Cross-Linguistic Generalization of Syntactic Attention in Transformers: Analysis of Romance Languages

Generalización Interlingüística de la Atención Sintáctica en Transformers: Análisis de Lenguas Románicas

Resumen: La arquitectura Transformer ha demostrado capacidad para capturar regularidades sintácticas mediante mecanismos de autoatención. Sin embargo, aún no está totalmente claro hasta qué punto estos patrones permanecen estables entre diferentes lenguas, especialmente cuando se utilizan modelos monolingües independientes en lugar de modelos multilingües con parámetros compartidos. Este trabajo investiga la generalización interlingüística de patrones de atención asociados a relaciones sintácticas definidas en el esquema Universal Dependencies, analizando distribuciones de atención relacionadas con las dependencias *nsubj*, *obj*, *case* y *amod* en siete lenguas: seis románicas (portugués, gallego, español, francés, italiano y rumano) y una lengua tipológicamente distinta (alemán), utilizada como condición de control. Utilizando exclusivamente modelos BERT monolingües independientes, evaluamos la generalización en dos niveles: micro (cabezas individuales) y macro (capas). Los resultados revelan una disociación clara: aunque la correlación a nivel de cabezas individuales es baja ($\rho \approx$ 0,07-0,15), la consistencia a nivel de capas es moderada ($\rho \approx$ 0,23-0,35), sobre todo para relaciones argumentales. El par portugués-gallego se destaca con correlaciones excepcionalmente altas ($\rho \approx$ 0,89-0,91).
Palabras clave: transformers, atención sintáctica, dependencias sintácticas, universal dependencies

Abstract: The Transformer architecture has shown the ability to capture syntactic regularities through self-attention mechanisms. However, the extent to which these patterns remain stable across languages is still not fully understood, particularly when using independent monolingual models rather than multilingual models with shared parameters. This study investigates the cross-linguistic generalization of attention patterns associated with syntactic relations defined in the Universal Dependencies framework. We analyze attention distributions related to the dependencies *nsubj*, *obj*, *case*, and *amod* across seven languages: six Romance (Portuguese, Galician, Spanish, French, Italian, and Romanian) and German as a typologically distinct control. Using exclusively independent monolingual BERT models, we evaluate generalization at two levels: micro (individual heads) and macro (layers). Results reveal a clear dissociation: while head-level correlation is low ($\rho \approx$ 0.07-0.15), layer-level consistency is moderate ($\rho \approx$ 0.23-0.35), especially for argumental relations. The Portuguese-Galician pair stands out with exceptionally high correlations ($\rho \approx$ 0.89-0.91).
Keywords: transformers, syntactic attention, syntactic dependencies, universal dependencies

1 Introduction

Transformer-based models have become the dominant paradigm in Natural Language Processing (NLP), achieving state-of-the-art results in tasks such as machine translation, textual inference, and syntactic analysis. A core component of this architecture is the self-attention mechanism, which enables the

modeling of dependencies between tokens regardless of how far apart they are in the sentence (Vaswani et al., 2017). This property facilitates the capture of complex linguistic regularities in large-scale models such as BERT (Devlin et al., 2019).

In recent years, there has been growing interest in understanding the extent to which the internal mechanisms of these models reflect linguistic structure. Pioneering studies have shown that certain attention heads exhibit systematic alignment with syntactic relations and other linguistic phenomena. Clark et al. (2019) identified attention patterns associated with syntactic dependencies, while Voita et al. (2019) observed functional specialization in some self-attention heads. Complementarily, Hewitt and Manning (2019) showed that the internal representations of neural models may encode syntactic information that can be recovered through structural probes.

Despite these advances, much of this line of research has focused on English data and has often relied on multilingual models such as mBERT, in which parameter sharing across languages confounds the analysis of genuinely linguistic generalization with architectural effects. More recent work has sought to extend this analysis to other languages and more specific phenomena. For example, Anonymous (2025a) investigate the identification of idiomatic expressions in Portuguese using fine-tuned Transformers, while Anonymous (2025b) analyze the specialization of attention heads associated with syntactic dependencies in Portuguese. These findings suggest that structural patterns may emerge consistently in neural architectures applied to different languages.

However, one central question remains open: to what extent do the syntactic patterns captured by attention mechanisms remain stable across different languages *when independent models are used*? The literature still presents gaps and opportunities regarding systematic evaluations capable of separating generalization arising from language properties from effects produced by parameter sharing across models (Rogers, Kovaleva, and Rumshisky, 2020). In this context, the present study investigates the cross-linguistic generalization of attention patterns associated with syntactic relations in independent monolingual Transformer models, focusing on four central dependencies from the *Universal Dependencies* framework (*nsubj*, *obj*, *case*, and *amod*), selected because of their fundamental role in the organization of sentence structure (Nivre et al., 2016).

The main contributions of this work are as follows: (i) a methodology for evaluating cross-linguistic generalization using exclusively monolingual models, allowing a clearer distinction between linguistic universality and parameter-sharing effects; (ii) empirical evidence of a dissociation between macro and micro level patterns, with moderate generalization across layers but limited correspondence across individual heads; and (iii) the identification of the Portuguese–Galician pair as an exceptional case of high correlation ($\rho \approx$ 0,89-0,91).

2 Related Work

The investigation of the linguistic capacity of Transformer-based models has received increasing attention, especially with regard to the presence of syntactic information in their internal representations. Part of this literature seeks to determine whether self-attention mechanisms capture structural regularities of language or whether the performance observed in linguistic tasks emerges mainly from statistical correlations present in the training data.

Building on these initial findings, one line of research has focused on characterizing the specialization of attention heads more precisely. Clark et al. (2019) showed that such specialization encompasses not only syntactic dependencies but also positional and punctuation-related patterns. Voita et al. (2019), in turn, quantified the degree of redundancy across heads, showing that only a subset contributes substantially to model performance.

Another relevant approach investigates the linguistic structure encoded in the internal representations of these models. Hewitt and Manning (2019) introduced the use of structural probes to identify syntactic information in contextualized embeddings, showing that internal representations may contain sufficient information to recover structural properties of syntax. This perspective is grounded in the tradition of dependency-based syntactic analysis (Tesnière, 2015).

More recently, a number of studies have explored the functional role of attention

heads from causal and structural perspectives. Elhelo and Geva (2025) propose a method for inferring attention-head functionality directly from model parameters, while Nam et al. (2025) introduce the *Causal Head Gating* method to estimate the causal contribution of different heads to model performance. In a different direction, Huang, Li, and Zhang (2024) investigate the explicit incorporation of syntactic information into the attention mechanism, showing that grammatical structure can improve the modeling of complex linguistic relations.

Broader studies have also emerged to synthesize accumulated knowledge on linguistic interpretability in Transformers. Rogers, Kovaleva, and Rumshisky (2020) provide a review of BERT's internal functioning, covering syntactic, semantic, and representational aspects across different layers, while Graichen, de Dios-Flores, and Boleda (2026) present a systematic review specifically focused on grammatical knowledge encoded by language models.

In the context of Portuguese, recent work has examined the behavior of Transformer models with respect to language-specific phenomena. Anonymous (2025b) analyze attention patterns associated with syntactic dependencies in models applied to Brazilian Portuguese, whereas Anonymous (2025a) explore the identification of idiomatic expressions based on contextualized representations.

Despite these advances, much of the literature remains centered on monolingual analyses or relies on multilingual models with shared parameters. In the latter case, when two languages exhibit similar attention patterns, it is not possible to determine whether this similarity arises from properties of the languages themselves or simply from the fact that both are processed by the same model. In this context, the present work investigates cross-linguistic generalization using exclusively independent monolingual models, analyzing attention distributions across multiple Romance languages and one typologically distinct language used as a control.

3 Methodology

This section describes the data, models, attention-extraction procedure, and evaluation metrics adopted to investigate the cross-linguistic stability of attention patterns associated with syntactic dependencies in Transformer-based models.

3.1 Models

The experiments were conducted using models based on the Transformer architecture (Vaswani et al., 2017), most of them monolingual BERT variants (Devlin et al., 2019). Unlike studies that rely on multilingual models such as mBERT, this work focuses on independent monolingual models, each trained exclusively on data from its respective language. This choice is central to the experimental design: by eliminating parameter sharing across languages, it reduces an important source of artificial similarity between representations. Even so, similarities between the observed attention patterns may still reflect, among other factors, structural properties of the languages, shared architectural constraints, tokenization differences, variation in pretraining corpora, and specific characteristics of the optimization process.

The models used are: neuralmind/bert-base-portuguese-cased (**Portuguese**), marcosgg/bert-base-gl-cased (**Galician**), dccuchile/bert-base-spanish-wwm-cased (**Spanish**), camembert-base (**French**), dbmdz/bert-base-italian-xxl-cased (**Italian**), dumitrescustefan/bert-base-romanian-cased-v1 (**Romanian**), and bert-base-german-cased (**German**). All models have 12 layers and 12 attention heads, totaling 144 $(\text{layer}, \text{head})$ positions. It should be noted, however, that the French model (camembert-base) belongs to the RoBERTa family rather than to the original BERT configuration, which constitutes a relevant methodological variable in the cross-linguistic comparison.

Although the experimental pipeline also supports comparisons involving multilingual models, all analyses reported in this article were obtained after filtering for the *monolingual* model family.

3.2 Data

The multilingual syntactic analysis used treebanks annotated according to the *Universal Dependencies* (UD) framework (Nivre et al., 2016), which provides a standardized and comparable syntactic representation across languages. The experiments include seven languages: **Portuguese** (UD_Portuguese-Bosque), **Galician** (UD_Galician-TreeGal), **Spanish** (UD_Spanish-AnCora), **French**

(UD_French-GSD), **Italian** (UD_Italian-ISDT), **Romanian** (UD_Romanian-RRT), and **German** (UD_German-GSD). The first six belong to the Romance family, whereas German was included as a typologically distinct control language.

In the final run reported in this article, we used up to 1000 sentences per split (*train*, *dev*, and *test*) for each treebank, with a maximum length of 512 subwords after tokenization. These values correspond to the final configuration used to generate the reported results. When the *dev* split is unavailable, as in the case of Galician, we use the *train* split as a fallback for that stage.

3.3 Syntactic Dependencies

The analysis focuses on four core syntactic relations from the UD framework: *nsubj* (nominal subject), *obj* (object), *case* (case marking), and *amod* (adjectival modifier). The first two correspond to central relations in sentence argument structure, whereas *case* and *amod* represent functional and modificational relations. The selection of these dependencies makes it possible to observe whether attention mechanisms behave differently when confronted with argumental, grammatical, and adjectival relations.

3.4 Attention Extraction and Alignment

For each sentence processed by a model, we extracted the attention matrices produced by all heads in all layers of the architecture. The alignment between UD-annotated tokens and the subwords generated by the tokenizer is performed through character-span intersection, based on the *offset mapping* provided by the fast tokenizer.

Since a syntactic token may correspond to multiple subwords, the attention between two UD tokens is obtained by aggregation over the corresponding subword spans. Thus, the adopted measure does not operate directly on atomic word tokens, but rather on subword–token projections reconstructed from the offsets.

The pipeline computes attention in two possible directions: *head_to_dep* (head - dependent) and *dep_to_head* (dependent - head). The results reported in this article use the *head_to_dep* direction, that is, attention emitted by the head token toward the dependent token.

Sentences for which token-subword alignment or model inference failed were discarded from the final aggregation. This exclusion policy was adopted to preserve processing consistency and to avoid introducing spurious values into the aggregated statistics.

3.5 Formalization

Consider a Transformer model with L layers and H attention heads per layer. For a sentence tokenized into subwords, each head produces an attention matrix $A_{l,h} \in R^{m\times m}$, where m denotes the number of subwords effectively processed by the model, and $A_{l,h}(u, v)$ denotes the attention weight from subword u to subword v at layer l and head h.

From the UD syntactic annotations, we define the set of arcs associated with a dependency d in a language ℓ as $\mathcal{E}^{(\ell,d)} = \{(x, y)\}$, where x and y correspond, respectively, to the head token and the dependent token. Let $\mathcal{S}(x)$ be the set of subwords aligned with token x and $\mathcal{S}(y)$ the set aligned with token y. The average attention between these two tokens is then computed as

$$\alpha_{l,h}(x, y) = \frac{1}{|\mathcal{S}(x)|\,|\mathcal{S}(y)|} \sum_{u\in\mathcal{S}(x)} \sum_{v\in\mathcal{S}(y)} A_{l,h}(u, v) \tag{1}$$

The average attention associated with dependency d in language ℓ is given by

$$a_{l,h}^{(\ell,d)} = \frac{1}{|\mathcal{E}^{(\ell,d)}|} \sum_{(x,y)\in\mathcal{E}^{(\ell,d)}} \alpha_{l,h}(x, y) \tag{2}$$

This procedure produces an aggregated matrix $M^{(\ell,d)} \in R^{L\times H}$, which is subsequently vectorized to yield the attention profile $\mathbf{a}^{(\ell,d)} = \mathrm{vec}(M^{(\ell,d)}) \in R^{LH}$.

3.6 Metrics

Cross-linguistic generalization is evaluated at two complementary levels.

Micro level (individual heads). For each pair of languages (ℓ_1, ℓ_2), we compute the Spearman correlation between the full vectors $\mathbf{a}^{(\ell_1,d)}$ and $\mathbf{a}^{(\ell_2,d)}$, with dimensionality $LH = 144$. This correlation measures the degree of fine-grained alignment between specific layer–head positions. The aggregated metric reported, denoted ρ_{micro}, is a weighted average of the correlations across all language

pairs, where the weight assigned to each pair is given by

$$w(\ell_1, \ell_2; d) = \min\big(n_{\ell_1,d},\, n_{\ell_2,d}\big) \qquad (3)$$

where $n_{\ell,d}$ denotes the total number of arcs observed for dependency d in language ℓ. Thus, pairs involving languages with greater empirical support contribute more to the final average, without allowing a single language with many arcs to dominate the pairwise weighting on its own.

Macro level (layers). For each language ℓ and dependency d, we construct a reduced vector $\mathbf{u}^{(\ell,d)} \in R^L$ by aggregating at the layer level, computing the simple mean of the H heads in each layer. The Spearman correlation between these reduced vectors defines ρ_{macro}, which measures whether the layers concentrating syntactic attention tend to coincide across languages, regardless of the exact correspondence between individual heads.

Normalized Entropy. For each attention vector $\mathbf{a}^{(\ell,d)}$, we define a discrete distribution $p_i = a_i / \sum_k a_k$. Normalized entropy is then computed as

$$H_{\text{norm}}(\ell, d) = \frac{-\sum_{i=1}^{LH} p_i \log_2 p_i}{\log_2(LH)} \qquad (4)$$

so that $H_{\text{norm}} \in [0, 1]$. Values close to 1 indicate a more uniform attention distribution across layer–head positions, whereas values close to 0 indicate greater concentration in a few positions.

Composite Generalization Score. In addition to the primary metrics, we use a composite score for comparative summarization across dependencies. This score should not be interpreted as a canonical measure, but rather as a ranking heuristic combining micro alignment, macro alignment, and attention concentration. The score is defined as

$$\begin{aligned} S = {} & w_1 \cdot f(\rho_{\text{micro}}) + w_2 \cdot f(\rho_{\text{macro}}) \\ & + w_3 \cdot (1 - \bar{H}) + w_4 \cdot (1 - \sigma_H / \tau) \end{aligned} \qquad (5)$$

where $f(x) = \frac{x+1}{2}$ rescales the correlations to the interval $[0, 1]$, $\bar{H}$ denotes the mean entropy across languages, σ_H is the standard deviation of entropy across languages, $\tau = 0.10$ is a normalization parameter, and the adopted weights are $w_1 = 0.35$, $w_2 = 0.45$, $w_3 = 0.10$, and $w_4 = 0.10$.

The greater weight assigned to ρ_{macro} reflects the central aim of this study, namely, to assess whether cross-linguistic stability emerges more clearly at the layer level than at the level of individual heads. Accordingly, the score S is used only as an auxiliary synthesis instrument, whereas the substantive interpretations in this article rely primarily on the correlation and entropy metrics.

4 Results

This section presents the results obtained with independent monolingual models, as described in Section 3. The reported analyses refer exclusively to the monolingual model family and to the *head_to_dep* direction.

4.1 Dissociation between the Micro and Macro Levels

Table 1 presents the global cross-linguistic generalization metrics for the seven languages considered. The central result is a systematic dissociation between the micro and macro levels of analysis. Across all dependencies, micro-level correlations, computed over full vectors of 144 layer-head positions, remain low. In contrast, macro-level correlations, computed from reduced 12-layer vectors, are consistently higher.

Rel.	ρ_{micro}	ρ_{macro}	$\bar{H}_{\text{norm}}$	S
nsubj	0.146	0.348	0.912	0.586
obj	0.112	0.332	0.903	0.587
case	0.107	0.306	0.834	0.586
amod	0.075	0.232	0.881	0.563

Table 1: Cross-linguistic generalization metrics (global split, 7 languages, monolingual models). ρ_{micro}: Spearman correlation between 144-position vectors; ρ_{macro}: Spearman correlation between 12-layer vectors; $\bar{H}_{\text{norm}}$: mean normalized entropy; S: composite generalization score.

These results indicate that distinct monolingual models do not tend to systematically associate the same syntactic relation with the same individual heads, as shown by the low values of ρ_{micro}. At the same time, the higher values of ρ_{macro} suggest that the concentration of syntactic attention along net-

work depth exhibits greater cross-linguistic convergence than the fine-grained correspondence between specific heads. Substantively, this means that different models may diverge with respect to the precise location of specialization in the layer-head space, while still converging on the layers in which a given type of syntactic information tends to emerge.

Among the dependencies analyzed, *nsubj* shows the highest values at both the micro and macro levels, suggesting greater cross-linguistic stability for subject-predicate relations. At the opposite extreme, *amod* shows the lowest correlation values, which is consistent with greater structural variability across languages in the domain of adjectival modification.

Table 2 complements this analysis by evaluating the statistical significance of the micro-level correlations.

Rel.	ρ_{micro}	**95% CI**	**z**	**p**
nsubj	0.172	[0.114; 0.232]	8.99	<0.001
obj	0.128	[0.073; 0.182]	6.56	<0.001
amod	0.090	[0.040; 0.141]	4.76	<0.001
case	0.067	[0.022; 0.111]	3.66	<0.001

Table 2: Statistical significance of ρ_{micro} (bootstrap and permutation test, 1000 iterations each). CI: confidence interval; z: z-score; p: p-value.

Although the absolute values of ρ_{micro} are modest, Table 2 shows that all correlations are statistically above chance ($p < 0.001$). The bootstrap confidence intervals do not include zero for any dependency, and the z-scores remain positive and substantial. These results do not warrant the conclusion that there is strong micro-level alignment across models, but they do indicate that the weak correlation observed is not merely the result of random noise. In other words, there is evidence of some shared structure across the monolingual models, although the degree of similarity among them remains low at the level of individual heads.

4.2 Stability across Splits

Table 3 compares the results obtained for the *train* and *test* splits. Overall, the metrics are highly stable across the two sets, with only small differences in ρ_{micro}, ρ_{macro}, and the composite score S.

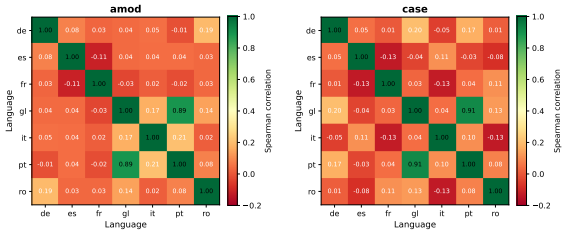
Rel.	ρ_{micro}		ρ_{macro}		S	
	train	test	train	test	train	test
nsubj	0.138	0.142	0.365	0.346	0.586	0.585
obj	0.115	0.111	0.339	0.327	0.586	0.586
case	0.110	0.111	0.287	0.307	0.578	0.586
amod	0.082	0.075	0.242	0.201	0.565	0.556
σ	0.003		0.016		0.005	

Table 3: Comparison of generalization metrics across the *train* and *test* splits (7 languages, monolingual models). The last row reports the mean standard deviation (σ) across splits.

The observed deviations are small, especially for ρ_{micro} and S, suggesting that the reported patterns do not strongly depend on the particular data partition. The largest relative variation occurs in ρ_{macro} for *amod*, but it still remains within a moderate range. Taken together, these results reinforce that the contrast between macro and micro level generalization is not a sampling artifact restricted to a single split.

4.3 Correlation between Language Pairs

Figure 1 presents the micro-level Spearman correlation matrices for all language pairs, for each dependency analyzed.



(a) *amod* (b) *case*

(c) *nsubj* (d) *obj*

Figure 1: Micro-level Spearman correlation between languages (144-position vectors, monolingual models). The Portuguese-Galician pair stands out with $\rho \approx 0.89$-0.91.

The most salient result is the exceptionally high correlation between Portuguese and

Galician, with values of $\rho = 0.89$ for *amod*, $\rho = 0.91$ for *case* and *obj*, and $\rho = 0.88$ for *nsubj*. In contrast, all other language pairs show substantially lower correlations ($\rho < 0.27$), including typologically close pairs such as Spanish–Italian and Spanish–Portuguese.

Another relevant finding is that German, used as a non-Romance control, exhibits correlations with the Romance languages that fall within the same general range observed for several Romance pairs. For *nsubj*, for example, the German-Portuguese correlation ($\rho = 0.10$) is numerically comparable to the Italian-Portuguese ($\rho = 0.19$) and Spanish-Portuguese ($\rho = 0.07$) correlations. This pattern indicates that, with the exception of the Portuguese-Galician pair, Romance pairs do not display a clear and consistent advantage over the non-Romance control at the micro level of individual heads.

Taken together, these results suggest that typological proximity may favor fine-grained alignment in cases of very high linguistic proximity, but it does not automatically translate into micro-level correspondence between attention heads in distinct monolingual models. At the same time, because the experimental design does not fully control for differences in corpus, tokenization, architecture, and training, this evidence should be interpreted as an empirical indication rather than as conclusive causal demonstration.

4.4 Layer-wise Attention Profiles

Figures 2-5 present the mean attention profiles by layer, contrasting the Romance-language mean with German.

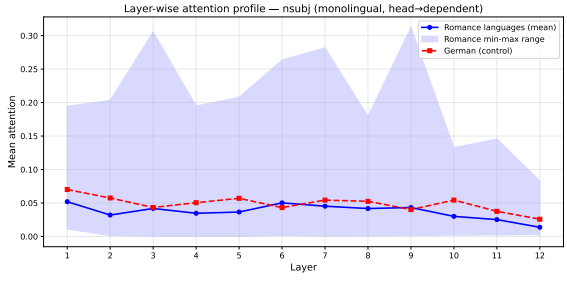


Figure 2: Mean attention by layer for *nsubj*. Blue: Romance mean; band: min-max; dashed: German.

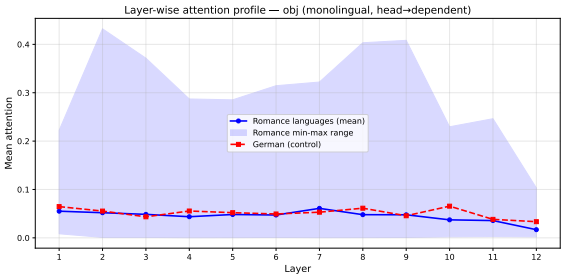


Figure 3: Mean attention by layer for *obj*.

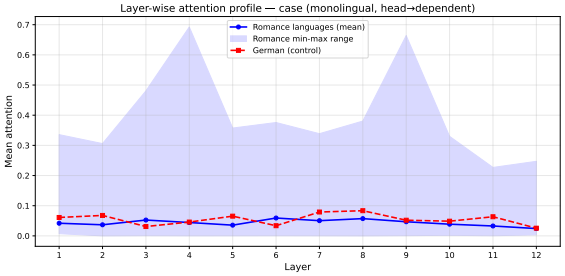


Figure 4: Mean attention by layer for *case*.

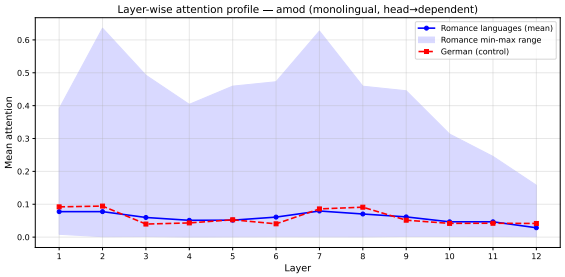


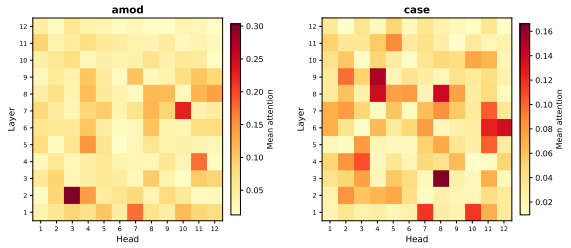
Figure 5: Mean attention by layer for *amod*.

The layer-wise profiles show that attention associated with the argumental relations *nsubj* and *obj* remains relatively distributed across network depth, without strong concentration in a single layer range. The broad min-max band observed in the Romance-language curves reflects variability across models, particularly due to the discrepant behavior of the Portuguese-Galician pair relative to the remaining pairs.

In the case of German, one observes a global distributional trend across layers that remains within the same order of magnitude as the Romance mean, without strongly localized peaks or pronounced structural deviations. This result is consistent with the moderate ρ_{macro} values reported earlier: although the models do not coincide finely at the head level, the overall organization of attention across layers shows noticeable similarities across languages.

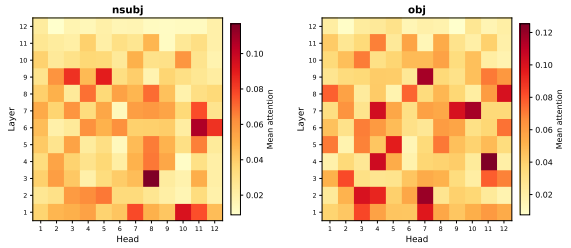
4.5 Distribution across Heads

Figures 6 and 7 show the global heatmaps of mean attention by layer–head position for each relation.



(a) *amod* (b) *case*

Figure 6: Mean attention by layer-head position for *amod* and *case* (monolingual models).



(a) *nsubj* (b) *obj*

Figure 7: Mean attention by layer-head position for *nsubj* and *obj* (monolingual models).

The heatmaps reinforce the contrast between more concentrated and more distributed relations. In *amod*, a relatively sharper focus is observed around position (layer 2, head 3), suggesting that part of the attention associated with adjectival modification falls more locally on a specific region of the architecture. In *case*, relatively stronger activation is observed in intermediate and upper layers, especially between layers 8 and 10. In contrast, *nsubj* and *obj* exhibit more diffuse patterns in the layer-head space, in line with their higher normalized entropy values.

These results do not imply that a single head is exclusively responsible for each relation, but they do indicate that some dependencies tend to concentrate attention more strongly in particular regions of the network, whereas others are distributed more broadly.

4.6 Entropy by Language

Table 4 presents the normalized entropy for each language and relation.

Language	amod	case	nsubj	obj
de	0.876	0.865	0.948	0.938
es	0.900	0.837	0.921	0.911
fr	0.863	0.819	0.871	0.892
gl	0.865	0.824	0.897	0.899
it	0.892	0.848	0.927	0.900
pt	0.892	0.835	0.892	0.891
ro	0.878	0.811	0.930	0.891
Mean	0.881	0.834	0.912	0.903
σ	0.014	0.019	0.027	0.017

Table 4: Normalized entropy by language and relation (global split). Values close to 1 indicate attention distributed across multiple layer-head positions; low σ values indicate relative consistency across languages.

All relations exhibit high entropy ($\bar{H} > 0.83$), indicating that syntactic attention is generally not concentrated in a very small number of heads, but is instead distributed across multiple positions in the architecture. The *case* dependency has the lowest mean entropy ($\bar{H} = 0.834$), suggesting a relatively greater degree of concentration when compared with the other relations. By contrast, *nsubj* and *obj* show the highest entropy values, consistent with the more diffuse patterns observed in the heatmaps.

The low standard deviations across languages indicate that this overall distributional pattern of attention is relatively stable in the analyzed set. Thus, entropy complements the correlation analyses by showing that, even when fine-grained alignment across models is weak, the global distribution of attention may remain broadly spread in a consistent way across languages.

5 Analysis

The results show that the cross-linguistic generalization of syntactic attention patterns in monolingual models depends strongly on the level of analysis: it manifests itself more limitedly at the level of individual heads and more clearly at the level of layers.

5.1 Generalization across Layers, Not across Heads

The main result of the study is the dissociation between the macro and micro levels. Fine-grained correspondence between syntactic relations and specific heads remains low across models ($\rho_{\text{micro}} \approx 0.07$-$0.15$), whereas the distribution of attention at the layer level shows moderate consistency ($\rho_{\text{macro}} \approx$

0.23-0.35), especially for argumental relations. This pattern is consistent with the hypothesis that the Transformer architecture favors a global hierarchical organization of syntactic information without rigidly determining which individual heads perform each function. This interpretation converges with Voita et al. (2019), who show functional redundancy across heads, and with Elhelo and Geva (2025), who show that head functionality can be partially inferred from model parameters without implying strict sharing across training instances.

5.2 Contrast across Language Pairs

The Portuguese-Galician pair constitutes an exceptional case, with correlations far higher than those of the remaining pairs ($\rho \approx 0.89$-0.91). This result suggests strong convergence under conditions of extreme linguistic proximity, but it should be interpreted with caution, since it may simultaneously reflect historical proximity between the languages, possible overlap in training data, and similarities in architecture and tokenization.

By contrast, the correlations between German and the Romance languages ($\rho \approx 0.01$-0.20) often fall within the same general range as those observed among distinct Romance pairs, excluding the Portuguese-Galician case. This weakens the idea that language-family proximity alone would suffice to produce greater micro-level alignment across models. Instead, the results suggest that, at the level of individual heads, factors related to training, tokenization, corpus, and architectural choices may exert an influence comparable to that of typological proximity.

5.3 Implications for Interpretability

These results suggest that the practice of identifying "the attention head" responsible for a given syntactic relation (Clark et al., 2019) should be treated with caution, especially when one intends to generalize such an attribution across models or languages. By contrast, more aggregated interpretability approaches, at the level of layers, functional regions of the network, or representational subspaces, tend to be methodologically more robust, in line with Hewitt and Manning (2019) and Elhelo and Geva (2025). The results are also compatible with the perspective of Huang, Li, and Zhang (2024), according to which the explicit incorporation of syntactic information may favor a more stable organization of linguistic relations.

6 Discussion

The results support the interpretation that self-attention mechanisms capture syntactic regularities with some degree of cross-linguistic stability, but they also show that this stability emerges more clearly at the layer level than at the level of individual heads. In particular, *nsubj* and *obj* exhibit greater relative consistency, suggesting that core argument-structure relations tend to emerge more stably across models and languages, whereas the greater variability observed in *amod* is compatible with typological and distributional differences in nominal modification.

The comparison with German reinforces this picture by showing that, excluding the Portuguese-Galician case, Romance pairs do not exhibit a clear and systematic micro-level advantage over the non-Romance control. This weakens the idea that typological proximity alone would lead to fine-grained alignment between attention heads and suggests that micro-level alignment across monolingual models depends on multiple factors, including corpus, tokenization, and architectural choices.

These findings are in dialogue with the literature that interprets self-attention as carrying structural information (Rogers, Kovaleva, and Rumshisky, 2020), but they also suggest that such structure should be analyzed at an appropriate level of abstraction. Rather than seeking direct correspondence between syntactic relations and isolated heads, the results point to the convenience of privileging broader analytical units, such as layers, functional regions of the network, or aggregated activation patterns. In this sense, the study reinforces a more cautious view of interpretability in Transformers: syntactic information appears to emerge in these models, but its internal distribution is neither rigidly fixed nor fully transferable across distinct monolingual instances.

7 Limitations

This study has some limitations. The analysis is restricted to four syntactic dependencies from the UD framework, so other relations

may exhibit different generalization patterns. In addition, the French model used (*CamemBERT*) is based on the RoBERTa architecture, which introduces architectural heterogeneity into the analyzed set. The experiment also does not fully control for differences in corpus, tokenization, and training conditions across models, which limits causal inferences about the isolated role of linguistic typology. The sample of up to 1000 sentences per split, although sufficient to reveal stable tendencies, may still be limited for smaller treebanks. Finally, as discussed in the literature (Rogers, Kovaleva, and Rumshisky, 2020; Graichen, de Dios-Flores, and Boleda, 2026), attention patterns should not be automatically interpreted as causal explanations of model behavior.

8 Conclusion and Future Work

This work investigated the cross-linguistic generalization of syntactic attention in independent monolingual Transformer models. The results showed a clear dissociation between the micro and macro levels of analysis: while correspondence across individual heads remains low ($\rho_{\text{micro}} \approx 0.07$–$0.15$), the distribution of attention at the layer level exhibits moderate consistency ($\rho_{\text{macro}} \approx 0.23$–$0.35$), especially for argument-structure relations.

Among the analyzed languages, the Portuguese-Galician pair constitutes an exceptional case, with correlations far higher than those of the remaining pairs, suggesting strong convergence under conditions of extreme linguistic proximity. By contrast, the other Romance pairs do not exhibit a clear and consistent micro-level advantage over German, which was used as a non-Romance control. Taken together, these results indicate that typological proximity, although relevant in very closely related cases, does not automatically translate into fine-grained alignment between attention heads in distinct monolingual models.

More broadly, the findings are compatible with the interpretation that the Transformer architecture favors a certain global organization of syntactic information across layers, without rigidly determining its localization in specific heads. This reinforces the value of interpretability approaches that are less centered on isolated heads and more attentive to aggregated patterns of internal organization.

As future work, promising directions include extending the analysis to additional languages and language families, incorporating other relations from the UD framework, applying causal intervention methods (Nam et al., 2025), and systematically comparing monolingual and multilingual models in order to quantify more directly the effect of parameter sharing on syntactic generalization.

References

Anonymous. 2025a. Fine-tuned model evaluation on Transformer Fragments for Identifying Idiomatic Expressions in Portuguese. In *Proceedings of STIL*. (anonymized for review).

Anonymous. 2025b. Syntactic Analysis in Transformers through Attention Heads. In *Proceedings of STIL*. (anonymized for review).

Clark, K., U. Khandelwal, O. Levy, and C. 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.

Devlin, J., M.-W. Chang, K. Lee, and K. 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*, pages 4171–4186, Minneapolis, MN, USA.

Elhelo, A. and M. Geva. 2025. Inferring Functionality of Attention Heads from their Parameters. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, pages 17701–17733, Vienna, Austria.

Graichen, N., I. de Dios-Flores, and G. Boleda. 2026. The Grammar of Transformers: A Systematic Review of Interpretability Research on Syntactic Knowledge in Language Models. *arXiv preprint arXiv:2601.19926*.

Hewitt, J. and C. 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, pages 4129–4138, Minneapolis, MN, USA.

Huang, X., Y. Li, and J. Zhang. 2024. Flexibly Utilizing Syntactic Knowledge in Transformer Models. *Information Processing and Management*, 61.

Nam, A., H. Conklin, Y. Yang, T. Griffiths, J. Cohen, and S.-J. Leslie. 2025. Causal head gating: A framework for interpreting roles of attention heads in transformers. Presented at NeurIPS 2025.

Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia.

Rogers, A., O. Kovaleva, and A. Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Tesnière, L. 2015. *Elements of Structural Syntax*. John Benjamins, Amsterdam. Translated by Timothy Osborne and Sylvain Kahane.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 5998–6008, Long Beach, CA, USA.

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