Explore Political Discourse with Transformers. Emergent Paradigmatic and Syntagmatic Representations.

Laurent Vanni, Damon Mayaffre
UMR7320: Bases, Corpus, Langage - CNRS - UniCA - France
laurent.vanni@univ-cotedazur.fr, damon.mayaffre@univ-cotedazur.fr

Abstract

Textual data analysis lies at the heart of inductive reasoning in corpus linguistics. Corpus-driven approaches place the corpus at the center of working hypotheses and use statistical processing as an exploratory tool. With deep neural networks, the training corpus is also crucial, but the objectives are less exploratory. Nevertheless, the performance of Transformers in automatic language processing suggests that self-attention is an effective means of extracting structural information from corpora. In this article, we present interdisciplinary work that uses Transformers descriptively to shed light on linguistic phenomena present in a learning corpus. We propose using two feature-based interpretation methods in a case study of political speeches applied to a text generation task. The first method is a global approach that uses attention scores to analyse the training corpus. The second is a local approach that uses gradient-based features to analyse predictions. These methods are compared to standard statistical techniques, providing empirical confirmation of the observed phenomena. We conclude on the potential of Transformers as a heuristic tool for corpus linguistics.

Keywords: Transformers, Features, Corpus, Linguistics, Interpretations

1. Introduction

Political discourse is constantly permeated by linguistic signs that shape the structure and meaning of our leaders' messages. The study of these signs, also known as semiotics, is therefore crucial to understanding the phenomena at play in language that lead to the expression of different political ideologies. The semiotic approach to textual data analysis is based on fundamental concepts proposed by Ferdinand de Saussure, the founding father of modern linguistics, who distinguishes two axes in discourse: a paradigmatic axis and a syntagmatic axis. The paradigmatic axis represents the relationships *in abstentia* (in memory) between words: associative relationships that allow them to be substituted in a given context (e.g., "social" and "political" in the example in Figure 1). Conversely, the syntagmatic axis represents the relationships *in praesentia* (in the chain of discourse) that affect the position and order of words and form comprehensible units such as sentences or passages.

Tools for detecting and analyzing syntagmatic or paradigmatic relationships traditionally use the distributional hypothesis (Harris, 1954; Firth, 1957), which posits that words that share similar contexts tend to have similar meanings. Some tools focus on the syntagmatic axis, such as Latent Semantic Analysis (Deerwester et al., 1990) or collocation analysis (Sinclair, 1991), while others focus on the paradigmatic axis, such as CBOW and SKIP-GRAM models (Mikolov et al., 2013). The specialization and performance of these models depend largely on the parameters used, particularly the contextualization window, as shown by (Lapesa et al., 2014) and (Sun et al., 2015). The introduction of Transformers (Vaswani et al., 2017), along with models such as GPT (Radford and Narasimhan, 2018) and BERT (Devlin et al., 2019), has expanded contexts and boosted performance. However, it is unclear whether Transformers focus specifically on the syntagmatic or paradigmatic axis, or whether they are even able to combine the two levels to perform the required tasks.

In this article, we will present a heuristic approach showing how Transformers can help to explore various types of relationships between words in texts using a model trained to generate political discourse. After a brief introduction to the theoretical framework of this study, we will turn to two exploration methods: a global method, using attention-based features extracted from the training corpus, and a local method using gradient-based features extracted from the model's predictions. The results will show that these methods highlight both paradigmatic and syntagmatic phenomena *in abstentia* and *in praesentia*. Finally, we will conclude by discussing the added heuristic value of Transformers in the context of text analysis applied to the study of political discourse.

2. Related work

Since the advent of the first computers, natural language processing has been a major tool for analysing political discourse. Its application has recently expanded to include deep neural networks (Mayaffre and Poudat, 2013; Marwala, 2023; Skubic and Fišer, 2024). The methods we investigate

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7534–7544
11–16 May 2026.

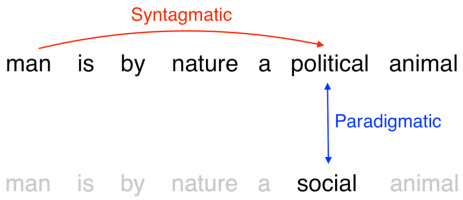

Figure 1: Paradigmatic and syntagmatic axis in modern linguistic

to explore linguistic signs are based on a corpus-driven approach (Tognini-Bonelli, 2002). These are heuristic methods in which statistics and deep neural networks serve an exploratory purpose. The value of Transformers for a corpus-driven exploratory approach has been demonstrated by (Jawahar et al., 2019), who highlighted the ability of BERT-type models to capture syntactic and semantic structural information. The performance of Transformers in semantic analysis of political discourse has also recently been demonstrated (Huguet Cabot et al., 2020; Sucu et al., 2025). This has been achieved through complex attribution tasks, such as the identification of framing, metaphor, emotion and stance in informal discussions. However, most studies on political discourse using Transformers focus on the predictive performance of models (e.g., detection of political lines (Jakob et al., 2024), or sentiment analysis (Abercrombie and Batista-Navarro, 2020)) rather than being directly useful for exploring and analysing discursive linguistic signs.

The exploratory approach that we propose is directly related to model interpretability. Interpretation methods can be classified into different categories (Lipton, 2018). Some depend on the model, such as BertViz (Vig, 2019), while others are agnostic, such as LIME (Ribeiro et al., 2016). Others are associated with game-theoretic methods, such as SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017). The methods chosen in this contribution are based on an analysis of two types of Transformer features: attention-based and gradient-based. Our study evaluates the added value of the feature extraction approaches by comparing them to the baseline statistics used in corpus linguistics. (Oakes, 1998).

The attention-based approach draws inspiration from work such as that of (Vig, 2019), which demonstrates the heuristic potential of attention weights in Transformers. Although the interpretability of attention scores is a much-debated topic (Serrano and Smith, 2019; Jain and Wallace, 2019), we hypothesise that they offer several avenues for exploration that we wish to investigate. Section 4.1 shows a method of global interpretation using attention-based features to explore word relationships across the entire training corpus. Attention scores are converted into word embeddings for visualisation, similar to the approach taken in (Molino et al., 2019). Although it is unclear what types of linguistic signs are captured by standard embeddings (Rogers et al., 2020; Poliak et al., 2018), recent work by (Han et al., 2024) shows that exploring vector representations of words provides valuable insights into discourse construction in text generation tasks.

The gradient-based approach is inspired by image processing (Simonyan et al., 2014; Zhou et al., 2016). These methods have been successfully applied to text in translation tasks (He et al., 2019) and question-answering tasks (Mudrakarta et al., 2018) tasks, achieving good results in the detection of salient features (typically words) responsible for a model's decision. Section 5 shows a local interpretation method that uses a gradient-based feature to explore the word relationships within a given prompt that are responsible for the model's prediction (text generation).

3. Miniature GPT to explore political corpora

The experimental protocol of this study is based on a minimalist Transformer architecture for text generation. The simplicity of the model is a response to the need for interpretability. All the technical choices presented in the following sections aim to simplify the exploration of the model's features. The training corpus was also selected to maximise the interpretability of the model's results. For this study, we selected a set of French political speeches using an interdisciplinary approach supported by linguistic expertise. Section 3.3 provides details of the corpus structure and the model's performance.

3.1. Model

The model is based on the implementation of (Nandan, 2020), who proposed a mini architecture based on Transformers for text generation. Its author illustrates the model using the entire IMDB sentiment classification dataset. We adapted the architecture to fit the French political discourse needs and trained the model on the corpus presented in Section 3.3. Technically, the final model is identical to the one proposed for review generation. It consists of a Transformer decoder block with multiple attention heads, to which a causal attention mask is applied. The Transformer takes a positional embedding representing the text, which is segmented into words[1], as input, and uses a simple dense layer with an output size corresponding

[1]Tokenization is manually adapted to the French political corpus we have chosen for the study (see section 3.3).

to the size of the text multiplied by the size of the vocabulary as output. The sequence-to-sequence task involves predicting the input text after shifting it by one additional word. The modifications we have made are detailed in the following section and mainly concern feature extraction at two points in the neural network: the attention scores and the Transformer output. All the hyperparameters and conditions required to reproduce the results are presented in the section 8.

3.2. Methods

The proposed methods for exploring linguistic signs are specific to the model and employ feature-based approaches. In particular, we focus on two interpretation methods associated with two types of feature related to self-attention calculation (see Figure 2) : i) a global method based on attention scores ; ii) a local, gradient-based method at the Transformer output.

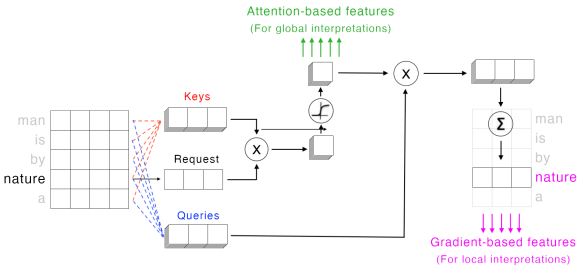

Figure 2: Feature extraction attention-based (used for global interpretations method) and gradient-based (used for local interpretations method)

3.2.1. Global method

As discussed in Section 2, analysing attention scores is not straightforward (Serrano and Smith, 2019; Jain and Wallace, 2019). When attention scores are used to interpret a model's predictions on text samples locally, the detected word associations do not always make linguistic sense. One of the major contributions of this work is shifting the focus from local to global analysis of the model, reusing the training corpus to highlight the associations identified by the model across the entire training data. This approach challenges the general representation of the training corpus obtained from the attention scores.

The main idea is to calculate an average attention score for each pair of co-occurrences within the corpus. The size of the text sequences determines the contextualisation window. We set this to 20 words, which is the average sentence length of the corpus presented in Section 3.3. The raw attention scores correspond to the dot product (plus activation function) of each *query* (word) with each *key* (co-occurrence), as provided by the Multihead-Attention layer. Our technique involves running all the training data through the model to extract the attention scores for each pair of co-occurrences found in the text sequences. The attention scores provided by the Multihead-Attention layer correspond to a tensor for each sequence. Summing the Multihead axis yields a 20×20 matrix representing the co-occurrence scores. These attention scores are not bijective; they represent directed links, i.e. word associations in which directionality is considered. While directionality information is important from a linguistic point of view, especially at the syntagmatic level, we have chosen to neutralise the directionality of the attention scores to simplify the analysis, encouraged by the work of (Saponati et al., 2025). We will discuss the potential impact of this choice in Section 6.

The attention scores for each pair of co-occurrences in a text sequence are added together to create a global $A_{m \times n}$ matrix, where m and n correspond to the size of the vocabulary. This matrix ultimately represents the total scores obtained for all $m \times n$ pairs of co-occurrences in the X_{train} data set. This matrix is triangular and represents all pairs of co-occurrences in the corpus. The total attention for each pair of co-occurrences, a_{mn}, is calculated as follows:

$$a_{mn} = \sum_{i=0}^{|X_{train}|} \frac{a_{i_{mn}} + a_{i_{nm}}}{2} \quad (1)$$

A vector is then extracted for each word from the $A_{m \times n}$ matrix (either a row or a column). This high-dimensional vector can be considered as an embedding representing the words in all the contexts captured by the Transformer (similar to a Word2Vec model, but calculated from attention scores). As with any embedding, the vectors encode a large amount of information, so the number of dimensions must be reduced to make them human-readable. We propose using principal component analysis (PCA) (Pearson, 1901) to achieve this. This method is widely used in statistics and also applied to corpus linguistics (Oakes, 1998; Redmond et al., 2023). The advantage of this technique is that it provides a general and interpretable overview of the vector representations calculated by the attention scores and the overall behaviour of the model.

3.2.2. Local method

As a complement to the global approach, we propose to use a local approach that focuses on feature extraction using a gradient-based method. Gradient-based methods are useful for identifying important features associated with predictions.

In text input to the network, an important feature (also known as Textual Saliency) is generally represented by a token (a word) or a group of tokens that influence the prediction.

We chose a method that uses the heatmap of the Transformer's output layer: a matrix $T_{i\times j}$ where each row i corresponds to a word in the text and each column j corresponds to the weights set by the layer. The sum of the values in a word's vector provides an approximation of its importance in the context of the prediction. However, this sum does not indicate whether the word contributed positively or negatively to the prediction. To overcome this limitation, we adjust the Transformer's heatmap according to the weights of the neurons in the final layer of the network. Our method is inspired by class activation map (CAM) techniques, such as CAM (Zhou et al., 2016) and Gradient-based CAM (Grad-CAM) (Selvaraju et al., 2017), which have produced good results with images and which we wanted to apply to text.

The weighted matrix $\bar{T}$ corresponds to the scalar product of matrix T and vectors V, which represents the weight of the neurons associated with the selected output word. The weighted representation of each word is calculated as follows:

$$\bar{T}_{i\times j} = T_{i\times j} V_j + b \tag{2}$$

Where i corresponds to the ith word in the text sample, j corresponds to the jth weight associated with the corresponding vector, and b corresponds to the bias applied in the layer in the last layer of the network.

The weighted matrix $\bar{T}$ provides a gradient-based representation of the textual saliences that the model uses to predict the next word. To evaluate the importance of the saliences with a score, we sum the values of the vector corresponding to each word. This gives a score we called t-score, which corresponds to the weight of the word w at the Transformer output calculated as follow:

$$t_{score}(w) = \sum_{\alpha=0}^{\alpha=j} t_{w\alpha} \tag{3}$$

As we will see in Section 4, the two proposed methods (global and local) complement each other when exploring paradigmatic phenomena *in praesentia* (i.e. in the memory of the corpus) and syntagmatic phenomena *in praesentia* (i.e. in the sequence of a prompt).

3.3. Corpus

The explainability of phenomena detected by Transformers is closely linked to training corpora. In our study, we analyse modern French political discourse — specifically that of the current president, Emmanuel Macron — with the help of linguists who specialise in this area. To achieve this, we collected all of the president's official speeches available on the *Élysée* French institutional website[2] for the period 2017–2024. This amount to 70 speeches containing 459,239 tokens. We divided these speeches into samples of 20 words using a sliding window, which provide a total of 426,503 samples. The total vocabulary consists of 17,634 tokens, including punctuation and all signs separated by spaces.

The model's task is to predict the next word in each sample using sequence-to-sequence encoding. The output corresponds to the input sample, but shifted by one word. We normalised the vocabulary by removing capital letters and limited it to 10,000 words to reduce the impact of isolated words.

The model evaluation uses 10% of the dataset to measure the precision of n-grams generated by the model (fixed at 20 new tokens), against the reference text. We compute BLEU and ROUGE scores, achieving the following results: BLEU=68.35 ; ROUGE-1 F1=0.74 ; ROUGE-L F1=0.73.

While the focus of the study is not on the model's performance in automatic generation, it is worth noting that the modest size of the corpus associated with the minimal model we propose allows us to achieve satisfactory results from both statistical and human perspectives (when reading the generated texts). We hypothesise that this performance is sufficient to use the model for exploratory purposes, querying the linguistic signs detected during the training process that led to these results.

4. results

This section shows the results obtained on the study corpus. Based on the attention scores calculated on the most frequent words in the corpus, we will first look at the main co-occurrence profiles identified by the model across the entire corpus. Then, using an example prompt, we will illustrate how gradient-based features can be used to identify local n-grams that are important for model prediction.

4.1. Exploring global syntagmatic and paradigmatic linguistic signs

The initial results presented focus on the attention matrix $A_{m\times n}$. As mentioned in section 3.2.1, the dimensionality of the obtained word vectors requires the application of a dimension reduction method to enable visualisation of the embeddings and examination of their interpretability. We opted for PCA,

[2] https://www.elysee.fr

which is widely used by the text data analysis community. This method projects the vectors onto two axes, making it possible to observe the similarities or differences between words based on their vector profiles.

The size of the raw $A_{m\times n}$ matrix corresponds to the size of the model's vocabulary, i.e. 10,000 words. Even when applying PCA, displaying the entire vocabulary would be difficult to visualise and interpret. In an initial experiment, we projected the 300 most frequent words in the corpus (based on absolute frequency across the entire training dataset). Figure 3 shows the result, with the strongest contributions to the principal axes highlighted (i.e. the important words).

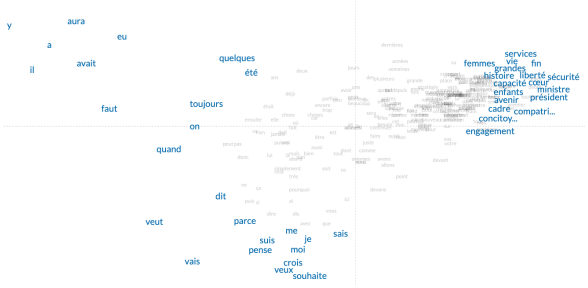

Figure 3: PCA of the 300 most frequent words in the corpus based on attention scores matrix

The result is striking. Axis 1 (the horizontal axis) shows a contrast between syntagmatic and paradigmatic relationships. The left side mainly comprises verbs such as "être" (to be), "avoir" (to have), "penser" (to think) and "croire" (to believe), conjugated with nearby pronouns such as "je", "il", "on" (I, they, it). The right-hand side of the graph highlights nouns that form themes such as "ministre/président" (minister/president), "enfants/avenir" (children/future), and "sécurité/liberté" (security/freedom). Axis 2 (vertical) adds a further contrast to the left of the graph, between discourse that is mainly focused on the first person (at the bottom) and more detached discourse that is focused on the third person (at the top). This result illustrates two well-known communication strategies in politics. The first is a discourse in which the head of state expresses a personal position: "je pense" (I Think), "je crois" (I believe), "je veux" (I want). The second is a detached discourse in which the president makes general statements such as "il y a/avait/aura'/...' (there is/are/was/will be/...), "il faut" (it is necessary), illustrated also by the gender-neutral third person "on". It is also important to note the natural placement of adverbs in the syntagmatic chain such as "quelques" (some/a few), "toujours" (always), "quand" (when) around verbs that indicate the speaker's position in space and time. The result therefore shows categories of words that oppose or attract each other, as well as phrases (particularly verbal) that are formed in speech. These results suggest that attention scores analysed globally across the entire learning corpus can provide information about relationships between words in terms of both their position in the text (syntagmatic axis) and their selection (paradigmatic axis).

To further explore the paradigmatic axis, we conducted a second analysis focusing on common nouns. We reproduced the PCA from the $A_{m\times n}$ matrix, this time retaining only the 300 most frequent nouns in the corpus (obtained via automatic annotation of the corpus). These were then projected onto the first two principal axes of the PCA. Figure 4 shows the resulting accentuated representation of semantic correlations.

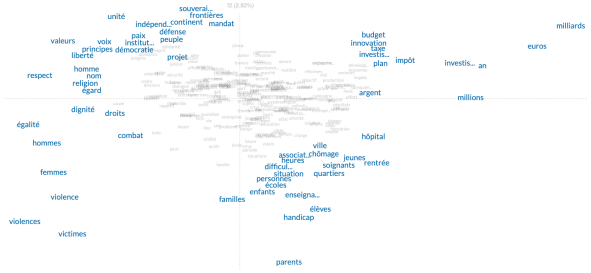

Figure 4: PCA of the 300 most frequent nouns in the corpus based on attention scores matrix

The PCA of names reveals several clusters of words at the ends of the graph (highlighted in the figure), which clearly illustrate the main themes of French presidential discourse. In the top right quadrant is vocabulary related to the country's economy:"milliards/millions" (billion/million), "euros", "investissement" (investment), "impôt/taxe" (taxation/tax), "budget", and so on. On the opposite side of the horizontal axis are words related to the defended values: "liberté" (freedom), "démocratie" (democracy), "respect", etc. On the secondary axis, there is a contrast between national discourse touching on social themes in the country: "homme(s)/femme(s)/violence(s)" (men violence against women), "écoles/enseignants/élèves" (schools/teachers/students), "famille/enfants" (family, children), "chômage/jeunes" (unemployment/young people), etc. , and an international discourse: "souveraineté" (sovereignty), "frontières" (borders), "continent", "paix" (peace), etc. This second result, which focuses on nouns, confirms the exploratory value of attention scores and how representations can adapt depending on the words selected.

From a methodological perspective, we also observe that the attention scores converge towards a standard representation of co-occurrence in texts. The representations provided by figure 3 and 4 are well known in classical statistics applied to

textual data analysis. These are usually derived from matrices similar to attention scores, but are based on absolute frequency and are much easier to compute. To calculate a similarity index between the two approaches (absolute frequencies and attention scores), we constructed a second co-occurrence matrix identical to the first one (based on the 300 most frequent words), but calculated using the absolute frequency of each pair observed in the corpus (with the same contextual window as the model, set at 20 words). By applying a Pearson correlation coefficient, we obtained a correlation rate of 82.32%. This indicates a tendency for attention scores to primarily reflect frequency phenomena (a tendency that is fairly expected but rarely measured). However, given the remaining performance gap of almost 20% associated with Transformers, we argue that attention scores provide additional exploratory value to complement traditional statistics.

This experiment has empirically demonstrated that attention scores provide a useful basis for analysing textual data, such as political discourse. Using the training corpus as the sole source of linguistic hypotheses, we have shown that it is possible to analyse the syntagmatic and paradigmatic relationships that structure the French president's speeches.

Note. Pearson's correlation coefficient test suggests that for an automatic text generation task, attention scores could be initialized or even replaced by a frequency matrix with constant-time computation in order to significantly reduce the computational cost of Transformers.

5. Exploring local syntagmatic and paradigmatic linguistic signs

The second part of our research into exploring political discourse using a Transformer focuses on analysing predictions on given samples. We use a gradient-based approach to study local textual salience (i.e. within a given text sequence), which plays an important role in the model's prediction. As a reminder, the miniature model that we implemented was trained on a corpus of political discourse and was tasked with predicting words that could complete the input text (the prompt). To illustrate the added value of the approach, we will show first the results obtained on a given exemple (see Figure 5).

To compute the t-score (gradient-based feature score) for each word in the prompt, we selected the five highest probabilities from the prediction. In our example, the prompt correspond to the text: "L'intelligence artificielle est une révolution" (Artificial intelligence is a revolution). And the prediction

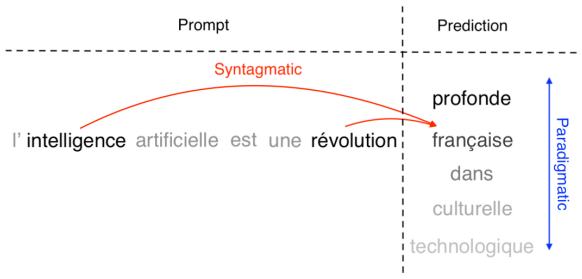

Figure 5: Syntagmatic and paradigmatic signs related to the example prompt : "L'intelligence artificielle est une révolution" (Artificial intelligence is a revolution)

corresponds, in descending order of probability, to the words: "profonde" (deep), "française" (French), "dans" (in), "culturelle" (cultural), and "technological" (technological). Four of the five words proposed by the model are therefore qualifying adjectives (often positioned after the noun in French). This category of words confirms the model's ability to detect associative relationships, as observed in Section 4.1. These relationships refer to *in abstentia* phenomena which are encoded in the model's memory and observed using the proposed global analysis of the training data.

From a linguistic point of view, the use of adjectives in politics often refers to technical discourse that aims to be precise. Here, the word "profound" (the most likely word predicted by the model) adds a dramatic, even hyperbolic dimension, an exaggeration of the facts that is indeed found in the contemporary political discourse of many leaders, such as in the discourse of the current US President Donald Trump (Abbas, 2019).

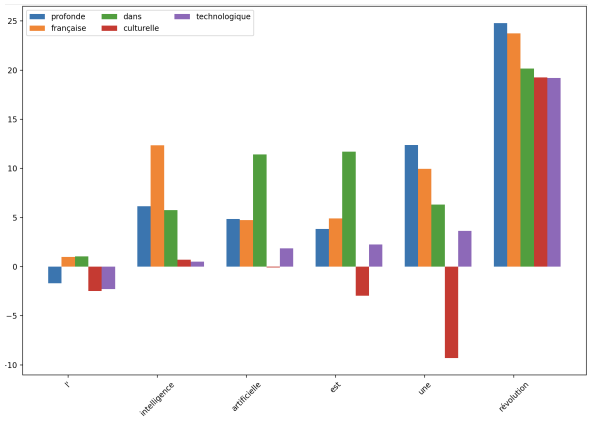

Figure 6: Gradient-based feature scores (t_{score}) applied on prompt "L'intelligence artificielle est une révolution" (Artificial intelligence is a revolution)

Figure 6 shows how the t-score (based on gradient calculation see section 3.2.2) varies depending

on the choice of one of the five words with the highest probability of being the output. The first observation from this graph is that the word “révolution” has the highest score of all five prediction candidates. This indicates in our example that the final word in the prompt is the most important feature for predicting the next word. From a syntagmatic point of view, this intuitive result can also be explained by Table 1, which shows that all bigrams formed with the word “révolution” are statistically salient in the training corpus (i.e. they have a low probability under the null hypothesis of distribution).

révolution...	occ	co-occ	probability
..profonde	241	2	3.92e-6
...française	1542	2	1.59e-4
...dans	3259	1	0.18
...culturelle	142	3	7.07e-10
...technologique	82	1	9.66e-4

Table 1: Bigram probability with word “révolution”

The graph also shows that the scores of preceding words can vary significantly depending on the word selected in the model output. This indicates the presence of other significant co-occurrences. For instance, the word “française” is associated with “intelligence” to form the bigram “intelligence française” (French intelligence), which the model seems to point statistical significance in Macron’s speech. It occurs seven times out of 116 instances of intelligence and 138 instances of French, giving a probability of 8.8e-15. Statistical confirmation here shows significant complementarity in validating the observations made using the gradient-based features.

In this example, the t-score’s correlation with the presence of statistically significant bigrams indicates that the local analysis of gradient-based features primarily reveals syntagmatic relationships. To confirm this hypothesis and extend the scope of the demonstration, we compare the t-score calculation with the z-score, a statistical baseline test for corpus linguistics (Oakes, 1998), applied to bigram calculations. Using all the prompt sequences in the test dataset (10% of the corpus), we created a t-score matrix representing the average gradient-based feature score for each source word in relation to the highest target word (prediction) score. After sorting the results, we retain the top 20 source words in the prompt and compare the score with the bigram z-score calculation as follows:

$$z = \frac{k - fp}{\sqrt{fpq}} \tag{4}$$

Where k is the number of occurrences of the bigram formed with the target word (the highest prediction) and the source word in the text, f is the number of occurrences of the target word in the corpus, $p = \frac{f}{T}$ where T corresponds to the number of words in the corpus with $q = 1 - p$. The z-score gives positive scores to words that are overused alongside the target word in bigrams. The test we conducted compares the variations in the z-score when the t-score deacrease.

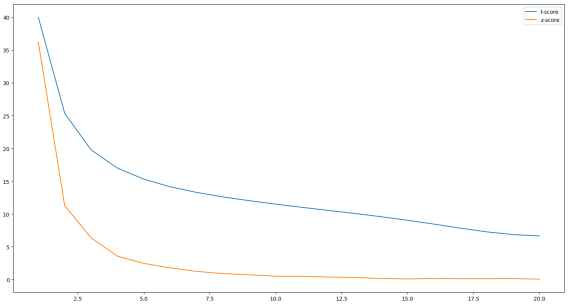

Figure 7: Variation in z-score relative to t-score based on the ranking of the 20 highest t-score.

Figure 7 shows the results. A comparison of the profiles of the two curves shows that the two measures are highly correlated. The bigram z-score perfectly tracks the decline in the t-score. This demonstrates the important role of gradient-based features in detecting statistical syntagmatic relationships. Combining this result with those observed using attention scores, Transformers offer numerous possibilities for analysing textual data, such as political discourse. As illustrated in the examples (Figures 6 and 5), Transformers can detect both *in praesentia* and *in abstentia* linguistic phenomena and create representations that can be used not only for prediction tasks, but also for exploring linguistic corpora.

6. Discussion

Self-attention, inspired by human visual attention, is bidirectional. The phenomena captured by Transformers use positional encoding to indicate the position of words in the text, allowing self-attention to distinguish the directionality of a co-occurring pair. In our work, we chose to neutralise directionality, reducing the attention score matrix to a symmetric matrix that is easier to analyse. Our approach is based on the recent work of (Saponati et al., 2025), which demonstrates that Transformers tend to induce symmetries in the features they detect. However, the authors validate their hypothesis using encoders, whereas we use decoders in our experiments. Therefore, exploring the variation in attention related to word position is important, given that word order is a significant syntactic component in linguistics.

The analysis of attention scores (Section 4.1) illustrates both paradigmatic and syntagmatic phenomena. Indeed, PCA on the entire vocabulary (Figure 3) seems to distinguish between paradigms and syntagms on axis 1, and PCA on nouns (Figure 4) clearly offers a paradigmatic representation of the corpus. Analysis of the feature gradient at the transformer output appears to favour syntagmatic relationships (n-grams). However, paradigmatic relationships, as observed with attention scores, play an important role in the model's predictions. Nevertheless, our method does not allow us to observe these directly in the text prompts at the network input. A potential solution to be explored in future studies would be to project the attention scores onto the text, as BertViz does, taking into account the various attention heads, and then compare these projections to the scores obtained in the corpus's global matrix.

A detailed linguistic analysis of the results obtained was conducted using a French political corpus. It gives good linguistic interpretations based on expert knowledge in French political discourse, but we cannot generalise the observations to other corpora. Additional studies involving different languages and types of discourse would enable to validate the results on a larger scale. Furthermore, in our work, the emergence of paradigmatic and syntagmatic relationships in Transformers only concerns text generation. It would be useful to compare these results with those of models trained on other tasks, such as automatic classification. It is uncertain whether attention scores would focus on syntagmatic and paradigmatic relationships or merely on statistical phenomena known in corpus linguistics, such as lexical specificities (Pincemin, 2024).

7. Conclusion

In this paper, we empirically demonstrate that Transformers can contribute to the exploration of linguistic signs in corpora such as political discourse. The features calculated from self-attention point to emerging linguistic phenomena that contrast *in praesentia* relationships (within the text sequence) of a syntagmatic nature with *in abstentia* relationships (within memory of the training corpus) of a paradigmatic nature. We focus on two methods of using Transformers to explore such linguistic signs. The first global method effectively captures word categories and associative relationships, enabling direct exploration of the training corpus as a linguistic object. The second method, which is local, highlights salient textual elements that illustrate the generative process, encouraging researchers to examine the syntagmatic relationships detected by the model. Transformers therefore offer an encouraging results for exploratory approach needed by corpus linguistics. The weights of the neurons in a trained model appear to be promising descriptive material that complements traditional statistical analysis of textual data.

8. Supplementary Materials

All the tokens in the dataset are lowercased and the Spacy Python library (model "fr_core_news_lg") for part-of-speech tagging is used to extract the most frequent nouns in Section 4.1. (labelled "NOUN"). Tokenisation was adapted to our study corpus using specific cleaning rules. All stop words and punctuation are retained to preserve all possible linguistic features in the text.

The miniature GPT model was implemented using the Keras/TensorFlow source code provided by (Nandan, 2020). The hyperparameters used to fit the dataset are: vocabulary size: 10,000 tokens; sequence size: 20 tokens; embedding size: 256; number of heads (multi-head attention): 8; dense size (last layer): 512. The model was trained using a batch size of 512 and a sparse categorical cross-entropy loss function (with the default Adam optimiser) on an RTX 4090.

The attention scores are given by the MultiHeadAttention layer (using the parameter "return_attention_scores"). The weights of the final dense layer are provided by the "get_weights" method of the Keras library, which we use to adjust the output representation of the Transformers block to compute the feature gradient-based score (t-score). Statistical analysis and PCA visualisations are provided by the Hyperbase web platform (`https://hyperbase.unice.fr`).

9. Bibliographical References

Ali Abbas. 2019. Super-hyperbolic man: Hyperbole as an ideological discourse strategy in trump's speeches. *International Journal for the Semiotics of Law*, 32:1–18.

Gavin Abercrombie and Riza Batista-Navarro. 2020. ParlVote: A corpus for sentiment analysis of political debates. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5073–5078, Marseille, France. European Language Resources Association.

S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R.A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41*, pages 391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

J. R. Firth. 1957. A synopsis of linguistic theory 1930-55. 1952-59:1–32.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. 2024. Word embeddings are steers for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1)*, pages 16410–16430, Bangkok, Thailand. Association for Computational Linguistics.

Zellig Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Shilin He, Zhaopeng Tu, Xing Wang, Longyue Wang, Michael Lyu, and Shuming Shi. 2019. Towards understanding neural machine translation with word importance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 953–962, Hong Kong, China. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4479–4488, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *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)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Charlott Jakob, Pia Wenzel, Salar Mohtaj, and Vera Schmitt. 2024. Augmented political leaning detection: Leveraging parliamentary speeches for classifying news articles. In *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and short papers*, pages 126–133, Vienna, Austria. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Gabriella Lapesa, Stefan Evert, and Sabine Schulte im Walde. 2014. Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 160–170, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Neural Information Processing Systems*.

Tshilidzi Marwala. 2023. *Natural Language Processing in Politics*, pages 99–115. Springer Nature Singapore, Singapore.

Damon Mayaffre and Céline Poudat. 2013. Quantitative approaches to political discourse: corpus linguistics and text statistics. In Kjersti Flottum, editor, *Speaking of Europe. Approaches to complexity in European political discourse*, pages 65–83. John Benjamins.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*.

Piero Molino, Yang Wang, and Jiawei Zhang. 2019. Parallax: Visualizing and understanding the semantics of embedding spaces via algebraic formulae. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 165–180, Florence, Italy. Association for Computational Linguistics.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1)*, pages 1896–1906, Melbourne, Australia. Association for Computational Linguistics.

Apoorv Nandan. 2020. Text generation with a miniature gpt. https://github.com/keras-team/keras-io/blob/

master/examples/generative/text_generation_with_miniature_gpt.py.

Michael P. Oakes. 1998. *Statistics for Corpus Linguistics*. Edinburgh University Press.

Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 1*, 2:559–572.

Bénédicte Pincemin. 2024. Specificities and other applications of the Fisher's exact test to textual data: What's the matter with lexical frequencies? In *JADT 2024 : Mots comptés, textes déchiffrés*, volume 2 of *Cahiers du Cental*, pages 703–712, Bruxelles, Belgium. Presses universitaires de Louvain.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Leslie Redmond, Denis Foucambert, and Lucie Libersan. 2023. *Language Corpora and Principal Components Analysis*, pages 117–132. Springer International Publishing, Cham.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Matteo Saponati, Pascal Josef Sager, Pau Vilimelis Aceituno, Thilo Stadelmann, and Benjamin F Grewe. 2025. The underlying structures of self-attention: symmetry, directionality, and emergent dynamics in transformer training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 52958–52994. PMLR.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

K. Simonyan, A. Vedaldi, and A. Zisserman. 2014. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track*.

John Sinclair. 1991. Corpus, concordance, collocation.

Jure Skubic and Darja Fišer. 2024. Parliamentary discourse research in political science: Literature review. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Arman Engin Sucu, Yixiang Zhou, Mario A. Nascimento, and Tony Mullen. 2025. Exploiting contextual information to improve stance detection in informal political discourse with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1097–1110, Vienna, Austria. Association for Computational Linguistics.

Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2015. Learning word representations by jointly modeling syntagmatic and paradigmatic relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1)*, pages 136–145, Beijing, China. Association for Computational Linguistics.

Elena Tognini-Bonelli. 2002. Corpus linguistics at work. *Computational Linguistics*, 28:583–583.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.