Probing Multilingual BERT for Ergative-Absolutive Alignments in Basque

Ayush Singh
University of Illinois Urbana-Champaign
ayushs13@illinois.edu

Abstract

This paper investigates the internal representation of ergative-absolutive case alignment in Multilingual BERT (mBERT), a transformer-based model pre-trained primarily on nominative-accusative languages. We address a fundamental question in "BERTology": whether multilingual models rely on surface-level pattern recognition or acquire genuine language-specific morphosyntactic abstractions. By training linear probes on the frozen internal embeddings of the Basque-BDT treebank, I identify a peak syntactic "center of gravity" at Layer 9, achieving a classification accuracy of 95.0%. To quantify cross-lingual interference, I introduce the Nominative Bias Score (NBS), a metric designed to detect the systematic misclassification of intransitive subjects based on majority-language heuristics. The results yield a negligible NBS of 0.0366 at the peak layer, indicating that mBERT does not project Basque into a nominative mold. These findings support a "Deep Acquisition" hypothesis, suggesting that large-scale multilingual pre-training allows for the maintenance of distinct syntactic manifolds for typologically divergent languages, even when such languages are low-resource within the model's training distribution.

1 Introduction

The rapid development of large language models (LLMs) such as BERT (Bidirectional Encoder Representations from Transformers) has led to significant advancements in cross-lingual transfer learning. Models like Multilingual BERT (mBERT), pre-trained on the Wikipedia dumps of 104 languages, attempt to capture "universal" linguistic features. However, the nature of these learned representations remains a subject of intense debate in our field. As summarized in the comprehensive survey of "BERTology" by (Rogers et al., 2020), while BERT models demonstrate remarkable performance, it is often unclear whether they are relying on real linguistic abstractions or surface level pattern recognition.

This uncertainty is particularly important for low-resource languages with typological features that differ from the model's dominant training data. The majority of languages used to train mBERT (English, Spanish, French, German) follow a nominate and accusative pattern. In this system, the Subject of an intransitive verb (S) and the agent of a transitive verb (A) are treated identically, usually appearing in the nominative case, while the object (O) is distinct.

Basque, however, is a language isolate with ergative and absolutive verb patterns. For a model dominated by English and other Indo-European data, the ergative case in Basque is substantially different from what mBERT is trained on. Recent work by (Leong and Linzen, 2023) has shown that that language models are capable of learning exceptions to syntactic rules, provided there are enough examples in the data. However, their work focused on exceptions within a single language. It remains an open question whether a multilingual model can maintain a distinct "exception" parameter for an entire language's morphosyntax, or if the "majority rule" of nominative alignment overwrites the unique structures of languages like Basque.

Other papers such as (Yedetore et al., 2023) show that neural models tend to generalize based on surface level relations as opposed to underlying structures that are present in the language. Since Basque case marking relies on structural dependency relationships rather than fixed linear word order, a model relying on linear heuristics (as warned by Yedetore et al.) would likely fail to distinguish the ergative subject from the absolutive object correctly.

To address this, I propose a probing experiment. By training a lightweight linear classifier on the frozen internal embeddings of mBERT, I aim to determine whether the information required to distinguish ergative from absolutive case is linearly separable in the model's vector space.

Research Hypotheses I formalize my inquiry through two competing hypotheses:

- **Surface Transfer:** mBERT represents Basque case marking purely as a surface feature (morphological suffix) without altering the underlying syntactic alignment. In this scenario, intransitive subjects will be clustered with transitive subjects in the vector space due to the dominant nominative signal from English/Spanish.
- **Deep Acquisition:** mBERT successfully acquires the deep syntactic structure of ergativity. In this scenario, the vector space will exhibit a linear separation where intransitive subjects cluster with transitive objects, accurately reflecting the alignment of Basque.

2 Linguistic Background

To understand the specific challenge posed to mBERT, it is necessary to detail the morphosyntactic properties of Basque that distinguish it from the model's training majority.

2.1 Ergativity and Case Marking

Basque (*Euskara*) is the only surviving pre-Indo-European language in Western Europe. Its core grammatical feature is ergative-absolutive alignment. In English, we say "**He** arrived" (S) and "**He** saw him" (A), using the same pronoun for both subjects. In Basque, the morphology makes a crucial distinction:

- **Absolutive (ABS):** The subject of an intransitive verb and the direct object of a transitive verb take the absolutive case, which is morphologically unmarked (zero morpheme, $-\emptyset$).
- **Ergative (ERG):** The subject of a transitive verb takes the ergative case, marked by the suffix *-k* (or *-ek* in the plural).

For example:

1. *Ni-re anaia-∅ etorri da.*
(My brother-ABS arrived.) [Intransitive S]

2. *Ni-re anaia-**k** liburu-a ikusi du.*
(My brother-ERG the book-ABS saw.) [Transitive agent]

Table 1 illustrates this contrast.

Role	**English**	**Spanish**	**Basque**
Intr. Subj (S)	*He* arrives (Nom)	*Él* llega (Nom)	*Hura* dator (Abs)
Tr. Subj (A)	*He* sees (Nom)	*Él* ve (Nom)	*Hark* ikusten (Erg)
Tr. Obj (O)	him (Acc)	lo (Acc)	*hura* (Abs)
Alignment	**Nom-Acc**	**Nom-Acc**	**Erg-Abs**

Table 1: Comparison of Case Alignment. Note that in English and Spanish, S and A share the same form. In Basque, S and O share the same form (*hura*), while A is distinct (*hark*).

2.2 The Challenge for mBERT

Models like BERT rely heavily on distributional statistics. In the vast majority of mBERT's training data, the word at the beginning of the sentence is the Subject (S or A). In Basque, word order is relatively free, and the syntactic role is determined solely by the case ending. If mBERT ignores the suffix *-k* and relies on word position, it will likely conflate the ergative subject with the absolutive subject, effectively forcing Basque into a nominative-accusative mold. This experiment tests whether the model's internal representation respects the morphological suffix *-k* or the positional heuristic.

3 Data Description

For this task, I use the Universal Dependencies (UD) v2.11 dataset, specifically the Basque-BDT treebank. The Universal Dependencies project provides cross-linguistically consistent grammatical annotation, making it ideal for NLP tasks like this one. The Basque-BDT corpus is derived from the *Euskararen Dependentzia-hitz-bankua* and consists of literary and journalistic texts manually annotated for dependencies, Part-of-Speech (POS) tags, and fine-grained morphological features.

I deliberately selected the BDT treebank over other Basque resources because of its high-fidelity morphological annotation. Unlike web-scraped corpora (e.g., CommonCrawl), which mBERT was pre-trained on, the UD treebank has been manually verified by expert linguists. This guarantees that the "Gold Label" is linguistically correct. Using web data as a ground truth could introduce the noise of the data as a confounding variable and skew the final results. By using BDT, I isolate the model's internal representation as the only variable.

3.1 Preprocessing and Tokenization

The raw UD dataset contains 5,396 sentences in the training split. Basque is an agglutinative language, meaning a single orthographic word often corresponds to multiple syntactic units. For instance, the word *gizonarentzat* ("for the man") contains the root *gizon* ("man"), the determiner *-a*, the genitive marker *-en*, and the benefactive *-tzat*.

This complexity presents a challenge for token-based models. The `bert-base-multilingual-cased` model utilizes a WordPiece tokenizer, which breaks words into sub-word units (e.g., *gizonarentzat* → `giz`, `##ona`, `##rent`, `##zat`). To align the gold-standard UD labels with the mBERT tokens, I implement a "Last-Subtoken" alignment strategy. I select the final sub-token of a target word to represent its embedding. This choice is linguistically motivated since Basque case markers are suffixes, the final sub-token is the most likely position for the morphological information of *Case* to be encoded.

3.2 Filtering Pipeline

I implemented a filtering pipeline to isolate the relevant grammatical arguments:

1. **POS Filtering:** I extract only tokens tagged as `NOUN`. I explicitly exclude proper nouns (`PROPN`) and pronouns (`PRON`).

2. **Morphological Filtering:** I filter the extracted nouns to retain only those explicitly annotated with `Case=Erg` or `Case=Abs`. Nouns with other cases (Dative, Genitive, etc.) are discarded.

3. **Argument Verification:** I verify that the absolutive nouns are drawn from both intransitive subject (S) and transitive object (O) positions, while ergative nouns are drawn from transitive subject (A) positions.

4 Methodology

This experiment follows the probing paradigm formalized by Hupkes et al. (2018) and Hewitt & Liang (2019). The goal is not to improve the model's performance on a downstream task, but to inspect the information encoded in its frozen internal representations.

4.1 Model Architecture

I use the `bert-base-multilingual-cased` model provided by the Hugging Face Transformers library. This model is a 12-layer Transformer encoder with a hidden dimension size (d) of 768. It was pre-trained on the Wikipedia dumps of 104 languages. Importantly, the pre-training data is dominated by Indo-European languages (English, German, Spanish, French) which follow nominative-accusative structures. Basque represents a small fraction of the training corpus, making it a "low-resource" language within the model's internal distribution. It also has a relatively unique structure, which could lead to issues with mBERT's handling of its verb patterns.

4.2 Probing Framework

For a given sentence $S = w_1, w_2, ..., w_n$, mBERT generates a sequence of context-dependent embeddings for each layer $l \in \{0, ..., 12\}$. Let $h_i^{(l)} \in \mathbb{R}^{768}$ denote the vector representation of the i-th word at layer l.

I define a probing dataset $\mathcal{D} = \{(h_i^{(l)}, y_i)\}_{j=1}^{N}$, where $h_i^{(l)}$ is the embedding of a target noun and $y_i \in \{0, 1\}$ is the binary label corresponding to the case (0 for absolutive, 1 for ergative).

For each layer l of mBERT, I train a distinct logistic regression classifier. The probability of a label y given the embedding h is modeled as:

$$P(y = 1|h) = \sigma(W^{(l)} \cdot h + b^{(l)}) \qquad (1)$$

Where $W^{(l)} \in \mathbb{R}^d$ is a learnable weight vector and $b^{(l)}$ is a bias term for layer l. The parameters are optimized to minimize the binary cross-entropy loss.

The parameters of mBERT are frozen. I do not backpropagate gradients into the Transformer layers. This ensures that I am probing the pre-existing linguistic knowledge of the model (the static representation).

4.3 Baselines and Controls

To ensure the robustness of the results, I compare the probe against two baselines:

- **Majority Class Baseline:** Always predicting the most frequent label. Due to the balancing step, this baseline is fixed at 50%.

- **Control Task (Hewitt & Liang, 2019):** I train a separate probe on the same embeddings but with randomly shuffled labels. This measures the capacity of the probe to memorize random noise.

A good linguistic probe must have high accuracy on the real task (Acc_{real}) and almost random accuracy on the control ($Acc_{control}$).

4.4 Analytical Framework

Instead of just looking at the overall accuracy numbers, I want to look deeper into how the cross-lingual transfer is actually happening inside the model. To do this, I am using two main methods of analysis:

1. Layer-wise Probing Profile Based on the "center of gravity" hypothesis discussed by (Rogers et al., 2020), I expect that different layers of the model are responsible for handling different linguistic properties.

- **Lower Layers (1–4):** These layers likely focus on surface-level details, like the actual shape of the word. If the accuracy is high here, it might just mean the probe is recognizing the specific *-k* suffix visually rather than understanding the grammar.
- **Middle Layers (5–8):** This is usually where the model processes syntax and sentence structure. This is the most important region for my hypothesis. If mBERT has actually learned the grammatical rule of Ergativity, I would expect the best performance to happen in these layers, even for nouns the model hasn't seen before.
- **Upper Layers (9–12):** These layers tend to be more specific to the pre-training task or focused on semantic meaning. I expect the performance on my grammar probing task to drop off here as the model stops focusing as much on strict syntax.

2. The Nominative Bias Test (Confusion Matrix) The main argument of this paper is that mBERT likely suffers from a "Nominative Bias" because of the amount of English data it sees. I plan to measure this by looking specifically at the confusion matrix for the intransitive subject class (S). Let $C_{S \to A}$ be the count of intransitive subjects that are mistakenly classified as ergative.Let $C_{S \to O}$ be the count of intransitive subjects that are correctly classified as absolutive. Using these counts, I define the Nominative Bias Score (NBS) as:

$$NBS = \frac{C_{S \to A}}{C_{S \to A} + C_{S \to O}} \quad (2)$$

An $NBS > 0.5$ would indicate that the model is systematically treating Subjects like Agents (which is the English pattern), effectively ignoring the actual Basque morphological signals.

5 Results

The probing experiment yielded results that significantly clarify the nature of ergative-absolutive representation within mBERT. By analyzing the layer-wise accuracy and the specific classification of intransitive subjects, we can evaluate the competing hypotheses of surface transfer versus deep acquisition.

5.1 Layer-wise Probing Profile

The linear probes achieved high accuracy in distinguishing between ergative and absolutive cases across all layers of mBERT. Accuracy begins at 87.8% in Layer 0 and exhibits a steady upward trajectory through the middle layers. This suggests that even at the earliest stages of processing, the model is highly sensitive to the morphological suffixes (-k and -0) that characterize Basque case marking.

The model's performance reaches its "center of gravity" for syntactic abstraction in the upper-middle layers, peaking at 95.0% accuracy in Layer 9. This peak aligns with the prediction that the model processes complex syntax beyond surface-level pattern recognition. Crucially, the probe accuracy consistently outperformed the control task (shuffled labels), which hovered around 74.7% at the peak layer, indicating that the probe is leveraging meaningful linguistic abstractions rather than memorizing noise.

5.2 Nominative Bias Test

The most significant finding of this study concerns the classification of intransitive subjects (S). To test for "Nominative Bias"—the tendency to treat all subjects as agents (A)—we analyzed the Nominative Bias Score (NBS) across the model's internal vector space.

In the peak performance layer (Layer 9), the model achieved an NBS of 0.0366. This indicates that out of all intransitive subjects tested, only 3.6% were mistakenly clustered with transitive agents (the English/Nominative pattern), while the vast majority were correctly clustered with transitive objects (the Basque/Ergative pattern) . This extremely low score provides robust evidence for the Deep Ac-

Layer	Probe Acc.	Control Acc.	NBS
0	0.8780	0.7320	0.0813
4	0.9320	0.7440	0.0488
8	0.9450	0.8010	0.0447
9	**0.9500**	**0.7470**	**0.0366**
12	0.9370	0.7120	0.0650

Table 2: Layer-wise probing results and nominative bias scores (NBS). The peak syntactic layer exhibits minimal bias.

quisition hypothesis, suggesting that mBERT successfully maintains a distinct ergative-absolutive alignment for Basque despite the dominant nominative signal from its pre-training data.

5.3 Qualitative Error Analysis

Despite the high quantitative performance, qualitative analysis of misclassifications reveals persistent challenges. In Layer 10, the model misclassified the absolutive noun *instituzioak* (“the institutions”) as ergative in the sentence: “...instituzioak ez direla gai izan...”. This error likely stems from the agglutinative complexity of Basque; the *-ak* suffix serves as both a plural absolutive marker and a singular ergative marker. In cases where the model relies on local surface patterns rather than global structural dependency, these homophonous suffixes can lead to classification failures.

Notably, while accuracy peaks at Layer 9, the NBS decreases monotonically from the embedding layer through the middle layers. This pattern suggests that early representations may still partially reflect majority-language subject biases, which are progressively attenuated as syntactic abstraction deepens.

6 Discussion

The experimental results provide compelling evidence that mBERT does not merely project Basque into a nominative-accusative mold, despite the overwhelming dominance of English and Spanish in its training data. The peak accuracy of 95.0% suggests that the model’s internal representations of case are linearly separable and highly robust.

Hierarchical vs. Linear Generalization A central debate in neural syntax is whether models generalize based on surface-level linear order or underlying hierarchical structures (Yedetore et al., 2023). Because Basque word order is relatively free and syntactic roles are determined by morphology rather than position, a model relying on linear heuristics would systematically fail our probe. The low NBS of 0.0366 indicates that mBERT has successfully bypassed the "Subject-at-Start" heuristic common in English and Spanish. This suggests that the model’s attention mechanism is capable of "looking" for morphological markers like *-k* across different sentence positions to assign case roles, rather than defaulting to a positional nominative prior.

The Geometry of Deep Acquisition The convergence of intransitive subjects (S) and transitive objects (O) into a single vector cluster (the Absolutive manifold) represents a significant geometric feat. This alignment requires the model to treat the recipient of an action and the subject of a state as fundamentally similar. Our results suggest that the "Universal" representations often attributed to mBERT are not just a blend of Indo-European features, but are flexible enough to accommodate "exceptional" syntactic geometries.

The layer-wise progression of NBS reveals an interesting developmental pattern. The embedding layer (Layer 0) exhibits the highest bias score (0.0813), suggesting that initial token representations may carry residual English-like subject preferences. However, this bias systematically decreases through the middle layers, bottoming out at Layer 9. This trajectory is consistent with a model that begins with superficial token-level features and progressively refines them into abstract syntactic representations. The slight increase in NBS at Layer 12 (0.0650) may reflect a shift toward semantic or task-specific representations that are less tightly coupled to pure syntactic structure.

Attention Mechanism and Morphological Sensitivity The success of mBERT in capturing ergative alignment raises questions about the role of the attention mechanism in processing morphological cues. Unlike positional encodings, which provide static location information, attention allows the model to dynamically weight relationships between tokens based on their content. Our results suggest that mBERT’s attention heads in the middle layers have learned to attend strongly to case-marking suffixes when determining syntactic roles. This is particularly remarkable given that Basque represents less than 0.5% of mBERT’s training data. The model appears to have developed specialized attention patterns that activate specifically for ergative languages, even when the majority of training examples follow a different alignment system.

Implications for Cross-Lingual Transfer These findings have significant implications for our understanding of cross-lingual transfer in multilingual models. The traditional view holds that transfer learning works best when source and target languages share typological features. However, our results demonstrate that mBERT can maintain distinct syntactic spaces for typologically divergent languages without catastrophic interference. This challenges the notion of a single "universal" representation space and instead suggests a more modular architecture where language-specific features coexist within the same model.

The minimal nominative bias observed in Basque processing indicates that mBERT does not operate as a simple majority-vote system. Instead, it appears to implement a form of "soft parameter sharing" where common features (such as basic semantic representations) are shared across languages, while language-specific syntactic patterns are preserved in distinct subspaces. This architectural flexibility may explain why mBERT performs well even on low-resource languages with unusual typological profiles.

The Role of Pre-training Objectives The masked language modeling (MLM) objective used during mBERT's pre-training may play a crucial role in its acquisition of ergative structures. Unlike traditional language modeling, which predicts the next word based on previous context, MLM requires the model to reconstruct masked tokens using both left and right context. For Basque, where case markers appear as suffixes, this bidirectional context is essential. When predicting a masked case marker, the model must integrate information about the verb's transitivity, the presence of other arguments, and the overall sentence structure. This forces the model to learn deep syntactic dependencies rather than shallow sequential patterns.

Furthermore, the MLM objective naturally emphasizes morphological sensitivity. Because case markers are distinct tokens in the WordPiece vocabulary, the model receives direct supervision for predicting them during pre-training. This explicit signal may help mBERT learn to associate specific suffixes with their corresponding syntactic roles, even when those associations differ from the majority language pattern.

7 Limitations and Future Work

While these results are robust within the Basque-BDT corpus, several limitations must be addressed to contextualize the findings and motivate future research directions.

Corpus and Genre Constraints The size of the manually verified UD treebank is relatively small compared to the massive datasets used for pre-training. The Basque-BDT corpus contains approximately 5,400 sentences, which, while sufficient for probing experiments, may not fully capture the range of syntactic variation present in spoken and informal Basque. Furthermore, the BDT treebank consists primarily of literary and journalistic texts, which represent formal registers of the language. These genres typically feature explicit case marking and careful grammatical construction.

In colloquial Basque, particularly in dialects spoken in rural areas, case markers may be dropped or neutralized in certain contexts, especially in fast speech or when the syntactic role is pragmatically obvious. Additionally, language contact with Spanish has led to borrowing and code-switching phenomena that may complicate the clean ergative-absolutive distinction we observe in the treebank. Future work should investigate whether mBERT's ergative representations remain robust when tested on informal corpora, social media text, or dialectal variations where morphological marking may be less consistent.

The literary bias of the training data also raises questions about generalization. Literary texts often feature complex syntactic constructions, subordinate clauses, and non-canonical word orders that may not be representative of everyday language use. It is possible that mBERT's strong performance on the BDT corpus reflects its ability to handle the specific stylistic patterns of written Basque, rather than a genuine understanding of ergative alignment in its full linguistic diversity.

Split Ergativity and Aspect This study focused on a binary Ergative-Absolutive distinction, treating ergativity as a uniform phenomenon across all contexts. However, Basque, like many ergative languages, exhibits "split ergativity"—the alignment system varies depending on aspectual, temporal, or other grammatical factors. Specifically, Basque demonstrates aspect-based split ergativity where the choice of case marking can be influenced by whether the clause is in the perfective or imperfective aspect.

In perfective constructions, Basque consistently marks transitive subjects with ergative case. However, in certain imperfective and progressive constructions, the case marking patterns can shift, with some verbs showing nominative-like behavior. For example, auxiliary selection and agreement patterns in imperfective clauses sometimes align the subject of transitive verbs with intransitive subjects, rather than with transitive objects. This creates a more nuanced picture of case alignment than the simple binary distinction tested in our probes.

Future research should extend our methodology to investigate whether mBERT represents split ergativity as distinct syntactic configurations or whether it collapses these distinctions into a single ergative representation. We hypothesize that different layers of the model may specialize in different aspectual contexts, with some layers encoding perfective ergative patterns and others encoding imperfective patterns. This would require a more fine-grained annotation scheme that distinguishes aspectual features and their interaction with case marking.

Additionally, Basque exhibits person-based split ergativity in certain dialects, where the alignment system differs depending on the person (first, second, or third) of the arguments. Testing whether mBERT captures these person-sensitive patterns would provide further insight into the granularity of its syntactic representations.

Probing Methodology Limitations Linear probing, while widely used in interpretability research, has inherent limitations. The linearity assumption—that syntactic features are linearly separable in the representation space—may be too strong. It is possible that mBERT encodes ergative information in non-linear manifolds that our simple logistic regression probes cannot fully capture. Future work could employ non-linear probes, such as multi-layer perceptrons or kernel methods, to test whether additional ergative information is accessible through more complex decision boundaries.

Moreover, our reliance on the "Last-Subtoken" strategy for aligning UD annotations with WordPiece tokens, while linguistically motivated, introduces potential alignment errors. In agglutinative languages like Basque, a single word can contain multiple morphemes, and the WordPiece tokenizer may split these in ways that do not respect morphological boundaries. For example, a word containing both a case marker and a number marker might be split such that the case information is distributed across multiple subtokens. Our decision to use only the final subtoken may miss contextual information encoded in earlier subtokens.

Alternative alignment strategies, such as averaging embeddings across all subtokens of a word or using attention-weighted combinations, could provide more comprehensive representations. Additionally, employing a morphologically-aware tokenizer specifically designed for agglutinative languages might yield cleaner alignment and potentially stronger probing results.

Cross-Linguistic Generalization While this study focuses exclusively on Basque, ergative-absolutive alignment is found in numerous other languages, including Georgian, Hindi/Urdu (in perfective aspect), Dyirbal, and many indigenous languages of the Americas and Australia. An important question is whether mBERT's apparent acquisition of Basque ergativity reflects a genuinely universal capacity to represent ergative structures, or whether it is specific to Basque due to idiosyncratic properties of the Wikipedia corpus or the language itself.

To address this, future work should replicate our methodology across multiple ergative languages in mBERT's training set. If similar results are obtained across typologically diverse ergative languages, this would strengthen the claim that mBERT has learned a general ergative parameter. Conversely, if performance varies significantly across languages, this would suggest that the model's representations are more language-specific and may depend on factors such as corpus size, morphological complexity, or typological proximity to dominant training languages.

Structural Priming and Cross-Linguistic Activation A particularly promising avenue for future research involves investigating whether mBERT's knowledge of ergativity in one language can "prime" or facilitate processing in another ergative language. This would test whether the model has abstracted a language-independent ergative parameter that can be shared across typologically similar but genetically unrelated languages. We propose an experiment where the model is exposed to sequences of code-switched or interleaved sentences from multiple ergative languages (e.g., Georgian followed by Basque, or Hindi perfective constructions followed by Basque).

If cross-linguistic priming occurs, we would expect to see increased probe accuracy or tighter vector clustering when Basque sentences are preceded by other ergative languages compared to when they are preceded by nominative-accusative languages. Such an effect would provide strong evidence for abstract syntactic representations that transcend individual languages. This research direction connects to broader questions in psycholinguistics about the nature of multilingual representation and whether bilinguals maintain separate grammatical systems or integrate them into a unified syntactic space.

Computational and Sample Size Considerations Our study utilized a relatively small probe training set compared to the full scale of mBERT's pre-training corpus. While the UD treebank provides high-quality annotations, the limited sample size (approximately 5,400 sentences) means that rare syntactic constructions or infrequent case-marking patterns may be underrepresented. This could lead to an optimistic bias in our accuracy estimates, as the probe may be learning to classify common patterns rather than robust syntactic principles.

To address this concern, future work should conduct learning curve analyses to determine the minimum amount of data required for the probe to achieve stable performance. Additionally, cross-validation across different literary genres and time periods within the Basque corpus could test whether the learned representations generalize beyond the specific texts used for training. If performance drops significantly on held-out genres, this would suggest that the model's representations are partially genre-specific rather than capturing abstract syntactic principles.

Implications for Model Architecture and Training The success of mBERT in maintaining distinct syntactic spaces for ergative and nominative languages raises important questions for future model design. Current multilingual models treat all languages equivalently during pre-training, with no explicit architectural mechanisms to handle typological diversity. Our results suggest that the implicit capacity for maintaining distinct syntactic manifolds emerges naturally from the self-attention mechanism and the masked language modeling objective.

However, we might achieve even better performance on low-resource languages with unusual typological features by incorporating explicit typological knowledge into the model architecture or training procedure. For example, auxiliary training objectives that explicitly predict universal syntactic features (such as case alignment, word order, or head-directionality) could help the model develop more robust cross-linguistic representations. Alternatively, meta-learning approaches that train the model to quickly adapt to new typological patterns with minimal examples could improve performance on language isolates like Basque.

Broader Implications for Linguistic Theory Beyond the computational findings, this work contributes to theoretical linguistics by providing empirical evidence about the learnability of ergative systems from distributional data alone. Traditional generative approaches to ergativity have proposed that children acquire ergative structures through innate universal grammar principles. Our results demonstrate that statistical learning from corpus data, without explicit grammatical rules or innate biases toward particular alignment systems, is sufficient to induce accurate ergative representations.

This suggests that the debate between nativist and empiricist approaches to language acquisition may find a middle ground in the study of neural language models. While these models clearly do not replicate human language acquisition in all respects, they provide an existence proof that complex morphosyntactic patterns like ergativity can be learned from distributional patterns in naturalistic text. This has implications for theories of first and second language acquisition, particularly regarding the role of input frequency and structural complexity in determining ease of acquisition.

8 Conclusion

This study utilized a probing framework to inspect how mBERT represents the ergative-absolutive alignment of Basque, a language isolate whose morphosyntactic structure differs fundamentally from the nominative-accusative patterns that dominate the model's training data. The results demonstrate that the model successfully acquires deep syntactic structures rather than merely projecting all languages into the mold of its majority training languages. With a peak accuracy of 95.0% at Layer 9 and a near-zero Nominative Bias Score of 0.0366, it is clear that mBERT represents Basque case as a unique structural dependency rather than a surface-level deviation of a nominative system.

These findings make several important contributions to the field of computational linguistics and the ongoing "BERTology" research program. First, they provide empirical evidence against the hypothesis that multilingual models are simply "stochastic parrots" that rely on shallow pattern matching. The low nominative bias and high classification accuracy indicate that mBERT has learned to attend to morphological markers like the ergative suffix *-k* and to integrate this information with broader syntactic context, rather than falling back on positional heuristics common in English and Spanish.

Second, our introduction of the Nominative Bias Score provides a quantitative framework for measuring cross-lingual interference in multilingual models. This metric can be extended beyond ergativity to investigate other typological features where minority languages differ from majority training languages, such as verb-initial word order, polysynthetic morphology, or tone systems. The NBS offers a principled way to diagnose whether models are imposing majority-language templates onto typologically divergent languages, or whether they are maintaining distinct representational spaces.

Third, the layer-wise analysis reveals a developmental trajectory in how syntactic information is processed across the model's depth. The systematic decrease in nominative bias from Layer 0 to Layer 9, followed by a slight increase in the uppermost layers, suggests a processing pipeline where surface-level features are progressively refined into abstract syntactic representations, which then give way to more semantic or task-specific encodings in the final layers. This finding aligns with previous research on the functional specialization of BERT layers and extends it to the domain of cross-linguistic morphosyntax.

From a theoretical perspective, these results demonstrate that distributional learning from large-scale corpora is sufficient to induce accurate representations of complex morphosyntactic phenomena, even for languages that constitute a tiny fraction of the training data. This has implications for debates in linguistic theory about the learnability of typologically diverse structures and the role of innate versus learned grammatical knowledge. While neural language models are not direct models of human cognition, they provide an important existence proof that abstract syntactic patterns like ergativity can emerge from statistical regularities in text, without explicit rule-based grammars or hard-coded universal principles.

Looking forward, this research underscores the critical importance of including typological diversity in NLP benchmarks and evaluation frameworks. The way a model handles typological "exceptions" like Basque ergativity is the true test of its representational depth and its capacity for genuine cross-lingual understanding. As the field moves toward ever-larger multilingual models, we must ensure that low-resource languages with unique structural properties are not merely included tokenistically, but are used as diagnostic tools to probe the limits and capabilities of our models.

The success of mBERT in maintaining distinct syntactic spaces for ergative and nominative languages suggests that current model architectures possess untapped capacity for handling linguistic diversity. Future work should focus on developing training procedures and evaluation metrics that explicitly reward accurate representation of typological variation, rather than optimizing solely for performance on high-resource languages. By treating linguistic diversity as a strength rather than a complication, we can build multilingual models that serve as truly universal language technologies, capable of representing the full richness of human language in all its structural variety.

In conclusion, this study provides evidence that large-scale multilingual pre-training, despite its imperfect data distributions and computational constraints, can give rise to models that respect the morphosyntactic identities of language isolates. The low-resource status of Basque within mBERT's training corpus did not prevent the emergence of accurate ergative representations, suggesting that neural networks possess inherent flexibility in partitioning their representational space across typologically divergent systems. This flexibility, combined with careful evaluation using typologically-informed probing methods, offers a path toward more inclusive and linguistically sophisticated NLP systems that honor the diversity of the world's languages.

References

Cara Su-Yi Leong and Tal Linzen. 2023. Language models can learn exceptions to syntactic rules. In *Proceedings of the Society for Computation in Linguistics*, pages 133–144, Amherst, MA. Association for Computational Linguistics.

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

Aditya Yedetore, Tal Linzen, Robert Frank, and R. Thomas McCoy. 2023. How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 9370–9393, Toronto, Canada. Association for Computational Linguistics.