# Do Multilingual Transformers Encode Paninian Grammatical Relations? A Layer-wise Probing Study

**Akshit Kumar, Dipti Sharma, Parameswari Krishnamurthy**
LTRC, International Institute of Information Technology, Hyderabad, India
akshit.kumar@research.iiit.ac.in
{dipti, param.krishna}@iiit.ac.in

## Abstract

Large multilingual transformers such as XLM-RoBERTa achieve impressive performance on diverse NLP benchmarks, but understanding how they internally encode grammatical information remains challenging. This study investigates the encoding of syntactic and morphological information derived from the Paninian grammatical framework—specifically designed for morphologically rich Indian languages—across model layers. Using diagnostic probing, we analyze the hidden representations of frozen XLM-RoBERTa-base, mBERT, and IndicBERT-v2 models across seven Indian languages (Hindi, Kannada, Malayalam, Marathi, Telugu, Urdu, Bengali). Probes are trained to predict Paninian dependency relations (by edge probing) and essential morphosyntactic features (UPOS tags, Vibhakti markers). We find that syntactic structure (dependencies) is primarily encoded in the middle-to-upper-middle layers (layers 6–9), while lexical features peak slightly earlier. Although the general layer-wise trends are shared across models, significant variations in absolute probing performance reflect differences in model capacity, pre-training data, and language-specific characteristics. These findings shed light on how theory-specific grammatical information emerges implicitly within multilingual transformer representations trained largely on unstructured raw text.

## 1 Introduction

Multilingual pre-trained transformer models such as XLM-RoBERTa (Conneau et al., 2019) have become foundational in natural language processing (NLP), achieving remarkable cross-lingual transfer capabilities. However, their internal representations remain only partially understood. Identifying where and how linguistic structures are encoded across the layers of these transformers improves their interpretability, efficiency, and adaptability across diverse languages and linguistic tasks (Belinkov and Glass, 2019).

A fundamental question is whether multilingual transformer models implicitly encode theory-specific grammatical information despite not being explicitly trained with linguistic annotations grounded in such formalisms.

Much prior work has analyzed transformer representations using probing tasks based mainly on Universal Dependencies (UD) (Hewitt and Liang, 2019; Tenney et al., 2019). While valuable, the UD framework aims for cross-lingual consistency and potentially abstracts away from linguistic phenomena better captured by language-specific or theory-specific grammars. Indian languages (IL), characterized by rich morphology and relatively free word order, are often analyzed using frameworks derived from the classical Paninian grammatical tradition (Bharati et al., 1995). This tradition introduces distinctive linguistic concepts such as Kārakas (syntactico-semantic roles) and emphasizes morphological cues such as Vibhakti markers as fundamental in signaling syntactic structures.

A detailed overview of the specific Paninian annotation scheme (Begum et al., 2007) relevant to this work is provided in Appendix A.

In this work, we carry out the first Paninian analysis of multilingual transformer representations across layers. We probe the hidden layers of XLM-RoBERTa, IndicBERT-v2, and mBERT across seven Indian languages annotated under a shared Paninian annotation scheme (Begum et al., 2007). To measure these linguistic representations, we employ edge probing (Tenney et al., 2019) for syntactic dependency structures and token-classification probes for morphosyntactic features (Universal POS, Vibhakti markers).

Specifically, the paper aims to answer the following research questions clearly:

1. Do multilingual transformer models implicitly

*Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025)*, pages 124–130
August 29, 2025 ©2025 Association for Computational Linguistics

represent Paninian dependency structures and morphosyntactic information?

2. If such representations exist, at which layer do they emerge distinctly?

3. How consistent are these observed patterns across different multilingual transformer architectures and different Indian languages?

Our key findings can be summarized as follows:

1. Paninian dependency structures are indeed implicitly encoded, predominantly emerging in middle layers of multilingual transformers (layers 6–9), in line with prior UD-based studies.

2. Lexical-level morphosyntactic signals (UPOS, Vibhakti markers) peak slightly earlier in the layers, highlighting differentiated storage of structural versus lexical linguistic information.

3. Despite shared general trends, we observe substantial variations across different architectures and languages, reflecting cross-linguistic diversity, data differences, and model-specific factors.

## 2 Experimental Setup

To investigate and compare the layer-wise encoding of Paninian grammar in different multilingual transformers, we conducted probing experiments using three base models: XLM-RoBERTa-base (Conneau et al., 2019), Multilingual BERT (Devlin et al., 2019) and IndicBERT-v2 (Doddapaneni et al., 2023). For each base model, parameters were kept frozen[1], and we extracted hidden states from all layers (including embeddings; Layer 0 up to Layer 12 for base models). Our analysis covers seven Indian languages: Hindi (hi), Kannada (ka), Malayalam (ml), Marathi (mr), Telugu (te), Urdu (ur), and Bengali (be). The datasets utilize annotations following an extended version of the Paninian dependency annotation scheme proposed by Begum et al. (2007).

**Probing Tasks and Models:** We designed diagnostic classifiers (Tenney et al., 2019) for three distinct tasks reflecting key aspects of the formalism. Following best practices for probing (Hewitt and Liang, 2019), we aimed for low-capacity probes:

- **Dependency Relations (Edge Probing):** To assess structural syntactic encoding, we trained probes using a biaffine attention/classifier architecture (Dozat and Manning, 2017). This probe incorporates nonlinearity through intermediate single-layer MLPs (dimensionality 256, GELU activation) applied to the input hidden states before the biaffine transformations for predicting heads and relations.

- **UPOS Tags:** To capture basic lexical categories, we trained strictly linear classifier probes mapping the hidden state directly to tag logits.

- **Vibhakti Features:** Similarly, we used linear classifier probes to predict grammatical case/postposition markers. Vibhakti labels were extracted from the FEATS column based on observed patterns; non-applicable tokens were ignored during accuracy calculation.

Separate probe models were trained for each task, base model, language, and layer combination.

**Data Handling:** Input sequences were processed using the respective tokenizer for each base model (XLM-R, mBERT, IndicBERT-v2). Sequences exceeding a tokenized length of 128 tokens were filtered out. Remaining sequences were padded/truncated to 128 tokens. Target labels were aligned to the first sub-word token, with others ignored. Detailed dataset statistics are provided in Appendix B.

**Training and Evaluation:** Probes were trained independently for 3 epochs using the AdamW optimizer (Loshchilov and Hutter, 2017) with CrossEntropyLoss. Learning rates and batch sizes were tuned per model and task type based on preliminary experiments (Edge Probes: LR=1e-4, BS=64; Token Probes: LR=1.5e-3, BS=256). The model state yielding the best validation performance (LAS for dependencies, Accuracy for features) was selected. Evaluations were performed on held-out validation sets.

## 3 Results

We evaluated the performance of probes trained on each layer's representation across the three base models (XLM-RoBERTa, mBERT, IndicBERT-v2),

[1] Only probe parameters were trained (Alain and Bengio, 2016; Hewitt and Liang, 2019).

seven languages, and three probing tasks (Dependency Relations, UPOS, Vibhakti). Key findings are presented below.

**Dependency Relations (LAS):** Figure 1 presents the best Labeled Attachment Score (LAS) achieved by dependency probes trained on each layer (0-12) for the three base models across the seven Indian languages. A consistent trend across all three models and most languages is the distribution of syntactic information across layers. LAS is low at the initial embedding layer (L0), increases through the lower layers, generally peaks in the middle-to-upper-middle layers (typically Layers 7-9), and then declines towards the final layer (L12). This confirms the widely observed phenomenon that structural syntactic information is most saliently represented in the intermediate representations of transformer models (Jawahar et al., 2019; Tenney et al., 2019).

Model-wise, XLM-RoBERTa (Fig. 1a) and mBERT (Fig. 1b) achieve comparable peak performance on higher-resource languages like Hindi (peak LAS 61.6% for XLM-R at L8; 59.1% for mBERT at L7) and Urdu (peak LAS 48.1% for XLM-R; 47.9% for mBERT, both at L7). IndicBERT-v2 (Fig. 1c), despite its Indic pre-training, does not consistently outperform the general multilingual models on this task for all languages. While reaching a high peak for Hindi (62.1% at L6), its LAS scores for several other languages (ka, ml, be, mr, te) are often lower than those achieved by XLM-R or mBERT, particularly in the upper layers. IndicBERT-v2 shows earlier peaks in several languages (e.g., L6 for Hindi compared to L8 and L7 for XLM-R and mBERT respectively).

**UPOS Tagging (Accuracy):** Figure 2 illustrates the layer-wise accuracy for predicting Universal Part-of-Speech (UPOS) tags. Generally, UPOS tagging accuracy is high, significantly exceeding LAS scores, confirming that basic lexical category information is more readily extractable. Performance tends to peak relatively early compared to LAS, often plateauing across several lower-to-middle layers (e.g., Layers 3-9 for many languages in XLM-R and mBERT) before potentially declining slightly in the final layers.

Model-wise, XLM-RoBERTa and mBERT show strong results, with peak accuracies often reaching 85-90%+ for languages like Hindi and Kannada. IndicBERT-v2 generally achieves somewhat lower peak accuracies on this task compared to the other two models (e.g., peaking around 75% for Hindi).

**Vibhakti Feature Prediction (Accuracy):** The accuracy for predicting Vibhakti features, a key morphological cue in Paninian grammar, is presented in Figure 3. Overall accuracy for this task is generally higher than LAS but can be lower than UPOS accuracy. Hindi and Urdu consistently yield the highest accuracies, frequently exceeding 80-90% across several layers in XLM-RoBERTa and mBERT, and maintaining high accuracy across almost all layers in IndicBERT-v2.

Compared to the sharper LAS peaks, strong Vibhakti prediction performance often extends across a broader range of middle and sometimes upper layers (e.g., Layers 2-10 for Hindi/Urdu in XLM-R/mBERT). The optimal layers for Vibhakti tend to overlap with or slightly precede the peak layers for dependency relations. Among the models, IndicBERT-v2 demonstrates particularly strong and stable Vibhakti prediction for Hindi and Urdu across nearly all layers.

**Cross-Lingual Variation:** Performance varies markedly across languages within each model (Figures 1, 2 and 3). While a comprehensive cross-lingual performance comparison is complicated by factors such as differing script representations and varying vocabulary coverage for each language within the pre-trained models, clear tiers of performance emerge. Hindi and Urdu consistently show the strongest LAS results, suggesting the models capture their Paninian structures relatively effectively. Kannada, Malayalam, Bengali, and Marathi form a mid-tier group, with peak LAS typically ranging from ~20% to ~40% depending on the model and language. Telugu consistently exhibits the lowest LAS scores across all models and layers (peak < 10% for XLM-R/mBERT, < 5% for IndicBERT-v2), a finding that strongly correlates with its significantly smaller probing dataset size (Appendix B) and potential underlying data sparsity in the models' pre-training.

## 4 Analysis and Discussion

Our multi-model, multi-task probing experiments reveal consistent layer-wise patterns for encoding Paninian grammar, alongside notable performance variations.

**Layer Specialization for Paninian Grammar:** Across all three models, a functional specialization

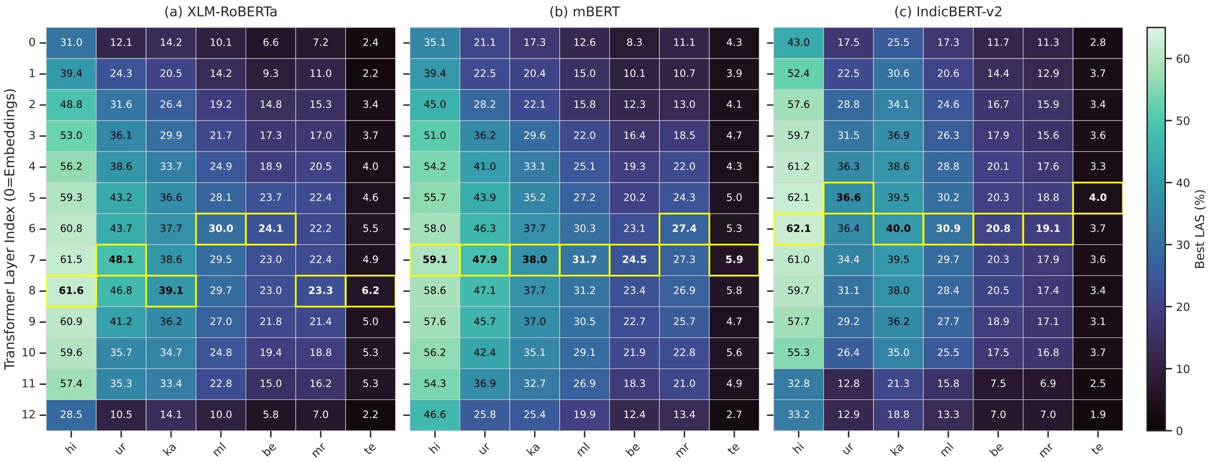

Figure 1: Layer-wise Dependency LAS (%) across Languages for (a) XLM-RoBERTa, (b) mBERT, and (c) IndicBERT-v2. Brighter colors indicate higher LAS. Yellow boxes highlight approximate peak performance regions for selected languages.

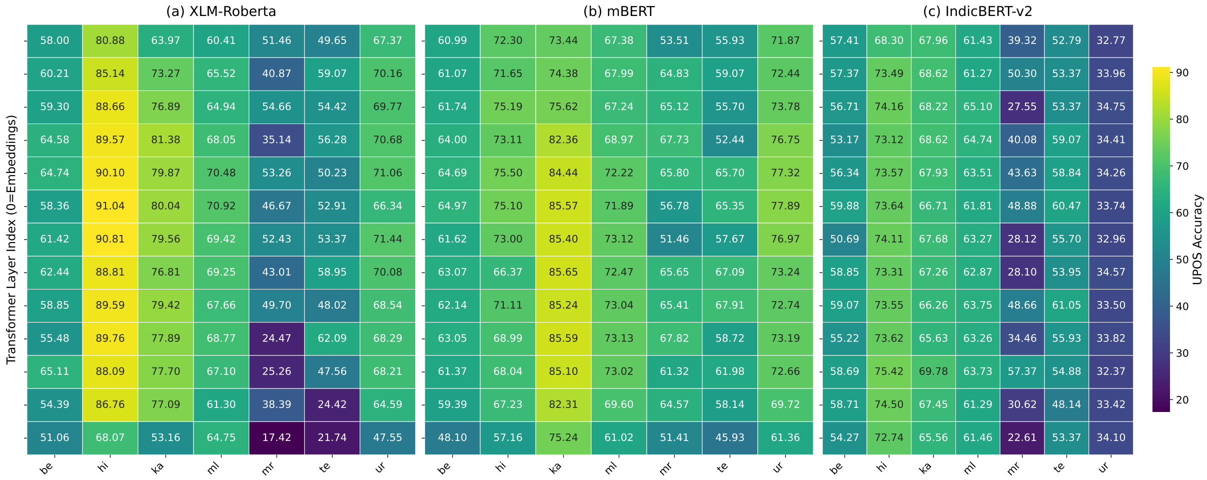

Figure 2: Layer-wise UPOS Accuracy (%) across Languages for (a) XLM-RoBERTa, (b) mBERT, and (c) IndicBERT-v2. Brighter colors indicate higher accuracy.

of layers aligns with prior probing studies on formalisms like Universal Dependencies (UD) (Jawahar et al., 2019; Tenney et al., 2019). Basic lexical information (UPOS tags) becomes accurately predictable in lower-to-middle layers. Morphological features (Vibhakti) also show strong representation across middle layers, their optimal encoding often overlapping with or slightly preceding layers most informative for syntactic structure. Crucially, complex Paninian dependency relations (LAS) consistently peak later, in the upper-middle layers (7-9), suggesting a hierarchical process where models integrate lexical/morphological cues to build syntactic representations. Performance generally degrades in final layers, possibly as representations specialize towards pre-training objectives.

**Model Architectures and Pre-training Influence:** While layer-wise trends are broadly similar, absolute performance and peak locations vary across models. XLM-RoBERTa and mBERT show comparable LAS capabilities on higher-resource languages like Hindi and Urdu. IndicBERT-v2, despite its Indic-focused pre-training, does not uniformly outperform general multilingual models in LAS for all tested languages, though it achieves strong LAS for Hindi. However, IndicBERT-v2 excels in stable Vibhakti prediction for Hindi/Urdu across most layers, potentially reflecting better morphological encoding due to its specialized training. This nuanced behavior suggests language-family specific pre-training can enhance morphological representation, but generalization to complex syn-

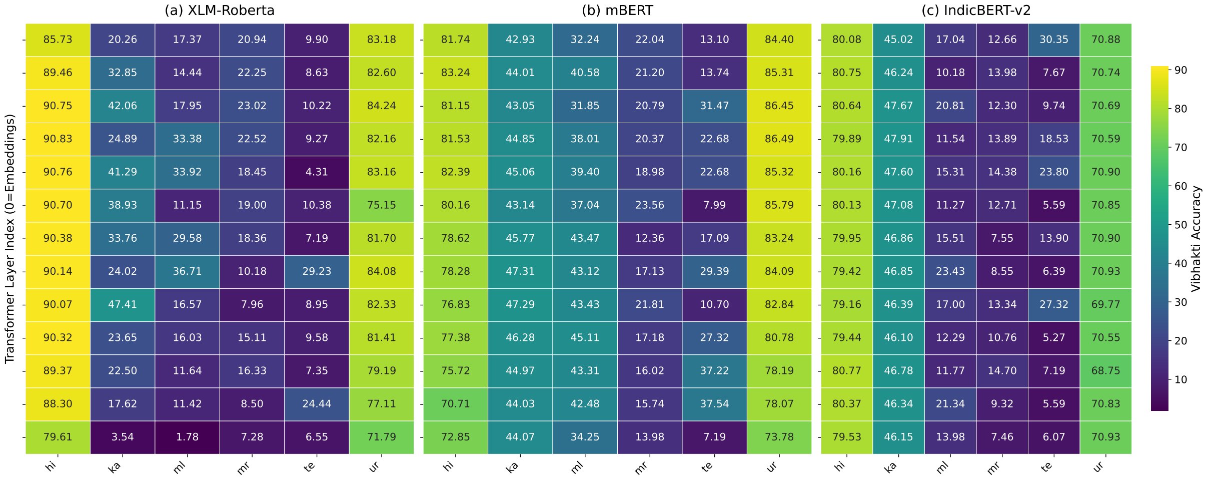

Figure 3: Layer-wise Vibhakti Accuracy (%) across Languages for (a) XLM-RoBERTa, (b) mBERT, and (c) IndicBERT-v2. Brighter colors indicate higher accuracy.

tax across diverse languages within that family remains challenging and interacts with other model properties.

**Cross-Lingual Consistency and Variation:** Significant cross-lingual variation in probing performance is evident for all tasks. While direct comparison is complicated by differing script representations and vocabulary coverage, Hindi and Urdu consistently yield the strongest results. The markedly lower performance for Telugu correlates with its smaller probing dataset size (Appendix B) and likely reflects underlying challenges from pre-training data sparsity or dataset quality. This underscores that probing performance reflects an interplay between information encoded by the base model and the characteristics of the probe training dataset.

**Encoding Paninian-Specific Information:** Our results demonstrate that diagnostic probes can successfully extract information pertinent to the Paninian grammatical framework – Kāraka-based dependency relations and Vibhakti features – from the frozen representations of these multilingual transformers. The ability to predict these suggests models implicitly learn representations sensitive to this distinct formalism, primarily consolidating structural knowledge in their middle layers.

## 5 Conclusion

We presented a layer-wise probing analysis comparing the encoding of Paninian grammatical information within XLM-RoBERTa, mBERT, and IndicBERT-v2 across seven Indian languages, examining dependency relations, UPOS tags, and Vibhakti features. Our findings reveal that Paninian dependency structure generally peaks in the upper-middle transformer layers, following lexical and morphological feature encoding in lower-to-middle layers, consistent with known patterns of linguistic representation.

Substantial cross-lingual and cross-model variations were observed. While IndicBERT-v2 showed strengths in Vibhakti prediction for core Indic languages, it did not uniformly surpass general multilingual models in representing Paninian dependency structures. Performance differences across languages correlate strongly with probing dataset sizes and likely reflect variations in pre-training data. Our results confirm that probing effectively reveals how theory-specific grammatical formalisms are represented within standard multilingual models.

## 6 Limitations

This study is limited to three base models and seven Indian languages; findings may not generalize broadly. Probe performance reflects both model representations and probe/dataset characteristics, with notable dataset size/quality variations (e.g., Telugu, Appendix B). The Paninian features probed (deprel, Vibhakti) are not exhaustive. Our analysis of frozen representations establishes correlations, not causal model mechanisms. Future work involves expanding model/language scope, probing more fine-grained Paninian features and studying the emergence of these representations.

## References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Rafiya Begum, Samar Husain, Arun Dhwaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2007. Dependency annotation scheme for indian languages. In *Proceedings of the Linguistic Annotation Workshop (LAW '07)*, pages 149–156, Prague, Czech Republic. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics (TACL)*, 7:49–72.

Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal. 1995. *Natural Language Processing: A Paninian Perspective*. Prentice-Hall of India, New Delhi.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. *Preprint*, arXiv:2212.05409.

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations (ICLR)*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

## A Background

Our probing analysis leverages the Paninian dependency annotation scheme, specifically developed for accurately representing linguistic phenomena within morphologically rich and relatively free word order ILs (Begum et al., 2007).

The Paninian grammatical tradition, originating from the ancient linguist Pāṇini, describes sentences primarily through modifier-modified dependency structures centered around the verb (Bharati et al., 1995). A crucial aspect of Paninian grammar is the use of Kārakas, specialized syntactico-semantic relations linking verbs and their arguments or modifiers. In the annotation scheme adopted here (Begum et al., 2007), six main Kārakas are identified:

1. *adhikaraṇa* (location)
2. *apādān* (source)
3. *sampradān* (recipient)
4. *karaṇa* (instrument)
5. *karma* (theme, loosely object-like)
6. *karta* (agent, loosely subject-like)

Importantly, Kārakas are not exact equivalents to purely semantic thematic roles (e.g., agent, patient). Instead, they encode a distinctly Paninian syntactico-semantic perspective. Consider for instance the English-like example 'key opened the door'. While semantically an instrument, the 'key' here would be annotated as the *karta* (loosely agent-like) Kāraka in the Paninian tradition (Bharati et al., 1995).

Identification of these Kārakas depends heavily on morphological indicators such as Vibhakti (case-endings, postpositions) and verb-based TAM markers (tense-aspect-modality) within sentences (Begum et al., 2007). The strong, systematic correlation between morphological features and syntactic dependency structures motivates our probing approach: we probe multilingual transformer models' representations for both the structural Kāraka relations (via dependency links and edge labels) and essential morphological cues (*Vibhaktis*, UPOS tags), analyzing explicitly how these linguistic representations are distributed layer-wise.

### A.1 Vibhakti

In the context of Paninian grammar and many modern Indian languages, *Vibhakti* refers to morphological markers, primarily case endings or postpositions, that are attached to nouns or noun phrases. These markers play a crucial role in signaling the grammatical function and syntactico-semantic role of the noun phrase within the sentence.

While often translated loosely as *case*, Vibhakti in the Paninian tradition is intimately linked to the concept of *Kārakas* (described in Appendix A). Specific Vibhaktis are typically associated with signaling specific Kāraka roles (e.g., a particular Vibhakti might commonly mark the *karta* 'agent-like' role, while another marks the *karma* 'theme-like' role, and others mark instrument, location, etc.). However, the mapping is not always one-to-one and can be influenced by other factors like verb semantics and sentence structure.

Essentially, Vibhaktis provide explicit surface cues about the underlying grammatical relationships in the sentence, making them particularly important in languages with relatively flexible word order where syntactic function is not solely determined by position.

## B Dataset Statistics

Table 1 provides statistics for the annotated datasets used in our probing experiments. Counts reflect the data *after* filtering sequences longer than 128 tokens but *before* any potential subsetting for development runs. Please note that the Bengali annotated data does not include Vibhakti features.

| Language | | Sentences | | Tokens | |
|---|---|---|---|---|---|
| Code | Name | train | val | train | val |
| be | Bengali | 3939 | 488 | 99504 | 12389 |
| hi | Hindi | 19855 | 2478 | 570772 | 71333 |
| ka | Kannada | 10388 | 1297 | 263827 | 32910 |
| ml | Malayalam | 7109 | 882 | 178754 | 21873 |
| mr | Marathi | 3608 | 451 | 91077 | 11461 |
| te | Telugu | 1514 | 185 | 19663 | 2219 |
| ur | Urdu | 4871 | 607 | 164129 | 21606 |

Table 1: Statistics of the Paninian–annotated datasets used for probing (post-filtering, pre-subsetting).