A Syntactic and Semantic Probe into Language Evolution based on Large Language Models

Hao Pang[1], **Changcheng Li**[1], **Yingxue Liu**[1]
[1]Dalian University of Technology
Correspondence: yxl281@dlut.edu.cn

Abstract

Language evolution is cognitively motivated by the reduction of communicative effort. Current research exploring this reported tendency has been constrained by the heavy reliance on manually annotated resources (e.g., dependency parsing) as well as a narrow focus (e.g., syntax as the single metric). To transcend these limitations, we propose two measures: **Attention-based Structural Distance** (ASD) and **Semantic Space Distance** (SSD). ASD is a parser-free measure of syntactic locality derived from the attention mechanism of pretrained large language models (LLM), while SSD is a measure of lexical distances that quantify the degree of separation between different parts of speech in the word vector space. Based on multiple diachronic and multilingual corpora, our experiments show a significant decrease of ASD while an increase of SSD, which implies a language developmental trend towards structural compactness and semantic divergence. Our research pioneers a novel lens grounded in LLM for studying language evolution, which has two major contributions. Linguistically, our study corroborates the hypothesized law of human language evolution by demonstrating that its development optimizes syntactic locality as well as functional semantic discriminability. Cognitively, our study shows that human and LLMs share common characteristics in language processing, lending support to the potential of employing LLMs in the study of human cognition.

1 Introduction

An important question in linguistics concerns the role of communicative efficiency in driving language evolution. Specifically, according to the Principle of Least Effort (Zipf, 1949), whereby individuals, in pursuing a goal, tend to select the path of shortest distance and least energy costs. On human languages, linguistic structures exhibit the tendencies towards efficiency to maximize information transmission while minimizing cognitive and productive costs, thereby striking a balance between the speaker's preference for expressive economy and the listener's demand for interpretive clarity.

Among the research examining this principle, a notable line is the recent use of dependency distance, the minimization of which is hypothesized as human language universal (Liu et al., 2016).

However, most existing work relies on manually annotated dependency treebanks, resources that are sparse or unavailable for historical texts and low-resource languages.

In this light, of particular importance for supporting such a hypothesized language universal is investigations on a large-scale cross-linguistic diachronic corpus.

Recent advances in pretrained language models offer a promising alternative. Transformer-based models have been shown to implicitly encode rich syntactic and semantic information within their internal representations and attention mechanisms (Vaswani et al., 2017; Htut et al., 2019; Clark et al., 2019; Li et al., 2020). These findings suggest that attention patterns can function as a data-driven measure of syntactic construction, enabling structural analysis directly from raw texts.

Building on this insight, we propose two measures: Attention-based Structural Distance (ASD) and Semantic Space Distance (SSD). ASD is a parser-free metric that quantifies syntactic locality derived from the attention patterns/mechanisms of pretrained large language models (LLM) by converting attention distributions into weighted inter-word distances. In contrast to traditional parsing-based measures, ASD allows scalable and comparable analysis across languages and historical periods without reliance on annotated syntactic resources. And SSD is a measure of lexical distances that quantifies the degree of separation between different parts of speech in the word vector space, which complements ASD from the semantic perspective.

Findings of the Association for Computational Linguistics: ACL 2026, pages 42228–42250

Our empirical large-scale analysis based on multiple diachronic and multilingual corpora (approximately 70 million tokens) reveals a clear trend toward a decrease in ASD and an increase in SSD, which shows syntactic compactness and lexical disambiguation, and their synergy promotes the development of more efficient and clearer langauge communication systems.

Taken together, our framework contributes a scalable, annotation-free approach to diachronic linguistic analysis and offers new empirical support for efficiency-driven language evolution. Our study also shows that human and large language models share common characteristics in language processing, lending support to the potential of employing LLM in the study of human cognition.

2 Related Work

2.1 Efficiency-driven Language Evolution from Syntactic and Semantic Perspectives

One line for language efficiency evolution is the dependency distance minimization (DDM) from the syntactic perspective.

DDM, the tendency for syntactically related words to appear close in linear order, has been widely proposed as a core principle of syntactic organization and a reflection of cognitive constraints on language processing. Given the limited capacity of human working memory (Miyake and Shah, 1999), shorter dependency distances are hypothesized to reduce integration cost and facilitate efficient comprehension (Yngve, 1960; Liu, 2008).

Large-scale cross-linguistic studies show that natural sentences exhibit significantly shorter mean dependency distances (MDD) than randomized baselines across diverse languages (Liu, 2008; Futrell et al., 2015). Diachronic and typological analyses further suggest that DDM may function as a near-universal pressure shaping word order preferences (Liu et al., 2016; Lei and Jockers, 2020). Psycholinguistic research corroborates these findings: longer dependencies incur higher processing costs, as predicted by dependency locality theory (Gibson, 1998), and are associated with increased reading difficulty in eye-tracking studies (Demberg and Keller, 2008).

Language efficiency has also been studied from semantic and information-theoretic perspectives. The Uniform Information Density (UID) hypothesis posits that speakers distribute information to avoid processing bottlenecks and reduce ambiguity (Jaeger, 2010). This line of work primarily addresses predictability and information flow, but it rarely examines the evolution of functional semantic distinguishability itself.

Notably, most of the existing diachronic evidence for language efficiency evolution relies on manually annotated dependency treebanks, severely limiting its applicability to historical corpora and low-resource languages. Most of the existing studies also do not utilize the recent advances in transformer and large language models.

2.2 Attention and Implicit Syntactic Structure in Transformer Models

Attention-based modeling shares a similar cognitive principle–efficient language processing under limited resources. Humans are limited by working memory, while models are shaped by positional encodings and the statistical dominance of local patterns in training data—both favoring short-range dependencies.

In Transformer-based language models without explicit structural bias, self-attention assigns higher weights to nearby tokens, exhibiting a strong locality preference. This behavior reflects both cognitive and computational constraints.

Models with relative positional encodings converge faster on short sequences (Press et al., 2022). Disrupting word order increases dependency length and degrades both syntactic coherence and model performance, paralleling human difficulties with unnatural word orders.

Recent studies demonstrated that pretrained Transformer models encode substantial syntactic knowledge despite lacking explicit structural supervision. Contextual representations from BERT (Devlin et al., 2019) captured subject–verb agreement and other grammatical constraints (Goldberg, 2019). Analysis showed that syntactic structure is often more strongly encoded than semantic content (Tenney et al., 2019). Structural probes further revealed that dependency tree distances can be approximately recovered from BERT embeddings, indicating latent tree-like organization (Hewitt and Manning, 2019). And recent work suggested that large language models implicitly rely on latent tree-structured representations during sentence processing (Liu et al., 2025).

Beyond representations, self-attention weights themselves have been shown to reflect syntactic relations. Individual attention heads reliably capture specific dependency types, such as subject–verb

and verb–object relations (Htut et al., 2019; Clark et al., 2019). However, no single head reconstructs full syntactic trees, and attention-based structure is distributed across heads. Analyses in neural machine translation similarly find that low-entropy, high-confidence heads often align with syntactic dependencies or rare lexical items (Voita et al., 2019).

While existing supervised probing methods provided strong evidence for latent syntax, they rely on external annotations or task-specific objectives, making them less suitable for large-scale diachronic comparison. This motivates parser-free approaches that operate directly on attention distributions. In this regard, we propose an attention-based structural distance, which applies to raw historical text without external annotations.

2.3 Semantic Embedding Space and Functional–Semantic Discriminability

In parallel, distributional semantics research has demonstrated that part-of-speech (POS) distinctions are inherently reflected in semantic space. Kutuzov et al. (2016) found that unsupervised embeddings trained on large corpora cause words of the same POS to cluster together, with each POS class tending to occupy a distinct region in semantic space. Even simple co-occurrence models can recover syntactic categories from word meaning alone: co-occurrence vectors carry "accurate information about syntactic category," such as noun versus verb (Westbury and Hollis, 2019). Analyses of contextual embeddings further reveal POS-sensitive organization—different word classes tend to occupy distinct regions of embedding space, which in turn affects their stability and variance. For example, Schulte im Walde and Frassinelli (2022) found that nouns and verbs behave differently in distributional measures.

Similarly, Gong et al. (2023) noted that incorporating POS cues can "determine and reinforce the semantics in sentence representation," underscoring the role of functional categories in shaping contextual meaning. Together, this body of evidence suggests that lexical categories are not arbitrarily mixed in semantic space but are separated by distributional features.

Motivated by these existing works, our work directly measures how POS categories diverge in semantic embedding space over time, thereby revealing the evolution of language efficiency through both syntactic compactness and semantic divergence within a unified representational framework, particularly from a diachronic and cross-corpus perspective.

3 Method

3.1 Attention-Based Structural Distance

3.1.1 Average Attention Matrix

To capture structurally informative attention patterns, we extract self-attention weights from the final layer of the model. Previous studies indicate that higher layers preferentially encode long-range semantic and discourse relations, whereas lower layers are dominated by local syntactic cues (Clark et al., 2019; Kovaleva et al., 2019). Using the top layer therefore aligns with our goal of measuring global structural organization rather than surface adjacency.

Not all attention heads contribute equally to structural representation. Similar to Voita et al. (2019), we observe substantial redundancy across heads, with only a subset exhibiting low-entropy, sharply peaked distributions indicative of meaningful relational focus. To aggregate attention signals, we weight each head by the inverse of its average entropy:

$$w_h^{(k)} = \frac{\exp\left(-\bar{H}_h^{(k)}\right)}{\sum_{h'} \exp\left(-\bar{H}_{h'}^{(k)}\right)},$$

$$\bar{H}_h^{(k)} = \frac{1}{m_k} \sum_{i=1}^{m_k} \left(-\sum_{j=1}^{m_k} A_{ij}^{(h,k)} \log A_{ij}^{(h,k)} \right).$$

where k indexes sentences, m_k denotes the number of real tokens in sentence k, and $A^{(h,k)} \in \mathbb{R}^{m_k \times m_k}$ is the attention matrix of head h for sentence k. The final attention matrix is then a weighted sum:

$$A^{(k)} = \sum_{h=1}^{H} w_h^{(k)} A^{(h,k)}.$$

This entropy-weighted aggregation emphasizes structurally informative heads while suppressing diffuse or redundant ones. We then remove special tokens and retain the submatrix corresponding only to real lexical items, yielding a clean intra-sentential attention representation.

3.1.2 Attention Structural Distance and Mean Attention Structural Distance

Intuitively, if attention mass is concentrated on nearby tokens, the resulting structure is compact;

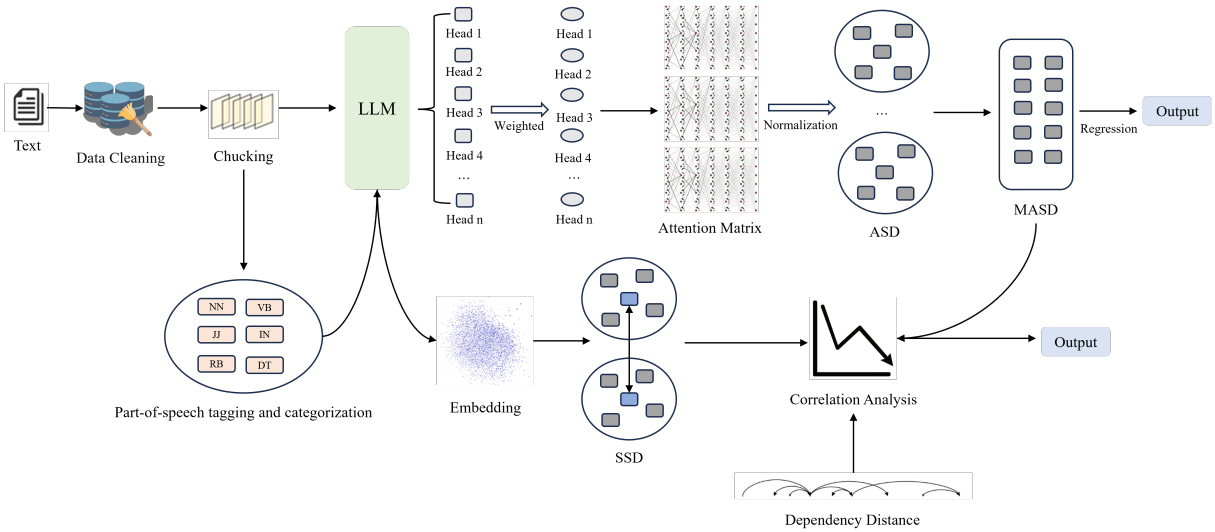

Figure 1: Overview of the method.

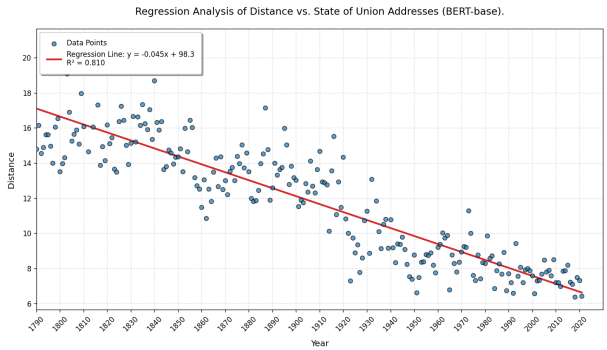

Figure 2: Experimental analysis of MASD in the State of the Union Address based on the BERT model.

frequent attention to distant tokens indicates greater reliance on long-range dependencies.

Formally, for a sentence with $m = L - L_1$ real tokens, where L_1 denotes the number of special tokens (e.g., [CLS], [SEP]) , let $A' \in \mathbb{R}^{m\times m}$ be its filtered and re-normalized attention matrix and $D \in \mathbb{R}^{m\times m}$ the absolute positional distance matrix with $D_{ij} = |i - j|$. For each token i, we compute its normalized attention structural distance:

$$\mathrm{ASD}_i^{(k)} = \frac{\sum_{j=1}^{m_k} A_{ij}^{\prime(k)} \cdot |i-j|}{\sum_{j=1}^{m_k} A_{ij}^{\prime(k)} + \varepsilon}, \qquad \varepsilon = 10^{-8}.$$

The denominator accounts for possible numerical deviations introduced by masking and token filtering. Here, $A'^{(k)}$ denotes the aggregated attention matrix after removing special tokens and re-normalizing the remaining rows.

The sentence-level ASD is the mean over all tokens:

$$\mathrm{ASD}_{\text{sent}}^{(k)} = \frac{1}{m_k} \sum_{i=1}^{m_k} \mathrm{ASD}_i^{(k)}.$$

For a document comprising N sentences, we define its **Mean Attention Structural Distance** (MASD) as:

$$\mathrm{MASD} = \frac{1}{N} \sum_{k=1}^{N} \mathrm{ASD}_{\text{sent}}^{(k)}.$$

Overall, the calculation process for mean attention structural distance is shown in the first row of Figure 1.

MASD thus operationalizes structural compactness from a processing-oriented perspective: it reflects the average span over which linguistic relations must be maintained to interpret a text. Compared to traditional dependency distance, MASD does not rely on explicit parses and instead captures how structure is implicitly realized in model-internal representations.

3.1.3 Regression Analysis

We perform linear regression with year as the independent variable and MASD as the dependent variable. We fit the model $y = \beta_1 x + \beta_0$ and report the coefficient of determination (R^2) to quantify the proportion of variance in structural compactness explained by time.

To account for possible serial correlation across years, we further apply Newey–West HAC robust standard errors to the slope estimate.

3.2 Semantic Space Distance

3.2.1 Functional Semantic Grouping and Distance Measures

We construct a diachronic semantic analysis framework to investigate the temporal evolution of functional semantic categories. Given a document, we first segment it into sentences, yielding a sentence set $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$. For each sentence s, we perform part-of-speech (POS) tagging using spaCy and categorize its tokens into six functional semantic groups based on linguistic typology: Nouns (NN), Verbs (VB), Adjectives (JJ), Adverbs (RB), Prepositions (IN), and Determiners (DT). Let $\mathcal{G} = \{\text{NN}, \text{VB}, \text{JJ}, \text{RB}, \text{IN}, \text{DT}\}$ denote the set of functional groups. For any $g \in \mathcal{G}$, the set of tokens in sentence s belonging to group g is defined as:

$$\mathcal{T}_s^{(g)} = \{w \in s \mid \text{POS}(w) \in \text{TagSet}(g)\}, \quad (1)$$

where $\text{TagSet}(g)$ is the set of POS tags associated with functional group g (e.g., $\text{TagSet}(\text{NN}) = \{\texttt{NN}, \texttt{NNS}\}$). In implementation, only tokens that can be successfully aligned to model subwords are retained, while special tokens, punctuation, and whitespace are excluded from further computation.

To represent lexical semantics, we employ contextualized embeddings from the pre-trained model to obtain an embedding vector $\mathbf{e}(w) \in \mathbb{R}^d$ for each token w, where d is the model's hidden dimension. Tokens without aligned subwords are excluded from centroid computation. Based on this, we compute the semantic centroid (i.e., mean embedding) of each functional group within sentence s:

$$\boldsymbol{\mu}_s^{(g)} = \begin{cases} \dfrac{1}{|\mathcal{T}_s^{(g)}|} \displaystyle\sum_{w \in \mathcal{T}_s^{(g)}} \mathbf{e}(w), & \text{if } |\mathcal{T}_s^{(g)}| > 0, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \quad (2)$$

Furthermore, the semantic distance between any two functional groups g_i and g_j is defined as the Euclidean distance between their centroids:

$$d_s(g_i, g_j) = \left\| \boldsymbol{\mu}_s^{(g_i)} - \boldsymbol{\mu}_s^{(g_j)} \right\|_2. \quad (3)$$

This distance quantifies the degree of separation between distinct grammatical roles (e.g., nouns vs. verbs) in semantic space within a local context and serves as the core metric for measuring discriminability along functional semantic axes.

In addition to individual group-pair distances, we define the overall functional semantic dispersion of sentence s as the mean centroid distance across all valid functional group pairs:

$$D_s = \frac{1}{|\mathcal{P}_s|} \sum_{(g_i, g_j) \in \mathcal{P}_s} d_s(g_i, g_j), \quad (4)$$

where $\mathcal{P}_s = \{(g_i, g_j) \mid g_i < g_j,\ |\mathcal{T}_s^{(g_i)}| > 0,\ |\mathcal{T}_s^{(g_j)}| > 0\}$ denotes the set of all unordered functional group pairs both present in sentence s.

3.2.2 Diachronic Variation in Functional Semantic Distances

For diachronic analysis, we aggregate sentence-level distances by year. Let $\mathcal{S}_y$ denote the set of all sentences from year y. The annual average semantic distance for functional group pair (g_i, g_j) is then:

$$\bar{d}_y(g_i, g_j) = \frac{1}{|\mathcal{S}_y^{(g_i, g_j)}|} \sum_{s \in \mathcal{S}_y^{(g_i, g_j)}} d_s(g_i, g_j), \quad (5)$$

where $\mathcal{S}_y^{(g_i, g_j)} \subseteq \mathcal{S}_y$ is the subset of sentences satisfying $|\mathcal{T}_s^{(g_i)}| > 0$ and $|\mathcal{T}_s^{(g_j)}| > 0$. We compute the yearly mean for each POS-pair distance and then further derive an overall SSD trend by averaging across all valid POS-pair distances within each year.

To eliminate scale effects and enable cross-pair comparison, we standardize the annual distances for each functional group pair via Z-score normalization:

$$z_y(g_i, g_j) = \frac{\bar{d}_y(g_i, g_j) - \mu_{g_i g_j}}{\sigma_{g_i g_j}}, \quad (6)$$

where $\mu_{g_i g_j}$ and $\sigma_{g_i g_j}$ are the mean and standard deviation of all valid $\bar{d}_y(g_i, g_j)$ values across the entire time span.

We apply the Mann–Kendall trend test to the time series $\{\bar{d}_y(g_i, g_j)\}_{y=y_{\min}}^{y_{\max}}$ to assess monotonicity and estimate a robust trend slope $\hat{\beta}$ with a 95% confidence interval using the Theil-Sen estimator.

3.2.3 Correlation between Syntactic Compactness and Semantic Distance

Additionally, we introduce mean dependency distance as a proxy for syntactic compactness. For sentence s, it is defined as:

$$\delta_s = \frac{1}{|\mathcal{W}_s|} \sum_{w \in \mathcal{W}_s} \left| \text{pos}(w) - \text{pos}(\text{head}(w)) \right|, \quad (7)$$

where $\mathcal{W}_s$ denotes the set of all non-root tokens in sentence s, $\text{pos}(w)$ is the linear index of token w in the sentence, and $\text{head}(w)$ is its syntactic head.

We compute the Spearman rank correlation coefficient ρ between δ_s and selected functional semantic space distances (e.g., $d_s(\mathrm{NN}, \mathrm{VB})$), as well as between δ_s and the overall sentence-level dispersion D_s, to examine the association between semantic differentiation and syntactic integration.

Moreover, we align the annually averaged overall functional semantic dispersion

$$\bar{D}_y = \frac{1}{|\mathcal{S}_y|} \sum_{s \in \mathcal{S}_y} D_s \tag{8}$$

with an externally provided attention-based structural distance a_y (computed independently from the same corpus) over shared years. Their cross-year association is quantified by Spearman correlation:

$$\rho_{\text{text-struct}} = \mathrm{Corr}\Big(\{\bar{D}_y\}_{y \in \mathcal{Y}_{\text{common}}}, \{a_y\}_{y \in \mathcal{Y}_{\text{common}}}\Big), \tag{9}$$

where Corr is the Spearman correlation, $\mathcal{Y}_{\text{common}}$ is the set of years common to both data sources. Overall, the analysis workflow based on the semantic space distance is shown in the second row of Figure 1.

This analysis aims to reveal the co-evolutionary relationship between macro-level discourse structure and micro-level functional semantic differentiation.

This methodological framework not only quantifies the diachronic divergence along functional semantic axes but also provides computational evidence that the increased semantic discriminability helps to reduce role-mapping ambiguity through multi-dimensional correlation analyses.

4 Experiments

4.1 Data Acquisition and Preparation

To ensure both diachronic continuity and stylistic comparability, we construct a curated corpus of formal texts spanning multiple centuries, languages, and institutional registers. Rather than maximizing genre diversity, our data design prioritizes register stability, which is crucial for isolating language-internal evolutionary trends from confounding stylistic variation.

Specifically, we integrate five institutionally grounded text collections, covering both SVO and SOV languages while maintaining relatively formal and constrained communicative settings. The Royal Society Corpus (RSC) consists of English scientific articles from the 17th to the 20th centuries and represents a form of early modern academic prose (Fischer et al., 2020). American Presidential Speeches and State of the Union speeches provide formal political discourse in English. Japan General Policy Speeches extend the dataset to SOV languages in comparable institutional contexts. Annual Christmas messages by Western monarchs and New Year's editorials of China's Xinhua News Agency further supplement the corpus with additional forms of publicly delivered, institutionally produced texts.

All texts are obtained from authoritative official sources, which cover formal registers such as science, governance, ritual communication, and state media. Notably, we use these materials as parallel instances of relatively stable, formal registers across languages and time periods, rather than assuming that any single source is broadly representative. This makes our corpus highly suitable for diachronic comparison, since the observed linguistic variations reflect internal structural differences rather than changes in genre or communicative purpose.

After data cleaning, the resulting formal-text corpus comprises approximately 70 million tokens, drawing from multiple publicly available and institutionally curated sources with a high degree of register consistency.

Moreover, the corpus span over 200 continuous years, encompassing at least six to eight generations of language users, which may leave discernible and statistically detectable traces on language structures.

Therefore, the corpus forms a robust empirical foundation for cross-linguistic and cross-temporal analysis of language efficiency.

Table 1: Token counts of the datasets used in our experiments.

Dataset	Tokens
The Royal Society Corpus	63237372
State of Union Addresses	1764432
American Presidential Speeches	1961286
New Year Editorial of Xinhua Daily	666357
Japan General Policy Speeches	765307

4.2 Attention-based Structural Distance Analysis

We begin by investigating whether syntactic organization exhibits systematic diachronic compression

when viewed through the internal representations of pretrained language models. To this end, we employ the proposed Attention-based Structural Distance (ASD), an attention-based metric that captures the average span of model-internal attentional dependencies. Lower ASD values indicate that attention is more locally concentrated, corresponding to more compact structures.

We first focus on bidirectional attention models, such as BERT, because their fully bidirectional attention mechanism can completely reveal the allocation of attention across contexts, thereby directly modeling global structural dependencies, which is crucial for tracing linguistic trends.

We then turn to causal attention models to better simulate incremental and predictive language processing, as their constrained context aligns with the real-time, left-to-right nature of auditory language comprehension. We include the Qwen family to test whether the observed diachronic trends hold under unidirectional attention.

We apply the procedure to four diachronic corpora: the Royal Society Corpus, American Presidential Inaugural Addresses, State of the Union speeches, and the New Year Editorials of the Xinhua Daily. All experiments are conducted on NVIDIA A100 GPUs. Processing time varies with corpus size, ranging from under one hour for smaller datasets to approximately two hours for the full RSC.

Across all corpora, MASD exhibits a clear downward trajectory over time, as illustrated in Figure 2, 5, and 6. This pattern indicates that syntactic structure becomes increasingly compact in the model's internal representation. Importantly, the trend is consistent across both bidirectional and unidirectional architectures, suggesting that the observed compression is not an artifact of a particular modeling assumption but a stable property of the underlying textual data.

To eliminate sentence length as a potential confounding factor, we selected two SVO corpora (the State of the Union Addresses and New Year Editorial of Xinhua Daily) and one SOV corpus (Japan General Policy Speeches) for our empirical investigation. Specifically, we partitioned sentences into bins based on length using a step size of 5 tokens. For the State of the Union and New Year Editorial of Xinhua Daily corpora, bins were defined over the interval [0,50] (i.e., [0,5], [5,10], . . . , [45,50]), which covers close to 90% of all sentences. For the Japanese corpus, we used the same bin width and extended the range to [0,70] to accommodate its longer sentence distribution. Within each bin, residual variation in sentence length is limited. We also applied Newey–West HAC standard errors (max lag = 4) to all regression analyses to address temporal auto-correlation.

The results indicate that for MASD, the negative trend reflecting syntactic simplification remains statistically significant ($p < 0.0001$) across nearly all sentence-length bins in the State of the Union Addresses, especially within the most prevalent range of 5 to 45 words. The only clear exception occurs in the sparsely populated shortest bin ([0,5]), where the trend is not significant. In the New Year Editorial of Xinhua Daily corpus, the downward trend in MASD is also pronounced and statistically significant for the vast majority of bins from 5 to 50 tokens, with the steepest declines observed in longer sentences such as the [40,45] and [45,50] ranges. Again, the shortest bin ([0,5]) exhibits a non-significant slope, consistent with the limited data and minimal syntactic complexity in that range. In the Japanese corpus, a similar monotonic pattern is also observed. Although weaker and less uniform in shorter and mid-length bins, significantly negative slopes still emerge in the longer-sentence ranges, particularly from 40 tokens onward. This suggests that the long-run reduction in structural distance is robustly preserved in more syntactically elaborate sentences. Detailed estimates for each bin are reported in Table 3, 4, and 5.

The decline in MASD aligns with earlier findings on dependency distance minimization and stylistic simplification in modern written language. However, unlike traditional surface-level metrics, MASD captures this change directly from learned attentional structure, without explicit syntactic annotation. This demonstrates that diachronic structural compression is not only reflected in observable syntax but also captured by the latent processing geometry induced by large language models.

4.3 Semantic Space Distance Analysis

To examine whether the observed patterns extend beyond a single corpus and reflect a more general diachronic tendency, we replicate the core analysis across the five historical datasets following an identical processing pipeline.

We obtain contextualized embeddings using two variants of the Qwen embedding model series—Qwen-Embedding-0.6B and Qwen-Embedding-4B. For each sentence, compute the

Table 2: Cross-corpus Theil-Sen Slopes (Z-scores) and p-values of Syntactic Pairs Based on Qwen-Embedding-0.6B.

	NN-VB		NN-JJ		VB-JJ	
	Slope	**p**	**Slope**	**p**	**Slope**	**p**
Royal Society Corpus	0.000069	0.7177	0.0006	< 0.0001	0.002083	< 0.0001
American Presidential Speeches	0.004947	< 0.0001	0.004780	< 0.0001	0.004390	< 0.0001
State of Union Addresses	0.004856	< 0.0001	0.004680	< 0.0001	0.004305	< 0.0001

semantic distance between POS group centroids, the average syntactic dependency distance, and the MASD derived from model attentions. All experiments are run on NVIDIA A100 GPUs. While most datasets require approximately three hours to process, the substantially larger Royal Society Corpus (RSC) takes about 45 hours to complete.

As shown in Figure 3, 8, 9, 10, 11, 12, 13, 14, 15 and 16, the semantic space distance between word classes increases significantly over time. This trend remains pronounced across specific part-of-speech pairings—such as noun-verb (NN-VB), noun-adjective (NN-JJ), and verb-adjective (VB-JJ)—whose correlation values are detailed in Table 2. This convergence across genres and historical periods suggests that the phenomenon is not corpus-specific, but reflects a stable trajectory of linguistic change.

Notably, this analysis involved multiple hypothesis tests. While corrections for multiple comparisons (e.g., Bonferroni) were not applied in the initial computation, the inferential validity of the observed trends remains robust for two primary reasons. First, the objective is not isolated discovery but the corroboration of a single theoretical prediction across distinct POS pairs and datasets. Second, the reported p-values fall predominantly below 0.0001, a threshold that retains statistical significance even under the most conservative adjustments.

We also plot the correlation between MASD and SSD. As shown in Figure 4, 20, 21 and 22, a significant negative correlation is observed, indicating a co-evolutionary relationship between the two measures.

For SSD, the analysis controlling for sentence length followed the same binning and estimation procedure described for MASD. In the State of the Union Addresses, Theil–Sen estimates indicate a positive trend that is statistically significant at the $p < 0.05$ level across all bins spanning 5 to 45 words, while the longest bin ([45,50]) yields no significant result. In the New Year Editorial of

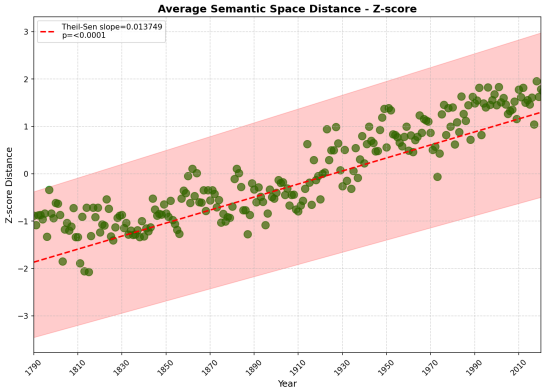

Figure 3: Temporal variation in semantic space distance in State of the Union Addresses and significance testing based on Qwen-Embedding-4B.

Xinhua Daily corpus, the positive trend in SSD is even more robust, with Theil–Sen slopes achieving statistical significance at the $p < 0.0001$ level across all bins from 5 to 45 tokens. The only exception is the [15,20] bin, where the positive slope does not reach significance. In Japan General Policy Speeches, positive slopes are also observed in most bins, with statistically significant increases concentrated in short, medium, and some long sentence ranges, although several bins do not reach significance and a small number show near-zero or negative estimates. Overall, the results remain consistent with an increase in semantic separability over time, albeit with greater heterogeneity than in the English corpus. Full bin-level estimates are provided in Table 3, 4, and 5.

This systematic expansion of POS-level semantic separation indicates a gradual increase in functional discriminability in lexical representations. From a psycholinguistic perspective, such a trend aligns closely with decades of evidence showing that human language comprehension is fundamentally predictive in nature. Neurocognitive studies have consistently demonstrated that the brain actively anticipates upcoming linguistic input at multiple representational levels, including lexical form, syntactic category, and semantic content (Wicha et al., 2004; DeLong et al., 2005; Van Berkum et al.,

2005; Federmeier, 2007; Dambacher et al., 2009; Laszlo and Federmeier, 2009; Dikker et al., 2010; Dikker and Pylkkänen, 2013; Lau et al., 2016; Willems et al., 2015). Prediction is now widely regarded as a core organizing principle of language processing (Van Petten and Luka, 2012; Hagoort and Indefrey, 2014; Huettig, 2015).

Within such a predictive system, clearer separation between functional categories directly reduces uncertainty. When semantic representations of different parts of speech overlap substantially, for example, in cases of a word used as different POS in one sentence, listeners must rely more heavily on extended context to resolve grammatical roles. By contrast, greater semantic divergence between POS categories narrows the hypothesis space of upcoming words, facilitating faster and more reliable online interpretation. From this viewpoint, the observed diachronic increase in POS group embedding distance can be interpreted as a gradual reduction in cognitive processing cost, consistent with the long-standing efficiency-based principle (Zipf, 1949).

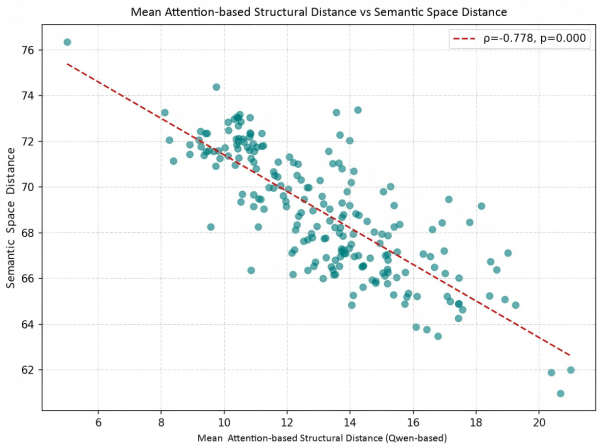

Figure 4: Trends in MASD and SSD on RSC data (Qwen3-Embedding-0.6B).

Taken together, our findings align with the long-standing hypothesis of syntactic compression and functional–semantic divergence. Importantly, our results bridge human cognitive processing and neural language models. Just as predictive mechanisms in the human brain benefit from shorter integration spans and clearer functional semantic discriminability, attention-based language models allocate processing resources more efficiently when dependencies are local and functional categories are well separated in representation space. The convergence of trends observed across human language data and model-internal representations thus provides converging computational and cognitive evidence for the existing efficiency-driven hypothesis.

5 Conclusion

This paper presents an annotation-free framework for diachronic linguistic analysis based on pre-trained language models. We introduce Attention-based Structural Distance (ASD) as a parser-independent measure of syntactic locality derived from attention patterns, together with Semantic Space Distance (SSD), which quantifies functional–semantic separation between parts of speech in embedding space.

Across multiple multilingual and diachronic corpora, we observed a consistent decrease in ASD alongside an increase in SSD. This co-evolutionary pattern suggests that the language change involves not only structural compression but also an increased functional–semantic discriminability. The negative association between these two dimensions aligns with efficiency-driven principles of language evolution, in which reduced integration cost is accompanied by clearer functional differentiation.

Overall, the consistency of these observed trends across diverse model architectures demonstrates that they stem from the inherent properties of the linguistic data, rather than from artifacts unique to specific models.

From a linguistic perspective, our results can therefore be integrated into the existing theoretical framework, limited typological samples though. Meanwhile, our study also demonstrates the potential of language models as scalable tools for studying long-term language change beyond annotated resources, and the cognitive foundation underlying our metrics gives us confidence in its applicability to more languages.

Also, our study finds that attention-based and semantic space modeling share a similar cognitive principle with the human mind, both maximizing language processing capacity under limited resources. Such a common characteristic implies the potential of employing LLMs in the study of human cognition in the future.

Limitations

Although this study has made certain progress in exploring language evolution patterns and analyzing part-of-speech attention through attention mechanisms, the spatiotemporal coverage and typological

diversity of research samples remain significantly constrained due to the fragmented and incomplete nature of historical corpus data. Furthermore, while observed attention patterns are suggestive of cognitive mechanisms, their correspondence to specific cognitive processes remains unresolved across models with distinct attention architectures.

To overcome these limitations, future research will focus on typological analysis and establish a multidimensional language evolution framework, including the construction of parallel diachronic cross-linguistic and cross-dialectal corpora, as well as hierarchical attention networks specifically designed for diachronic linguistic analysis. Additionally, we will further compare how different attention mechanisms reflect varied cognitive aspects, and clarify the interpretative links between model-internal computations and human language processing.

Ethical Considerations

All experiments in this study were conducted using publicly available datasets and newly created datasets based on official public facts. These datasets contain neither offensive content nor information with negative societal implications. The objective of this research is to investigate the objective laws of linguistics through the integration of artificial intelligence technologies. We hereby confirm that this study fully complies with ethical review guidelines.

Acknowledgements

We are deeply grateful to the anonymous reviewers for their careful reading of our work and for offering such thoughtful and constructive feedback. Their suggestions truly helped us strengthen the manuscript. The second author acknowledges support from the Xiaomi Foundation.

References

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Michael Dambacher, Martin Rolfs, Karsten Göllner, Reinhold Kliegl, and Arthur M. Jacobs. 2009. Event-related potentials reveal rapid verification of predicted visual input. *PLoS ONE*, 4(3):e5047.

Katherine A. DeLong, Thomas P. Urbach, and Marta Kutas. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8):1117–1121.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

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 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Suzanne Dikker and Liina Pylkkänen. 2013. Predicting language: MEG evidence for lexical preactivation. *Brain and Language*, 127(1):55–64.

Suzanne Dikker, Hugh Rabagliati, Thomas A. Farmer, and Liina Pylkkänen. 2010. Early occipital sensitivity to syntactic category is based on form typicality. *Psychological Science*, 21(5):629–634.

Kara D. Federmeier. 2007. Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, 44(4):491–505.

Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. The Royal Society Corpus 6.0: Providing 300+ years of scientific writing for humanistic study. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 794–802, Marseille, France. European Language Resources Association.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287.

Peizhu Gong, Jin Liu, Yurong Xie, Minjie Liu, and Xiliang Zhang. 2023. Enhancing context representations with part-of-speech information and neighboring signals for question classification. *Complex Intell. Syst.*, 9:6191–6209.

Peter Hagoort and Peter Indefrey. 2014. The neurobiology of language beyond single words. *Annual Review of Neuroscience*, 37:347–362.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do attention heads in bert track syntactic dependencies? *ArXiv*, abs/1911.12246.

Falk Huettig. 2015. Four central questions about prediction in language processing. *Brain Research*, 1626:118–135. Predictive and Attentive Processing in Perception and Action.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. 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 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2016. Redefining part-of-speech classes with distributional semantic models. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 115–125, Berlin, Germany. Association for Computational Linguistics.

Sara Laszlo and Kara D. Federmeier. 2009. A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language*, 61(3):326–338.

Ellen F. Lau, Kirsten Weber, Alexandre Gramfort, Matti S. Hämäläinen, and Gina R. Kuperberg. 2016. Spatiotemporal signatures of lexical-semantic prediction. *Cerebral Cortex*, 26(4):1377–1387.

Lei Lei and Matthew L. Jockers. 2020. Normalized dependency distance: Proposing a new measure. *Journal of Quantitative Linguistics*, 27(1):62–79.

Bowen Li, Taeuk Kim, Reinald Kim Amplayo, and Frank Keller. 2020. Heads-up! unsupervised constituency parsing via self-attention heads. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 409–424, Suzhou, China. Association for Computational Linguistics.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *The Journal of Cognitive Science*, 9:159–191.

Haitao Liu, Chunshan Xu, and Junying Liang. 2016. Dependency length minimization: Puzzles and promises. *Glottometrics*, 33:35–38.

W. Liu, M. Xiang, and N. Ding. 2025. Active use of latent tree-structured sentence representation in humans and large language models. *Nature Human Behaviour*.

Akira Miyake and Priti Shah, editors. 1999. *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Cambridge University Press, Cambridge.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Sabine Schulte im Walde and Diego Frassinelli. 2022. Distributional measures of semantic abstraction. *Frontiers in Artificial Intelligence*, Volume 4 - 2021.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Jos J. A. Van Berkum, Colin M. Brown, Pienie Zwitserlood, Vibeke Kooijman, and Peter Hagoort. 2005. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3):443–467.

Cyma Van Petten and Barbara J. Luka. 2012. Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, 83(2):176–190.

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.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

C. Westbury and G. Hollis. 2019. Conceptualizing syntactic categories as semantic categories: Unifying part-of-speech identification and semantics using co-occurrence vector averaging. *Behav Res*, 51:1371–1398.

Nicole Y. Y. Wicha, Eva M. Moreno, and Marta Kutas. 2004. Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in spanish sentence reading. *Journal of Cognitive Neuroscience*, 16(7):1272–1288.

Roel M. Willems, Stefan L. Frank, Annabel D. Nijhof, Peter Hagoort, and Antal van den Bosch. 2015. Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516.

Victor H. Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444–466.

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, Boston.

A More Experimental Results

Table 3: Controlling for sentence length on regression results of MASD and SSD across length bins on State of the Union addresses with Qwen3-4B and Qwen3-Embedding-4B.

Length Bin	**MASD**			**SSD**		
	Slope	p	R^2	**Theil-Sen Slope**	p	R^2
[0,5]	-0.00342	0.1651	0.010	0.008142	0.0244	0.1881
[5,10]	-0.00221	< 0.0001	0.097	0.004132	0.0254	0.0422
[10,15]	-0.00186	< 0.0001	0.100	0.012622	< 0.0001	0.3888
[15,20]	-0.00192	< 0.0001	0.144	0.013478	< 0.0001	0.6029
[20,25]	-0.00346	< 0.0001	0.259	0.013846	< 0.0001	0.6714
[25,30]	-0.00430	< 0.0001	0.298	0.012197	< 0.0001	0.5660
[30,35]	-0.00540	< 0.0001	0.342	0.010925	< 0.0001	0.4538
[35,40]	-0.00622	< 0.0001	0.306	0.009420	< 0.0001	0.4201
[40,45]	-0.00639	< 0.0001	0.193	0.006817	< 0.0001	0.2873
[45,50]	-0.00445	0.0428	0.064	0.004449	0.1137	0.1631

Table 4: Controlling for sentence length on regression results of MASD and SSD across length bins on New Year Editorial of Xinhua Daily with Qwen3-4B and Qwen3-Embedding-4B.

Length Bin	**MASD**			**SSD**		
	Slope	p	R^2	**Theil-Sen Slope**	p	R^2
[0,5]	-0.001379	0.6746	0.0053	0.257137	0.0080	0.1365
[5,10]	-0.010494	< 0.0001	0.4522	0.117528	< 0.0001	0.4685
[10,15]	-0.006486	0.01586	0.1669	0.050608	< 0.0001	0.2347
[15,20]	-0.011298	0.01879	0.2034	0.033882	0.1437	0.0870
[20,25]	-0.013595	0.02548	0.1606	0.084742	< 0.0001	0.4320
[25,30]	-0.016752	0.03139	0.1761	0.097623	< 0.0001	0.3552
[30,35]	-0.018389	0.05171	0.1255	0.122794	< 0.0001	0.4020
[35,40]	-0.036187	0.02072	0.2180	0.156061	< 0.0001	0.3172
[40,45]	-0.041223	0.00251	0.1494	0.145910	< 0.0001	0.2860
[45,50]	-0.038806	0.08975	0.0669	0.160182	0.0034	0.1197

Table 5: Controlling for sentence length on regression results of MASD and SSD across length bins on Japan General Policy Speeches with Qwen3-4B and Qwen3-Embedding-4B.

Length Bin	**MASD**			**SSD**		
	Slope	p	R^2	**Theil-Sen Slope**	p	R^2
[0,5]	-0.009000	0.2813	0.0344	1.669117	< 0.0001	0.5595
[5,10]	0.000169	0.9650	< 0.0001	0.126810	0.0175	0.1111
[10,15]	0.005439	0.2652	0.0279	0.064115	0.0261	0.0398
[15,20]	0.008770	0.0019	0.0891	-0.070973	0.0795	0.0803
[20,25]	-0.000176	0.9601	< 0.0001	0.049979	0.1436	0.0429
[25,30]	-0.001668	0.6027	0.0030	-0.028223	0.2899	0.0197
[30,35]	-0.008445	0.1199	0.0355	0.046680	0.1948	0.0340
[35,40]	-0.006416	0.6512	0.0053	0.006486	0.7903	0.0011
[40,45]	-0.050053	< 0.0001	0.2930	-0.038107	0.1367	0.0244
[45,50]	-0.079376	0.0021	0.3123	0.054221	0.0853	0.0552
[50,55]	-0.063707	0.0082	0.1706	0.075720	0.0393	0.0561
[55,60]	-0.110334	0.0009	0.2399	0.026792	0.4694	0.0060
[60,65]	-0.146010	< 0.0001	0.3356	-0.024806	0.4990	0.0051
[65,70]	-0.091559	0.0346	0.1318	0.003251	0.3797	0.0056

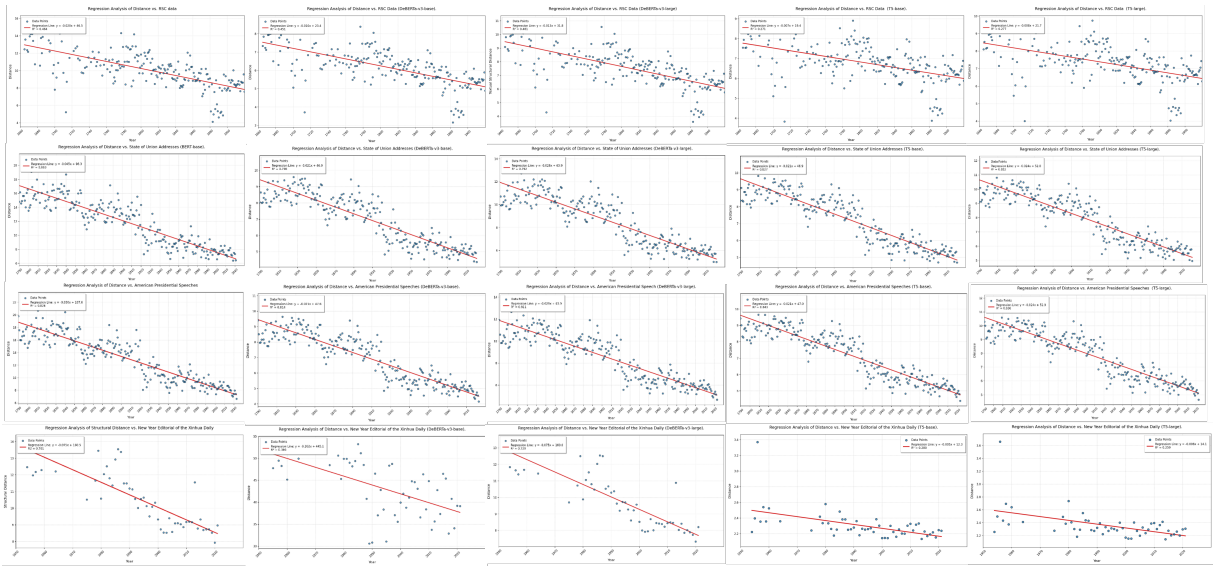

Figure 5: Experiments on Multiple Bidirectional Attention Models.

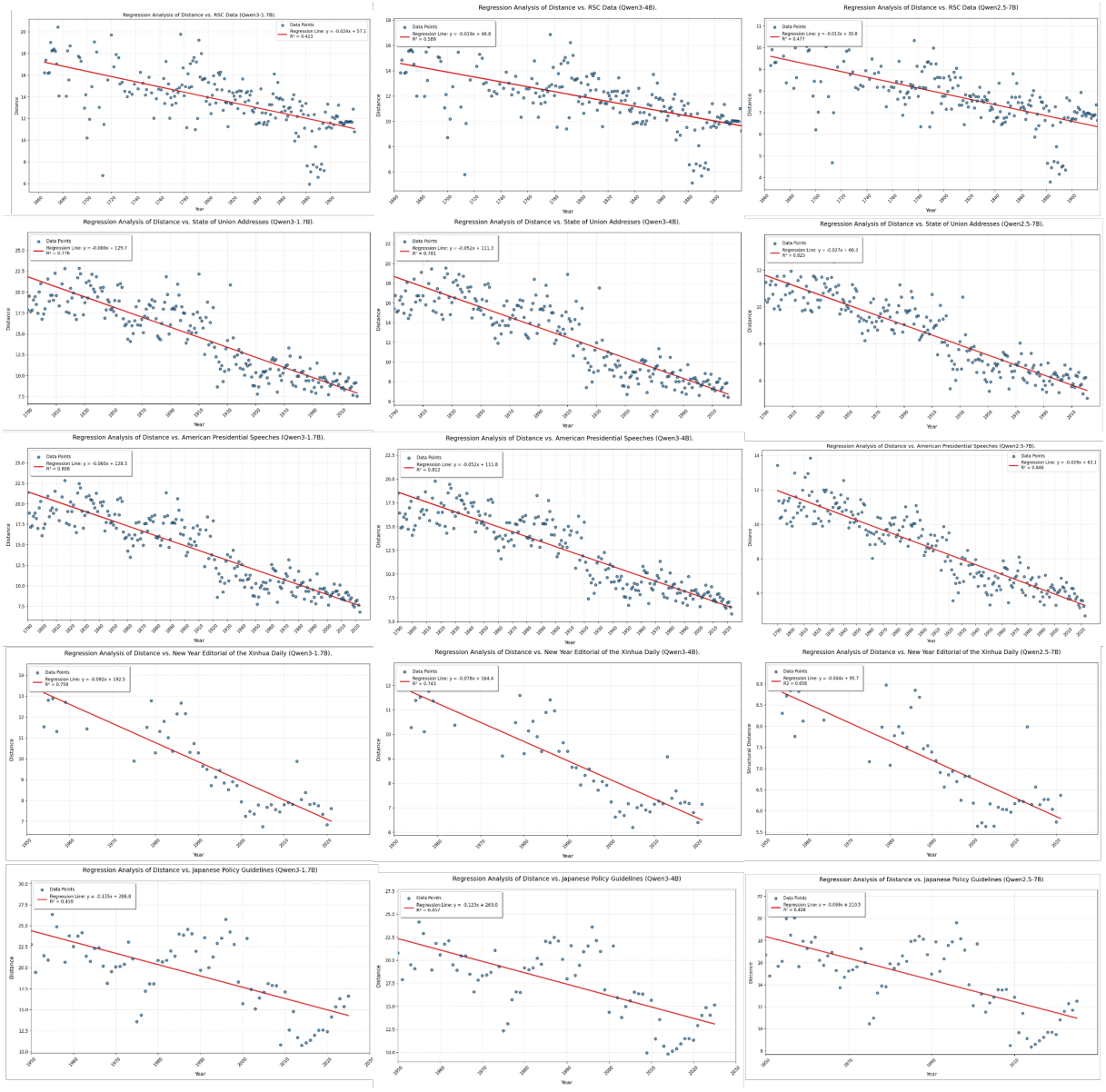

Figure 6: Experiments on Multiple Causal Attention Models.

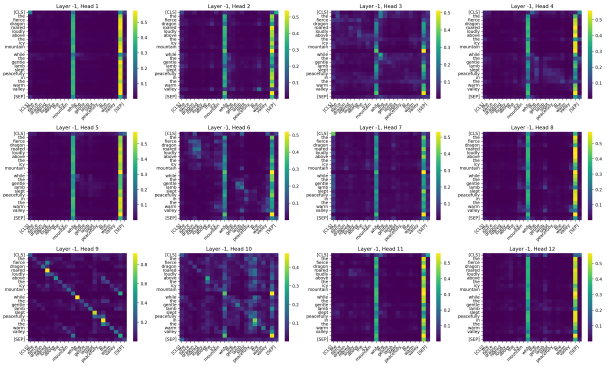

Figure 7: Schematic illustration of the attention distribution from the final layer of BERT on the example sentence.

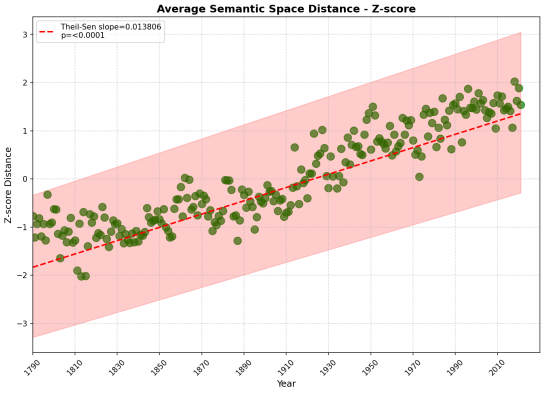

Figure 8: Temporal variation in semantic space distance in State of the Union Addresses and significance testing based on Qwen-Embedding-0.6B.

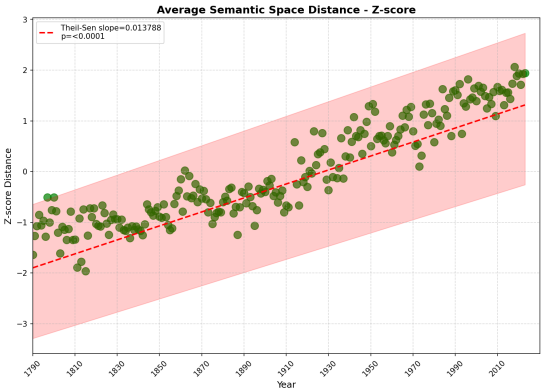

Figure 9: Temporal variation in semantic space distance in state of the American Presidential Speeches and significance testing based on Qwen-Embedding-0.6B.

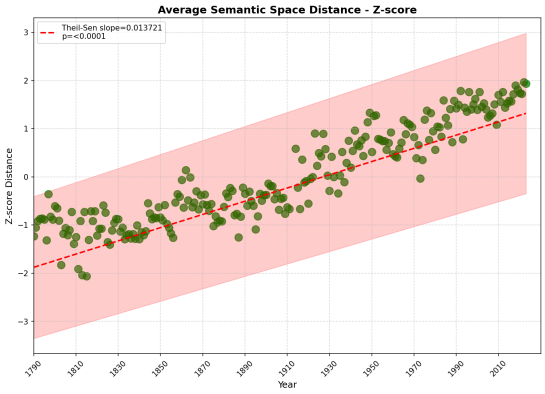

Figure 10: Temporal variation in semantic space distance in state of the American Presidential Speeches and significance testing based on Qwen-Embedding-4B.

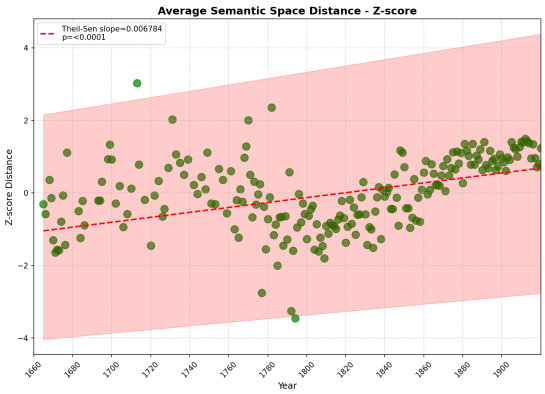

Figure 11: Temporal variation in semantic space distance in RSC data and significance testing based on Qwen-Embedding-0.6B.

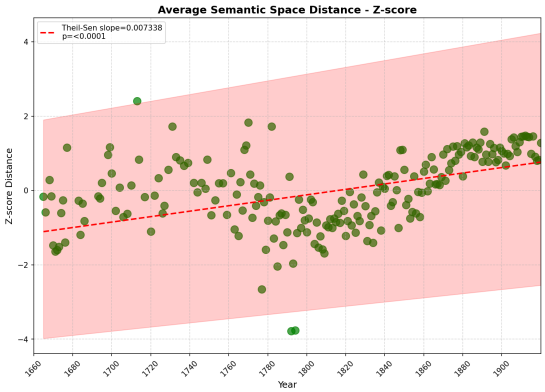

Figure 12: Temporal variation in semantic space distance in RSC data and significance testing based on Qwen-Embedding-4B.

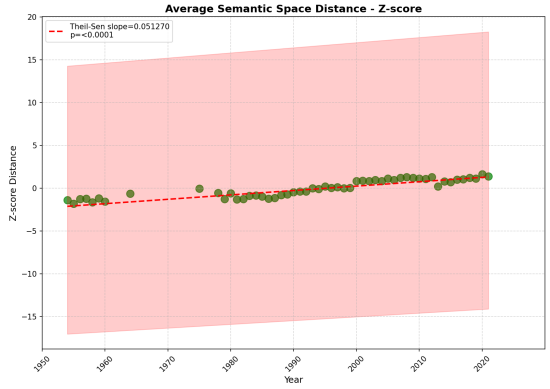

Figure 13: Temporal variation in semantic space distance in New Year Editorial of Xinhua Daily and significance testing based on Qwen-Embedding-0.6B.

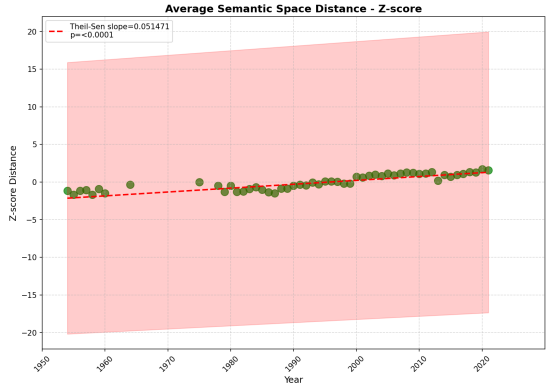

Figure 14: Temporal variation in semantic space distance in New Year Editorial of Xinhua Daily and significance testing based on Qwen-Embedding-4B.

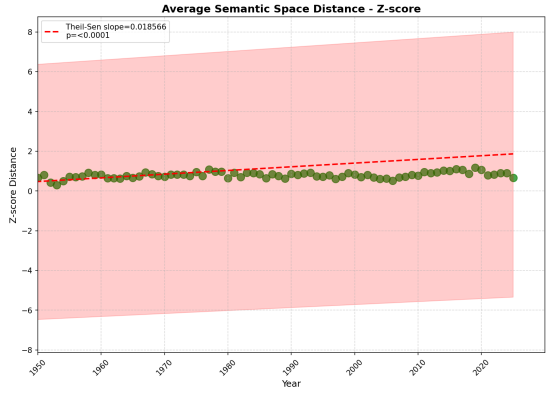

Figure 15: Temporal variation in semantic space distance in Japan General Policy Speeches and significance testing based on Qwen-Embedding-0.6B.

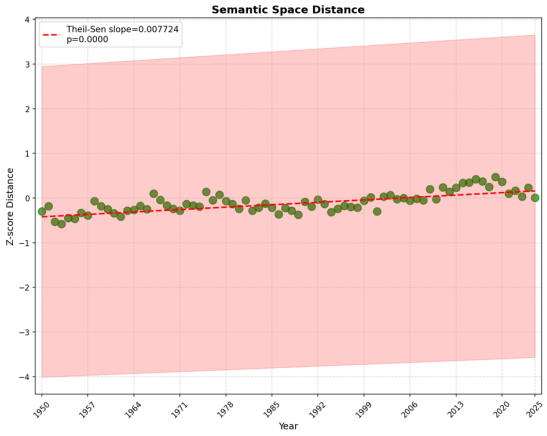

Figure 16: Temporal variation in semantic space distance in Japan General Policy Speeches and significance testing based on Qwen-Embedding-4B.

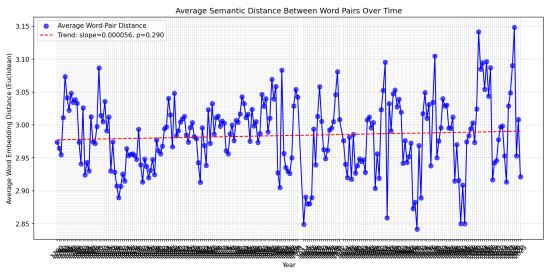

Figure 17: Diachronic changes in distances of all words (regardless of POS) computed on the American presidential speeches dataset exhibit substantial fluctuations with no clear trend.

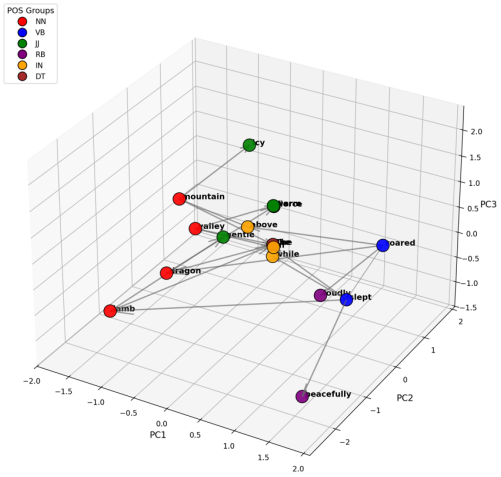

Figure 18: Visualization of semantic space distance and dependency distance in example sentences (Qwen3-Embedding-0.6B).

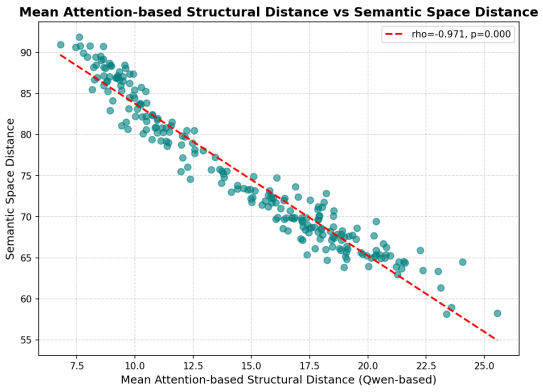

Figure 19: Trends in MASD and SSD on American presidential speeches (Qwen3-embedding-0.6B).

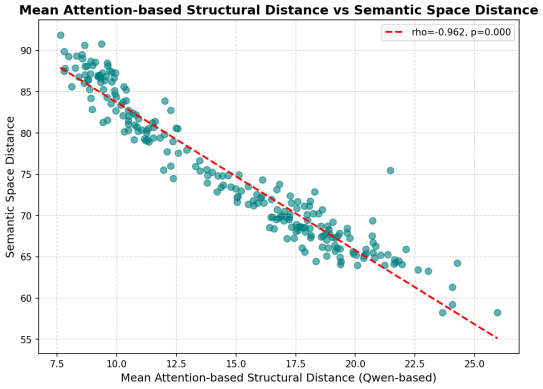

Figure 20: Trends in MASD and SSD on State of the Union Address (Qwen3-Embedding-0.6B).

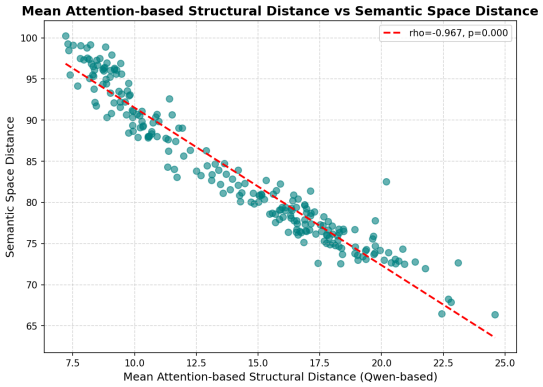

Figure 21: Trends in MASD and SSD on State of the Union Addresses (Qwen3-Embedding-4B).

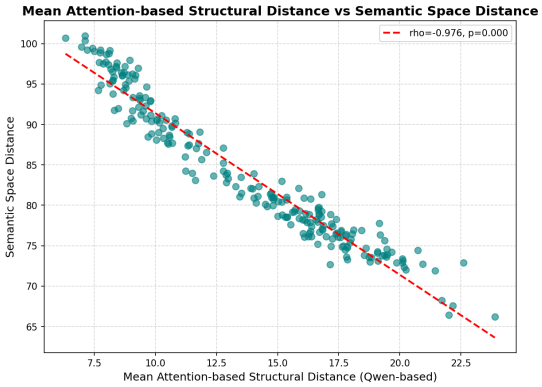

Figure 22: Trends in MASD and SSD on American presidential speeches over the years (Qwen3-Embedding-4B).

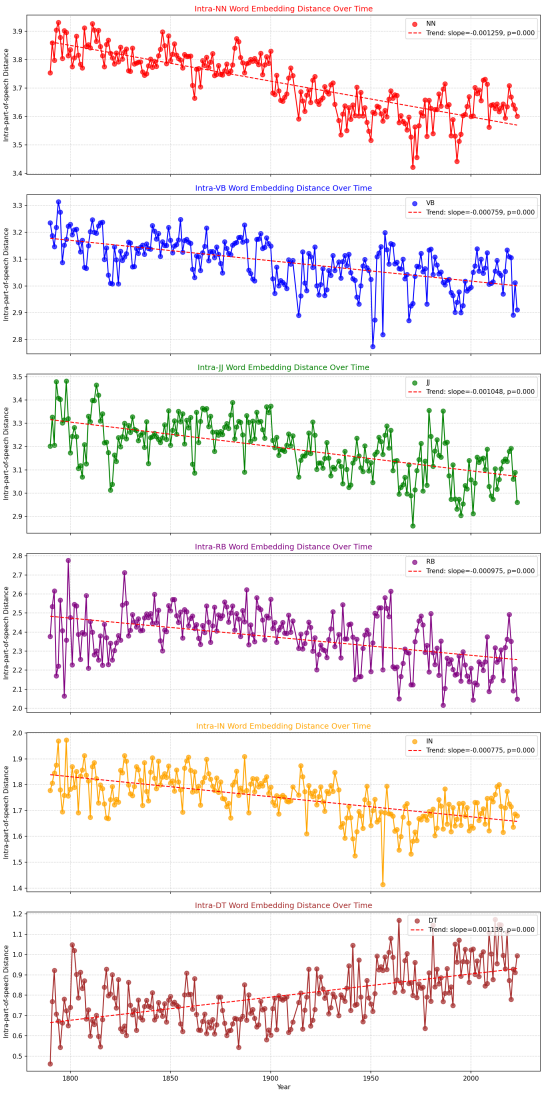

Figure 23: Diachronic changes in embedding distance among words of the same POS, computed on the American presidential speeches dataset, show a general decrease in within-POS group distances.

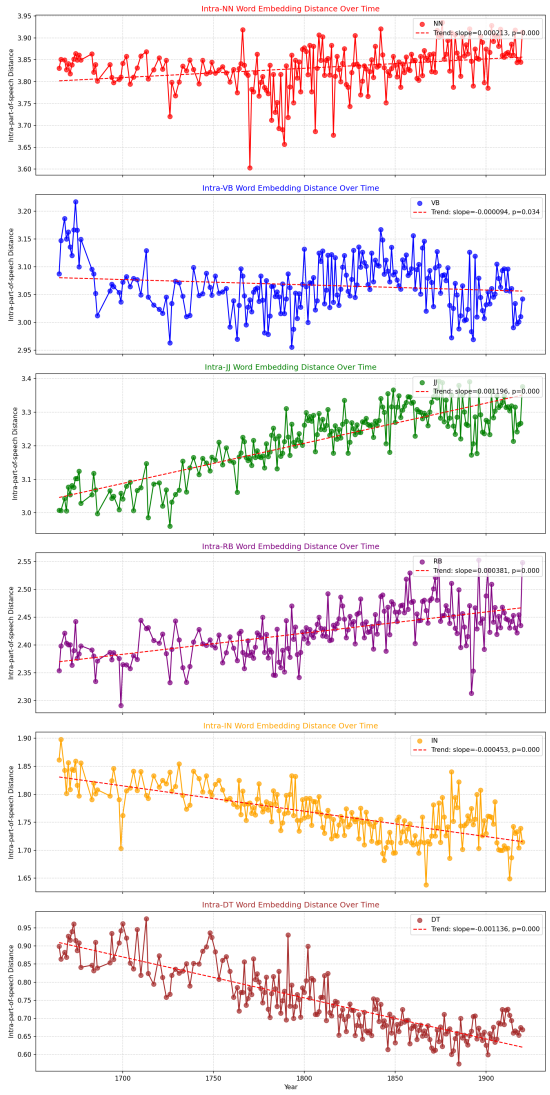

Figure 24: Diachronic changes in embedding distance among words of the same POS, computed on the RSC dataset, show that within-POS group distances increase for some parts-of-speech and decrease for others.

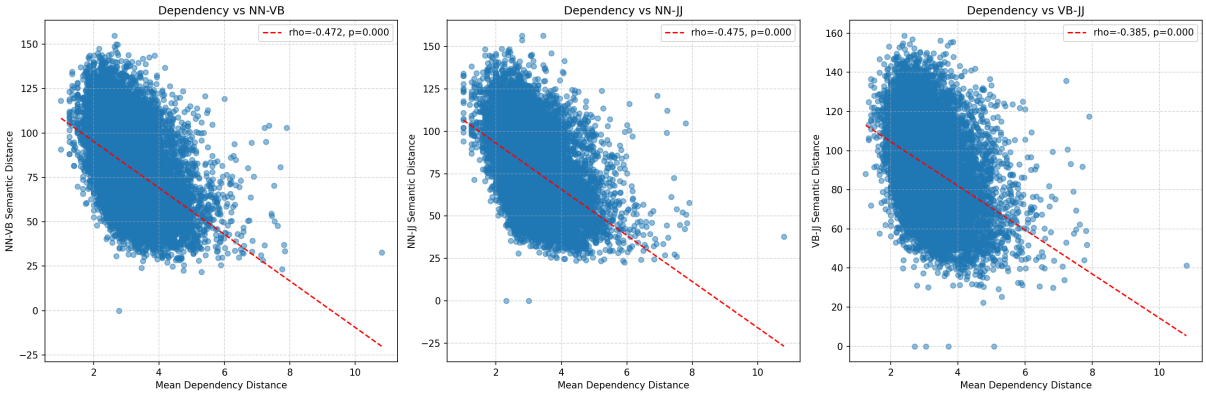

Figure 25: Trends in dependency distance and the main semantic space distance on state of the union addresses over the years (Qwen3-Embedding-0.6B) .

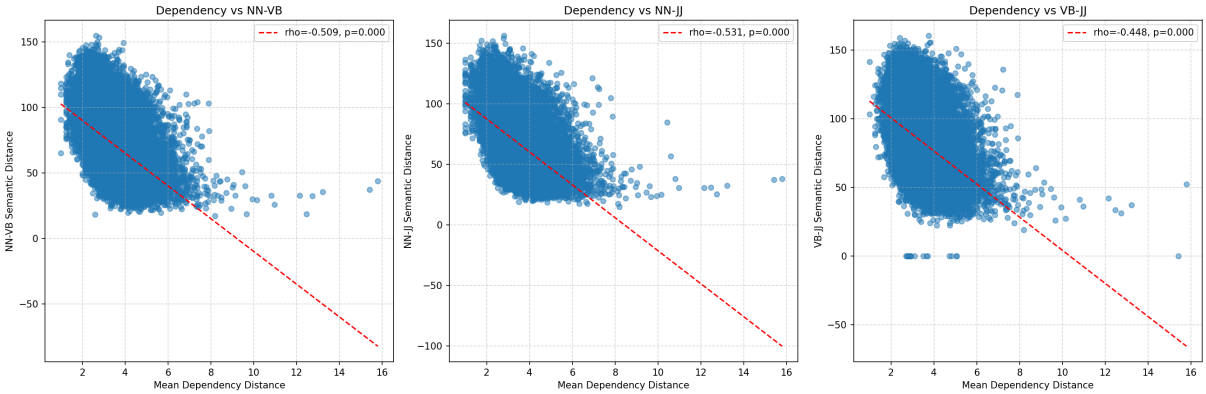

Figure 26: Trends in dependency distance and the main semantic space distance on American presidential speeches over the years (Qwen3-Embedding-0.6B).

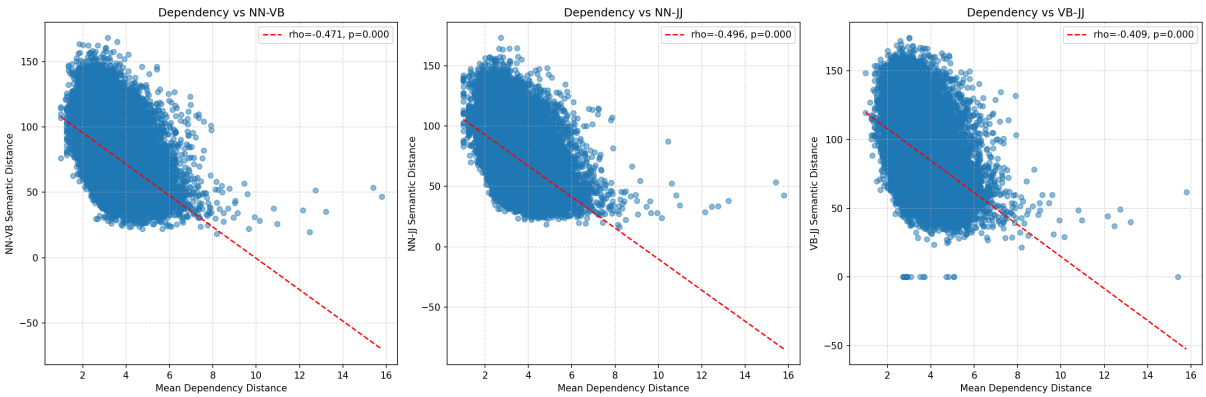

Figure 27: Trends in dependency distance and the main semantic space distance on American presidential speeches over the years (Qwen3-Embedding-4B).

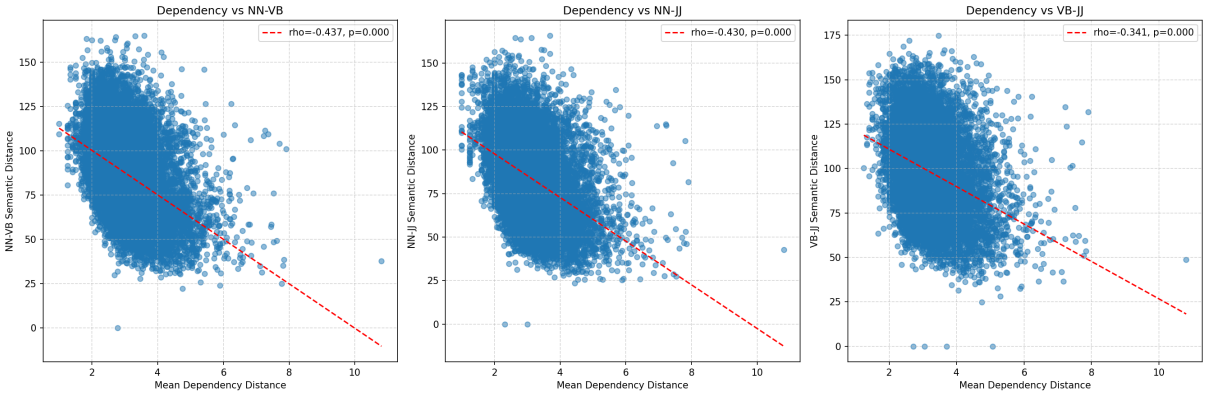

Figure 28: Trends in dependency distance and the main semantic space distance on State of Union Addresses (Qwen3-Embedding-4B).

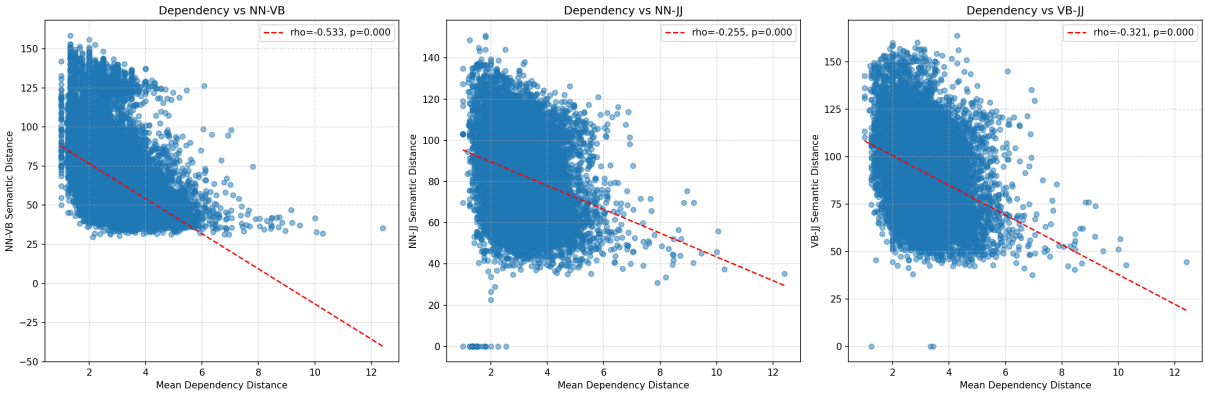

Figure 29: Trends in dependency distance and the main semantic space distance on New Year Editorial of Xinhua Daily (Qwen3-Embedding-0.6B).

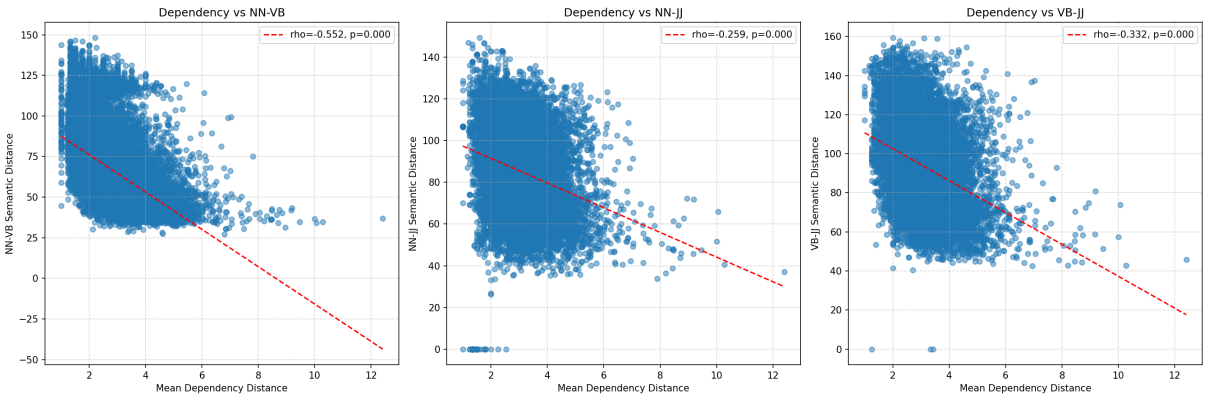

Figure 30: Trends in dependency distance and the main semantic space distance on New Year Editorial of Xinhua Daily (Qwen3-Embedding-4B).

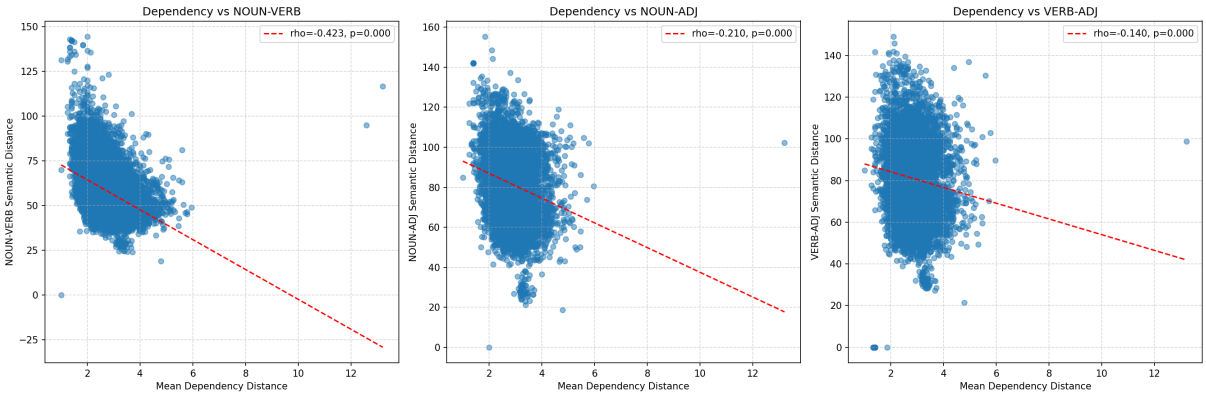

Figure 31: Trends in dependency distance and the main semantic space distance on Japan General Policy Speeches (Qwen3-Embedding-0.6B).

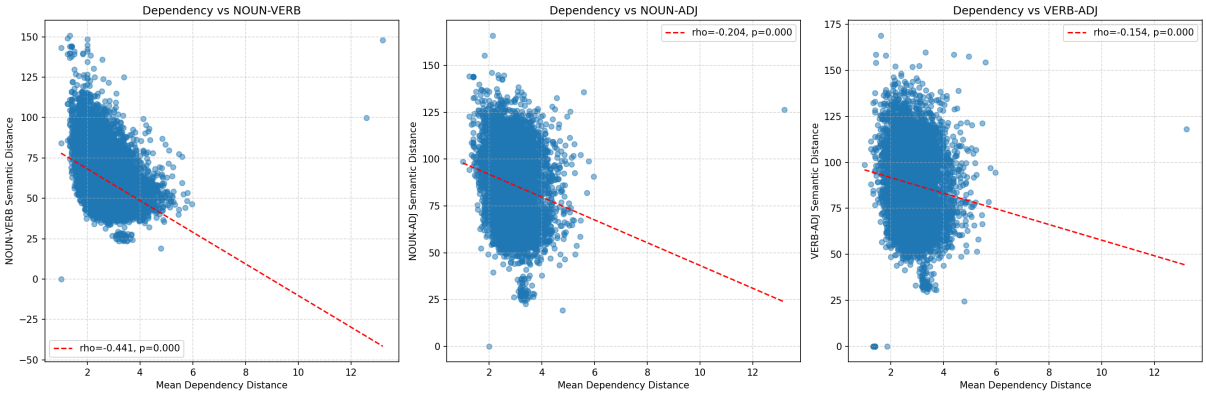

Figure 32: Trends in dependency distance and the main semantic space distance on Japan General Policy Speeches (Qwen3-Embedding-4B).

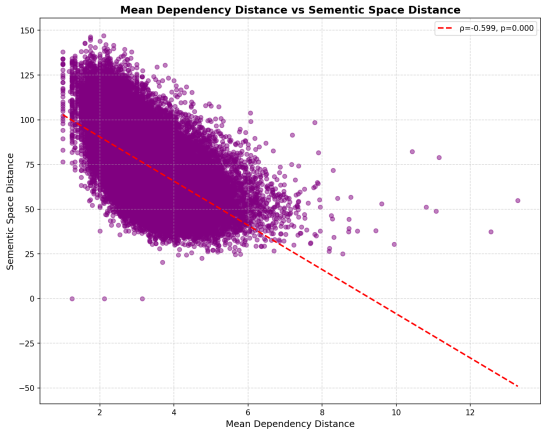

Figure 33: Trends in dependency distance and SSD on State of the Union Addresses (Qwen3-Embedding-0.6B).

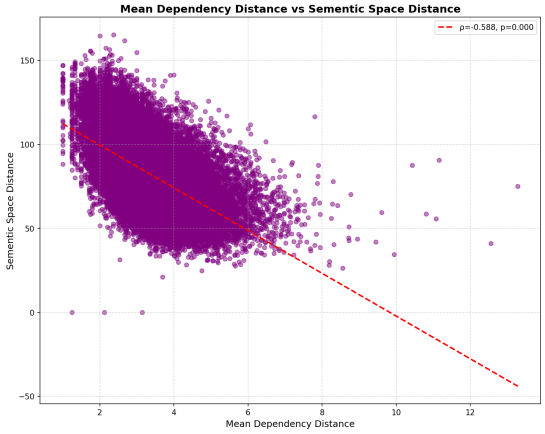

Figure 34: Trends in dependency distance and SSD on State of the Union Addresses (Qwen3-Embedding-4B).

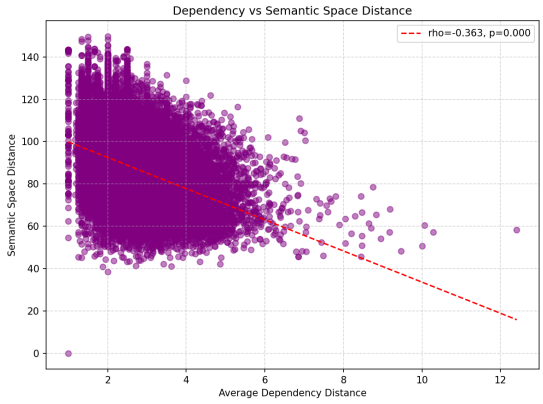

Figure 35: Trends in dependency distance and SSD on New Year Editorial of Xinhua Daily (Qwen3-Embedding-0.6B).

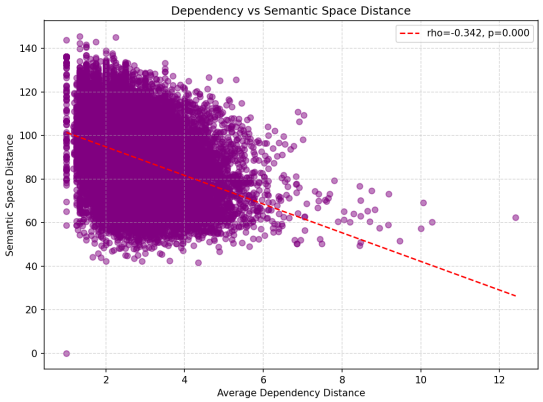

Figure 36: Trends in dependency distance and SSD on New Year Editorial of Xinhua Daily (Qwen3-Embedding-4B).

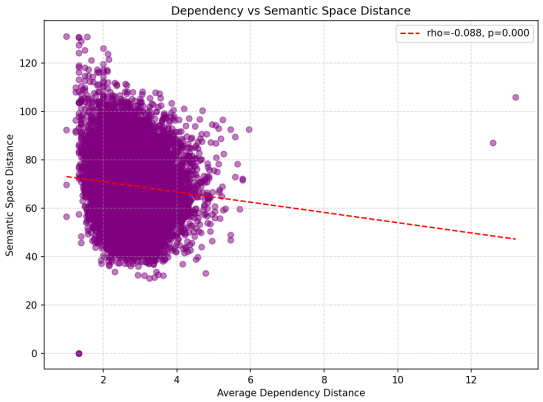

Figure 37: Trends in dependency distance and SSD on Japan General Policy Speeches (Qwen3-Embedding-0.6B).

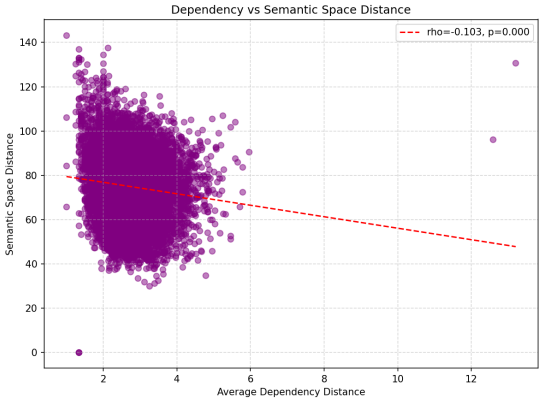

Figure 38: Trends in dependency distance and SSD on Japan General Policy Speeches (Qwen3-Embedding-4B).