Lost in Translation, and Found: Detecting and Interpreting Translation Effects

Shira Wein[1], **Anna Serbina**[2], **Jiyuan Ji**[3], **Nathan Wolf**[4]
Jason DeGraaff[5], **Prajakta Kini**[6], **Maria Leonor Pacheco**[6]
[1]University of South Florida, [2]University of Southern California, Information Sciences Institute
[3]Amherst College, [4]University of Massachusetts Amherst, [5]Proof School, [6]University of Colorado Boulder
[1]shirawein@usf.edu

Abstract

Translationese refers to the statistical patterns that distinguish translated texts from original texts, which are often subtle and imperceptible to human readers. When translated texts appear in either training or testing data, these patterns can negatively affect model performance or warp model evaluation. We approach the task of discerning whether a text was originally written in English or translated into English by fine-tuning contemporary foundation models at distinct item lengths and achieve state-of-the-art performance (94% Macro F1). Given that these linguistic cues are subtle and often imperceptible to humans, we analyze the features which enable our model's high performance. Employing a suite of interpretability-based techniques, we find that: (1) our high accuracy is enabled by a collection of linguistic features, a number of which correspond with linguistic theories of translationese, and (2) pretrained neural models are adept at picking up these features without any fine-tuning.

1 Introduction

Translationese (or translation effects) describes the underlying statistical patterns, including certain lexical and grammatical characteristics, that distinguish translated texts from texts originally written in the language. These characteristics can be caused by imprints from both the source language and the translation process itself (Koppel and Ordan, 2011). Although these cues are subtle and do not reflect a low-quality translation, the presence of translationese can affect downstream applications when translated texts are incorporated into model training or evaluation stages (Zhang and Toral, 2019; Freitag et al., 2019; Graham et al., 2020; Riley et al., 2020; Artetxe et al., 2020; Yu et al., 2022; Ni et al., 2022; Wang et al., 2023; Doshi et al., 2024). Therefore, translationese classification, the task of determining whether a text was originally written in the language or translated into that language, is a critical step in ensuring that these texts do not appear in training or test sets.

While humans can only perform this classification with approximately 50% accuracy (Wein, 2023), prior research has achieved greater success, with feature-based approaches achieving an accuracy of over 80% (Ilisei et al., 2010), and most recently a pretrained BERT-based translationese classifier achieving an accuracy of up to approximately 92% (Amponsah-Kaakyire et al., 2022). In this work, we investigate which foundation models are able to be most successfully fine-tuned for English translationese classification by fine-tuning nine foundation models on translated and original English data. Further, we assess the role of context in the models' ability to classify translation by measuring performance across *chunk* sizes: the number of sentences included in each training and testing item. We find that the model's performance is highly dependent on the choice of the foundation model, with ELECTRA achieving state-of-the-art results (over 94% Macro F1 score) and that incorporating contextual information via moderate chunk sizes (5-10 sentences) is crucial for maximizing classification performance.

Given our model's remarkably high performance, a stark contrast with both human performance and feature-based classifiers on similar data, we analyze the underlying patterns that the model is using to distinguish original from translated English texts. Using a suite of interpretability techniques, including feature analyses, LIME token-level saliency (Ribeiro et al., 2016), SHAP values (Lundberg and Lee, 2017), discourse-based linguistic analysis, Integrated Gradients saliency, and linear probing, we reveal key insights into both the model and the characteristics of translated language.

We find that no one linguistic cue enables high classifier performance; rather, a mosaic of features promotes accuracy, and we find evidence that supports literature on translationese from linguistics

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17172–17187

about suspected features of translated texts. Interestingly, our experiments reveal that the best-performing pretrained foundation model is already adept at discerning original English, before any translationese-focused fine-tuning.

Our contributions include:[1]

- fine-tuned translationese classification models based on nine underlying foundation models (Section 4.1);
- an investigation into chunking strategies–the number of sentences included in each training item for translationese (Section 4.2); and
- interpretability-based analyses of the features that distinguish original from translated English (Section 4.3).

2 Related Work

Effects of Translationese. While translationese has some linguistic markers (Volansky et al., 2013) such as source language shining-through, reduced lexical richness, and pragmatic explicitation (Osmelak et al., 2025), the visible effects of the translation process are subtle and largely undetectable to humans (Tirkkonen-Condit, 2002; Wein, 2023). In spite of these subtle cues, translationese has a marked impact on downstream performance and on model evaluation.

The presence of translationese in machine translation (MT) test data results in a gap between automatic and human evaluations, which is more pronounced when systems for translating into the target language are weaker (Zhang and Toral, 2019). Ni et al. (2022) further show that these evaluation differences can cause reversed MT system rankings; therefore, they recommend that translationese not be included as gold references in the test data. Translationese also has an impact on MT model performance, as discrepancies in translation direction between training and test sets can harm both model learning and performance across the dataset (Zhu et al., 2024).

Translationese in summarization test sets also leads to a gap in automatic and human evaluations, and the presence of translationese in training data harms cross-lingual summarization performance (Wang et al., 2023). Similarly, including translationese in model training negatively impacts model performance in multilingual question answering, but can be mitigated by reducing the presence of translationese in training data (Yu et al., 2022).

Finally, Artetxe et al. (2020) demonstrate that translationese in test data impacts the validity of cross-lingual transfer learning for multilingual natural language inference, and Doshi et al. (2024) finds that including machine translations as synthetic data in small language models reduces downstream performance across a range of tasks (such as headline generation and multilingual natural language inference).

Translationese Classification. Prior work on translationese has shown that it can be reliably detected using text classification techniques. Early methods use feature-based models, such as SVMs with handcrafted features (e.g., lexical richness, sentence length, and grammatical-to-lexical word ratios (Volansky et al., 2013)), achieving over 80% accuracy (Ilisei et al., 2010).

More recent efforts focus on neural models, with Amponsah-Kaakyire et al. (2022) finding that their most accurate model, a pretrained BERT model fine-tuned on European Parliament data (Amponsah-Kaakyire et al., 2021), achieves 92% accuracy. We also use a pretrained BERT-based classifier fine-tuned on European Parliament data and treat this as our baseline model in this work. Amponsah-Kaakyire et al. (2022) analyze their models by feeding an SVM BERT-based representations, finding that neural classifiers perform better because of the features they learn, and handcrafted features only provide a subset of the translationese patterns that BERT picks up on. In this work, we set out to uncover these features which enable our neural classifier's high performance.

2.1 Translationese Mitigation

Rather than focusing on detection, several studies have focused on mitigating translationese effects to improve downstream NLP tasks. For example, Riley et al. (2020) use zero-shot translation between translationese and original texts; Jalota et al. (2023) apply self-supervised style transfer models; Wein and Schneider (2024) use Abstract Meaning Representation as an intermediate representation to newly generate the text; Kunilovskaya et al. (2024) prompt GPT-4 (OpenAI, 2024) to rewrite the translated text; and Li et al. (2025) propose training-aware mitigation strategies to remove biases introduced in the fine-tuning of large language models. These mitigation techniques attempt to

[1]Our models and code are available at https://github.com/ACNLPlab/translationese-classification.git

make the transformed texts indistinguishable from original text, which is evaluated by classifier accuracy (dropping to chance-level performance), and suffer from the prospect of unnecessarily changing the text itself, while classification offers an opportunity to simply remove the translated text.

3 Methods

We fine-tune foundation models to build translationese classifiers. Here we describe the training and test data (Section 3.1), the models used (Section 3.2), and the interpretability methods applied to analyze our top-performing model (Section 3.3).

3.1 Data

We use the ENNTT corpus (Nisioi et al., 2016), a parallel dataset containing original and translated speeches from the European Parliament. Our dataset comprises two main subsets: (1) **Original English**: utterances originally spoken in English by native English speakers, and (2) **Translated English**: utterances translated into English from a non-English European language. The translations are carried out by professional human translators working for the European Parliament, ensuring high-fluency and semantically faithful translations.[2] We balance the dataset to include equal amounts of original and translated text, with the translated portion also evenly distributed across the seven source languages, encouraging source language-agnostic detection of translationese. Our final dataset comprises 116,340 original English sentences and 116,340 translated English sentences. The translated subset includes 16,620 sentences each from the following source languages: Dutch, Greek, German, French, Spanish, Italian, and Portuguese. Each entry in the final dataset consists of the sentence text, the source language (`src`), and a binary label (`label`, translated=1 and original=0) indicating whether it is translated.

We use a 70/30 train-test split, where the number of sentences in each split varies slightly based on the chunk size (Section 4.2). While prior work primarily reports accuracy, we report Macro F1 score; given our perfectly balanced classes, these metrics are equivalent, ensuring direct comparability with previous benchmarks.

3.2 Models

To evaluate recent neural approaches that learn features directly from data, we fine-tune a suite of pretrained transformer models for translationese classification. All models are trained as binary classifiers and evaluated in chunks of four sentences.

First, we employ a variety of BERT-based models, including **BERT** (Devlin et al., 2019), **RoBERTa** (Liu et al., 2019) with dynamic masking and more data, **ALBERT** (Lan et al., 2020) with parameter sharing and reduced memory use, **ELECTRA** (Clark et al., 2020) trained via replaced token detection, **DeBERTa** (He et al., 2021) with disentangled attention and positional encoding. We also use another Transformer-based model, **XLNet** (Yang et al., 2019), which uses permutation-based autoregressive pretraining. These models are chosen for their strong performance on sentence-level classification tasks and their ability to capture complex syntactic and semantic features that may distinguish original text from translated text.

Next, we employ two types of Kolmogorov-Arnold Networks (KANs). KANs were recently introduced as alternatives to traditional Multi-Layer Perceptrons (Liu et al., 2024). They replace linear weights with learned univariate functions applied on edges, often enabling superior performance on various tasks (Liu et al., 2024). We use the **KA-Transformer** (Yang and Wang, 2024), which we adapt for textual tasks, and **KAN-GPT**, a KAN-based generative transformer (Ganesh, 2024).[3]

Finally, we explore a non-Transformer-based model: **Mamba** (Dao and Gu, 2024), a state space model that replaces the attention mechanism with a learned linear recurrence, enabling faster and more memory-efficient modeling of long sequences.

3.3 Interpretability Methods

To unveil why the model achieves such high accuracy in translationese classification, capturing what is possible beyond hand-engineered features, we apply a suite of interpretability methods. All experiments are conducted using our best-performing model (which is our fine-tuned ELECTRA model, see section 4). We begin by examining how classification outcomes relate to established indicators of translationese, including **pronoun count**

[2]While ENNTT also contains text uttered originally in English by non-native speakers, we exclude this data to clearly isolate translationese effects from general second-language interference. This ensures that the classifier is trained to detect patterns introduced by the translation process itself, rather than those caused by non-native fluency.

[3]We extend KAN-GPT by adding a simple classification head: a dropout layer followed by a linear layer that maps the final hidden state to the binary output classes.

and type-token ratio (TTR), following Wein and Schneider (2024), as well as **sentence length and Flesch–Kincaid readability scores** (Kincaid et al., 1975). Note that the Flesch–Kincaid score estimates the U.S. school grade level required to understand a sentence; higher values indicate more complex, harder-to-read text.

For further linguistic analysis, we leverage **Rhetorical Structure Theory** (RST; Mann and Thompson, 1987). RST proposes that sentences can be parsed into tree structures with leaf nodes, called "elementary discourse units" (EDUs), and internal nodes that are labeled with a coherence relation. Following Kim et al. (2024), we extract the discourse relations in original and translationese texts by parsing the texts into discourse motifs, which are recurring subgraph patterns that correspond to RST relations. Specifically, we investigate the smallest unit, a single triad motif (M3), and calculate the motif frequency-inverse document frequency (mf-idf) scores across the datasets to obtain the most prevalent motifs. We consider motifs with mf-idf scores that exceed at least one standard deviation above the mean.

Next, we perform a **SHapley Additive exPlanations (SHAP) analysis** (Lundberg and Lee, 2017), a feature-based interpretability method that quantifies feature importance by SHAP values. SHAP values are computed from averaging the Shapley regression values obtained from sampled permutations of the feature order, in which a higher SHAP value indicates the feature has a larger marginal contribution to the model prediction. We randomly choose 40 examples, 10 for each of the following quadrants: (1) true original and predicted original, (2) true original and predicted translationese, (3) true translationese and predicted original, and (4) true translationese and predicted translationese.

Then, we use two saliency-based measures. **LIME saliency maps** (Ribeiro et al., 2016) enable us to visualize which tokens most influence the model's predictions. Following Amponsah-Kaakyire et al. (2022), we employ **Integrated Gradients** (IG) for token-level feature attribution (Sundararajan et al., 2017). For an input text, IG computes the gradient of each class's predicted probability with respect to each input feature of the text (i.e., each embedding dimension of each token). These attributions are then aggregated within each token embedding to estimate the overall contribution of each token to the model's output. Unlike SHAP or LIME, IG does not treat the model as a black box, which allows for insight into the model's internal representations. Further, because IG targets the predicted probability of a specific class, it yields one saliency score per class per token, allowing for direct comparison of a token's influence on the prediction of each class.

Finally, we perform **Linear Probing** (Alain and Bengio, 2016). We adopt a two-stage probing approach to analyze how syntactic, lexical, and discourse features are encoded in transformer representations and their relationship to translationese. We train independent linear classifiers (probes) on frozen layer representations to assess what linguistic information is accessible at different depths of the model, probing the ELECTRA model's encoding of linguistic features across its 13 layers. We compare pretrained ELECTRA and our fine-tuned ELECTRA model. We train and test on the same splits as our fine-tuning experiments, and report Macro F1 scores as our primary metric due to feature imbalance.

We probe the RST motifs as well as three syntactic features: passive voice, pre-verbal complexity, and non-subject fronting, using Universal Dependencies annotations via UDPipe (Straka et al., 2016). A chunk is labeled as passive if the ratio of passive sentences to total sentences in the chunk meets or exceeds a threshold $\tau \in \{0.3, 0.4, 0.5\}$.[4] We measure pre-verbal dependency length as the number of syntactic tokens preceding the main verb in the dependency tree, because longer pre-verbal spans often indicate translated English due to source-language word order interference (Volansky et al., 2013). Fronted constituents where non-subject elements precede the standard SVO word order pattern of English (non-subject fronting) are reflective of source interference. Sentences with pre-verbal length ≥ 15 are marked as complex, which is the average maximum pre-verbal length for any sentence within a chunk.

4 Results

This section presents results on translationese classification across foundation models (Section 4.1) and chunk sizes (Section 4.2), along with findings from our interpretability analyses (Section 4.3).

[4]These three thresholds are chosen to capture varying degrees of discourse-level overuse of the passive voice.

Metric	BERT	RoBERTa	ALBERT	ELECTRA	DeBERTa	XLNet	KATransformer	KAN-GPT	Mamba
Macro F1	93.36 ± 0.23	93.57 ± 0.20	91.30	**94.46 ± 0.24**	92.11	93.11	79.44	80.40	84.95

Table 1: Macro F1 score (% out of 100) for each model on chunk size of four at epoch five. To obtain a more reliable estimate of the performance of the top three models (ELECTRA, RoBERTa, and BERT), we use ten-fold cross-validation (also on chunk size four, for five epochs). ELECTRA achieves highest Macro F1.

4.1 Foundation Model Comparison

We compare overall model performance on four-sentence chunks (Table 1). We select four-sentence chunks following prior work (Wein, 2023) uses approximately a paragraph of text or texts of 100-300 tokens, corresponding to approximately four sentences in our dataset. Among the BERT-based models, ELECTRA achieves the best results, with the highest test Macro F1 score (94.46%), demonstrating robust ability in identifying translationese. RoBERTa and BERT follow closely. BERT remains a strong baseline, attaining a competitive Macro F1 score of 93.36%. Our ten-fold cross-validation reveals that the ELECTRA model consistently outperforms BERT and RoBERTa on this task. DeBERTa and XLNet perform comparably, while ALBERT lags behind. Notably, the performance difference between the best and worst BERT-based models reaches approximately 5%, underscoring the impact of architectural and pretraining differences even under standardized input conditions. The KAN-based models also trail the BERT-based models, suggesting that while KAN's theoretical grounding in functional approximation is promising, its application to translationese classification may require further adaptation. Finally, while Mamba does not match the performance of top BERT-based models, it outperforms the KAN-based models and ALBERT, highlighting the potential of non-Transformer architectures for long-sequence modeling in translationese classification. Overall, these results suggest that pretraining strategies (e.g., ELECTRA's replaced token detection) and architectural nuances (e.g., ALBERT's parameter sharing) influence model performance. Models with stronger contextual understanding appear to be more capable of distinguishing translated text from original text, likely due to their ability to capture subtle stylistic and syntactic differences.

As our models achieve remarkably high classification accuracy, we now turn towards analysis of these results, with regard to chunk size and interpretability.

Chunk Size	Macro F1
1	81.82
2	88.74
3	92.53
4	93.73
5	94.40
6	**94.68**
7	94.48
8	94.07
9	93.95
10	94.21
11	93.57
12	93.11
13	92.92
14	93.80
15	93.25
16	92.32

Table 2: ELECTRA performance (Macro F1, % out of 100) across chunk sizes. Chunk size 6 yields the best performance.

4.2 Chunk Size Experiments

To further examine the role of the amount of data necessary to perform this classification, we evaluate ELECTRA classification performance on **chunks**: a concatenation of multiple consecutive sentences. For chunk-level data, we concatenate n consecutive sentences sharing the same label, with n ranging from 1 to 16 in different experiments.[5]

The performance trend (Table 2) reveals that moderate chunk sizes (5-10) generally outperform smaller or excessively large chunks. Notably, ELECTRA at chunk size 6 sets a new state-of-the-art, achieving a maximum test Macro F1 of 0.9468. Performance deteriorates at chunk sizes above 12, likely because larger chunks reduce the number of training examples, thereby limiting the model's exposure to diverse inputs during training. To account for the impact of the number of training items (as the number decreases with larger chunks), we repeat this experiment using a fixed number of training samples across all chunk sizes, and find that when dataset size is held constant, chunk sizes between 9 and 15 yield the best perfor-

[5]To create the chunks, we split the entire ordered dataset into fixed-size chunks. We then shuffle these chunks and divide them into training and test sets (same size chunks for both). This typically results in sentences from the same speech, though continuity across sentences is not guaranteed.

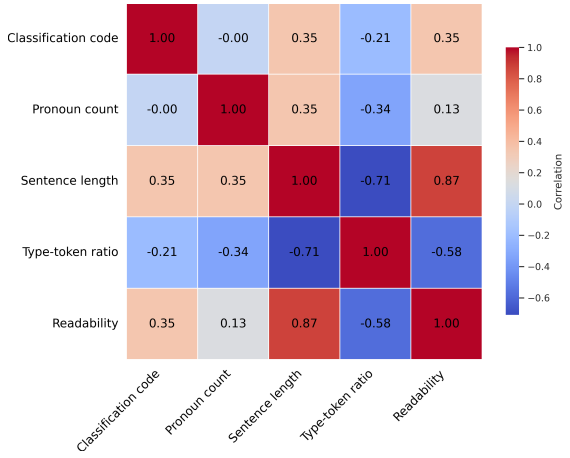

Figure 1: Correlation matrix illustrating the relationship between the classification code and key linguistic features, where the classification code denotes translated entries.

mance, indicating that longer passages offer more stable contextual cues that aid classification.

4.3 Interpretability Analyses

Highly fluent translations are indistinguishable to human readers from original texts, and feature-based classifiers perform far worse than ours, which begs the question: how does our fine-tuned ELECTRA model classify translated texts so accurately? To better understand what linguistic features enable neural translationese classification, we conduct a series of interpretability analyses. All experiments are conducted using our best-performing model, ELECTRA fine-tuned on chunk size six.

Our six interpretability techniques (Section 3.3) coalesce into two core findings. First, they provide empirical support for established linguistic accounts of translationese, showing that high classification performance arises from a mosaic of complementary features rather than a single dominant cue. Second, we find that the pretrained model picks up on many of these cues without the need for fine-tuning, suggesting that ELECTRA is already very good at discerning original English.

Finding 1: support for linguistic theories of translationese. We begin by examining how classification outcomes relate to features of translationese as described by linguistics literature, including measures of **lexical richness and complexity**. As shown in the correlation matrix (Figure 1), sentence length and Flesch–Kincaid score emerge as the strongest positive correlates with classification code (both 0.35).[6] Type-token ratio (TTR), in contrast, is negatively correlated ($r = -0.13$), suggesting that translated texts tend to be more lexically repetitive–as indicated by prior work. When examining the TTR scores for the translationese versus original classes overall, the original data has (only slightly) higher TTR than the translationese data, with an average original TTR of 0.0111, and an average translationese TTR of 0.0104. Category-level breakdowns reveal that correctly classified translated sentences tend to have longer sentence lengths, higher Flesch–Kincaid scores (indicating greater reading difficulty), and slightly lower TTR compared to those that were misclassified. However, all correlations are relatively weak, indicating that none of these surface-level features strongly predict whether a text is translated or original.

Examining the LIME results, the model often highlights **function words** and high-frequency lexical items when classifying texts as translated. When classifying original English sentences, LIME frequently highlights pronouns, contractions, or atypical phrasing. To further investigate the role of individual tokens in classification, we apply SHAP to 40 examples and find that function words and **cohesive markers**[7] have higher feature importance in texts predicted as translationese. Although the normalized frequency counts of function words and cohesive markers do not differ much between translated and original texts, function words, such as "this" and "that the," have the highest SHAP values (i.e., the features of most importance in the example) in nine out of 20 examples predicted to be translationese. "Therefore" and "however" are the most important features in two of the translationese examples. This importance placed on function words indicates a decrease in lexical richness in translated texts since the model is focusing on syntactic clues instead of semantically varied vocabulary, which is confirmed by the TTR analysis. On the other hand, none of the most important features in the original texts are function words.

Supporting the SHAP and LIME analyses, the IG token-level saliency scores also show an association between function words and translationese, with the function word tokens having a mean trans-

[6] Considering that label 1 corresponds to translated entries, positive values in the correlation matrix indicate stronger associations with translationese, while negative values suggest features more common in original texts.

[7] We use the list of function words and cohesive markers provided by Volansky et al. (2013).

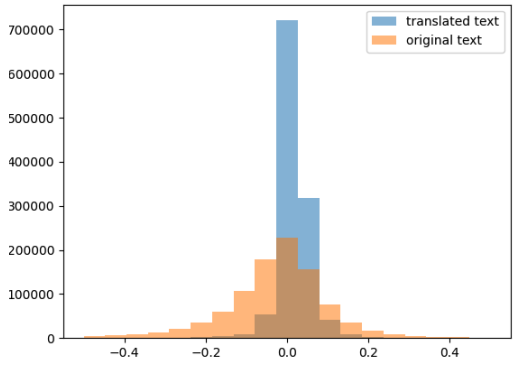

Figure 2: Distributions of saliency scores toward predictions of translated text for tokens in original text and tokens in translated text.

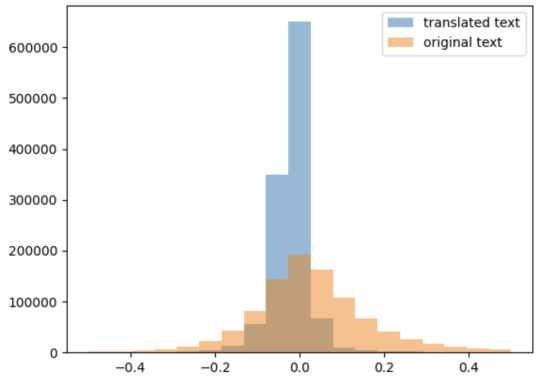

Figure 3: Distributions of saliency scores toward predictions of original text for tokens in original text and tokens in translated text.

lated text saliency score of 0.00615, higher than all tokens' mean translated text saliency score of -0.00606.

Through the IG token saliency scores, we observe an interesting pattern in the **distribution of token saliency scores**. Tokens in original text tend to receive higher magnitude attributions towards both classes, as shown in Figures 2 and 3.[8] Further, in translated text, 74% of tokens have a positive score towards translated text and a negative original English score; in original English text, only 58% of tokens have a positive score towards original English and a negative translated text score. Only 2% of tokens have the same sign for both IG class saliency scores. In short, the tokens in original text tend to have more extreme scores, with just over half attributed to original text, while the tokens in translated text tend to have low-magnitude scores, with most of them attributed to translated text (Figures 7 and 8, in Appendix A). These findings further support our understanding of lexical diversity as an indicator of original text, showing that the predicted probability of a text being translated is driven by many tokens with low-saliency translated text attributions, while the predicted probability of a text being original is driven by a smaller number of tokens with high-saliency original text attributions.

From our SHAP analysis, four of the 20 predicted translationese examples have highest SHAP values for **punctuation** (commas and periods). Investigating this through our RST discourse analysis, the most prevailing single triad discourse relations in translationese are the two types of same-unit relations, which have a mf-idf score difference of 0.002 and 0.001, respectively. The same-unit relation describes when a larger EDU is interrupted by a relative clause, e.g.:

> Mr President$_{\text{[same-unit]}}$, in relation to the second Alyssandrakis report$_{\text{[elaboration]}}$, I must mention my daughter$_{\text{[same-unit]}}$.

This offers a plausible explanation for punctuation marks identified as significant features of translationese, because commas are often used in place of same-unit relations. Indeed, Volansky et al. (2013) finds that the usage of the comma is 1.3 times more frequent in translationese compared to the original text.

Contrary to Volansky et al.'s (2013) hypothesis that passive voice is more excessively used in English than translated texts, we find that five out of 20 examples predicted to be translationese assign high SHAP values to passive voice. This may be due to the fact that Volansky et al. grounds their hypothesis primarily on a German-English dataset, whereas the true translationese examples with passive voice usage of high SHAP values are from Italian and Portuguese originally.

Our correlation analysis reveals that **syntactic features** show positive associations with translationese: namely passive voice ($\beta = +0.380$), preverbal complexity ($\beta = +0.495$), and non-subject fronting ($\beta = +0.201$) (performance for each features at the best threshold of 0.3 can be seen in Table 3, with full results in Table 4). The high encoding accuracy of the fine-tuned model for passive voice (81.41% MF1) compared to pre-verbal complexity (60.28% MF1) indicates that while both

[8] Amponsah-Kaakyire et al. (2022) investigate interpretability for translationese only through IG and like us found that there are some spurious correlations with location names, e.g., topic-related info. They also suggest that there may be punctuation-based spurious correlation in the data, but our investigation reveals that this is likely due to relations that are in fact indicative of translationese.

Feature	PT MF1	FT MF1	Corr. Coef. (β)	Predictive MF1
PV	81.27	81.41	+0.380	**53.7**
PVC	61.42	60.28	+0.495	**41.3**
NSF	78.29	78.25	+0.201	**51.9**

Table 3: Linear probing results for syntactic features (PV: Passive Voice, PVC: Pre-verbal Complexity, and NSF: Non-Subject Fronting) at a ratio of 0.3, for the pretrained ELECTRA model (PT) with all layers frozen and only the classification head trained on the ENNTT corpus, and our fine-tuned (FT) ELECTRA model. MF1: Macro F1 score at the highest scoring layer for each feature. Corr. Coef. (β): Logistic regression log-odds coefficient measuring the direction and strength of association between the syntactic feature and translationese. Predictive MF1: Macro F1 score for predicting translationese given syntactic feature labels as input.

features correlate with translationese, passivization is more consistently encoded in model representations, possibly because it involves clearer morphological markers.

Finally, the aforementioned individual lexical features (density of function words, pronouns, and cohesive markers) achieve translationese prediction performance up to a Macro F1 of 58.7% (Table 5). Combining all linguistic features via multi-feature logistic regression into a 9-dimensional classifier yields 59.58% Macro F1 (Table 7), above chance but well below our ELECTRA model performance. This performance reveals that our linguistically motivated features capture important but incomplete aspects of translationese. Neural classification models appear to leverage distributional patterns beyond interpretable linguistic features, potentially including not only surface-level features but also subtle fluency and syntactic cues.

Finding 2: pretrained ELECTRA is already a good translationese classifier. Our probing analysis reveals at which layers fine-tuned ELECTRA encodes the linguistic features we have encountered so far, and how this compares with feature encoding in pretrained ELECTRA, without task-specific fine-tuning. The layer-wise probing results (full results in Appendix A) reveal two distinct encoding patterns across feature types. The high Macro F1 scores achieved across most features, ranging from 61% for pre-verbal complexity to 93% for pronouns, demonstrate that ELECTRA encodes these linguistic properties well. For syntactic and discourse features (passive voice, pre-verbal complexity, non-subject fronting, and RST motifs), both pretrained and fine-tuned models exhibit similar trajectories: Macro F1 increases sharply from the embedding layer through layers 3–4, peaks at layers 5–7, maintains a relatively stable plateau through layer 11, then shows degradation at layer 12 (see e.g., the probing results for pre-verbal complexity, in Figure 4). This progression reflects the hierarchical nature of syntactic and discourse processing, where surface forms are gradually composed into higher-level structural representations.

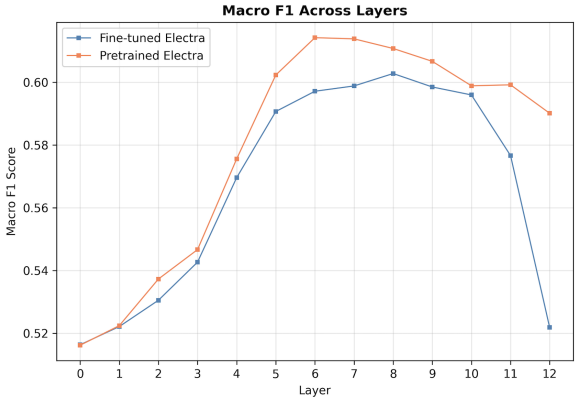

Figure 4: Layer-wise probing results for **Pre-verbal Complexity** with Ratio@0.3 aggregation. The plot shows Macro F1 scores across layers 0–12 for both pretrained (orange) and fine-tuned (blue) ELECTRA models. Best MF1: Pretrained: 61.42%, Fine-tuned: 60.28%.

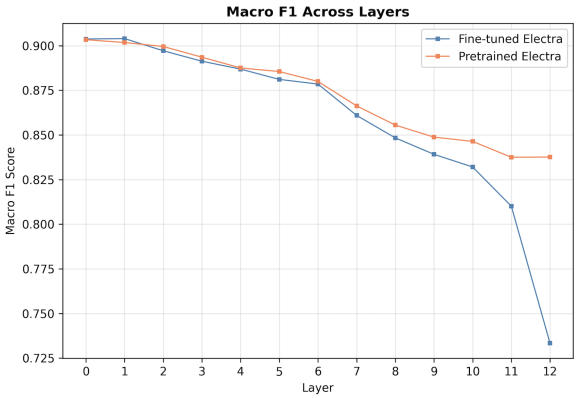

Figure 5: Layer-wise probing results for **Cohesive Markers** with 50th Percentile Density aggregation. The plot shows Macro F1 scores across layers 0–12 for both pretrained (orange) and fine-tuned (blue) ELECTRA models. Best MF1: Pretrained: 90.34%, Fine-tuned: 90.40%.

However, lexical features (function words, pronouns, cohesive markers) show markedly different behavior with encoding accuracy starting high at layer 0 (90–93% Macro F1) and steadily declining through deeper layers (see e.g., the probing results for cohesive markers in Figure 5). This behavior in-

dicates that lexical density patterns are most accessible in surface-level representations and become progressively compressed in higher layers.

Critically, while pretrained and fine-tuned models show similar encoding patterns through layer 11, they diverge dramatically at layer 12, with the fine-tuned model exhibiting sharp performance drops (e.g., passive voice: 78% to 71%, pre-verbal complexity: 59% to 52%) while the pretrained model remains stable. This final-layer collapse suggests that task-specific supervision through model fine-tuning prioritizes translationese classification in the final layers by compressing away auxiliary linguistic features that the pretrained model preserves through its general-purpose abstractions. This consistent encoding through intermediate layers, despite the models' different training objectives, explains why the pretrained ELECTRA model with only a trained classification head achieves 88.83% Macro F1 on translationese classification, with a slightly lower Macro F1 score than the 94.68% of the fully fine-tuned model.

To further examine the role of fine-tuning in ELECTRA's translationese classification performance, we employ Dodrio (Wang et al., 2021), an attention visualization tool, for a qualitative analysis on our test set. When comparing the attention heads producing the highest magnitude attention scores in the pretrained and fine-tuned ELECTRA models, we find that the ranking of attention heads by magnitude remains almost entirely the same before and after fine-tuning. In addition, many of the fine-tuned high-magnitude heads draw nearly identical connections (attention scores above a certain threshold) to their pretrained counterparts. This supports our finding that pretrained ELECTRA is able to capture many of the features necessary for translationese classification.

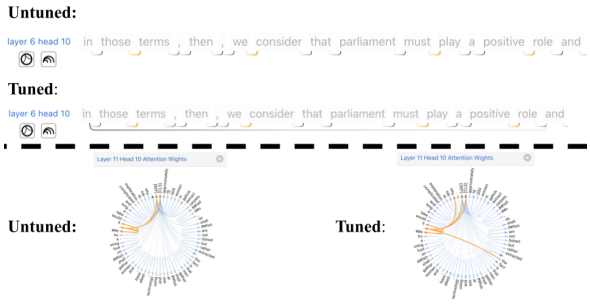

Figure 6: Dodrio visualizations of fine-tuned attention heads drawing additional long connections between function words and punctuation.

On the other hand, we observe one key area in which fine-tuned attention heads differ from pretrained attention heads: fine-tuned heads often relate function words and punctuation to other function words and punctuation, even when very far apart in the sentence (Figure 6). These observations further underscore the importance of function words and punctuation for translationese classification.

5 Conclusion

In this work, we perform translationese classification, the task of distinguishing whether English texts are translated or originally written in English, which has widespread impacts on NLP tasks broadly (any model that is pretrained or tested on translated data). We advance the state-of-the-art by achieving over 94% Macro F1 score with our best performing model, and explore the experimental variables of chunk size and foundation model. The presence of translationese impacts model performance and evaluation when included in model training or test data, and our high-accuracy classifier can be leveraged to filter for data in the desired original language, and translation direction if applicable.

Our interpretability analyses expand on statistically grounded analyses, providing critical insight into translationese classification. Our findings reveal that linguistically-grounded translationese features do play a role in translationese classification, but many subtle cues contribute to high classification performance, which is likely why humans are not able to perform this classification task well. Further, we find that the model picks up the majority of these cues on its own from pretraining, rather than task-specific fine-tuning, revealing that the model is able to encode translationese-related features well without adaptation.

Limitations

Our study presents strong results on translationese classification, but it is subject to several limitations. First, we follow related work by conducting all experiments on European Parliament data, which while high-quality and well-annotated, is a specific and formal domain.

Second, our chunking strategy, while effective in practice, is relatively coarse: it concatenates n consecutive sentences sharing the same label without enforcing that they originate from the same speech.

As a result, some chunks may combine sentences spoken by different individuals or in different contexts, potentially introducing noise, but improving performance regardless of speech consistency.

Third, although we identify chunk size six as optimal in our setting, it is unclear whether this generalizes to other datasets or domains. The interaction between model architecture, chunk size, and domain remains under-explored and should be investigated further.

Acknowledgments

We thank anonymous reviewers and members of the ACNLP lab for their helpful feedback. Thank you also to Ryan Hughes and Biance Delgado for their involvement in early discussions of this work. This work is supported by the Amherst College HPC, which is funded by NSF Award 2117377.

References

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

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Cristina España-Bonet, and Josef van Genabith. 2021. Do not rely on relay translations: Multilingual parallel direct Europarl. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 1–7, online. Association for Computational Linguistics.

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Josef Genabith, and Cristina España-Bonet. 2022. Explaining translationese: why are neural classifiers better and what do they learn? In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 281–296, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*.

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.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. Pretraining language models using translationese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5843–5862, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Aditya Nalgunda Ganesh. 2024. Kan-gpt: The pytorch implementation of generative pre-trained transformers (gpts) using kolmogorov-arnold networks (kans) for language modeling. Release 1.0.0, 9th May 2024.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In *Computational Linguistics and Intelligent Text Processing*, pages 503–511, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rricha Jalota, Koel Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2023. Translating away translationese without parallel data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7086–7100, Singapore. Association for Computational Linguistics.

Zae Myung Kim, Kwang Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. 2024. Threads of subtlety: Detecting machine-generated texts through discourse motifs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5449–5474, Bangkok, Thailand. Association for Computational Linguistics.

J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report 8-75, Naval Technical Training Command Millington TN Research Branch.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Maria Kunilovskaya, Koel Dutta Chowdhury, Heike Przybyl, Cristina España-Bonet, and Josef Genabith. 2024. Mitigating translationese with GPT-4: Strategies and performance. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 411–430, Sheffield, UK. European Association for Machine Translation (EAMT).

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yafu Li, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025. Lost in literalism: How supervised training shapes translationese in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12875–12894, Vienna, Austria. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. 2024. Kan: Kolmogorov-arnold networks. *Preprint*, arXiv:2404.19756.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

William C Mann and Sandra A Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical report, University of Southern California, Information Sciences Institute Los Angeles.

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Original or translated? a causal analysis of the impact of translationese on machine translation performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics.

Sergiu Nisioi, Ella Rabinovich, Liviu P. Dinu, and Shuly Wintner. 2016. A corpus of native, non-native and translated texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

OpenAI. 2024. Text generation and prompting.

Doreen Osmelak, Koel Dutta Chowdhury, Uliana Sentsova, Cristina España-Bonet, and Josef van Genabith. 2025. Pragextra: A multilingual corpus of pragmatic explicitation in translation. *Preprint*, arXiv:2511.02721.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365.

Sonja Tirkkonen-Condit. 2002. Translationese—a myth or an empirical fact?: A study into the linguistic identifiability of translated language. *Target. International Journal of Translation Studies*, 14(2):207–220.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Jiaan Wang, Fandong Meng, Yunlong Liang, Tingyi Zhang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2023. Understanding translationese in cross-lingual summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3837–3849, Singapore. Association for Computational Linguistics.

Zijie J. Wang, Robert Turko, and Duen Horng Chau. 2021. Dodrio: Exploring transformer models with interactive visualization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 132–141, Online. Association for Computational Linguistics.

Shira Wein. 2023. Human raters cannot distinguish English translations from original English texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12266–12272, Singapore. Association for Computational Linguistics.

Shira Wein and Nathan Schneider. 2024. Lost in translationese? reducing translation effect using Abstract Meaning Representation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 753–765, St. Julian's, Malta. Association for Computational Linguistics.

Xingyi Yang and Xinchao Wang. 2024. Kolmogorov-arnold transformer. *Preprint*, arXiv:2409.10594.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32.

Sicheng Yu, Qianru Sun, Hao Zhang, and Jing Jiang. 2022. Translate-train embracing translationese artifacts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 362–370, Dublin, Ireland. Association for Computational Linguistics.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. 2024. Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 388–409, Miami, Florida, USA. Association for Computational Linguistics.

A Additional Figures

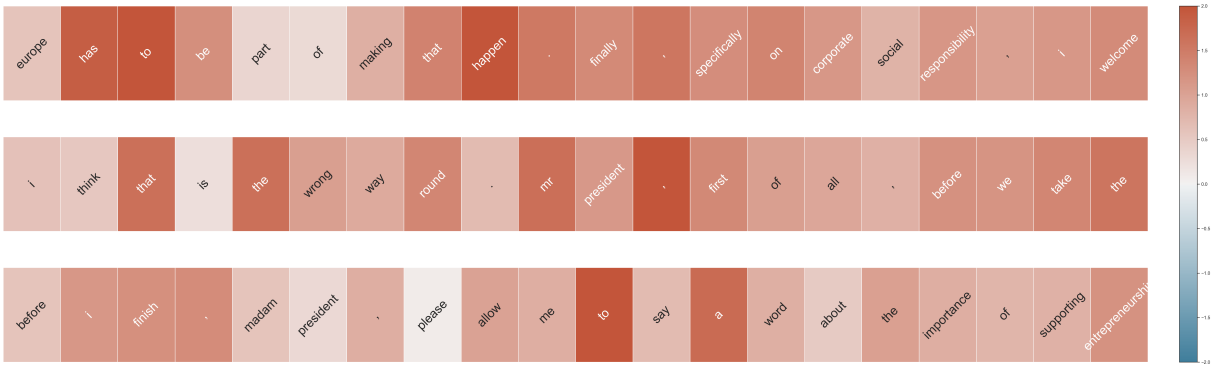

Figure 7: Saliency scores toward predictions of original text in the 3 sentences with the highest average original text saliencies.

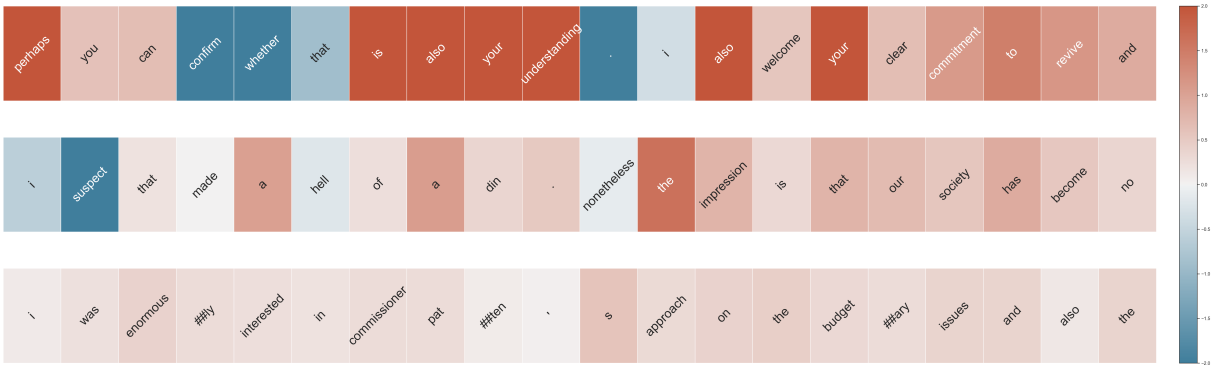

Figure 8: Saliency scores toward predictions of translated text in the 3 sentences with the highest average translated text saliencies. In most of the text, scores are lower-magnitude than their Original Text counterparts.

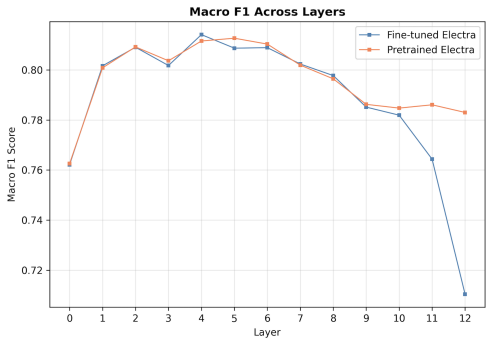

Figure 9: Layer-wise probing results for **Passive Voice** with Ratio@0.3 aggregation. The plot shows Macro F1 scores across layers 0–12 for both pretrained (orange) and fine-tuned (blue) ELECTRA models. Best MF1: Pretrained: 81.27%, Fine-tuned: 81.41%.

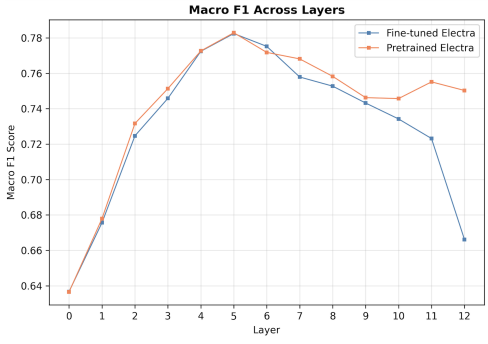

Figure 10: Layer-wise probing results for **Sentence Initial Non-Subject Fronting** with Ratio@0.3 aggregation. The plot shows Macro F1 scores across layers 0–12 for both pretrained (orange) and fine-tuned (blue) ELECTRA models. Best MF1: Pretrained: 78.29%, Fine-tuned: 78.25%.

Feature	Aggregation	Pretrained MF1↑	Fine-tuned MF1↑	Corr. Coef. (β)	Predictive MF1↑
Passive Voice	Ratio@0.3	81.27	81.41	+0.380	**53.7**
	Ratio@0.4	77.48	77.34	+0.429	**48.4**
	Ratio@0.5	75.82	77.34	+0.367	**46.6**
Pre-verbal Complexity	Ratio@0.3	61.42	60.28	+0.495	**41.3**
	Ratio@0.4	53.97	54.14	+0.577	**35.4**
	Ratio@0.5	53.86	52.95	+0.597	**34.9**
Sentence Initial Non-Subject Fronting	Ratio@0.3	78.29	78.25	+0.201	**51.9**
	Ratio@0.4	75.41	75.85	+0.312	**47.8**
	Ratio@0.5	74.04	74.15	+0.244	**46.2**

Table 4: Linear probing results for syntactic features across different density thresholds. Performance compared for the pretrained ELECTRA model with all layers frozen and only the classification head trained on the ENNTT corpus, or our fine-tuned ELECTRA model. MF1: Macro F1 score at the highest scoring layer for each feature. Corr. Coef. (β): Logistic regression log-odds coefficient measuring the direction and strength of association between the syntactic feature and translationese. Predictive MF1: Macro F1 score for predicting translationese given syntactic feature labels as input.

Feature	Aggregation	Pretrained MF1↑	Fine-tuned MF1↑	Corr. Coef. (β)	Predictive MF1↑
Function Words	50th Percentile	92.06	91.98 (−0.08)	−0.133	**52.6**
	70th Percentile	91.98	91.82 (−0.16)	−0.337	51.7
Pronouns	50th Percentile	93.35	93.47 (+0.12)	−0.649	**58.7**
	70th Percentile	92.55	92.74 (+0.19)	−0.753	55.8
Cohesive Markers	50th Percentile	90.34	90.40 (+0.06)	+0.429	**53.0**
	70th Percentile	90.02	90.10 (+0.08)	+0.340	52.9

Table 5: Linear probing results for **Lexical Features** across different density thresholds. Column descriptions are the same as in Table 4. Negative coefficients indicate the feature is *less* prevalent in translationese.

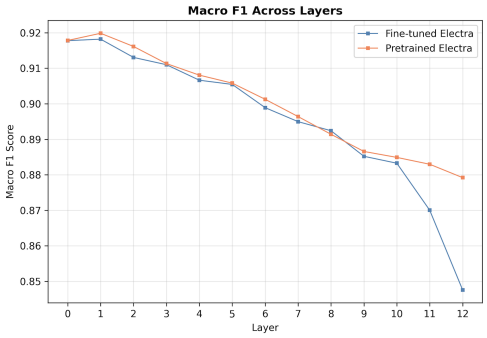

Figure 11: Layer-wise probing results for **Function Words** with 50th Percentile Density aggregation. The plot shows Macro F1 scores across layers 0–12 for both pretrained (orange) and fine-tuned (blue) ELECTRA models. Best MF1: Pretrained: 92.06%, Fine-tuned: 91.98%.

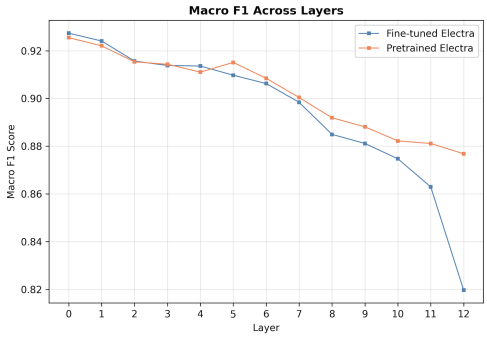

Figure 12: Layer-wise probing results for **Pronouns** with 50th Percentile Density aggregation. The plot shows Macro F1 scores across layers 0–12 for both pretrained (orange) and fine-tuned (blue) ELECTRA models. Best MF1: Pretrained: 93.35%, Fine-tuned: 93.47%.

Feature	Aggregation	Pretrained MF1↑	Fine-tuned MF1↑	Corr. Coef. (β)	Predictive MF1↑
M3 Single Triads	Sum 50th Percentile	85.21	84.85 (−0.36)	+0.459	**56.9**
	Sum 70th Percentile	84.76	84.69 (−0.07)	+0.584	55.5
	Max 70th Percentile	85.49	84.78 (−0.71)	+0.563	56.0
M6 Double Triads	Sum 50th Percentile	83.16	83.02 (−0.14)	+0.414	**56.0**
	Sum 70th Percentile	83.37	83.37 (+0.00)	+0.518	54.6
	Max 70th Percentile	73.23	73.23 (+0.00)	+0.273	52.0
M9 Triple Triads	Sum 50th Percentile	73.47	73.33 (−0.14)	+0.245	**53.4**
	Sum 70th Percentile	73.98	73.19 (−0.79)	+0.328	52.3
	Max 70th Percentile	71.90	71.19 (−0.71)	+0.262	51.9

Table 6: Linear probing results for **Discourse Motifs Features** across different aggregation methods and density thresholds. Column descriptions are the same as in Table 4. All coefficients are positive, indicating that higher discourse complexity is associated with translationese. M3, M6, and M9 refer to 3-node, 6-node, and 9-node subgraph motifs, respectively.

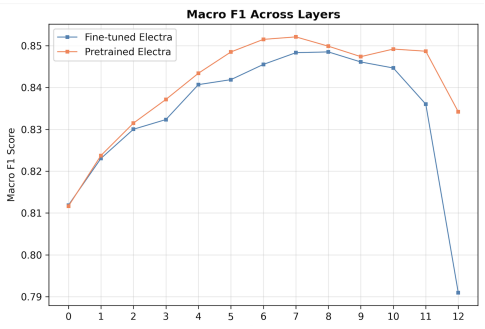

Figure 13: Layer-wise probing results for **M3 (Single Triads)** with Sum 50th Percentile Density aggregation. The plot shows Macro F1 scores across layers 0–12 for both pretrained (orange) and fine-tuned (blue) ELECTRA models. Best MF1: Pretrained: 85.21%, Fine-tuned: 84.85%.

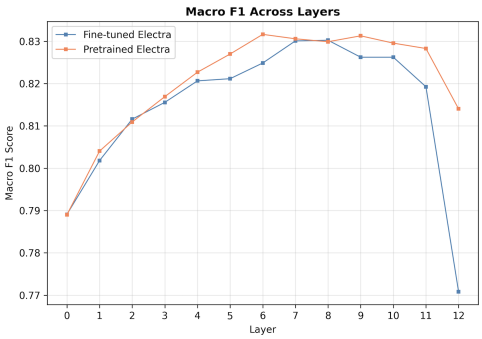

Figure 14: Layer-wise probing results for **M6 (Double Triads)** with Sum 50th Percentile Density aggregation. The plot shows Macro F1 scores across layers 0–12 for both pretrained (orange) and fine-tuned (blue) ELECTRA models. Best MF1: Pretrained: 83.16%, Fine-tuned: 83.02%.

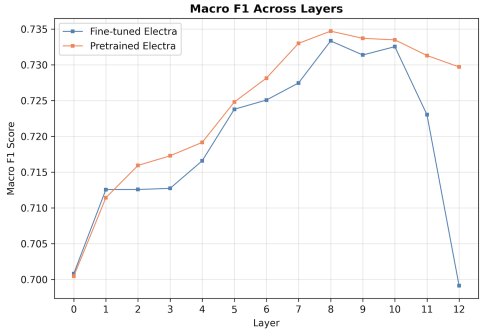

Figure 15: Layer-wise probing results for **M9 (Triple Triads)** with Sum 50th Percentile Density aggregation. The plot shows Macro F1 scores across layers 0–12 for both pretrained (orange) and fine-tuned (blue) ELECTRA models. Best MF1: Pretrained: 73.47%, Fine-tuned: 73.33%.

Metric	**Multi-Feature Prediction (%)**
Accuracy	59.64
Balanced Accuracy	59.66
Macro F1	59.58
Macro Precision	59.72
Macro Recall	59.66

Table 7: Multi-feature logistic regression performance on translationese classification combining all linguistic features. The classifier uses 9 binary features as input, where each feature's aggregation method was selected based on the highest Predictive MF1 score across all aggregation types: **Syntactic features** – passive voice (Ratio@0.3), pre-verbal complexity (Ratio@0.3), sentence initial non-subject fronting (Ratio@0.3); **Lexical features** – function words (Median Density), pronouns (Median Density), cohesive markers (Median Density); **Discourse motifs** – M3 single triads (Sum Median Density), M6 double triads (Sum Median Density), M9 triple triads (Sum Median Density).