\mathbf{x}, \hat{\mathbf{x}}) \quad (5)$$

where α is a hyperparameter. Throughout, we opted to follow O'Neill et al. (2024) and set $k_{aux} = 2k$ and $\alpha = \frac{1}{32}$, in an effort to limit the number of tunable hyperparameters.

4.2 Training Details

We train all configurations of our k-SAEs for a fixed number of training steps (20,000), using a batch size of 1,024, fixed learning rate (1e-4), and the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon =$1e-8. All CWEs are standardized prior to training.

Because our ultimate goal is not to minimize SAE reconstruction loss on held-out data, but rather to efficiently identify interpretable features, we train on all embeddings in our dataset. However, in an attempt to prevent potential problems caused by dataset imbalance with respect to publication date, we randomly sample a fixed number of CWEs from each time period every epoch for training. During prototyping, we experimented with different levels of per-period downsampling, setting the number of CWEs sampled per time period to a fraction of the size of the smallest time period. In practice, because we did not observe any measurable advantage to higher levels of downsampling, we randomly sample the total size of our smallest time period (4,942) from each time period, resulting in 24,710 training samples per epoch. While this ensures that the our k-SAE does not drastically underfit to time periods with fewer instances, it also means that only a small percentage of embeddings from other time periods are seen during each epoch, potentially leading to underfitting on these larger bins. We compensate for this by training for a large number of total training steps, increasing the probability that all CWEs will ultimately be seen during training. Importantly, though, this random sampling does not affect how often we account for dead latents, which are tracked over a fixed number of training samples equal to the total number of samples in our dataset.

4.3 Automated Feature Labelling and Validation

To generate and validate natural language descriptions for features in each of our trained k-SAE models, we adapt the interpretability technique outlined in Bills et al. (2023) and O'Neill et al. (2024). This method involves two stages: first, an *Interpreter* uses max-activating sentences and zero-activating sentences to generate a description for each feature. Max-activating sentences are sentences whose associated CWE maximally activates a given feature, whereas the CWE associated with zero-activating sentences does not activate that feature at all. Then, a *Predictor* is tasked with generating a confidence score between -1 and 1 reflecting whether or not a label generated by the Interpreter applies to a sentence, which is either non-zero activating or zero-activating (Figure 1). From these confidence scores, we compute F_1 and Pearson correlation for each feature, which indicate the feature's interpretability.

For our Interpreter, we use GPT-4o and a prompt based on O'Neill et al. (2024), which we translated to French and significantly modified to reflect the word-level and domain-specific characteristics of our task. We noted in early testing that GPT-4o and GPT-4o-mini were inclined to generate overbroad, generic, highly moralizing or theoretical responses, which we discouraged in our prompt to encourage greater specificity in feature descriptions. To minimize API fees, we use 5 max-activating and 5 zero-activating sentences for each feature. Each feature is interpreted in a zero-shot, chain-of-thought fashion.

Likewise, for our Predictor, we adapt our approach from O'Neill et al. (2024), translating their prompt to French and editing to reflect our task. Due to financial considerations, we randomly sample 3 non-zero activating and 3 zero-activating sentences for each feature; we also use GPT-4o-mini to generate confidence scores.[2] We prompt for each sentence-label pair in a zero-shot, chain-of-thought fashion. Using these confidence scores generated by the Predictor and the true scores for non-zero activating (1.0) and zero-activating sentences (-1.0), we compute accuracy, F_1, and Pearson correlation for each feature. Finally, given these feature-level

[2] We incurred approximately $280 in OpenAI API fees related to this project, which included prototyping and prompt engineering. The cost of feature interpretation and prediction for a single k-SAE ranged in price from approximately $15 to $60, depending on the latent dimensionality of the model (n).

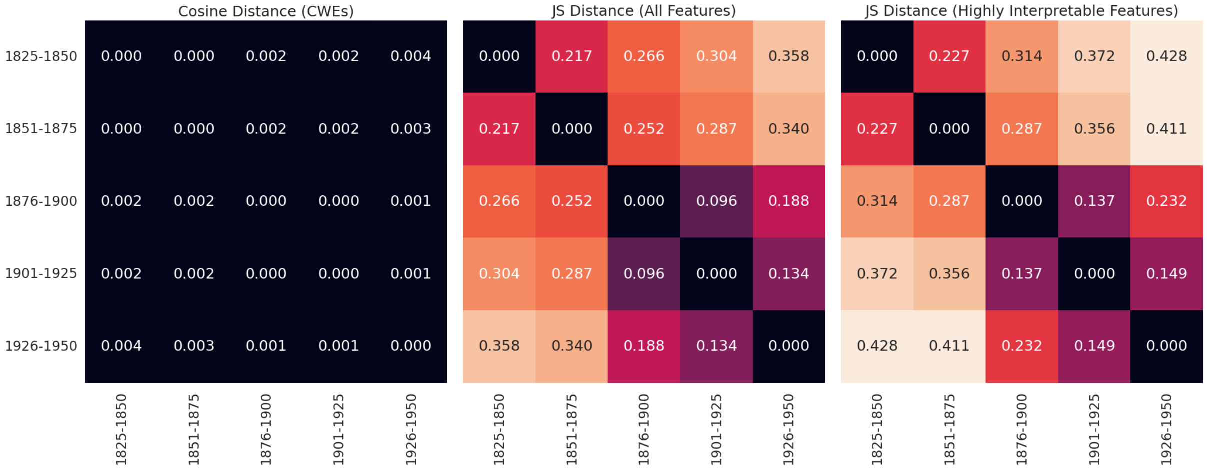


Figure 3: We compare pairwise JS distance of feature distributions across time periods ($n = 4d$, $k = 8$) (center, right) and cosine distance of averaged CWEs (left). For all features (center) and only highly interpretable features (right), periods that are further apart in time have more distant activation distributions as measured by JS distance. By contrast, averaged CWEs are all highly self-similar and do not show pronounced temporal variation (left).

scores, we are then able to identify "highly interpretable features" as those with F_1 and correlation above a threshold value (0.8). We emphasize that, in contrast to the Interpreter, the Predictor must correctly attribute the generated feature label to *non-zero activating* sentences for each feature, not just max-activating sentences. This avoids assigning high interpretability scores to feature labels which only apply to highly-activating examples. In other words, the Predictor expects that a feature label generated by the Interpreter should apply to sentences with any level of feature activation above zero.

5 k-SAE Performance and Interpretability

5.1 Training Metrics

Following prior work (O'Neill et al., 2024), we train six k-SAEs with $k \in \{8, 16\}$ and $n \in \{d, 2d, 4d\}$, where d is our CWE dimensionality (1,024). We report total loss and number of dead latents for each model during training in Figure 2. Despite our random sampling procedure used to balance samples temporally during training, we observe that all models converge, though some variants (those with high n and/or low k) still have a small number of dead latents (Table 1). We note that higher values of k and n give better reconstruction accuracy (lower total loss), but also require more training steps to remove dead latents.

5.2 Feature Interpretability

We report results of our interpretability procedure (Section 4.3) in Table 1. All trained k-SAEs have zero or very few dead latents, and we do not observe a clear relationship between reconstruction quality or final number of dead latents and any measures of feature interpretability. Indeed, our best performing model across all interpretability measures ($n = 4d$, $k = 8$) has a non-zero number of dead latents after 20,000 training steps, and exhibits worse reconstruction accuracy compared to all SAEs with $k = 16$. In general, SAEs with $k = 8$ produce more interpretable features on average than those with $k = 16$, and higher n is associated with more modest interpretability gains.

6 Feature Activation Analysis

6.1 Highly Interpretable Features

With our best-performing k-SAE ($n = 4d$, $k = 8$), we are able to identify 281 highly interpretable features from our CWEs. The feature interpretations produced by the Interpreter are short, often highly-specific descriptions of subject-level aspects of our sentence data. Reflecting the diversity and breadth of our historical dataset, feature descriptions encompass an extremely wide range of topics and contexts, ranging from agriculture ("Native vegatable species, comparative botanical context") to stereotyped descriptions ("Exploration, danger, mystery, 'some native'") to religious affiliation and ethnicity ("Indigenous muslims in French administrative context"). We provide examples of several highly-interpretable features and translations of their generated descriptions in Figures 4 and 5.

6.2 Temporal Variation in Feature Activations

Because we are primarily interested in how different uses and meanings of the term "*indigène(s)*" evolve over time, we explore how feature activation distributions change across 25-year time periods. To determine to what extent all features change across time periods, we first convert our feature activations for each time period to probability distributions over features p_t. To compare these distributions across time periods, we compute the pairwise Jensen-Shannon (JS) distance (Lin, 1991) between each time period, which in our case is given by

$$m = \frac{p_i + p_j}{2} \tag{6}$$

$$JS(p_i, p_j) = \sqrt{\frac{D(p_i \| m) + D(p_j \| m)}{2}} \tag{7}$$

where i and j are time periods and D is Kullback-Leibler divergence (Kullback and Leibler, 1951).

We report our results as heatmaps in Figure 3. Across time periods, we find that feature activation distributions from more temporally-distant periods show correspondingly higher JS distance. This holds true for distributions over all features as well as for distributions over only highly interpretable features, which suggests that our k-SAE is able to capture underlying diachronic variation present in CWEs. For reference, we also report pairwise cosine distances of average CWEs for each time period. In contrast to activation distributions, though, averaged CWEs are extremely self-similar, and are effectively identical when compared with cosine (Figure 3, left). This suggests that k-SAEs are able to effectively identify meaningful, feature-level differences between embeddings, even when these embeddings may all be virtually identical to one other in terms of cosine distance. We also include a standardized similarity analysis in Appendix B.

While JS distance provides an overall measure of how word meaning and use vary according to publication date, we also explore how individual features or subsets of feature activations evolve across time periods. From our 281 highly interpretable features identified by our best-performing model, we find the top 5 features with the largest absolute change in activation frequency between our first time period (1825-1850) and our last time period (1926-1950). We visualize the results of this analysis in Figure 4. What we find supports our earlier hypotheses: features associated with botanical use become less prominent in the period 1876-1900, during which time the *Code de l'indigénat* began to take form. In the same way, features associated with legal and administrative structures ("French colonial administrative and indigenous policy", "Administrative or legal institutional indigenous context") increase in activation frequency over time.

6.3 Qualitative Evaluation of Highly-interpretable Features

A key advantage of our approach is the ability to identify and generate labels for specific semantic features in our dataset without supervision. While our best performing k-SAE identifies too many highly interpretable features to analyze here, we briefly highlight several interesting features and their feature activation frequencies in Figure 5. Some feature descriptions, such as "'The indigenous question' in colonial/debate context", seem to describe an n-gram which includes "*indigène(s)*". However, this particular feature activates not only for sentences containing that specific n-gram ("*question indigène*"), but also for semantically similar subsequences like "*problème indigène*" and even for those with OCR errors or textual noise ("*ques- tion indigènes*"). In other words, this feature captures a meaningful, highly granular topic that is often but not always associated with an n-gram, and would potentially be missed by lexical topic modeling approaches. We explore this aspect of our method further in Appendix C.

In addition, other feature descriptions are not only highly specific, but were unanticipated prior to model training. Inspecting sentences which activate "Locally produced sugar, economic-commerical context" (Figure 4), we find discussion of beet sugar production in France, which is described as "indigenous" in contrast to the "exotic" or "foreign" sugars produced abroad or in colonial territories. The activation frequencies over time may also reflect investment and taxation strategies for domestic sugar production, which had begun in the Napoleonic era and flourished in 1825-1850 (Griffin, 1902). Similarly, "Local priests in Chinese Catholic missions" activates across all time periods, though occasionally seems to activate for more generic contexts containing references to priests and Catholic institutions, but which lack specific

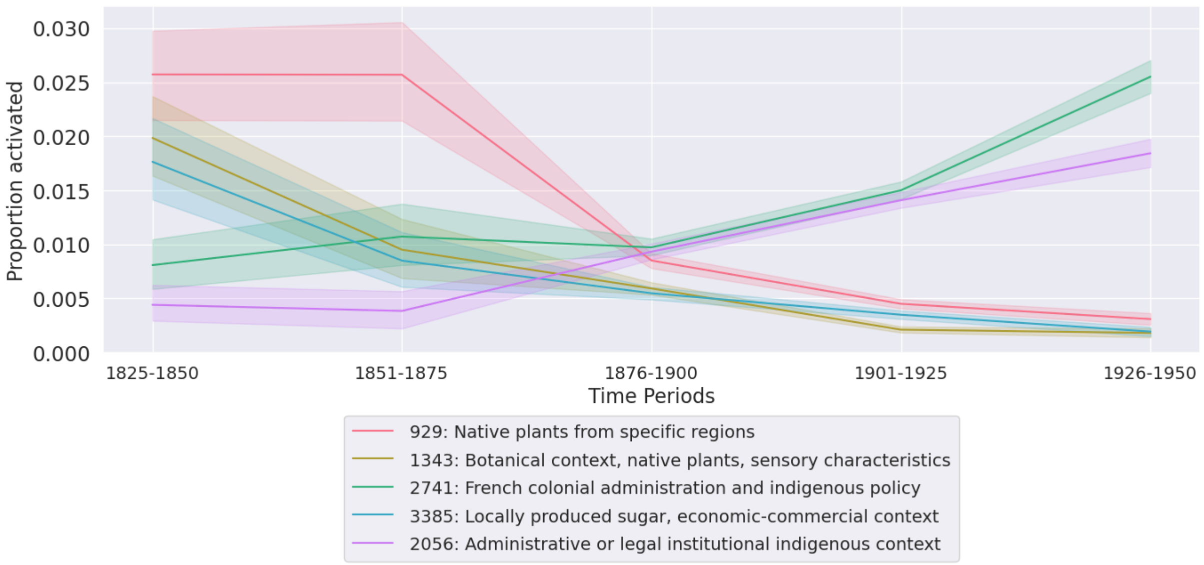


Figure 4: Highly interpretable features with greatest change in activation frequency between 1825-1850 and 1926-1950. The shaded region represents the 95% CI. Translations of generated feature descriptions suggest a steady decrease in botanical or specific agricultural use of "*indigène(s)*" after 1875 (929, 1343, 3385), whereas colonial, administrative, and policy-oriented uses increase sharply after 1876-1900 (2741, 2056).

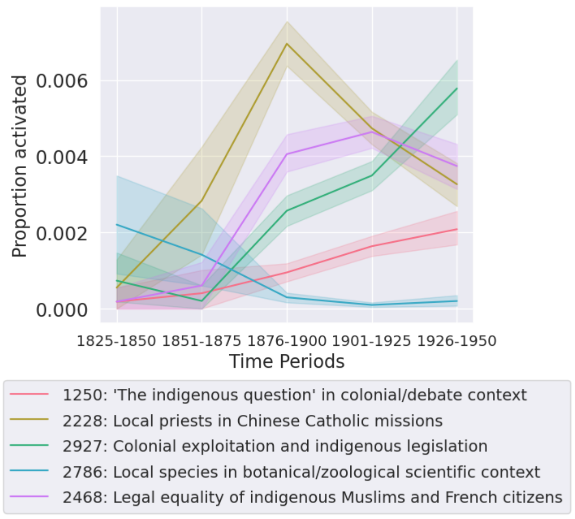


Figure 5: Selected highly interpretable features with activation frequency by time period. Some features, like 2228, reveal aspects of our data of which we were unaware prior to training. Historical activation trends for these selected features align with those identified in Figure 4.

mention of geographical location, or to locations historically associated with China, but outside its contemporary geographic boundaries. Following our identification of this feature, we were able to locate historical research attesting to French Catholic missionary presence in China since the 17th century (Dorais, 2005), demonstrating the potential utility of our feature descriptions for navigating large-scale historical data.

We also find qualitative evidence of feature "splitting", which has been documented in prior work (O'Neill et al., 2024), as well as overly generic or broad feature descriptions. Some feature descriptions, though highly interpretable, are essentially duplicates of one another, such as "Indigenous plants, specific botanical context" and "Native plants, detailed botanical descriptions". To ensure that the trend we observe in Figure 4 is not contradicted in other, similar features, we examined all features containing the word "plant" in their description, and observe that these features have either decreasing or very low overall activation frequency over time (see Appendix D). In this respect, it seems that many split features are simply redundant versions of other, more dominant features identified by our model, and as such reflect the same underlying historical trends.

7 Discussion

Our method allows for unsupervised identification, description, and validation of features from CWEs, as well as diachronic analysis of these features. We demonstrate the successful application of our method on a novel French sentence dataset, sourced from texts across 125 years of French and Francophone history. Through our analysis, we find that the meanings of "*indigène(s)*" not only evolved over time, but changed relative to one another in a manner consistent with our original hypotheses: the features with the greatest decrease in activation frequency reflect highly specific botanical contexts, whereas we observe an increase in features associated with colonial, legal, and administrative use beginning in the late 19th century. As a historical research tool, we intend to employ our k-SAE and generated feature descriptions to address additional qualitative concerns related to the concept of indigeneity in French history.

More broadly, our findings demonstrate the potential of SAEs for analyzing word-level language model representations, which can complement or potentially replace similarity-based methods. In line with recent work using LLMs to find and describe concepts in unstructured text datasets (Lam et al., 2024), our approach demonstrates how embeddings can be meaningfully described by generative models, adding interpretable insights to nuanced and difficult tasks in text analysis.

8 Ethics

Our sentence data is composed of historical source material related to French colonial treatment of indigenous populations, and as such includes highly offensive or inaccurate descriptions, characterizations, and attitudes related to race, religion, and ethnicity. We in no way endorse irresponsible, discriminatory, or defamatory use of this data or our analyses.

9 Limitations

While we made reasonable efforts to avoid bias when extracting sentences for our dataset, we cannot guarantee that our sample is entirely representative of the French language as a whole. Similarly, though our method attempts to account for our dataset's temporal imbalance, we cannot rule out the possibility that our analyses are still affected by the low number of samples in early time periods. Because we rely on our Predictor to generate measures to gauge feature interpretability, generated feature descriptions may not accurately reflect the content of every sentence for which a feature is activated. Because of financial constraints, our Predictor uses a limited number of randomly selected positive and negative examples. Finally, though we are interested in the overall change in pejorative or racial connotation associated with "*indigène(s)*" over time, this may only be indirectly inferred from extracted features or combinations of features, and may also require additional inspection of source text.

10 Acknowledgements

This project was conducted in the Humanities Lab at Cornell and was made possible by a special grant awarded by the French Embassy in the United States to the Cornell French Studies Program, with additional support (seed grant) from the Cornell Center for Social Sciences. Data acquisition was supported by grant NEH HAA-290374-23 (to MW). We are also grateful for the extremely valuable feedback we received from the *How to Say "Race" in French* 2024 workshop at Cornell University, the *Code Meets Culture* 2024 symposium at Villanova University, as well as the Cornell NLP group.

References

Archives Nationales d'Outre-Mer. 1863. Dossiers du Sénatus-consulte du 22 avril 1863. Accessed 10/09/2024.

Emile Benveniste. 2017. *Dictionary of Indo-European concepts and society*. HAU, Chicago, IL.

Homi K. Bhabha. 2004. *The Location of Culture*. Routledge, London and New York.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

Georges-Henri Bousquet. 1961. *Réflexions sur le mot "indigène"*, pages 396–402. Persée.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *ArXiv*, abs/2309.08600.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Louis-Jacques Dorais. 2005. Stratégies missionnaires des jésuites français en Nouvelle-France et en chine au XVII siècle. par shenwen li. (paris et québec, L'Harmattan et presses de l'université laval, 2001. pp. xvi + 379, ill., ISBN 2-7637-7792-9). *Ethnologies (Que)*, 27(1):335.

Lin Du, Brandon Le, and Edouardo Honig. 2024. Probing historical image contexts: Enhancing visual archive retrieval through computer vision. *J. Comput. Cult. Herit.*, 16(4).

Laurent Dubreuil. 2013. *Empire of Language*. Cornell University Press, New York and London.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Baker Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *ArXiv*, abs/2209.10652.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Conference on Empirical Methods in Natural Language Processing*.

France. 1870. Bulletin des lois de la République française. Accessed 10/09/2024.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. *ArXiv*, abs/1907.12009.

Leo Gao, Tom Dupr'e la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *ArXiv*, abs/2406.04093.

Charles S. Griffin. 1902. The sugar industry and legislation in europe. *The Quarterly Journal of Economics*, 17(1):1–43.

Wenqian Guo and Aylin Caliskan. 2020. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.

Madeline Hamilton and Marcus Pearce. 2024. Trajectories and revolutions in popular melody based on u.s. charts from 1950 to 2023. *Scientific Reports*, 14(1).

Katharina Hämmerl, Alina Fastowski, Jindich Libovický, and Alexander M. Fraser. 2023. Exploring anisotropy and outliers in multilingual language models for cross-lingual semantic sentence similarity. In *Annual Meeting of the Association for Computational Linguistics*.

Robin J. Hayes. 2021. *A Free Black Mind: Un Esprit Noir Libre Akili ya Bure Nyeusi*, pages 44–68. University of Washington Press.

Valentin Hofmann, Janet B. Pierrehumbert, and Hinrich Schütze. 2020. Dynamic contextualized word embeddings. In *Annual Meeting of the Association for Computational Linguistics*.

Julia Jakus. 2017. Laïcité and laiklik: When did the comparability of assertive laicism in france and turkey dissolve? *Sociology and Anthropology*, 5(11):923940.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Solomon Kullback and R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alireza Makhzani and Brendan J. Frey. 2013. k-sparse autoencoders. *CoRR*, abs/1312.5663.

Gregory Mann. 2009. What was the indigénat? the empire of law in french west africa. *The Journal of African History*, 50(3):331353.

Louis Martin, Benjamin Muller, Pedro Ortiz Suarez, Yoann Dupont, Laurent Romary, Eric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. In *Annual Meeting of the Association for Computational Linguistics*.

Jacob A. Matthews, John R. Starr, and Marten van Schijndel. 2024. Semantics or spelling? probing contextual word embeddings with orthographic noise. *ArXiv*, abs/2408.04162.

Charles O'Neill, Christine Ye, Kartheik G. Iyer, and John F. Wu. 2024. Disentangling dense embeddings with sparse autoencoders. *ArXiv*, abs/2408.00657.

Francesco Periti, Alfio Ferrara, Stefano Montanelli, and Martin Ruskov. 2022. What is done is done: an incremental approach to semantic shift detection. In *Workshop on Computational Approaches to Historical Language Change*.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11).

Edward Said. 1977. *Orientalism*. Penguin, London.

David Stap and Ali Araabi. 2023. ChatGPT is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167, Toronto, Canada. Association for Computational Linguistics.

The White House. 2021. A proclamation: Indigenous Peoples' Day, 2021. Accessed 10/12/2024.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. *ArXiv*, abs/2109.04404.

Trésor de la Langue Française informatisé. 2024. Entry for 'indigène'. `https://www.cnrtl.fr/definition/indig%C3%A8ne`. Accessed 10/1/2024.

Linda Wiechetek, Flammie A. Pirinen, Børre Gaup, Trond Trosterud, Maja Lisa Kappfjell, and Sjur Moshagen. 2024. The ethical question – use of indigenous corpora for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15922–15931, Torino, Italia. ELRA and ICCL.

A Dataset Construction

Following a request to the French National Library (BNF) for nearly all monographs in their digital collection, we received an initial batch of 88,298 documents, in no particular order, in the form of XML outputs from prior OCR processing. After converting the raw OCR information to usable text, we filtered documents to only include monographs with publication dates between 1825 and 1950, leaving 71,931 documents containing 3.9 billion whitespace-delimited words in total. Following manual inspection, we noted that OCR quality could be quite poor, particularly for texts from early time periods. Early on, this informed our decision not to use purely lexical approaches to topic modeling or SSD, as preliminary analyses revealed that textual noise would potentially introduce significant interpretive challenges. OCR quality, scale, and temporal imbalance also posed challenges for using part-of-speech tagging models to identify adjectival and substantive uses of our target term, since we have reason to believe that OCR quality is partially a function of publication date. However, reliable information regarding trends in adjectival versus substantive use would be valuable, as the substantive is almost always associated with humans.

From this filtered set of 71k documents, we extracted all sentences containing the word "*indigène*" or "*indigènes*" and excluded any sentence with less than 10 whitespace delimited words. This resulted in 210,305 total sentences. We report the sentence distributions with respect to 25-year time period in Figure 6. Sentence distributions are highly imbalanced over time, which we speculate is at least partially due to historical trends in publication volume, as well as internal BNF policies concerning text digitization. We also report the distribution of these sentences with respect to the genre label applied by the BNF in Figure 7. Many documents did not contain any genre information, either due to missing metadata or no genre assignment, which is included in the "Other" category.

B Standardized Similarity

Following Timkey and van Schijndel (2021), we repeated the similarity analyses in Section 6.2 after standardizing CWEs. We report our results in Figure 8. Unlike in Figure 3, which shows pairwise cosine distances between average CWEs for each time period, we instead visualize our standardized results in terms of cosine similarity. This is because the resulting similarity values in this particular analysis range from -1.0 to 1.0, which are less intuitive to visualize with cosine distance, where negative similarity values result in distances between 1.0 and 2.0. While the range of similarity values is much wider for standardized than for unstandardized CWEs, these results are difficult to interpret in terms of language change: many similarity values are negative, and similarity does not increase with temporal proximity.

C Lexical Feature Analysis

To ensure that feature descriptions generated by the Interpreter are reasonable, we used a simple lexical method to find keywords associated with each feature. We first compute unigram TF-IDF

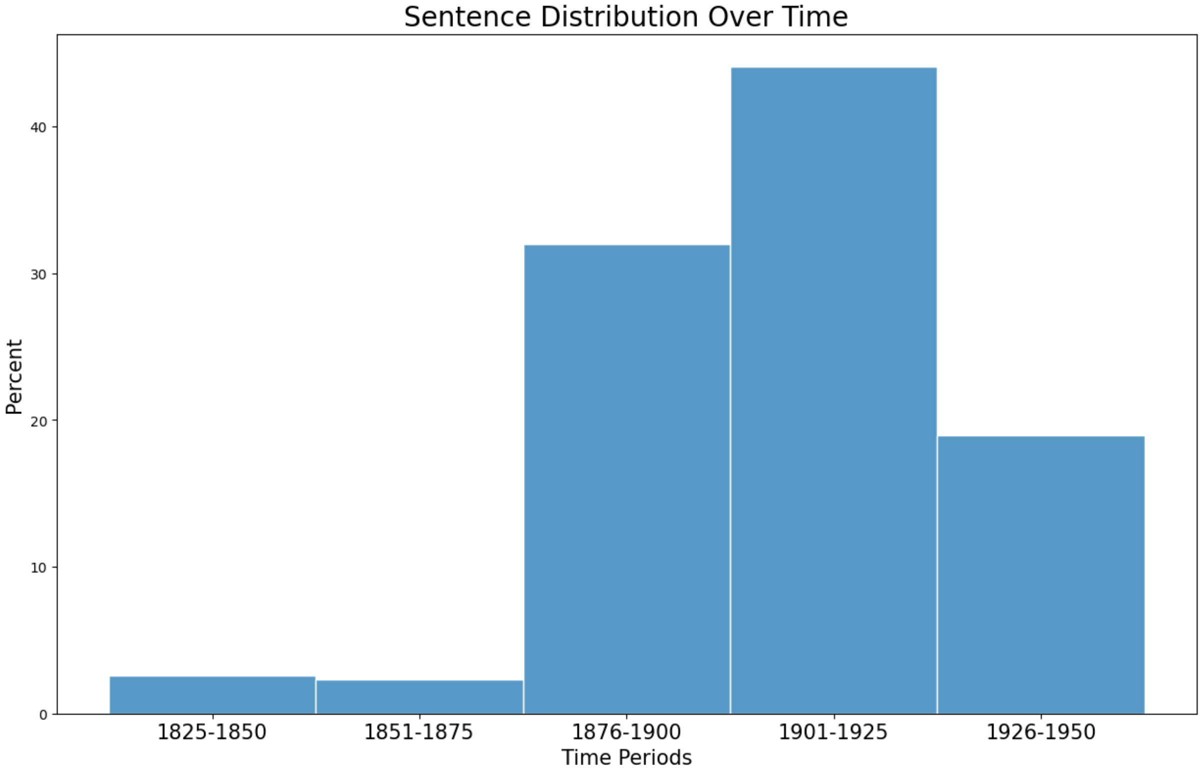


Figure 6: Distribution of sentences by time period.

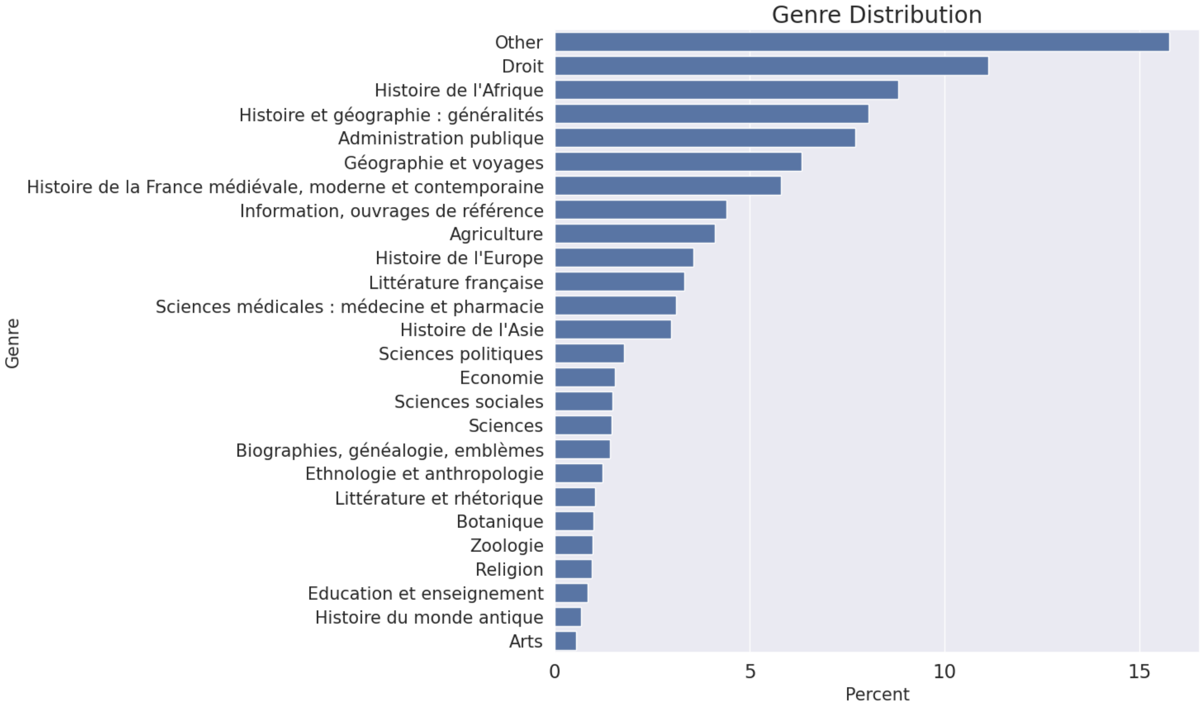
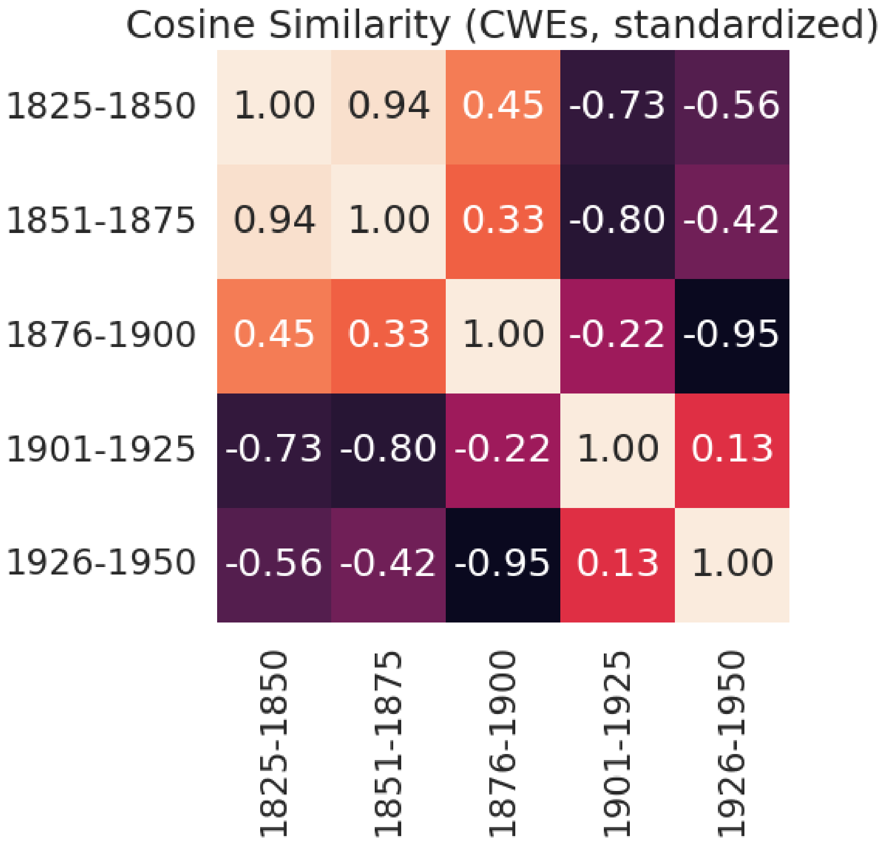


Figure 7: Distribution of genres for sentences in our dataset by time period.

vectors for each sentence, using a vocabulary size of 10,000 and a list of French stopwords. We then compute the Spearman correlation of each feature's activations and TF-IDF scores for each word. In Table 2, we show a sample of highly-interpretable feature descriptions generated by the Interpreter, along with three terms with the highest Spearman correlation to that feature. In general, we find that this lexical method associates keywords with similar meanings or associations to the short descrip-



Cosine Similarity (CWEs, standardized)

	1825-1850	1851-1875	1876-1900	1901-1925	1926-1950
1825-1850	1.00	0.94	0.45	-0.73	-0.56
1851-1875	0.94	1.00	0.33	-0.80	-0.42
1876-1900	0.45	0.33	1.00	-0.22	-0.95
1901-1925	-0.73	-0.80	-0.22	1.00	0.13
1926-1950	-0.56	-0.42	-0.95	0.13	1.00

Figure 8: Cosine similarities of standardized CWEs, by time period.

tions generated by the Interpreter. However, we also occasionally found examples of word fragments and numerical values with high correlations to certain features. In these cases, the Interpreter appears to synthesize a meaningful feature description despite this textual noise. For this reason, we believe a key advantage of the Interpreter-Predictor paradigm is the ability to generate and validate a human-readable description of a feature, whereas lexical approaches may be more sensitive to dataset cleanliness issues and be more difficult to interpret.

D Feature Activation Analysis

We report our feature activation frequency analysis for all features containing the word "plant" in Figure 9. We note that all features either decrease in activation frequency between 1825-1850 and 1926-1950 or are relatively flat, with low overall activation frequency. Our primary concern in conducting this analysis is to ensure that highly interpretable features with similar descriptions do not exhibit opposing diachronic trends. We also note here that the strongest trends are associated with more frequent features.

Description	1	2	3
Bois indigènes pour menuiserie et construction	bois	chêne	meilleur
Noms locaux pour éléments culturels/naturels	appellent	nomment	nom
Justice indigène dans contexte colonial administratif	tribunaux	jugements	décret
Produits textiles locaux, contexte économique industriel	coton	tissus	laines
Militaires indigènes, législation pensions et retraites	militaires	pensions	veuves

Table 2: A sample of feature descriptions generated by the Interpreter with the top three most correlated words for that feature. Generated feature descriptions align with the words identified using TF-IDF and Spearman correlation. For example, *"Bois indigènes pour menuiserie et construction"* ("Indigenous woods for carpentry and construction") has associated keywords *"bois"* (wood), *"chêne"* (oak), and *"meilleur"* (best/better).

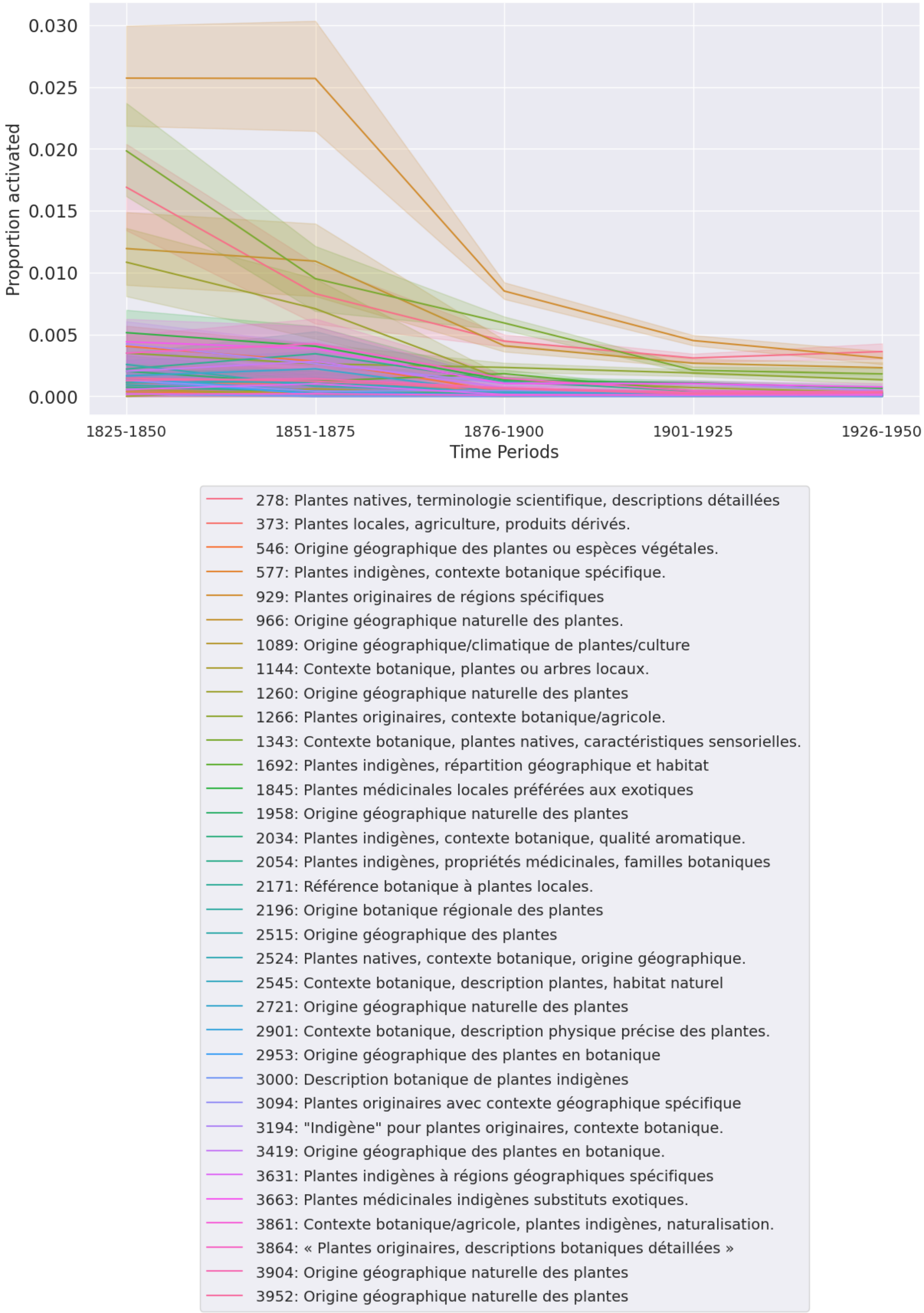


Figure 9: Feature activation frequencies for features with descriptions containing the word “plant”.