Syntactic Analysis in Transformers through Attention Heads

Ricardo Gomes de Oliveira[1], **Daniela Barreiro Claro**[1], **Rerisson Cavalcante**[2]

[1]FORMAS - Research Center on Data and Natural Language
Institute of Computing – Federal University of Bahia (UFBA)

[2]FORMAS - Research Center on Data and Natural Language
Institute of Letters – Federal University of Bahia (UFBA)

{gomesricardo,dclaro,rerisson}@ufba.br

Abstract. *The advances in Natural Language Processing (NLP) have led to the development of Transformer architectures, such as BERT. One of the most prominent features of this architecture is the attention mechanism. However, the application of attention mechanisms to languages other than English, like Brazilian Portuguese, remains underexplored. This work analyzes the attention heads in a Transformer architecture, considering the syntactic relations in a sentence. We analyze how attention patterns align with syntactic dependencies involving phenomena such as transitive verbs, reflexive pronouns, and subordinate clauses, as realized in Brazilian Portuguese. Results described the existence of specialized heads within arcs, such as subject–verb and verb–object. These findings open up new research opportunities to evaluate syntactic sensitivity in Transformer models and to contribute to the development of more linguistically informed models for Brazilian Portuguese.*

1. Introduction

Transformer-based architectures have emerged as a cornerstone of contemporary NLP, primarily due to their effectiveness in modeling contextual dependencies via self-attention mechanisms [Vaswani et al. 2017]. BERT-based models [Devlin et al. 2019] have demonstrated the capacity of attention mechanisms to effectively capture both syntactic structures and semantic relationships within text. Empirical studies [Clark et al. 2019, Voita et al. 2019, Tenney et al. 2019] have shown that certain attention heads are specialized in capturing specific linguistic dependencies, such as subject–verb and coreference. These findings have grown the interest in learning the internal representations by Transformer models.

However, this behavior remains underexplored in languages such as Brazilian Portuguese. Even with adaptations (e.g., fine-tuning, LoRA [Melo et al. 2024]), the internal representation of syntax in languages requires further investigation. Multilingual models such as mBERT [Pires et al. 2019, Wu and Dredze 2020] offer generalization but often miss language-specific syntactic cues.

Even though other BERT-based models, such as ALBERTina [Rodrigues et al. 2023], have achieved high performance in Portuguese, we focus on BERTimbau [Souza et al. 2020] due to its pre-training exclusively on Brazilian Portuguese data, primarily sourced from the brWaC corpus. This linguistic specificity ensures closer alignment with the morphosyntactic structures of the target language. Furthermore, the choice of BERTimbau is both strategic and methodological: it is a model linguistically adapted to Portuguese, architecturally aligned with the original BERT, and widely adopted in the NLP community—facilitating reproducibility, interpretability, and experimental validity in the context of syntactic specialization analysis.

This study investigates whether BERTimbau exhibits attention head specialization aligned with core syntactic relations, with a particular focus on subject–verb dependencies across varied syntactic constructions. Our contributions are twofold:

- Identification of attention heads that align with subject–verb dependencies across grammatical contexts.
- Visual mapping of attention behaviors based on linguistic annotation.
- Comparison with syntactic structures of mBERT.

The structure of this paper is as follows. Section 2 details the grammatical constructions. Section 3 discusses related work. Section 4 outlines our methodology. Section 5 presents our experimental setup. Section 6 discusses our findings. Section 7 describe our estimates energy consumption, and Section 9 concludes and envisage future work.

2. Grammatical Patterns and Linguistic Dependencies

Portuguese displays a range of grammatical structures involving diverse syntactic configurations. Central to this study is the subject–verb dependency, which serves as a basis for evaluating head-level attention behavior across varying constructions [Pagano et al. 2024].

2.1. Grammatical Constructions

Grammatical constructions in this study are defined as structural configurations that modulate the realization of subject–verb (SV) dependencies, optionally followed by an object [SV(O)] when present. This notation allows us to represent both transitive and intransitive verbs, as well as cases with optional complements, avoiding ambiguity regarding verb valency. Our analysis focuses exclusively on the syntactic relation between the subject and its governing verb, even in complex sentences such as those containing subordinate clauses. Each construction provides a controlled context for evaluating whether specific attention heads align with the SV(O) relation. The constructions covered in our experimental dataset include:

- **Canonical Word Order (SV(O))**: Sentences with standard subject–verb–(object) structure, such as “O Congresso aprovou a reforma tributária.” (in English: “Congress approved tax reform”) or “Felipe comprou um carro novo.” (in English: “Felipe bought a new car.”)
- **Non-Canonical Word Order**: Sentences involving dislocations or topicalization, including verb-initial or object-initial orders, e.g., “Chegou João” (in English: ”John arrived.”) and “Esse livro, o professor recomendou enfaticamente.” (in English: ”That book, the professor strongly recommended.”)
- **Voice Alternation**: Pairs of active and passive constructions to test whether attention shifts follow syntactic role changes. For example, “O garçom ofereceu uma bebida” (in English: ”The waiter offered a drink.”) versus “Uma bebida foi oferecida pelo garçom.” (in English: ”A drink was offered by the waiter.”)
- **Reflexive Sentences**: Constructions with reflexive clitics like “se”, where the subject and internal argument coincide, e.g., “Rodrigo se machucou durante a corrida.” (in English: ”Rodrigo hurt himself during the race.”) and “João se feriu.” (in English: ”João wounded himself.”)
- **Subordinate Clauses**: Embedded clauses under verbs of saying or thinking, as in “A professora disse que João chegou” (in English: ”The teacher said that John arrived.”) and “Jorge acha que o novo diretor vai demiti-lo.” (in English: ”Jorge thinks that the new principal is going to fire him.”) — in these cases, only the subject–verb relation of the matrix clause is considered.

- **Subject Predicatives with Agreement**: Sentences where the subject is followed by a predicative expression in agreement, e.g., “João voltou entusiasmado.” (in English: ”John came back enthusiastic.”)
- **Other Constructions**: A smaller set of sentences with indirect transitive verbs (“Alice acreditou em Pedro”) or pronominal references (“O delegado terminou o seu relatório”) (in English: ”The police chief finished his report.”), included for exploratory purposes but not analyzed systematically.

3. Related Work

The growing interest in interpretability of Transformer models has led to extensive investigations into how internal mechanisms, especially attention heads, capture linguistic structure. A central line of inquiry concerns whether attention distributions correlate with syntactic dependencies, even in the absence of explicit supervision for grammar.

Authors in [Clark et al. 2019] demonstrated that certain attention heads in BERT align with syntactic dependencies such as `nsubj`, `obj`, and also reflect discourse-level relations like coreference. These results reinforced the hypothesis that hierarchical syntactic structures can emerge from self-supervised objectives. Authors in [Voita et al. 2019] showed that head specialization can be functionally distinct and measurable. The work from [Michel et al. 2019] further revealed that many heads are redundant and can be pruned without significant performance loss.

In multilingual scenarios, [Pires et al. 2019] highlighted that mBERT’s syntactic performance is less consistent in morphologically rich languages, motivating the development of monolingual alternatives. BERTimbau [Souza et al. 2020], trained exclusively on Brazilian Portuguese, has shown competitive results in tagging and classification tasks. However, few studies have examined how its attention heads behave with respect to specific syntactic dependencies under varying grammatical conditions.

Other investigations explore the attention-based analyses in parsing [Lin et al. 2019] and probing [Raganato and Tiedemann 2018], including languages with rich inflections. Nonetheless, the behavior of attention heads in Brazilian Portuguese—particularly regarding consistent alignment with core dependencies such as `nsubj` across different constructions—remains underexplored.

This study addresses this gap by investigating whether subject–verb dependencies are encoded by specific attention heads in a BERT-based model for Brazilian Portuguese, namely, BERTimbau. The analysis spans a range of syntactic constructions, including reflexive clauses, subordinate embeddings, and passive alternations. Rather than modeling syntactic phenomena as output labels, we conceptualize them as structural conditions under which attention patterns may vary. Our objective is to evaluate the syntactic consistency of head-level attention, thereby contributing to a deeper understanding of grammatical encoding in monolingual Transformer architectures.

4. Methodology

This study employs a contrastive analysis focused on capturing the Subject–Verb (SV(O)) syntactic dependency across diverse grammatical constructions in Brazilian Portuguese. Rather than surveying multiple phenomena, the analysis is restricted to a single dependency arc—`nsubj`—which enables controlled comparison across structurally distinct contexts. The SV(O) notation explicitly accounts for intransitive verbs and optional complements, ensuring consistency in the interpretation of verb valency across all constructions. In subordinate clauses, we consider only the subject–verb relation of the matrix clause, excluding internal relations within the embedded clause.

A curated set of 42 sentences was manually constructed with the support of a linguist, each exemplifying one syntactic construction (e.g., canonical SV(O), subordinate clause, reflexive construction, passive voice). Lexical choices were controlled to isolate morphosyntactic variation while minimizing confounds. This design allows us to assess attention consistency under agreement, embedding, dislocation, and cliticization.

Table 1. Example sentences by construction type, with subject and verb annotation (SV(O) notation).

Construction	**Sentence**	**Subject**	**Verb**
Canonical (SV(O))	João chegou cedo.	João	chegou
	O Congresso aprovou a reforma tributária.	Congresso	aprovou
	Maria convidou os amigos para a festa.	Maria	convidou
Non-Canonical	Chegou João.	João	chegou
	Esse livro, o professor recomendou enfaticamente.	professor	recomendou
Passive Voice	Uma bebida foi oferecida pelo garçom.	bebida	foi oferecida
Reflexive	João se feriu.	João	se feriu
	Rodrigo se machucou durante a corrida.	Rodrigo	se machucou
Predicative	Fernanda saiu de casa enfurecida.	Fernanda	saiu
Subordinate Clause	A professora disse que João chegou.	professora	disse
	Jorge acha que o novo diretor vai demiti-lo.	Jorge	acha
Indirect Transitive	A política duvidou do depoimento da testemunha.	política	duvidou
Coreference	O delegado terminou o seu relatório.	delegado	terminou

4.1. Attention Head Analysis

We extracted attention weights from the 144 attention heads in the BERTimbau Base model (12 layers × 12 heads) for each sentence. The Transformer architecture employs scaled dot-product attention [Vaswani et al. 2017], as shown in Equation 1, enabling the model to weigh the relevance between tokens in a sentence:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{1}$$

Where Q, K, and V are query, key, and value matrices from token embeddings, and d_k is the key dimension. This mechanism allows the model to highlight relevant syntactic relations by assigning high weights to specific token pairs.

We quantify the attention strength between subject and verb tokens (SV(O)) and compare these values across the different grammatical contexts. This analysis allows us to identify attention heads that exhibit high sensitivity to the SV(O) relation, regardless of whether the object is present and independent of syntactic surface variation.

The overarching goal is to assess whether BERTimbau's internal mechanisms exhibit syntactic consistency when subjected to morphosyntactic perturbations characteristic of Brazilian Portuguese.

5. Experimental Setup

We focus on how BERTimbau's attention heads capture syntactic structure, specifically the **Subject–Verb (SV(O))** dependency, evaluated across grammatical constructions that challenge surface alignment in Brazilian Portuguese. The SV(O) notation accounts for both transitive and intransitive verbs, as well as optional complements, ensuring consistent interpretation of verb valency across all contexts.

The experiment was conducted in three main stages:

- **Attention Extraction**: For each of the 42 selected sentences, attention maps from all 144 heads (12 layers × 12 heads) were extracted using the Hugging Face ‘BertModel‘. These maps encode pairwise token attention, allowing detailed SV(O) dependency analysis under varied structures, where the object may be absent or optional.
- **High-Scoring Head Detection**: For each sentence, the head assigning the highest attention to the SV(O) pair was recorded. Aggregating these results by construction type (e.g., canonical, reflexive, passive, embedded) enabled the identification of consistently specialized heads. The SV(O) notation ensures that both transitive and intransitive verbs, as well as optional complements, are consistently represented.
- **Relevance Metrics**:
  - *Relative Attention Strength (RAS)*: Normalized attention weight directed from subject to verb per sentence.
  - *Layer Specialization Index (LSI)*: Distribution of top-RAS heads across layers, indicating vertical concentration of syntactic encoding.

5.1. Attention Distribution in the Subject–Verb Relation

The analysis focused exclusively on the syntactic relation between the subject and the main verb (SV(O)), following the `nsubj` dependency label in the Universal Dependencies scheme. For each sentence, attention values were extracted from the subject token to the main verb token across all 144 attention heads of the BERTimbau model. The SV(O) notation accounts for both transitive and intransitive verbs, as well as optional complements, ensuring consistency in the interpretation of verb valency. The objective was to identify which heads consistently assigned high attention to this syntactic link across different grammatical constructions.

Table 2. Top-3 heads per grammatical construction, ranked by average SV(O) attention.

Construction	Layer	Head	Mean Attention
Canonical (SV(O))	2	3	0.8329
	7	12	0.3692
	7	10	0.3579
Non-Canonical (VSO / topicalization)	6	11	0.9999
	3	8	0.9975
	7	8	0.8599
Pronominal / Referential	2	3	0.9993
	7	12	0.6840
	7	10	0.5386
Indirect Transitive	2	3	0.9996
	7	12	0.7432
	1	7	0.3631
Subject Predicative	2	3	0.9999
	7	10	0.8792
	4	11	0.5405
Subordinate Clause	2	3	0.9999
	7	10	0.9997
	4	11	0.9224
Passive Voice	4	6	0.4353
	4	5	0.4257
	1	7	0.2308

Note: The SV(O) notation accounts for both transitive and intransitive verbs, as well as optional complements.

Among all heads, **Layer 2, Head 3 (L2_H3)** was the most stable and recurrent in capturing the SV(O) dependency. It consistently ranked highest in average attention across canonical SV(O) structures and remained highly responsive in embedded, reflexive, and predicative constructions.

Other heads exhibited construction-specific specialization:

- **L7_H12** and **L7_H10** were consistently strong in canonical SV(O) contexts, suggesting upper-layer convergence for core dependencies.
- In **non-canonical word orders** (VSO, OSV, topicalization), **L6_H11**, **L3_H8**, and **L7_H8** achieved near-perfect alignment, indicating robustness to syntactic reordering.

- In **pronominal or referential constructions**, such as those involving possessives or anaphoric elements, **L7_H12** and **L7_H10** complemented the behavior of **L2_H3**.
- For **indirect transitive verbs**, attention was also concentrated in **L1_H7** and **L7_H12**, likely due to prepositional complexity.
- In **subject–predicative** structures, **L4_H11** and **L7_H10** assisted **L2_H3**, maintaining stable attention despite intervening modifiers.
- In **subordinate clauses**, attention from the matrix subject to the embedded verb was effectively tracked by **L7_H10** and **L4_H11**.
- In **passive constructions**, although average attention levels were lower, **L4_H6**, **L4_H5**, and **L1_H7** showed relatively stronger focus on SV(O) alignment.

These findings indicate that BERTimbau does not rely solely on surface patterns but develops **syntactic sensitivity** and **context-dependent head specialization**. While **L2_H3** serves as a general-purpose SV(O) tracker, other heads, particularly from deeper layers, contribute selectively in structurally complex environments.

5.2. Accuracy of SV(O) Dependency Across Contexts

We computed the average attention score from the subject token to the main verb token across all sentences to assess how reliably BERTimbau's attention heads capture the Subject–Verb (SV(O)) syntactic relation. Scores were grouped by grammatical construction and head index (Layer–Head). This allows us to identify which heads consistently align with the SV(O) dependency under different syntactic configurations.

Figure 1 presents a heatmap showing the top-3 heads per construction, ranked by normalized mean S→V attention. Each cell represents the average attention strength from subject to verb for a specific head within that construction.

Layer 2, Head 3 (L2_H3) maintained high attention scores across nearly all constructions—including canonical SV(O), subordinate clauses, reflexive forms, and predicative structures—indicating its general reliability in modeling SV(O) alignment.

Other heads showed selective contextual sensitivity:

- **L6_H11** was more prominent in subordinate clauses and sentences with embedded structures, where SV(O) distance increases.
- **L5_H5** performed better in predicative contexts and modal constructions, such as "João chegou entusiasmado" and "João deve estar cansado", suggesting sensitivity to modifiers and auxiliary verbs.
- **L7_H10** was most active in non-canonical word orders (e.g., "Chegou João"), where the surface position of the subject differs from typical SV(O) alignment.

These results indicate that BERTimbau distributes syntactic processing across multiple attention heads, with **L2_H3** acting as a general-purpose SV(O) encoder, while heads such as **L6_H11** and **L5_H5** exhibit context-specific specialization depending on structural variation.

6. Structural Alignment Analysis

While attention from subject to verb (SV(O)) reveals head-level sensitivity to syntactic dependencies, it remains limited to token pairs. To assess whether attention encodes broader syntactic structure, we applied a tree-based analysis using the *Undirected Unlabeled Attachment Score* (UUAS) [Hewitt and Manning 2019].

For each attention head, we constructed dependency trees from attention matrices using the Chu–Liu/Edmonds algorithm [Chu and Liu 1965]. Each attention matrix $A \in \mathbb{R}^{n \times n}$, with A_{ij} denoting attention from token i to j, defines a directed graph where edges are weighted by attention scores. From this graph, an MST rooted at a pseudo-ROOT node was extracted and compared to the gold-standard UD tree to compute UUAS.

To balance structural alignment and attention focus, we adopted a composite score:

$$\text{score} = \text{UUAS} - \alpha \cdot \text{entropy}$$

with $\alpha = 0.2$ following [Voita et al. 2019]. This penalizes diffuse heads while rewarding structural fidelity.

Table 3 shows the five top-ranking heads by syntactic score. **Layer 2, Head 3** emerged as the most consistent in recovering UD-like structure, with low entropy and high UUAS. These results confirm that BERTimbau's internal attention mechanisms partially reconstruct hierarchical syntax beyond local dependencies.

Table 3. Top-5 attention heads ranked by syntactic score.

Layer	Head	UUAS	Entropy	Score
2	3	0.4579	0.1288	0.4321
3	8	0.4579	0.1319	0.4315
6	11	0.4579	0.1591	0.4261
7	10	0.4558	0.2321	0.4094
4	11	0.4568	0.3132	0.3942

6.1. Comparison Between BERTimbau and mBERT on S–V Dependency Encoding

We compared attention head behavior between **BERTimbau** (monolingual) and **mBERT** (multilingual) using the same sentence set to evaluate the impact of model training on syntactic specialization. The evaluation computed UUAS, entropy, and a composite syntactic score (higher is better) for all 144 heads in both models, focusing on the `nsubj` dependency.

Tables 4 and 5 summarizes the top-10 heads for each model, ranked by syntactic score. Both models share similar high-performing heads—particularly **Layer 2, Head 3 (L2_H3)** and **Layer 6, Head 11 (L6_H11)**, with some important differences.

- Both models achieve identical UUAS values (0.4579) in their top heads, notably in **L2_H3** and **L6_H11**.
- BERTimbau shows lower entropy overall in top heads, suggesting more focused attention distributions.
- mBERT presents slightly higher UUAS in some heads (e.g., L6_H9 = 0.5803), but this comes at the cost of high entropy, reducing syntactic score.
- BERTimbau's top-5 heads all have syntactic scores above 0.39, while mBERT's drop below 0.37 after the fourth position.

These results suggest that while both models encode subject–verb structure, BERTimbau achieves more compact and consistent alignment, likely due to its monolingual exposure during pretraining. This reinforces prior evidence that multilingual models like mBERT may underperform in syntax-sensitive tasks for morphologically variable languages such as Portuguese.

Table 4. Top-10 attention heads in BERTimbau for `nsubj`.

Layer	Head	UUAS	Entropy	Score
2	3	0.4579	0.1288	0.4321
3	8	0.4579	0.1319	0.4315
6	11	0.4579	0.1591	0.4261
7	10	0.4558	0.2321	0.4094
4	11	0.4569	0.3132	0.3942
2	5	0.4040	0.2985	0.3443
7	8	0.4524	0.6288	0.3267
6	9	0.5439	1.1999	0.3040
3	4	0.4401	0.6908	0.3019
4	2	0.4519	0.7813	0.2956

Table 5. Top-10 attention heads in mBERT for `nsubj`.

Layer	Head	UUAS	Entropy	Score
2	3	0.4579	0.1598	0.4259
6	11	0.4579	0.1729	0.4233
7	10	0.4562	0.2931	0.3976
3	8	0.4579	0.4616	0.3655
6	9	0.5803	1.1742	0.3454
4	11	0.4537	0.5615	0.3414
7	8	0.4487	0.6770	0.3133
2	5	0.4411	0.7761	0.2859
3	12	0.4798	1.0288	0.2740
6	4	0.4525	0.9133	0.2699

7. Discussion

Our results demonstrate that BERTimbau exhibits attention head specialization aligned with syntactic functions, particularly in encoding the Subject–Verb (SV(O)) dependency across distinct grammatical constructions. Unlike analyses that classify constructions by surface lexical features (e.g., reflexives, modals), we focused on syntactic arcs defined by Universal Dependencies (UD), such as `nsubj`, `acl`, and `mark`. The SV(O) notation accounts for both transitive and intransitive verbs, as well as optional complements, foregrounding grammatical role over category membership.

Layer 2, Head 3 (L2_H3) emerged as the most robust SV(O) encoder across the corpus, maintaining high attention scores in canonical, embedded, and even passive constructions. For instance, it preserved `nsubj` alignment in 83.3% of subordinate sentences, even when subjects were distanced from verbs. This supports earlier claims by [Clark et al. 2019] that attention heads in intermediate layers develop role-tracking behaviors, though our findings extend them to Portuguese.

Other heads exhibited more focused, context-dependent specialization:

- **L6_H11** consistently attended from verbs to subordinating complementizers (e.g., *disse* → *que*), aligning with `mark` and `ccomp` arcs.
- **L7_H12** showed selective behavior in reflexive structures, targeting clitic-to-verb alignments (`expl:pv`, `obj`), even under ambiguity.
- **L5_H5**, though less consistent, peaked in contexts involving subject predicatives and auxiliaries, suggesting mild sensitivity to verbal periphery.

These findings reinforce that attention head specialization reflects **syntactic function** more than lexical category. For example, heads did not merely respond to the presence of ”reflexive” or ”modal” tokens per se, but to their grammatical role within a dependency chain. This aligns with recent insights from probing studies [Voita et al. 2019, Raganato and Tiedemann 2018] suggesting that intermediate heads carry hierarchical relational information, particularly in monolingual settings.

It is important to note that our analysis did not address more complex word-order alternations such as raising constructions, frequently observed with verbs like *parecer* (“to seem”), where raising to the matrix clause may or may not occur. For example, in “A Maria parece que viajou” and “Parece que a Maria viajou” (“It seems that Maria traveled”), subject positioning and clause embedding can differ substantially, potentially altering attention patterns. We leave the investigation of such phenomena for future work.

Another limitation concerns the size of our dataset. The set of 42 sentences was deliberately kept small to reduce computational cost and environmental impact (see Section 8), but we acknowledge that this scale constrains the generalizability of our findings. Future work will expand the corpus to enable more robust statistical validation.

While attention is a useful proxy for identifying syntactic dependencies, it does not guarantee causal interpretation. High attention weights may correlate with syntactic structure without being the mechanism by which the model encodes it. Future analyses could incorporate ablation studies or causal probing experiments to directly test whether disabling these heads degrades performance on syntax-sensitive tasks.

7.1. Visual Patterns and Functional Alignment

Figure 1 presents a heatmap of the top-3 heads per construction, based on mean attention from subject to verb. The visualization confirms that **L2_H3** is active across nearly all construction types, while **L6_H11** and **L5_H5** exhibit strong specialization for subordination and predication, respectively.

For example, in the non-canonical sentence “Chegou João” (“John arrived”), **L6_H11** maintained strong focus on the subject–verb pair despite the VSO ordering, illustrating robustness to syntactic reordering. Similarly, in the predicative sentence “João voltou entusiasmado” (“John came back enthusiastic”), **L5_H5** concentrated attention from the subject to the main verb, even with intervening modifiers.

- **L2_H3** was the only head to appear among the top-3 in *all* constructions except passive voice, where attention from subject to verb decreased globally.
- **L6_H11** and **L3_H8** dominated in non-canonical orders (e.g., VSO), showing positional flexibility in dependency tracking.
- **L5_H5** reached peak performance in constructions with predicative complements or modal stacking (e.g., *João deve estar cansado*).

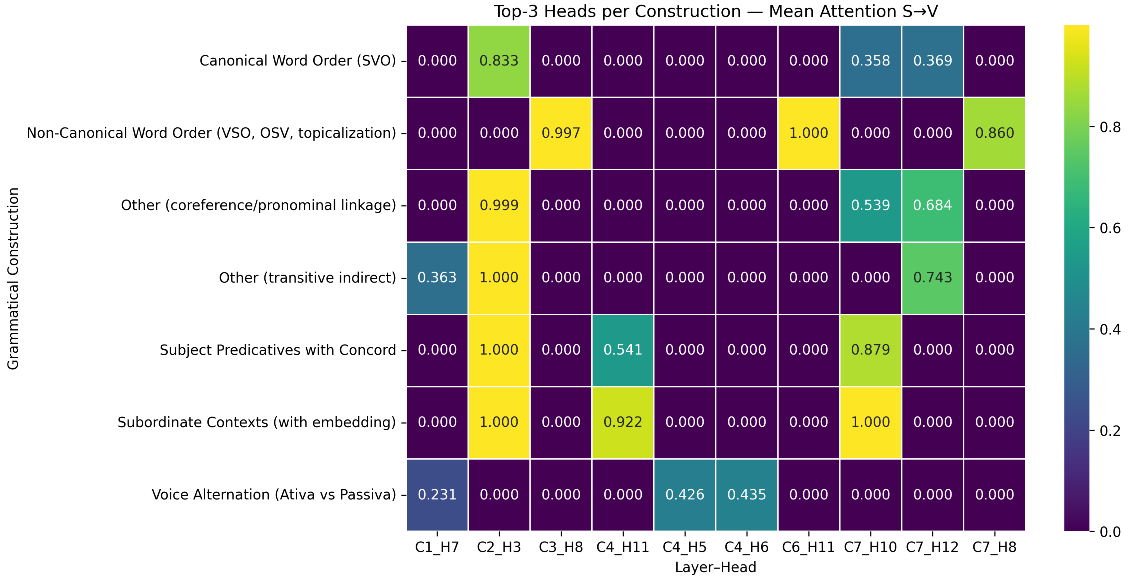

Figure 1. Heatmap of top-3 attention heads per construction, based on mean $S \to V$ attention. L2_H3 appears across all contexts; L6_H11 and L5_H5 show context-specific specialization.

In contrast, categories such as speaker-oriented adverbs (e.g., *infelizmente*) elicited diffuse and inconsistent attention, suggesting that semantic or pragmatic modifiers are weakly captured by syntactic heads.

7.2. Implications and Model Comparison

Compared with mBERT, BERTimbau shows more focused, interpretable syntactic attention. Although both share strong heads (**L2_H3**, **L6_H11**), BERTimbau attains lower entropy in top heads and higher syntactic scores (Tables 4, 5), reinforcing that monolingual pretraining yields more selective syntactic specialization [Pires et al. 2019]. Transformer heads capture fine-grained Portuguese dependencies with layer-wise specialization, validating UD-guided attention analysis as a framework for studying grammatical behavior in pre-trained language models.

8. Estimated Energy Consumption and Carbon Emissions

We estimated the carbon footprint of our experiments based on total GPU processing time and average hardware consumption. The analysis covered attention extraction and visualization tasks over 42 sentences, repeated across 15 cycles, summing approximately 6.36 hours of GPU usage.

Experiments were conducted on a local machine (Intel i7-13620H, 32 GiB RAM, NVIDIA RTX 3050, Ubuntu 24.04), with an estimated average power draw of $60W$. Total energy usage was depicted in Equation 2.

$$\text{Energy (kWh)} = \frac{60 \times 6.36}{1000} = 0.3816 \tag{2}$$

Using Brazil's emission factor (0.1 kg CO_2/kWh), estimated emissions totaled is depicted in Equation 3.

$$\text{CO}_2 = 0.3816 \times 0.1 = \textbf{0.038 kg} \tag{3}$$

This reflects a low environmental impact, consistent with sustainable computing practices in renewable-based energy grids.

9. Conclusions and Future Work

Our study analyzed the attention heads in BERTimbau concerning syntactic structures, focusing on the Subject–Verb (SV(O)) relation (`nsubj`) across varied grammatical constructions. Rather than treating sentence types as independent phenomena, we adopted a dependency-centric approach to evaluate alignment between attention and syntax.

Key findings on **head specialization**: **Layer2 Head3** consistently tracks `nsubj` across contexts; other heads (e.g., L6_H11) target embedded and non-canonical patterns. BERTimbau shows tighter, more stable attention than mBERT, supporting monolingual training for syntax-sensitive tasks. Attention degrades in semantically diffuse cases (speaker-oriented adverbs, passives), indicating limits for non-canonical or non-structural cues. Future work: analyze complex word-order alternations (e.g., raising), expand beyond 42 sentences, and probe causal mechanisms via ablations and probing.

Acknowledgment

This work is partially supported by the FAPESB under grant TIC 0002/2015, CCE 0022/2023 and INCITE PIE0002/2022.

References

Chu, Y.-J. and Liu, T.-H. (1965). On the shortest arborescence of a directed graph. In *Science Sinica*, volume 14, pages 1396–1400.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 276–286.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Hewitt, J. and Manning, C. D. (2019). Structural probe: Finding syntax in word representations. In *Proceedings of NAACL-HLT 2019*, Minneapolis, USA.

Lin, Y., Tan, Y. C., and Frank, R. (2019). Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Melo, A., Cabral, B., and Claro, D. (2024). Scaling and adapting large language models for portuguese open information extraction: A comparative study of fine-tuning and lora. In *Anais da XXXIV Brazilian Conference on Intelligent Systems*, pages 427–441, Porto Alegre, RS, Brasil. SBC.

Michel, P., Levy, O., and Neubig, G. (2019). Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, pages 14014–14024.

Pagano, A. S., Rassi, A., and Pagano, A. C. S. (2024). A ordem e a função das palavras em uma sentença: Sintaxe. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 6. BPLN, 2 edition.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Raganato, A. and Tiedemann, J. (2018). An analysis of encoder representations in transformer-based machine translation. In Linzen, T., Chrupała, G., and Alishahi, A., editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Rodrigues, T., Zeni, R., Souza, F., Bonatelli, I., and Fancellu, F. (2023). Albertina pt: State-of-the-art monolingual deberta models for brazilian and european portuguese. *arXiv preprint arXiv:2306.02741*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*.

Tenney, I., Das, D., and Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. (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.

Wu, S. and Dredze, M. (2020). Are all languages created equal in multilingual bert? In *Proceedings of the 5th Workshop on Representation Learning for NLP (RepL4NLP-2020)*, pages 120–130. Association for Computational Linguistics.