Embedding Style Beyond Topics: Analyzing Dispersion Effects Across Different Language Models

Benjamin Icard, Evangelia Zve, Lila Sainero, Alice Breton, Jean-Gabriel Ganascia

Embedding Style Beyond Topics: Analyzing Dispersion Effects Across Different Language Models

Benjamin Icard[1], **Evangelia Zve**[1,2], **Lila Sainero**[1], **Alice Breton**[1], **and Jean-Gabriel Ganascia**[1]

[1] LIP6, Sorbonne University, CNRS, France
[2] Infopro Digital, France

Abstract

This paper analyzes how writing style affects the dispersion of embedding vectors across multiple, state-of-the-art language models. While early transformer models primarily aligned with topic modeling, this study examines the role of writing style in shaping embedding spaces. Using a literary corpus that alternates between topics and styles, we compare the sensitivity of language models across French and English. By analyzing the particular impact of style on embedding dispersion, we aim to better understand how language models process stylistic information, contributing to their overall interpretability.

1 Introduction

In recent years, large language models (LLMs) have shown advanced natural language processing capabilities across diverse tasks, making their explainability an important area of research (Zhao et al., 2024). A key aspect of these models is their ability to generate meaningful text representations through *vector embeddings*, that encode semantic information. Although topic modeling along embedding representations has been widely studied (Peinelt et al., 2020), the influence of writing style on these representations has received less attention (Terreau et al., 2021; Chen et al., 2023). By leveraging sophisticated neural architectures (Achiam et al., 2023; Jiang et al., 2023), current large-scale models, such as those developed by OpenAI and Mistral, provide new investigative paths in that respect.

This paper aims to provide deeper insights into how different language models encode writing style and study their sensitivity to stylistic features. Specifically, we seek to examine the relative impact of style versus topic on the spatial dispersion of embedding vectors, across different language models in both French and English. Our primary focus is on the particular influence of writing style, with an emphasis on explainability.

To conduct this analysis, we designed an experimental study that systematically interchanged topic and style dimensions using text generation techniques. We selected two established literary works as raw material: Raymond Queneau's *Exercices de Style* (Queneau, 1947) and Félix Fénéon's *Nouvelles en trois lignes* (Fénéon and Halperin, 1970). *Exercices de Style* is a highly original piece of experimental literature in which Queneau writes numerous stylistic variations of a single narrative (a brief confrontation between bus passengers) while maintaining the same topic across all versions. By contrast, *Nouvelles en trois lignes* covers a wide range of topics (e.g., political events, crime, nature) while keeping a consistent style marked by a vivid and ironic tone. To enrich this material, we employed text generation techniques. We created a corpus where Queneau's style aligns with Fénéon's unique style, and another corpus where Fénéon's style varies in line with Queneau's plurality of styles. This design aimed to effectively assess the impact of topic and style on embedding dispersion.

Section 2 reviews existing work on the computational analysis of writing style, contrasting it with topic modeling, with an emphasis on vector embedding techniques. Section 3 describes the QUENEAU-FENEON dataset, a collection of textual documents compiled for topic and style experimentation, employing a specific text generation methodology. Section 4 describes the experimental tasks and results obtained on this dataset, including clustering to assess alignment with predefined classes, analysis of how style and topic influence embedding dispersion, and the identification of key linguistic features that may explain this dispersion. Finally, Section 5 concludes our investigation and outlines directions for future work.

2 Related Work

Embedding vector representations have been primarily studied in the context of topic modeling, building on BERT studies (Devlin et al., 2018). Techniques using word embeddings or sentence vectors, such as SBERT (Reimers and Gurevych, 2019), were developed to extract topics from textual documents, outperforming traditional statistical topic modeling methods like Latent Dirichlet Allocation (LDA) (Blei et al., 2003). A notable advancement in that respect is BERTopic, which refines these different methods (Grootendorst, 2022). Recent progress has focused on combining traditional methods like LDA with word embeddings, resulting in improved topic quality metrics and interpretability (Dieng et al., 2020).

Currently, research on embedding vector representations of writing style remains relatively underexplored compared to topic modeling (Dai et al., 2019). Existing computational and statistical approaches to writing style (Herrmann et al., 2021), including stylometry, focus on features such as word frequency, part-of-speech tags, N-grams (Ríos-Toledo et al., 2022), specific lexical entries and punctuation (Faye et al., 2024; Icard et al., 2024), TF-IDF (Bui et al., 2011), and vector embeddings (Chen et al., 2023). From a literary perspective, these approaches consider writing style as a manifestation of an author's unique voice and aesthetic choices (e.g. Verma and Srinivasan, 2019; Mani, 2022).

In continuation of computational stylometry, and enabled by transformer architecture (Hao et al., 2021), recent studies have examined stylistic features using embedding techniques (Liu et al., 2024), with particular focus on literary texts (Maharjan et al., 2019), but also in other domains, like news media and Generative AI (Bevendorff et al., 2024). However, a comprehensive approach that fully captures the entire spectrum of writing style remains underdeveloped. Terreau et al. (2021) proposed a novel evaluation framework for author verification embedding methods based on writing style, quantifying whether the embedding space effectively captures a set of stylistic features as the best proxy of an author's writing style. In addition to enhancing explainability, their work revealed that recent models are mostly driven by the inner semantics of authors' production and are outperformed by simple baselines on several linguistic axes. Chen et al. (2023) proposed a writing style embedding method based on contrastive learning for multi-author writing style analysis, achieving promising results in detecting style changes in multi-author documents. Addressing the challenge of content-independent style representations, Wegmann et al. (2022) introduced a variation of the authorship verification training task that controls for content using conversation or domain labels, finding that representations trained by controlling for conversation are better at representing style independent from content.

While a handful of studies have explored style embedding representations from a comprehensive perspective, most existing research has primarily focused on analyzing how specific models encode targeted stylistic features, such as syntactic and lexical embeddings. This paper proposes a structured methodology to compare how style versus topic variation influences the embedding dispersion of various language models in both French and English.

3 Dataset

To conduct this study, we compiled a corpus named QUENEAU-FENEON, consisting of 584 textual documents, with 292 texts in French and 292 texts in English. The corpus was developed in two stages: first, we created a *reference corpus* using extant literary works by Raymond Queneau and Félix Fénéon; second, we created a *generated corpus* by tranforming these original texts. The generated corpus was created using `GPT-4o`,[1] with a prompting methodology described below.

3.1 Reference corpus

We began by compiling a *reference corpus* of 146 texts in each language by uniting two symmetric classes with respect to topic and style variation. The first class, named QUENEAU_REF, contains 73 texts extracted from Raymond Queneau's *Exercices de style*. These texts are written in 73 different styles but all deal with the same topic of a bus journey. The second class, named FENEON_REF, also contains 73 texts, this time extracted from Félix Fénéon's *Nouvelles en trois lignes*. The specificity here is that these texts include numerous topics but all share the consistent style of Feneon.

We used the original French versions of the Queneau and Fénéon's corpus (Queneau, 1947; Fénéon and Halperin, 1970) to form the French QUENEAU_REF and FENEON_REF classes. For

[1] https://platform.openai.com/docs/models/gpt-4o

English, we used the extant English translations of both author's corpus (Queneau, 2013; Fénéon, 2007) to form the English QUENEAU_REF and FENEON_REF classes. Each of these classes contains exactly 73 texts.

Note that Raymond Queneau's *Exercices de style* originally contained 99 texts; however, we retained only 73 texts to ensure a balanced comparison with Félix Fénéon's *Nouvelles en trois lignes*, creating not only equally sized classes but also ensuring comparable text lengths between classes.

3.2 Generated corpus

We obtained a *generated corpus* by applying `GPT-4o` text generation on QUENEAU_REF and FENEON_REF, respectively, in both French and English. We began by generating a class named QUENEAU_GEN, prompting `GPT-4o` to rewrite all 73 stories of QUENEAU_REF in the uniform style of FENEON_REF. Then, we generated a complementary class named FENEON_GEN by prompting `GPT-4o` to rewrite each of the 73 stories of FENEON_REF into one of the 73 different writing styles of QUENEAU_REF. As in the reference class, the generated class contains exactly 146 texts equally divided into 73 texts for QUENEAU_GEN and 73 texts for FENEON_GEN. Table 1 gives a general overview of the QUENEAU-FENEON corpus, obtained with the French and English prompts given in Figure 1.

	Reference Corpus 146 texts per language	
FENEON_REF ↓	QUENEAU_REF *same topic, various styles* 73 texts per language	FENEON_REF *various topics, same style* 73 texts per language
	QUENEAU_GEN *same topic, same style* 73 texts per language	FENEON_GEN *various topics, various styles* 73 texts per language
	Generated Corpus 146 texts per language	

QUENEAU_REF ↓

Table 1: Overview of the QUENEAU-FENEON corpus involving text generation with `GPT-4o`.

4 Experiments

4.1 Corpus validation

We performed *k-means* clustering on the embeddings of the French and English QUENEAU-FENEON corpus. Twelve embedding models were selected in that respect, based on the following criteria: diversity, representativeness, dimensionality, multilingual capability, computational efficiency, explainability, and high performance according to the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2022) at time of the paper submission (September 16, 2024).[2] The full list of tested models is presented in Table 2 (see section of supplementary materials indicating dimensionality per model and URLs).

QUENEAU_GEN:

- French version: `"Ré écris ce texte : \n" + exercice.read() + "\n En copiant le style de Fénéon dans les 'nouvelles en trois lignes'"`
- English version: `"Re write this text in strictly less than 30 words and using only 1 to 3 sentences: \n "+exercice.read() +"\n Copying Feneon's style in 'novels in three lines'";`

FENEON_GEN:

- French version: `"Ré écris ce texte :\n" + nouvelle.read() +"\n En copiant le style de ce deuxième texte :" + exercice.read();`
- English version: `"Re write this text: \n" + nouvelle.read() +"\n Copying the style of this second text: "+exercice.read()`

Figure 1: French and English `GPT-4o` prompts used for generating QUENEAU_GEN and FENEON_GEN, based on QUENEAU_REF and FENEON_REF.

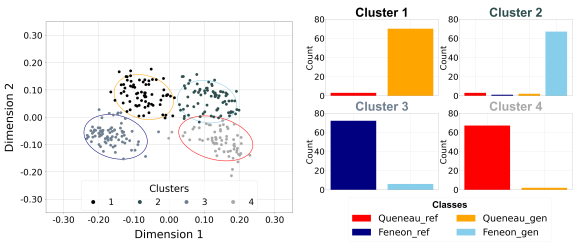
The clustering task aimed to assess the methodology used to build the QUENEAU-FENEON dataset. Here *k-means* measures the ability of the embedding models to effectively capture, and distinguish, the different topics and styles of our corpus. Notice that our goal was not to identify the optimal number of clusters for this dataset but, this number k being set to 4, to determine whether the texts in the four clusters align with the four classes of the QUENEAU-FENEON corpus.

We combined two popular evaluation metrics to assess the clustering task: *Purity* (Manning, 2008) and *NMI* (Normalized Mutual Information) (Danon et al., 2005). Ranging from 0 to 1, *Purity* and *NMI* are external cluster evaluation metrics based on the a priori knowledge of our dataset (Soni and Dwivedi, 2024). To facilitate model comparisons across languages and dimensions, we define a qualitative score $\bar{S}^{\mathrm{D}}(m)$ which, for a given model m, given a specific dimensionality D, averages the *Purity* and *NMI* scores of the model m for D:

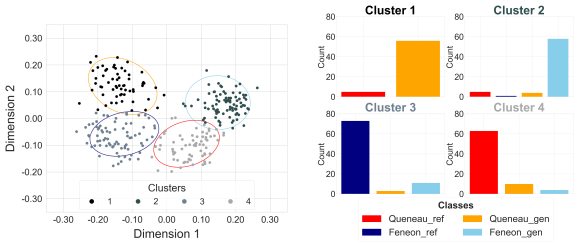
$$\bar{S}^{\mathrm{D}}(m) = \frac{Purity^{\mathrm{D}}(m) + NMI^{\mathrm{D}}(m)}{2} \quad (1)$$

We first applied clustering on the full-dimensional embeddings obtained with the twelve selected models on the French and English

[2] https://huggingface.co/spaces/mteb/leaderboard



(a) Projection for French QUENEAU-FENEON (left) with indication of majority class for each cluster (right).



(b) Projection for English QUENEAU-FENEON (left) with indication of majority class for each cluster (right).

Figure 2: 2D PCA projection of the 4 clusters obtained with `mistral-embed` on the QUENEAU-FENEON corpus and distribution of texts per cluster, for French (a) and for English (b).

QUENEAU-FENEON corpus. Subsequently, we employed Principal Component Analysis (PCA) (Wold, 1987) to determine the most effective dimensionality for an aligned comparison of performances across models and languages. PCA mitigates the so-called *"curse of dimensionality"* (Shlens, 2014; Jolliffe, 2016) that may arise with high-dimensional text embeddings, while preserving well the global structure of the vector space. Also, various studies have shown that *k-means* clustering performance is notably improved by PCA (e.g., Holland et al., 2020; Aliakbar et al., 2022).

To begin with, we calculated the $\bar{S}^{\mathrm{D}}(m)$ score of each model m for each of the PCA dimensionality D considered: 2D, 3D, 5D, and 10D. Then, we calculated the mean $\bar{S}^{\mathrm{D}}$ score of the 12 models for each of those dimensionalities to identify the best PCA combining reduction with information retention. Both languages considered, the ranking order obtained from best to worst is: 2D PCA, 3D PCA, 10D PCA, 5D PCA, followed by the FullD specific to each model. This ranking approach ensures an optimal balance between data compression and preservation of relevant information in the embedding vector space. Table 2 provides the mean and median $\bar{S}^{\mathrm{D}}$ scores for the top 3 best-reduced dimensionalities using PCA (2D, 3D, 10D) on French and English QUENEAU-FENEON, compared to FullD (now specific to each model).[3]

Mean $\bar{S}^{\mathrm{D}}$ scores obtained for 2D, 3D and 10D PCA are consistent, with all mean scores equal or higher than .6. For each dimension, the closeness of medians and means indicate that $\bar{S}^{\mathrm{D}}(m)$ scores per model m are balanced, with low or no skewness. Validation results turned out to be slightly better for French but consistent across languages and dimensions. For both French and English, the best dimension overall was 2D PCA with `mistral-embed` and `solon-embeddings-large-0.1` obtaining the best performances, then followed by `multilingual-e5-large` and closely by `voyage-2` with one rank difference. Figure 2 presents the 2D PCA projections of the clustering obtained with `mistral-embed` and the correspondence with the four known classes of the QUENEAU-FENEON corpus.

Projections in 2D PCA of the embeddings obtained with `mistral-embed` reveal four dense clusters in French and English, with clear separation of the four groups delineated. Distributions of texts per cluster also indicated in Figure 2 shows the existence of a majority class

FRENCH				
	2D PCA	**3D PCA**	**10D PCA**	**FullD**
`mistral-embed`	0.8522	0.8579	0.6309	0.6713
`solon-...-large-0.1`	0.8454	0.8630	0.7067	0.6312
`multilingual-e5-large`	0.8285	0.7935	0.6204	0.6202
`e5-base-v2`	0.7406	0.7326	0.7179	0.6510
`voyage-2`	0.6996	0.6912	0.5766	0.5919
`xlm-roberta-large`	0.6666	0.6485	0.7704	0.6500
`sentence-camembert-base`	0.6194	0.5476	0.6639	0.5147
`all-roberta-large-v1`	0.5960	0.5189	0.6074	0.6363
`distilbert-base-uncased`	0.5632	0.5622	0.5668	0.5679
`text-embedding-3-small`	0.5335	0.5578	0.5005	0.5087
`-multi...-mpnet-base-v2`	0.5146	0.4575	0.4724	0.4691
`all-MiniLM-L12-v2`	0.3915	0.4222	0.5178	0.4117
Mean	**0.6623**	**0.6117**	**0.6748**	**0.6633**
Median	**0.6354**	**0.5744**	**0.6917**	**0.6597**
ENGLISH				
`solon-...-large-0.1`	0.7779	0.5654	0.5350	0.6625
`mistral-embed`	0.7491	0.5497	0.5682	0.6547
`multilingual-e5-large`	0.6887	0.5696	0.5915	0.5959
`voyage-2`	0.6636	0.7045	0.5772	0.7519
`text-embedding-3-small`	0.6447	0.6551	0.5111	0.4577
`all-roberta-large-v1`	0.6429	0.5272	0.5270	0.5203
`distilbert-base-uncased`	0.6291	0.6712	0.6780	0.5335
`all-MiniLM-L12-v2`	0.5976	0.4966	0.4726	0.5355
`e5-base-v2`	0.5464	0.5280	0.5215	0.5335
`-multi...-mpnet-base-v2`	0.5211	0.4716	0.4697	0.4348
`sentence-camembert-base`	0.5132	0.5661	0.5312	0.5375
`xlm-roberta-large`	0.4693	0.6918	0.7027	0.6094
Mean	**0.6260**	**0.5977**	**0.5995**	**0.6168**
Median	**0.6289**	**0.5646**	**0.5605**	**0.6119**

Table 2: Details of $\bar{S}^{\mathrm{D}}$ scores per model for the three PCA reduction methods receiving the best mean $\bar{S}^{\mathrm{D}}$ scores in both languages, compared to FullD. Models are ordered based on their 2D PCA scores.

[3] All the individual *Purity* and *NMI* scores of each model are detailed for each specific dimensionality in the GitHub of supplementary materials.

matching fairly well with exactly one of the four initial classes of the QUENEAU-FENEON corpus, in line with the $\bar{S}^{2D}$ scores reported in Table 2 concerning `mistral-embed` (for French: 0.8522, for English: 0.7491). It can be observed in Table 2 that the good correspondence obtained with `mistral-embed` also transfers to other models that show high $\bar{S}^{D}$ scores (greater than .65) in both languages, such as e.g. `solon-embeddings-large-0.1`, `multilingual-e5-large`, and `voyage-2`. Support by other models is more moderate to low, with inconsistency across languages for e.g. `e5-base-v2` and `xml-roberta-large`.

4.2 Dispersion within classes

To gain a deeper understanding of how topic and style impact embedding representations, we analyzed embeddings dispersion within the QUENEAU-FENEON corpus. Specifically, we aimed to determine whether variations in topic and in style lead to increase or decrease dispersion. Additionally, we aimed to evaluate the relative contribution of style versus topic to this effect.

To conduct this analysis, we employed the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) technique for dimensionality reduction of our embedding vector space. UMAP was chosen over other methods, like PCA previously used for clustering, due to its superior ability to preserve both local and global structures within the embedding space. Additionally, t-distributed Stochastic Neighbor Embedding (t-SNE) is effective at preserving local structures, but it often distorts the global structure (Anowar et al., 2021), making it less suitable for our specific analysis. Since we aimed to ensure that local information was primarily preserved while also maintaining an accurate global structure, UMAP was the most appropriate choice, as particularly well-suited for distance-based analysis of high-dimensional text embedding spaces (Cox et al., 2021).

To account for the non-deterministic nature of UMAP (McInnes et al., 2018), we performed 30 applications of the model (iterations) with different random seeds, for dimensionality reductions of 2D, 3D, 5D, and 10D. The adoption of different random seeds ensures that the results are stable and not sensitive to specific initial conditions, providing a more robust estimate of embedding dispersion.

For the j-th iteration, we define $d_X^{(i,j)}$ as the Euclidean distance of the i-th embedding vector from the centroid $c_X^{(j)}$ of class X as follows:

$$d_X^{(i,j)} = \|v_X^{(i,j)} - c_X^{(j)}\| \tag{2}$$

where $v_X^{(i,j)}$ is the i-th embedding vector of class X in the j-th iteration, $c_X^{(j)}$ is the centroid vector for class X in the j-th iteration, and $\|\cdot\|$ is the Euclidean norm.

To capture the spatial distribution of high-dimensional embeddings, we calculate the mean Euclidean distance from the centroid of each class across all iterations, written $\bar{d}_X(i)$:

$$\bar{d}_X(i) = \frac{1}{30}\sum_{j=1}^{30} d_X^{(i,j)} \tag{3}$$

Finally, the overall mean distance $\bar{d}_X$ for class X across all embeddings is:

$$\bar{d}_X = \frac{1}{N}\sum_{i=1}^{N} \bar{d}_X(i) \tag{4}$$

where $\bar{d}_X(i)$ is the averaged Euclidean distance of the i-th embedding vector for class X and N is the total number of embedding vectors in the class.

In order to analyze the influence of topic variation, we compared the difference in embedding dispersion between classes presenting topic homogeneity versus topic heterogeneity, i.e. by comparing QUENEAU_REF with FENEON_GEN, and also QUENEAU_GEN with FENEON_REF. To analyze the influence of style, we compared the difference in embedding dispersion between classes showing style homogeneity versus style heterogeneity, i.e. by comparing FENEON_REF with FENEON_GEN, and also QUENEAU_GEN with QUENEAU_REF.

Using the metric defined in (4), we predict that both topic and writing style influence embeddings dispersion, as in the *local hypotheses* (T) and (S):

$$\text{Topic} \left| \begin{array}{ll} \bar{d}_{\text{FENEON\_GEN}} > \bar{d}_{\text{QUENEAU\_REF}} & \text{(T')} \\ \bar{d}_{\text{FENEON\_REF}} > \bar{d}_{\text{QUENEAU\_GEN}} & \text{(T'')} \end{array} \right. \text{(T)}$$

$$\text{Style} \left| \begin{array}{ll} \bar{d}_{\text{QUENEAU\_REF}} > \bar{d}_{\text{QUENEAU\_GEN}} & \text{(S')} \\ \bar{d}_{\text{FENEON\_GEN}} > \bar{d}_{\text{FENEON\_REF}} & \text{(S'')} \end{array} \right. \text{(S)}$$

Besides hypotheses (T) and (S), we expect the topic to have a greater impact on embeddings dispersion than style, as defined in the *global hypothesis* (T-S):

$$\bar{d}_{\text{FENEON\_REF}} > \bar{d}_{\text{QUENEAU\_REF}} \tag{T-S}$$

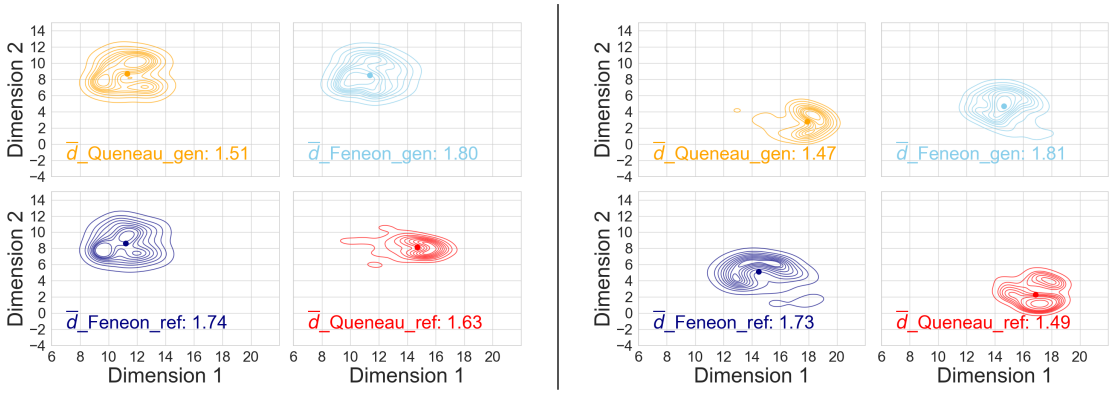


Figure 3: 2D UMAP contour plots of the embedding dispersion obtained on the QUENEAU-FENEON corpus with model `all-MiniLM-L12-v2`, for French (left) and for English (right). In each subplot, the overall spread of the embeddings around centroid (for the last seed) is represented by the external contour line, the isolines represent differences in densities of embedding vectors, the centroid is indicated by a dot, and $\bar{d}_X$ corresponds to the mean centroid distance for the targeted corpus X.

DISPERSION HYPOTHESES	**French 2D UMAP**					**English 2D UMAP**				
	Local				**Global**	**Local**				**Global**
	Topic (T)		**Style** (S)		(T-S)	**Topic** (T)		**Style** (S)		(T-S)
Are the predictions checked?	(T')	(T'')	(S')	(S'')		(T')	(T'')	(S')	(S'')	
`mistral-embed`	✓**	✓**	✗**	✓**	✓**	✓**	✓**	✗**	✓**	✓**
`solon-...-large-0.1`	✓**	✓**	✗	✓**	✓**	✓**	✓**	✗**	✓**	✓**
`multilingual-e5-large`	✓**	✓**	✓**	✓**	✓**	✓**	✓**	✗**	✓**	✓**
`e5-base-v2`	✓**	✓**	✓	✓**	✓**	✓**	✓**	✗	✓**	✓**
`voyage-2`	✓**	✓**	✗	✓**	✓**	✓**	✓**	✗**	✓**	✓**
`sentence-camembert-base`	✓**	✓**	✓**	✓**	✓**	✓**	✓**	✓**	✓**	✓
`all-MiniLM-L12-v2`	✓**	✓**	✓**	✓**	✓**	✓**	✓**	✓	✓**	✓**
`text-embedding-3-small`	✓**	✓**	✓**	✓**	✓**	✓**	✓**	✗*	✓**	✓**
`-multi...-mpnet-base-v2`	✓**	✓**	✗**	✓**	✓**	✓**	✓**	✓	✓**	✓**
`xlm-roberta-large`	✓**	✓**	✓**	✓**	✗**	✓**	✓**	✓**	✓**	✗**
`all-roberta-large-v1`	✗	✓**	✓**	✓**	✗**	✓**	✓**	✓**	✓**	✓**
`distilbert-base-uncased`	✓**	✓**	✓**	✓**	✗**	✓**	✗**	✗**	✓**	✓**
Mean	✓**	✓**	✓**	✓**	✓**	✓**	✓**	✗**	✓**	✓**

Table 3: Results of validation for hypotheses (T), (S) and (T-S) based on 2D UMAP projection for French and English. "✓" indicates that the prediction is verified, "✗" indicates that the opposite prediction is verified, with * and ** reporting p-value $< .05$ and $< .01$, respectively.

According to (T-S), the greater dispersion of FENEON_REF compared to QUENEAU_REF implies that topic variation influences more embedding dispersion than style variation, as the topic shifts to pluriform from QUENEAU_REF to FENEON_REF, whereas style, in contrast, moves toward uniformity.

Optimal hypotheses validation was obtained with 2D UMAP for both French and English, followed by weaker but still highly consistent results with 5D and 10D UMAP. To help visualize dispersion around centroids, Figure 3 shows the contour plots of the 2D UMAP projected embeddings obtained with `all-MiniLM-L12-v2` on the QUENEAU-FENEON corpus. Detailed 2D UMAP results per hypothesis (T), (S) and (T-S) of the twelve models are given in Table 3.

For both English and French, mean centroid distances reported in Figure 3 result in the following order: $\bar{d}_{\text{FENEON\_GEN}} > \bar{d}_{\text{FENEON\_REF}} > \bar{d}_{\text{QUENEAU\_REF}} > \bar{d}_{\text{QUENEAU\_GEN}}$. All pairwise comparisons of this order using two-sided t-tests proved to be significant at the .01 level. Locally, this validates (T')-(T''), supporting hypothesis (T) on the increasing effect of topic variation on embedding dispersion. Moreover, pairwise comparisons also significantly validate (S')-(S''), now verifying prediction (S) on the positive effect of writing style on vector dispersion. Globally, validation of (T-S) reveals a stronger effect of topic on this dispersion compared to style.

Looking at results individually given in Table 3, we observe that a vast majority of models validates hypotheses (T), (S) and (T-S) in both languages. Concerning results for French

QUENEAU-FENEON, the dispersion effect predicted by (T')-(T") for topic is consistently confirmed across models, with strong significance except in one case (`all-roberta-large-v1`). The dispersion effect of style predicted by (S')-(S") was also significantly verified, except in 4 cases for condition (S'): `mistral-embed`, `solon-embeddings-large-0.1`, `-multilingual-mpnet-base-v2`, and `voyage-2`. In support of the global hypothesis (T-S), varying the topic resulted in greater dispersion than varying the style for 9 (out of 12) models, with the 3 models showing the opposite direction.

The results obtained for English are largely similar to the results for French. Except in one single case (`distilbert-base-uncased`), the topic hypotheses (T) are significantly verified across all tested models. Concerning the style hypotheses (S), results were more mixed for (S') with 6 models significantly invalidating the prediction but still strongly verified for (S") across all models.

In conclusion, the models applied on the French corpus outperformed their English counterparts in 2D UMAP dimension. The models that consistently satisfied both the topic and style hypotheses (T) and (S) in both languages were `sentence-camembert-base`, `all-MiniLM-L12-v2`, and `xlm-roberta-large`. That said, the global hypothesis (T-S) is verified more on the English corpus than on the French corpus, with only one model rejecting the hypothesis in English (`xlm-roberta-large`).

4.3 Style embedding interpretability

In the previous section, we observed that style variation, in addition to the more common topic variation, also influences embedding dispersion. Here we attempt to identify the key stylistic features related to style hypothesis (S), that may drive embedding vectors to exhibit greater or lesser dispersion.

To conduct this analysis, we used a framework developed by Terreau et al. (2021) to evaluate how embedding vectors represent writing style.[4] This framework generates stylometric reports to assess how well embedding models recognize writing styles in alignment with stylistic features identified by well-known Python modules (e.g. spaCy, NLTK, Counter). Terreau et al. (2021) select eight groups of stylistic features as predictive targets for French and English regression models. These features include: the relative frequency of *function words* (e.g., prepositions, conjunctions, auxiliary verbs) compared to the total word count in the text, the average values of *structural features* (e.g., word length, word frequency, syllables per word), *indexes* of lexical complexity (e.g., Yule's K constancy measure (Yule, 2014), Shannon Entropy (Shannon, 1948)) and text readability metrics (e.g., Flesch-Kincaid Grade Level (Kincaid, 1975)), the relative frequency of *punctuation marks* (e.g., periods, commas) compared to total text length, *numbers* (i.e., numerical digits), the average frequency of *named entities* (i.e., NER: persons, locations, organizations) per sentence, and *part-of-speech tags* (i.e., TAG: nouns, verbs, adjectives).

Our main focus in this section is to analyze how embedding dispersion responds to variations in these stylistic features when style varies across classes, while the topic remains constant. In more detail, we focus on examining the interaction between differences in the frequencies of the eight stylistic features and the difference in dispersion between QUENEAU_GEN and QUENEAU_REF. This comparison is the only case where styles are expected to differ significantly between classes, while the topic remains exactly the same. As a control, we also compared QUENEAU_GEN and FENEON_REF, where the topics are expected to differ significantly, while the style remains the same.

We first measured mean ground frequencies of the eight stylistic features, written $\bar{f}^s$ with s a stylistic feature, by applying Terreau et al. (2021)'s extraction module on the three classes of interest: QUENEAU_GEN, QUENEAU_REF and FENEON_REF. For each of the two targeted comparisons, a pairwise t-test was conducted for each feature variation to assess statistical significance. Table 4 presents the changes in feature frequencies from QUENEAU_GEN to QUENEAU_REF, and from QUENEAU_GEN to FENEON_REF.

For both French and English, we observed that structural features were significantly more frequent in QUENEAU_REF than in QUENEAU_GEN, and that indexes were significantly less frequent in QUENEAU_REF compared to QUENEAU_GEN. Opposite tendencies were observed for function words and punctuation between French (where both significantly increase) and English (where both significantly decrease). Additionally, some variations observed as non-significant in French (e.g. for letters, NER, TAG) were found to be significant in English. Moving to the comparison between

[4] https://github.com/EnzoFleur/style_embedding_evaluation/

From QUENEAU_GEN	$\bar{f}^s$	to QUENEAU_REF		to FENEON_REF	
		French	**English**	**French**	**English**
	Function words	↗**	↘**	↘**	↘*
	Indexes	↘**	↘**	↗	↗
	Letters	↗	↘**	↗	↘
	NER	↗	↗*	↗**	↗**
	Numbers	↗	↗	↗**	↗*
	Punctuation	↗**	↘**	↗	↘
	Structural	↗**	↗**	↗**	↗**
	TAG	↗	↗*	↗**	↗**

Table 4: Evolution of mean ground frequencies ($\bar{f}^s$) per text for the 8 stylistic features from QUENEAU_GEN to QUENEAU_REF and FENEON_REF, with * and ** reporting p-value $< .05$ and $< .01$, respectively.

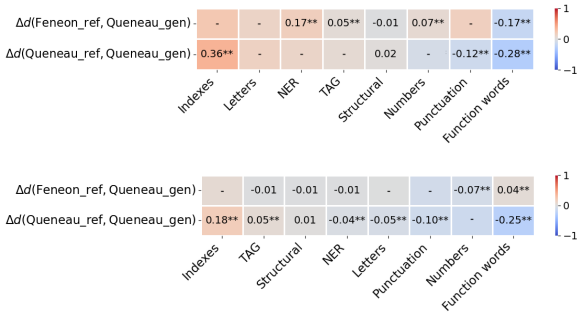


Figure 4: Correlation matrices between differences in dispersion (Δd) and differences in frequencies of the eight stylistic features (Δf^s) for the two comparisons of interest, for French (top) and English (bottom). Here "-" corresponds to correlations that were intentionally omitted, as they correspond to differences in features previously observed as non significant (see Table 4).

QUENEAU_GEN and FENEON_REF, we observed that function words were significantly less frequent in FENEON_REF compared to QUENEAU_GEN, but that NER, numbers, structural and TAG became significantly more frequent in FENEON_REF. The differences observed for indexes, letters and punctuation were not significant.

From now on, we restrict our attention only to differences *observed as significant* for ground frequencies (see Table 4, cases marked with * or ** only). To see how these differences interact with differences in dispersion, we use the non-averaged Δ values per class, denoted as Δd and Δf^s respectively, and defined as follows:

$$\Delta d(X, Y) = d_X(i) - d_Y(j) \quad (8)$$

$$\Delta f^s(X, Y) = f^s_X(i) - f^s_Y(j) \quad (9)$$

where Y = QUENEAU_GEN, X is either QUENEAU_REF or FENEON_REF depending on the targeted comparison, i is the i-th vector of class X, and j is the j-th vector of class Y.

Figure 4 reports the Pearson correlation coefficients measuring the interaction between the difference Δd (i.e., the difference in embedding dispersion per text) and the difference Δf^s (i.e., the difference in the frequency of stylistic features per text, for each stylistic feature s).

For QUENEAU_GEN and QUENEAU_REF intended to differ in style, we observed moderate positive correlation in French between dispersion and differences in frequencies for indexes ($r = 0.36^{**}$), and weak negative correlation with function words ($r = -0.28^{**}$) and punctuation ($r = -0.12^{**}$). No significant correlation was observed for structural ($r = 0.02$). In English, for the same comparison, we observed only weak positive correlation with indexes ($r = 0.18^{**}$), and weak negative correlation with function words ($r = -0.25^{**}$) and punctuation ($r = -0.10^{**}$). No correlation was observed in English with other features (e.g. TAG, NER, letters and structural).

Concerning QUENEAU_GEN and FENEON_REF now intended *not* to differ in style, we observed only weak positive correlation in French with NER ($r = 0.17^{**}$), and weak negative correlation with function words ($r = -0.17^{**}$). No correlation was found with numbers ($r = 0.07^{**}$), TAG ($r = 0.05^{**}$) and structural ($r = -0.01$). In English, no correlation was observed at all.

To summarize, when corpora were expected *not* to differ in writing style, such as QUENEAU_GEN and FENEON_REF, we observed weak or no correlation with dispersion for French, and no correlation at all for English. These findings align with expectations. By contrast, when corpora were expected to differ in style, such as QUENEAU_GEN and QUENEAU_REF, significant differences were observed for indexes, punctuation and function words. Also aligned with expectations, this result may account for the differences observed in dispersion between QUENEAU_GEN and QUENEAU_REF, indicating sensitivity of embedding vectors to these specific features in both French and English.

While both languages exhibited sensitivity in the comparison between QUENEAU_GEN and QUENEAU_REF, the correlations were stronger for French than for English. Similarly, when comparing QUENEAU_GEN and FENEON_REF, the same variations in frequency were observed in both lan-

guages (i.e., function words, NER, numbers, structural and TAG), but language models were insensitive to these features in English, contrary to French. Finally, for variations significant in English only (i.e., letters, NER and TAG), no correlation with dispersion was observed. That said, Table 4 offers a possible explanation for these discrepancies. Rather than reflecting an inherent limitation of language models with English, the directions of variations suggest that the translation from French to English has notably reduced the presence of function words, letters, and punctuation, while only slightly increasing the significance of features like NER and TAG. Accordingly, a plausible explanation for the reduced sensitivity in English is that translation may have diminished the first set of features to a degree that embedding vectors struggle to capture, without sufficiently amplifying the second set of features to balance this effect.

5 Conclusion and perspectives

This paper provides evidence that writing style influences embedding dispersion, though topic variation has a stronger effect. This result is supported across different language models in both French and English. Attempt at interpretability suggests that specific linguistic features, particularly readability and complexity indexes, function words and punctuation (to a lesser extent), partially explain embedding representations.

In the short run, two steps of investigation emerge. Firstly, some models we tested (e.g., `sentence-camembert-base`, `all-Mini LM-L12-v2`) showed greater responsiveness to stylistic variations leading to increased dispersion. Other models (e.g., `voyage-2`, `solon-embeddings-large-0.1`) were less or not affected. This observation, including cases where dispersion decreased contrary to expectations (e.g., `mistral-embed`), suggests that different architectures process stylistic features in unique ways. Secondly, models generally responded more to stylistic variation in French compared to English. This was also reflected in the weaker stylistic correlations observed for English, suggesting that factors such as translation or language-specific characteristics could play a role. Replicating the study on a larger French-English corpus would provide further insight into these differences and the detailed stylistic features controlling them.

In the long run, we aim for generalizability across other models and genres, also hoping to inspire further stylistic studies on languages that are more typologically diverse than French and English. Regarding models, we aim to compare the sensitivity of the architectures considered, with a focus on open-weight models to enhance explainability. Concerning genres, we aim to apply our methodology to news articles, as a genre responding to stylistic conventions other than literary conventions, associated to a great variety of topics and a potential for high scalability. We leave these investigations for future work.

Limitations

Our corpus methodology is validated by clustering. However, the corpus is limited to 292 textual documents in each language (a total of 584 documents), with 73 documents per class. A larger dataset should be used for replication in order to mitigate biases due to sample size and verify our hypotheses further.

While UMAP dimensionality reduction to 2D produced significant results in French, comparable results were also observed with higher UMAP dimensions. We omitted these in the paper for brevity but they are available on our GitHub repository (see supplementary materials below).

Additionally, our study is focused on eight main stylistic features, but these features are driven by subfeatures that should be considered to gain a clearer understanding of their impact on embedding dispersion. Other areas like journalism and scientific writing, which follow different stylistic conventions, are not explored either. Testing our hypotheses on other types of textual documents and assessing relevance of domain-specific features will be essential for assessing the generalizability of our results.

Lastly, open-weight LLMs that we tested, like DistilBERT and RoBERTa, offer potential for deeper explainability and customization, in contrast to proprietary models, like OpenAI embeddings, that lack transparency due to their closed-weights. Large open-weight models like LLaMA-2 and Mistral-7B do exist, but their use requires significant computational resources.

Ethical considerations

Our research adheres to the following ethical principles: open science, transparency, inclusiveness, and

sustainability. As academic researchers, we adhere to open science guidelines, with a concern for the reproducibility of experiments and the accessibility of our results. The dataset we provide contains no textual documents from the original sources, only vector embeddings derived from those documents, aligning with the concept of "transformative fair use". This approach ensures compliance with intellectual property and data protection regulations. Transparency is upheld through raw data availability, code and prompts sharing on a dedicated GitHub repository, and a comprehensive documentation to ensure reproducibility by others. We are also guided by inclusiveness, as we expect our research project to contribute to advancing educational and cultural AI literacy on the interplay between writing style and embedding representations. To promote sustainability, the GitHub of the study actually supports net-zero carbon initiatives by others based on our framework. We also prioritize using smaller, open-source pre-trained language models alongside larger ones, to reach a balance between carbon footprint and resource consumption.

Supplementary materials

Large language models used in the experiments included the 1536-dimensional model text-embedding-3-small[5] by OpenAI, and the 1024-dimensional models mistral-embed[6] by Mistral, voyage-2[7] by Voyage, and the RoBERTa-based models xlm-roberta-large,[8] all-roberta-large-v1,[9] and multilingual-e5-large.[10] Smaller models included the 768-dimensional models e5-base-v2,[11] distilbert-base-uncased,[12] all-MiniLM-L12-v2,[13] the SBERT model sentence-camembert-base[14] and the multilingual model paraphrase-multilingual-mpnet-base-v2.[15] We also included solon-embeddings-large-0.1[16] (1024D) by Solon as one of the best performing French embedding models (see MTEB leaderboard on HuggingFace,[17] at time of the submission: September 16, 2024).

All the code (including prompts for generating the extended corpus in French and English), data, and analytical results are available at the following URL link: https://github.com/evangeliazve/topic_style_embeddings_dispersion/tree/main

Acknowledgements

We thank three anonymous reviewers for helpful comments and feedback. EZ acknowledges Infopro Digital for supporting her PhD research, alongside her work. BI acknowledges the program THEMIS (grant agreements n°DOS022279400 and n°DOS022279500) for funding.

Declaration of contribution

BI and JGG conceptualized the research problem and designed the experiment with EZ. LS and AB managed the data collection and generation processes. EZ was responsible of coding, optimizing, integrating existing frameworks and testing selected models. EZ and BI analyzed and discussed the results. BI and EZ wrote the paper, which all authors read and revised together. BI and EZ share first authorship. Correspondence: benjamin.icard@lip6.fr, evangelia.zve@lip6.fr, jean-gabriel.ganascia@lip6.fr.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

A. Aliakbar et al. 2022. Enhancing k-means clustering performance using pca. *Journal of Data Science*.

Farzana Anowar, Samira Sadaoui, and Bassant Selim. 2021. Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Computer Science Review*, 40:100378.

Janek Bevendorff, Xavier Bonet Casals, Berta Chulvi, Daryna Dementieva, Ashaf Elnagar, Dayne Freitag, Maik Fröbe, Damir Korenčić, Maximilian Mayerl,

[5] https://platform.openai.com/docs/guides/embeddings

[6] https://docs.mistral.ai/capabilities/embeddings/

[7] https://docs.voyageai.com/docs/embeddings

[8] https://huggingface.co/FacebookAI/xlm-roberta-large

[9] https://huggingface.co/sentence-transformers/all-roberta-large-v1

[10] https://huggingface.co/intfloat/multilingual-e5-large

[11] https://huggingface.co/intfloat/e5-base-v2

[12] https://huggingface.co/distilbert/distilbert-base-uncased

[13] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

[14] https://huggingface.co/dangvantuan/sentence-camembert-base

[15] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[16] https://huggingface.co/OrdalieTech/solon-embeddings-large-0.1

[17] https://huggingface.co/spaces/mteb/leaderboard

Animesh Mukherjee, et al. 2024. Overview of pan 2024: multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative ai authorship verification. In *European Conference on Information Retrieval*, pages 3–10. Springer.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Quang Anh Bui, Muriel Visani, Sophea Prum, and Jean-Marc Ogier. 2011. Writer identification using tf-idf for cursive handwritten word recognition. In *2011 International Conference on Document Analysis and Recognition*, pages 844–848. IEEE.

Haoyang Chen, Zhongyuan Han, Zengyao Li, and Yong Han. 2023. A writing style embedding based on contrastive learning for multi-author writing style analysis. In *CLEF 2023 Working Notes*. CEUR Workshop Proceedings.

Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian Von Der Weth, and Brian Y. Lim. 2021. Directed diversity: Leveraging language embedding distances for collective creativity in crowd ideation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–35.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. *arXiv preprint arXiv:1905.05621*.

Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. 2005. Comparing community structure identification. *Journal of statistical mechanics: Theory and experiment*, 2005(09):P09008.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Géraud Faye, Benjamin Icard, Morgane Casanova, Julien Chanson, François Maine, François Bancilhon, Guillaume Gadek, Guillaume Gravier, and Paul Égré. 2024. Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 62–72, Malta.

Félix Fénéon. 2007. *Novels in Three Lines*. New York Review of Books. Translated by Luc Sante.

Félix Fénéon and Joan U Halperin. 1970. *Œuvres plus que complètes*, volume 2. Librairie Droz.

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

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971.

J Berenike Herrmann, Arthur M Jacobs, and Andrew Piper. 2021. Computational stylistics. *Handbook of Empirical Literary Studies*, pages 451–486.

P. Holland et al. 2020. Improving clustering interpretability with pca. *Machine Learning Journal*.

Benjamin Icard, François Maine, Morgane Casanova, Géraud Faye, Julien Chanson, Guillaume Gadek, Ghislain Atemezing, François Bancilhon, and Paul Égré. 2024. A multi-label dataset of french fake news: Human and machine insights. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 812–818.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

I. T. Jolliffe. 2016. *Principal Component Analysis*. Springer.

JP Kincaid. 1975. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for navy enlisted personnel.

Chang Liu, Zhongyuan Han, Haoyang Chen, and Qingbiao Hu. 2024. Team liuc0757 at pan: A writing style embedding method based on contrastive learning for multi-author writing style analysis. *Working Notes of CLEF*.

Suraj Maharjan, Deepthi Mave, Prasha Shrestha, Manuel Montes, Fabio A. González, and Thamar Solorio. 2019. Jointly learning author and annotated character n-gram embeddings: A case study in literary text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*.

Inderjeet Mani. 2022. *Computational modeling of narrative*. Springer Nature.

Christopher D Manning. 2008. Introduction to information retrieval.

L. McInnes, J. Healy, and J. Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tbert: Topic models and bert joining forces for semantic similarity detection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7047–7055.

Raymond Queneau. 1947. Exercices de style: Edition gallimard. *Collection Folio*.

Raymond Queneau. 2013. *Exercises in Style*. New Directions Publishing. Translated by Barbara Wright.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Germán Ríos-Toledo, Juan Pablo Francisco Posadas-Durán, Grigori Sidorov, and Noé Alejandro Castro-Sánchez. 2022. Detection of changes in literary writing style using n-grams as style markers and supervised machine learning. *Plos one*, 17(7):e0267590.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

J. Shlens. 2014. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Urvashi Soni and Sunita Dwivedi. 2024. Clutching of clustering validation criteria. *International Journal of Future Computer and Communication*, 13(1).

Enzo Terreau, Antoine Gourru, and Julien Velcin. 2021. Writing style author embedding evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 84–93. Association for Computational Linguistics.

Gaurav Verma and Balaji Vasan Srinivasan. 2019. A lexical, syntactic, and semantic perspective for understanding style in text. *arXiv preprint arXiv:1909.08349*.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268. Association for Computational Linguistics.

H. Wold. 1987. Principal component analysis. *Technometrics*, 38(3):235–238.

C Udny Yule. 2014. *The statistical study of literary vocabulary*. Cambridge University Press.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.