Decoding Self-Attention : How frequency matrices drive representation learning ?

Laurent Vanni[1], Dominique Longrée[2], Damon Mayaffre[3]

[1]UMR 7320: BCL - CNRS - UnivCA - France – Laurent.Vanni@univ-cotedazur.fr
[2]L.A.S.L.A. - UR Mondes Anciens - ULiège - Belgique - Dominique.Longree@uliege.be
[3]UMR 7320: BCL - CNRS - UnivCA - France – Damon.Mayaffre@univ-cotedazur.fr

Abstract

The interpretability of deep learning models remains a pressing challenge in contemporary computational linguistics. While the latest advancements in automatic language processing tools have achieved remarkable performance, they also raise significant questions for researchers accustomed to the transparency and predictability of classical statistical methods. In recent years, a variety of tools and methodologies have emerged to enhance the interpretability of AI-driven predictions. Both architecture-dependent and agnostic approaches have uncovered key insights into neural network behavior — such as word embeddings and long-range dependencies — however, these findings still resist the formalization of new knowledge about texts. To assess the unique contributions of AI to text analysis, this study investigates the relationship between traditional statistical markers and AI-derived indicators. Based on prior research and ongoing debates surrounding the interpretation of attention mechanisms in Transformer-based models, we propose a new approach of treating neural weights as statistical indices emerging from the training corpus. Using the texts submitted for model training, we construct a triangular attention matrix, which we then subject to principal component analysis (PCA) for further examination. Our findings reveal that the learned representations align closely with those predicted by traditional text analysis methods from the 1990s. This alignment is quantitatively supported by Pearson correlation coefficients, which indicate that attention scores predominantly capture frequency-based co-occurrence patterns. These results carry two key implications: i) Attention scores could be initialized or even replaced with word frequency data, thereby reducing the computational overhead of training Transformer models. ii) Trained models refine these representations further, offering exploratory insights that surpass traditional co-occurrence analysis methods. We demonstrate our approach through a small-scale GPT model trained on corpora spanning three languages: English, French, and Latin.

Keywords: Deep Learning, Transformers, Interpretability, Attention scores, Cooccurrences, Corpus, PCA

1. Introduction

Analyzing hidden layers in deep neural networks is one of the major challenges in computational linguistics. Researchers have been intrigued by the abstract representations produced by neural architectures, particularly since the advent of Transformers-based models and the seminal article "Attention Is All You Need" Vaswani et al. (2017). The unmatched performance of the *Self-attention* underlying mecanism in natural language processing (NLP) is based on calculating the weight of local and distant word associations in training corpus texts. These weights are relative to the trained model's task (e.g., classification, translation, or generation) and can be analyzed as attention scores, which are represented by a square matrix. Each cell in this matrix corresponds to the attention that word A pays to word B. Although attention scores reflect the model's behavior, their interpretability remains a subject of debate, as discussed in Jain and Wallace (2019); Serrano and Smith (2019). Traditional visualization methods, such as BertViz Vig (2019), suggest a multitude of links between words in texts that are difficult

to prioritize or interpret, either linguistically or statistically. Furthermore, the relationship between local phenomena in the input texts and global phenomena in the training corpus is unclear Meister et al. (2021); Škrlj et al. (2021). However, we postulate that understanding the linguistic phenomena captured by self-attention would be an important step forward for developing new computer algorithms and exploring the deep interweaving words in texts. The field of interpretability of deep learning models is rich in methods and tools. There are two types of approaches: global and local. The global approach aims to determine the model's general behavior. Methods based on word embeddings fall under this category, including static embeddings, such as Word2Vec Mikolov et al. (2013), and context-dependent embeddings, such as Transformers Dar et al. (2023). The local approach questions the model's decision-making based on the given input text. Local methods can be agnostic, like LIME Ribeiro et al. (2016), or model-dependent, like BertViz. Some tools focus on large-scale information visualization, such as Molino et al. (2019) based on Principal Component Analysis (PCA). However, no method or study has established a direct link between statistical measures and neuron weights, and much of the information contained in the model's features remains inaccessible to humans. In this paper, we propose a method to overcome the limitations on the interpretability of Transformers by comparing attention scores with traditional textual statistics from Lebart et al. (1998). This study aims to define the proportion of information overlap between a square matrix of classic frequency co-occurrences and a matrix of attention scores obtained from a Transformer trained on an automatic text generation task. Our objective is to evaluate the added value of attention scores for exploratory text analysis, considering a descriptive use of Transformers rather than just a predictive one, as suggested by their applications in NLP. After presenting the corpora used in this study briefly, we will present the model and method used to align attention scores with standard frequency matrices. Next, we will present the results in the form of factorial correspondence analyses, followed by a table showing the correlation between frequencies and attentions. We will conclude then with a discussion presenting the limitations and prospects of our studies

2. Compare traditional statistical approaches and Transformers

The study we propose is based on a cross-analysis of results obtained from both classical statistical methods and deep learning applied to the same dataset. The scope of our study covers three corpora representing three languages, different discourse genres, and different time periods.

2.1. Learning corpora

The first corpus consists of French presidential speeches that are available as *elysee* Base on the Hyperbase web platform: `https://hyperbase.unice.fr/elysee`. We chose to focus on President E. Macron, covering the period from 2017 to 2025. This corpus contains 77 complete speeches, which were either collected from the official *Élysée* website or transcribed from videos of official speeches and television interviews. The second corpus is a collection of classical Latin texts. Extracted from the LASLA Dataverse[1] and available as L.A.S.L.A. *cicero* Base operable with Hyperbase Web: `https://hyperbase.unice.fr/cicero`. It contains al the Cicero's speeches and a collection of his treatises. All texts have been lemmatised and morphosyntactically annotated according to the L.A.S.L.A. methods. The selection of this corpus stems from the desire to combine different genrees, eras, and languages

[1] `https://dataverse.uliege.be/dataverse/lasla`

to test the behavior of a model trained in automatic text generation. Finally, the last English-language corpus represents a sample of classical literary genre. It is a corpus of W. Shakespeare's plays, comprising 37 plays[2] also available as *shakespeare* Base with Hyperbase Web : `https://hyperbase.unice.fr/shakespeare`. This corpus places fictional discourse at the heart of the discussion. Complementary to political discourse, which is often spontaneous, we postulate that literature, and in this contribution theater, offers a privileged field of observation of texts and texture. This dialogic corpus can also be confronted to the Latin treatise which are also artificial dialogical texts. The corpora are divided into segments of 20 words using a sliding window on the texts. The structure of the training data collected in this way is described in the following table:

Corpus	**#words**	**#vocab**
French	503.482	18.385
Latin	554.727	43.617
English	1.215.407	33.585

Table 1: Corpora description

2.2. Miniature generative model

To interpret the behavior of deep neural networks on text, we opted for a generative model based on the Transformers architecture. This implementation is derived from the one proposed by Nandan (2020) and adapted for extracting attention scores. (figure 1). The architecture uses a decoder-type Transformer block[3] with a causal attention mask. The model takes a 20-word text as input and outputs the most probable words from a 10,000-word vocabulary. The model has nearly 6 million parameters, including two attention heads and embeddings containing 256 numerical values. This number is sufficient for the generation task targeted by our study corpus while minimizing the energy cost of training this type of model[4]. To preserve data integrity, which is necessary for interpreting machine outputs, no preprocessing is applied; the entire model is trained using the training corpus.

Although generating text is not the goal of our study, it is worth mentioning that each model can correctly generate around ten words with an accuracy of 60% to 70%[5]. These results suggest that our modestly sized models can identify text-encoded phenomena sufficient to predict the next word with relatively high accuracy. These phenomena are directly related to the attention scores calculated by the Transformer. In this study, we wish to analyze and compare these scores with more traditional frequency and statistical measures.

[2] The corpus contains the raw text of the plays, which were copied and pasted from the website `http://shakespeare.mit.edu/`.

[3] Transformers typically use encoders for classification tasks and decoders for generation tasks. The difference lies in the attention mask, which restricts the model to unidirectional attention from each word to its predecessors.

[4] The models were trained in an average of 30 minutes using a consumer RTX 4090 card.

[5] The accuracy score was measured using the French corpus and the B.L.E.U (Bilingual Evaluation Understudy) ; R.O.U.G.E (Recall- Oriented Understudy for Gisting Evaluation) reference algorithms on a test set corresponding to 10% of the corpus

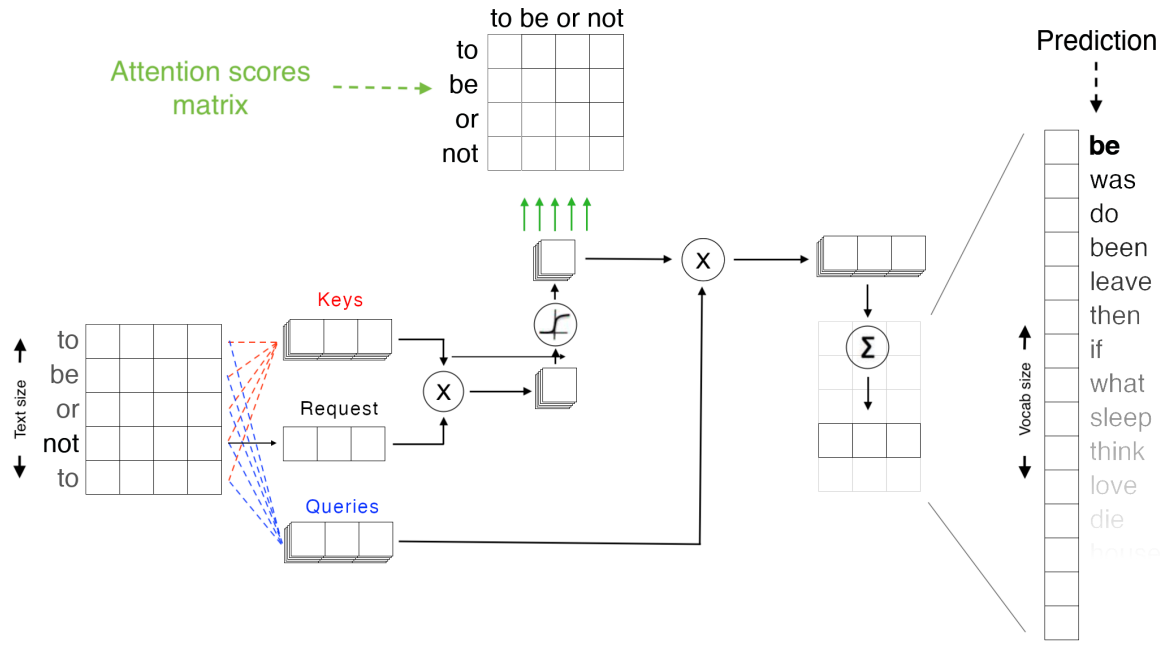

Figure 1: Miniature generative model with attention scores extraction.

2.3. Compare attention with frequency.

The attention scores calculated by Transformers measure the relationships detected by the model between each word in the training corpus texts. These scores are obtained at the output of the *Multihead-attention* layer. These scores are neural weights, also called features, that do not require approximation for projection and analysis, as is the case with convolution/deconvolution. An attention score directly represents the strength of the relationship between two words that the model uses to make a decision, such as choosing the next word. For a given text input, the resulting attention matrix (or tensor, in the case of multiple attention heads) is a square matrix, where each cell represents the attention score between two words. This matrix representation is reminiscent of the frequency tables that have been analyzed for decades by classical statistics, particularly those used to explore generalized co-occurrence in texts Viprey (1997). However, frequency tables and attention scores represent different objects. While frequency tables represent the entire corpus, attention scores only represent the texts submitted to the model's analysis, providing a very local interpretation. The separation between the training corpus and the analyzed texts creates uncertainty about the observed phenomena, which we postulate is at the root of many debates about the interpretability of attention scores. To align the analysis of attention scores with standard frequency tables, we propose shifting from a local analysis of the text used as a prompt to a global analysis of the training corpus. To accomplish this, we will use the training corpus to both train the model and collect attention scores for each pair of words, A and B, encountered in the corpus. Although attention is unidirectional in a decoder, at the corpus level, we neutralize directionality by summing all values to simplify comparison with classical statistical co-occurrence. The overall attention score between words A and B is obtained by scanning all texts in the training corpus X_{train} according to the following calculation:

$$\alpha_{AB} = \sum_{i=0}^{|X_{train}|} \frac{\alpha_{i_{AB}} + \alpha_{i_{BA}}}{2} \tag{1}$$

The global attention matrix M_α is thus symmetric and comparable to a classic frequency co-occurrence matrix M_{freq} which represents the absolute frequency of each pair of words A and B in a given context window. By setting the diagonal (co-occurrence of a word with itself) and the lower part of the matrices (symmetric to the upper part) to 0, we obtain two structurally identical matrices:

$$M_\alpha = \begin{bmatrix} 0 & \alpha_{12} & \cdots & \alpha_{1n} \\ & 0 & \cdots & \alpha_{2n} \\ & \mathbf{0} & \ddots & \vdots \\ & & & 0 \end{bmatrix} \qquad M_{freq} = \begin{bmatrix} 0 & freq_{12} & \cdots & freq_{1n} \\ & 0 & \cdots & freq_{2n} \\ & \mathbf{0} & \ddots & \vdots \\ & & & 0 \end{bmatrix}$$

The overall matrix projection of attention scores enables direct comparison of the two co-occurrence detection methods: frequency-based and attention-based. We chose Correspondence Analysis (CA) to analyze these large tables. CA is derived from Principal Component Analysis (PCA) Pearson (1901) and reduces the number of dimensions, projecting the information onto two axes and maximizing the inertia of the scatter plot. AFC is commonly used to analyze the frequencies of co-occurring pairs. We tested it for this purpose using the overall attention scores obtained from the training corpora. We hypothesize that this method provides an exploratory dimension that enables empirical analysis of word relationships in texts using deep learning of Transformers. In our contribution, we propose statistically measuring the similarity between the frequency and attention matrices by calculating their Pearson correlation coefficient. This calculation complements the observations made by the AFC and enables us to identify the contextual windows in which the two methods are most correlated. We hypothesize that the correlation score is an important indicator of the generative model's overall functioning on our training corpora.

3. Results

This section presents the results obtained from each corpus used in our study. We used CA to analyze and interpret attention scores and Pearson's coefficient to measure correlation with the absolute frequencies of co-occurring pairs. To improve readability, our analyses are limited to the 300 most frequent words in each corpus.[6].

3.1. Interpretation using correspondence analysis

The CA results for each of the three corpora (see Figures 2, 3 and 4) converge on the same general observation: attention scores primarily identify classes of words that are grouped by grammatical category and/or able of forming syntagms. The results illustrate the model's dual sensitivity, which appears to consider both the syntagmatic and paradigmatic (or associative) axes. However, variations in the CA are present in each corpus, reflecting their unique structures and offering a new interpretative approach that we compare to traditional textual statistics.

[6] Excluding outliers (about 5 words) that distort the CA, such as "Monsieur" and "Madame" in the French corpus, which mainly mark highly formatted political discourse.

3.1.1. Verbal syntagms versus topic in political discourse

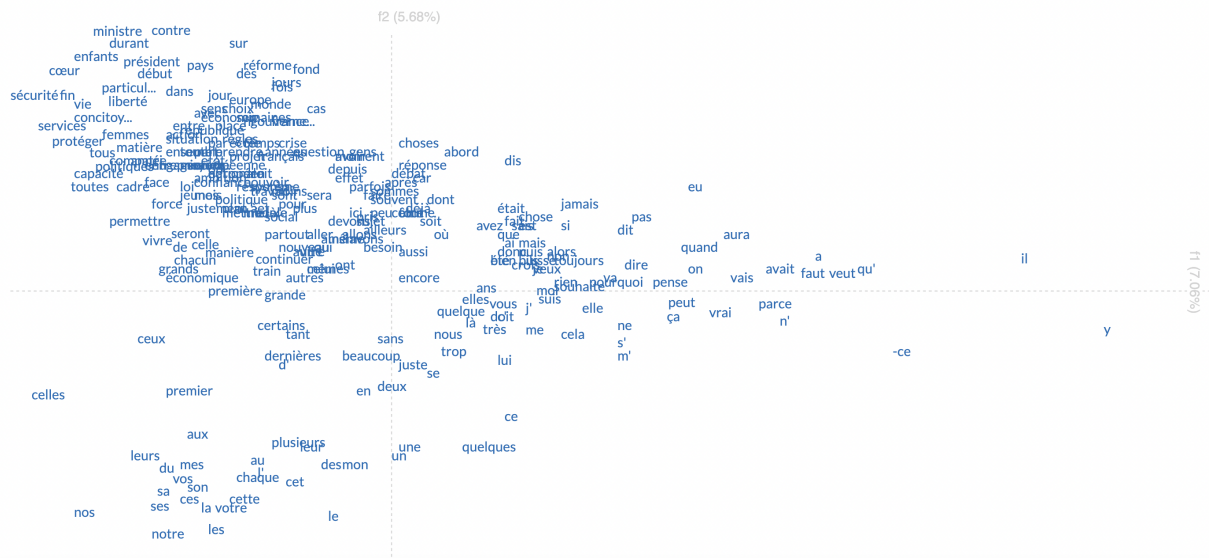

Figure 2: CA of attention scores for the 300 most frequently used words by E. Macron

The CA of attention scores on Macron's corpus (Figure 1) shows strong opposition on axis 1 between verbal phrases on the right and topics on the left. Verbal phrases are represented by syntagmatic markers made from personal pronouns, such as "je" (I), "nous" (we), "il" (he), "on" (it), and auxiliaries, as well as enunciative or modal verbs, such as "penser" (think), "falloir" (must), and "pouvoir" (can). Topics are modeled mainly by nouns such as "ministres" (ministers), "enfants" (children), "sécurité" (security) and "services' (services), along with some infinitive verbs such as "protéger" (protect), "continuer" (continue), and "permettre" (allow). These verbs are reminiscent of certain rhetorical choices made by the French president Mayaffre and Vanni (2023) and mark a tendency to use the verb as a timeless substance, devoid of any conjugation. The second vertical axis of the AFC separates another grammatical category related to nouns, determiners, which are on the left side of the graph. Definite, indefinite, demonstrative, and possessive articles mainly cluster in the lower left. While this category of words has no direct interpretive value, it indicates that it is a discriminating class necessary for the model's automatic speech generation. Note that the third axis, not shown in our examples, reveals a separation between first-person and third-person discourse. This result was previously observed in study Vanni, L. and Mayaffre, D. (2026).

3.1.2. Cicero words categories of words and syntagms

Similar to the French corpus, the experiment conducted on Cicero's speech reveals the acquisition of major word categories in the speech (Figure 3). On the left side of the AFC, a group of words shows the prepositions *ab*, *de*, *in* and *ex* and a litlle bit further *per*, *a*, *contra*, *inter*, *ad*, *sine*, *apud*, *cum*, *post*. The right side of the graph contains deictic words, possesive words or personnal pronouns, as well as some more plain words that point to polemic parts of the speeches (*prouincia*, *Sicilia*, *republica*, *senatu*...). The vertical axis contrasts verbal and nominal phrases. The bottom of the graph shows fixed expressions related to the Roman institutions *"populi Romani" "populus Romanus"*, *"populo Romano"*, *auctoritate*, *ciuitate*, *uirtute*, characteristic of the expository parts of the speeches. At the top of the graph, we see phraseological expressions as *"nullo modo"*, *"fieri non potest"*, as well as an expression with longer-distance links, such as

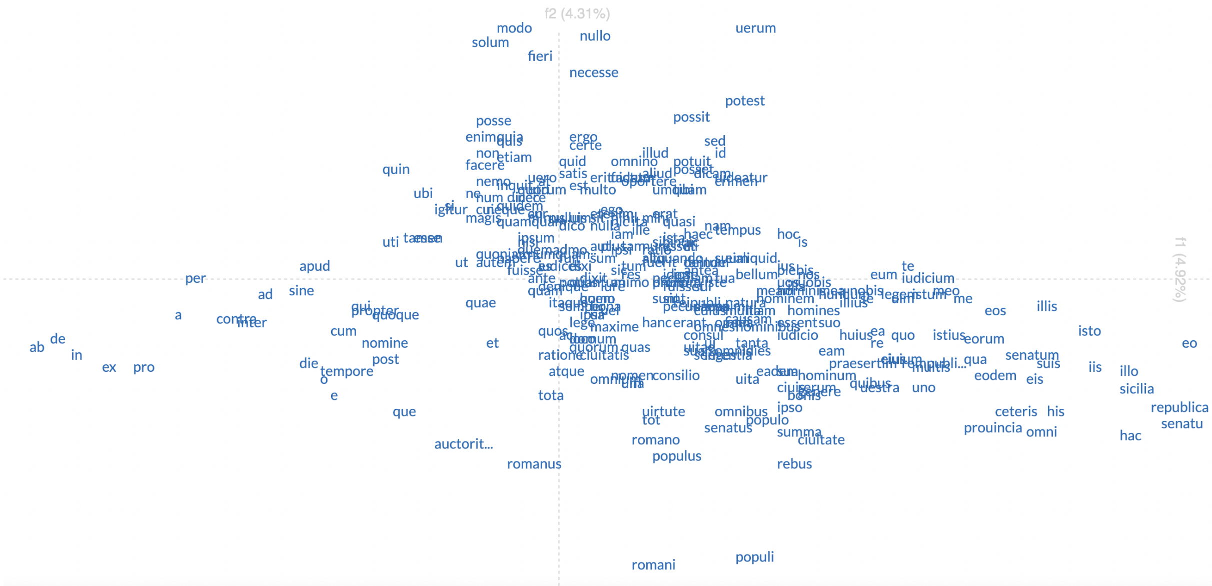

Figure 3: CA of attention scores for the 300 most frequent words in Cicero

"non solum... uerum etiam", various structures which could be analysed as *"textual motives"*.

3.1.3. Structure of Shakespearean theater

The final corpus of our study focuses on Shakespeare's plays. As expected, the attention scores also reveal word classes while specializing in the detection of the theatrical genre. The horizontal axis contrasts function words (with little semantic value but characteristical of dialogic texts) such as pronouns and determiners, with substantives and verbs. The vertical axis separates personal pronouns (top right) and possessive determiners (bottom right). Verbs, like pronouns, logically occupy the upper part of the graph on the left side, while nouns are grouped at the bottom-right space with prepositions as "before", "with", "by", "on', "upon", "from", "against', "into", "of" (partly similar whith the one we notice in Cicero's corpus). Unlike the French and Latin corpora, few syntagms are identified. Since the corpus is larger than previous ones, we hypothesize that the model can generalize information better than rote learning fixed phrases, as seen with Cicero or E. Macron, which shows a tendency for the model to overfit.

Projecting attention scores using the global method we just described and Correspondence Analysis on the three corpora shows good interpretive value and paves the way for descriptive uses of generative models in corpus linguistics. Our results also provide valuable insights into the behavior of generative models. Rather than revolutionizing the statistical analysis of textual data, these machine outputs confirm decades of work on analyzing statistical co-occurrences, particularly generalized co-occurrences Viprey (1997). Indeed, the co-occurrence profiles analyzed in the corpus using frequency tables, such as those presented in Section 2, show similar results when applied to all words (see Hyperbase web databases *elysee*, *cicero*, *shakespear* in section 2.1.). In the final part of this study, we will statistically measure the correlation between the two approaches and draw conclusions about the use of neural networks for textual data analysis and the performance of generative models.

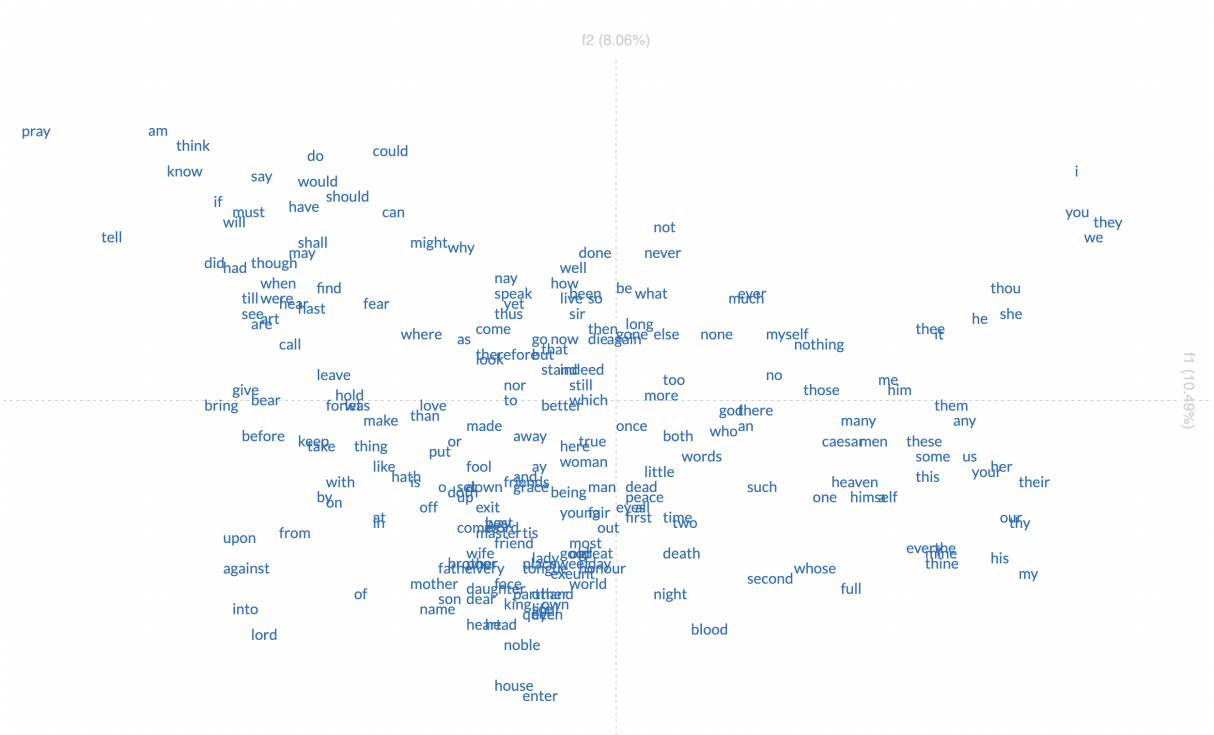

Figure 4: CA of attention scores for the 300 most frequent words in Shakespeare

3.2. Correlation between frequency and attention

Interpreting attention scores remains a significant challenge for the scientific community Charpentier et al. (2024); Seo et al. (2025). However, simple frequency phenomena, such as the detection of classic statistical co-occurrences, are rarely considered in attention score analyses. In this final section, we present a correlation test between the frequencies of co-occurring pairs and the Transformer's attention scores. To accomplish this, we compared the two vectorized matrices presented in Section 2.3. using a Pearson coefficient calculation. To test the scope of self-attention, we formed our comparison frequency matrices using several context sizes, ranging from a 40-word window around the target word (i.e., 80 words total) to a one-word window (i.e., bigrams formed with the target word before or after it). The results are presented in Table 2:

Corpus	Context 40	Context 20	Context 10	Context 5	Context 2	Context 1
French	81.50	80.15	87.17	92.04	**95.72**	86.46
Latin	85.53	87.60	90.76	94.99	**97.02**	86.88
English	82.90	84.13	86.08	88.91	93.83	**94.70**

Table 2: Pearson correlation test between attention-based and frequency-based co-occurrence calculations.

The table shows a correlation rate of over 80%, regardless of the corpus or context window. This result confirms the observations in Section 3.1.: the vector representations resulting from attention score calculations are similar to those manipulated by classical textual statistics. Conversely, the correlation rate appears to be highest when the context window is narrow, favoring

the detection of short-distance relationships. The reduced context size indicates the syntagmatic axis's predominance in predicting the next word in a generative model based on Transformers. More generally, Table 2 mainly shows the importance of traditional frequency co-occurrences in learning automatic text generation tasks with Transformers. The near-perfect correlation between the two matrices obtained for Latin also suggests a hybrid Transformer architecture in which the attention scores are static and frequency-based to optimize the computational cost of model learning. Transformers are among the most resource-intensive architectures; however, a significant portion of their calculations (Self-Attention) converge under certain conditions toward classic statistical calculations.

4. Conclusions

Transformers are proven architectures for many natural language processing tasks. In our study, we only considered miniature models adjusted to the selected study corpora. While these models do not reflect the performance of large language models (LLMs), they demonstrate interesting descriptive capabilities for corpus linguistics. We conducted a comparative study of traditional (frequency-based) co-occurrence analysis and the interpretation of attention scores derived from Transformers. Using three different corpora, we empirically demonstrated that Transformers provide meaningful representations for text analysis. Specifically, the study showed that deep neural networks can represent associative and syntagmatic relationships between words in a training corpus. A Pearson correlation coefficient confirmed the link with frequency-based co-occurrence analysis, showing that most of the information in the attention scores corresponds to the absolute frequency of co-occurring pairs in our corpora. This study suggests that Transformers could be used in an exploratory manner in corpus linguistics to complement traditional statistical tools, although the added value of the heuristic remains to be demonstrated. Thus, further studies could expand this field of research by adjusting text size, model depth, and word selection based on frequency, grammar, or semantics. This would place neural networks in line with historical work in textual statistics.

References

Charpentier F., Cugliari J., and Guille A. (2024). Exploring Semantics in Pretrained Language Model Attention. In Bollegala D. and Shwartz V., editors, *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pp. 326–333, Mexico City, Mexico. Association for Computational Linguistics.

Dar G., Geva M., Gupta A., and Berant J. (2023). Analyzing Transformers in Embedding Space. In Rogers A., Boyd-Graber J., and Okazaki N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16124–16170, Toronto, Canada. Association for Computational Linguistics.

Jain S. and Wallace B. C. (2019). Attention is not Explanation. In Burstein J., Doran C., and Solorio T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Lebart L., Salem A., and Berry L. (1998). *Exploring Textual Data*, volume 4 of *Text, Speech and Language Technology*. Springer Netherlands, Dordrecht.

Mayaffre D. and Vanni L. (2023). Sémantique de corpus numérique. Emmanuel Macron, président thaumaturge (2017-2023). *Espaces Linguistiques*, 6.

Meister C., Lazov S., Augenstein I., and Cotterell R. (2021). Is Sparse Attention more Interpretable?

In Zong C., Xia F., Li W., and Navigli R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 122–129, Online. Association for Computational Linguistics.

Mikolov T., Sutskever I., Chen K., Corrado G. S., and Dean J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pp. 3111–3119.

Molino P., Wang Y., and Zhang J. (2019). Parallax: Visualizing and Understanding the Semantics of Embedding Spaces via Algebraic Formulae. In Costa-jussà M. R. and Alfonseca E., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 165–180, Florence, Italy. Association for Computational Linguistics.

Nandan A. (2020). Text generation with a miniature GPT. Publication Title: GitHub repository.

Pearson K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 1*, 2:559–572.

Ribeiro M. T., Singh S., and Guestrin C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA. Association for Computing Machinery.

Seo S., Yoo S., Lee H., Jang Y., Park J. H., and Kim J.-N. (2025). A Sentence-Level Visualization of Attention in Large Language Models. In Dziri N., Ren S. X., and Diao S., editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 313–320, Albuquerque, New Mexico. Association for Computational Linguistics.

Serrano S. and Smith N. A. (2019). Is Attention Interpretable? In Korhonen A., Traum D., and Màrquez L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, Florence, Italy. Association for Computational Linguistics.

Vanni, L. and Mayaffre, D. (2026). Explore political discourse with transformers. Emergent paradigmatic and syntagmatic representations. In *15th edition of the Language Resources and Evaluation Conference (LREC 2026)*, Palma de Mallorca, Spain.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., and Polosukhin I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Vig J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Costa-jussà M. R. and Alfonseca E., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 37–42, Florence, Italy. Association for Computational Linguistics.

Viprey J.-M. (1997). *Dynamique du vocabulaire des Fleurs du mal,*. Honoré Champion, Paris.

Škrlj B., Sheehan S., Eržen N., Robnik-Šikonja M., Luz S., and Pollak S. (2021). Exploring Neural Language Models via Analysis of Local and Global Self-Attention Spaces. In Toivonen H. and Boggia M., editors, *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pp. 76–83, Online. Association for Computational Linguistics.