Investigating Features of Translationese Using AMR and Dependency Parsing

Parker Boyle, Marci Cordon, and Olivia Feldman
Amherst College

Abstract

Translated English often differs from text originally written in English in subtle but consistent ways, a phenomenon commonly referred to as translationese. In this project, we explore whether adding syntactic and semantic structure can help distinguish native English, non-native English, and translated English. Building on prior work on translationese detection, we compare models using contextual embeddings from BERT with features derived from dependency parsing (DP) and Abstract Meaning Representation (AMR), both on their own and in combination. We find that BERT alone performs strongly, but adding structural and semantic information does not generally lead to improvements. Models trained only on dependency or AMR graphs produce better-than-random classification, suggesting that there are both semantic and structural differences between translated, native, and non-native texts. However, the lack of improvement when these representations are combined with tokenized text suggests that language models already detect and analyze these patterns. Overall, our results suggest that structural, semantic, and lexical choices all play a role in translationese, and that richer linguistic representations are not automatically beneficial for translationese classification.

1 Introduction

The language of native speakers, the language of non-native speakers, and text translated from other languages exhibit properties that distinguish them from one another (Nisioi et al., 2016). The differences between translated and native language are collectively known as "translationese" (Gellerstam, 1986). While these differences seem to be undetectable by humans, text with features of translationese can inhibit the learning of neural language models when included in training data (Wein et al., 2022). Determining the features that distinguish translated, native, and non-native English can help identify text to remove from training sets and elucidate the linguistic properties of translated language.

Although research has identified general differences, it is unclear whether these differences are primarily lexical, semantic, or syntactic. It also remains to be explored whether, if strong contextual embeddings already achieve high accuracy, there is still value in adding explicit linguistic representations. In particular, do structured formalisms like Universal Dependencies or Abstract Meaning Representation provide complementary information, or do they introduce noise when combined with neural embeddings?

In this project, we explore these questions by comparing models that use BERT alone with models that incorporate syntactic features from dependency parsing, semantic features from AMR, and combinations of the two. We frame the task as a three-way classification problem distinguishing native English, non-native English, and translated English, and evaluate each feature configuration under the same training setup. This format allows us to determine the usefulness of semantic and syntactic features in predicting translationese both alone and combined with word embeddings.

2 Related Work

Researchers have previously studied the differences between native, non-native, and translated English through the lens of hand-picked textual characteristics. They found that the three text types differed in lexical diversity, mean word rank, collocations, cohesive markers, and personal pronoun use. They also looked at similarities between translations and non-native English with the same origin language group (Rabinovich et al., 2016). Early research on the application of machine learning to translationese has revealed that learning systems rely on a variety of features, including lexical richness,

proportion of grammatical words to lexical words, sentence length, word length, and some morphological attributes to distinguish translated texts from native texts (Ilisei et al., 2010). Source language is also relevant: previous research has shown that the source language affects what kind of errors are found in translated texts. A learned classifier can identify the source language when various source languages are translated into the same target language, in addition to whether a given text is translated or original (Koppel and Ordan, 2011).

More modern research has focused on the application of modern natural language processing techniques to translation and the detection of translationese. One study comparing neural models to traditional machine learning models in classifying translated text found that neural architectures far outperform other approaches, and it is unclear whether learned neural features can be explained by the hand-selected features in traditional models (Pylypenko et al., 2021). Researchers have also tried to reduce the prevalence of translationese by using abstract meaning representations during translation, which may help translation systems focus on meaning rather than surface wording. This use of AMR was found to preserve meaning while reducing translation artifacts (Wein and Schneider, 2024).

AMR cannot be seen as a neutral representation of meaning that is completely detached from source language, however. Research has also shown that source language significantly affects the AMR of translated text, which stores traces of the original language's grammar and semantic structure (Wein et al., 2022). While significant progress has been made in classifying the lexical features of translationese and incorporating semantic structure into translation, it remains unclear whether translationese is primarily lexical, syntactic, or semantic, and whether incorporating new types of representation can help classify translated text. This motivates our approach, which compares the predictive contribution of each linguistic layer. Our work aims to extend prior research by evaluating these representations side-by-side within a unified structure that provides a controlled comparison of where translationese and interlanguage effects actually arise.

3 Data

3.1 Dataset

Our research is based on the Europarl corpus of native, non-native and translated texts (ENNTT). ENNTT is composed of sentences uttered in European Parliamentary Proceedings. It contains 3 groups of sentences: English sentences uttered by native speakers (determined based on country of origin), English sentences uttered by non-native speakers, and sentences originally uttered in other languages, then translated into English by native English speakers (Nisioi et al., 2016).

3.2 Preprocessing

We compiled these three types of text into a list of sentences with labels indicating their origins. The dataset was initially quite unbalanced across groups, with 738,596 translated sentences, 116,341 native English sentences, and only 29,734 non-native sentences. To prevent bias towards any one group, we uniformly split the three types of text. We downsampled by randomly shuffling the translated and native sentences and then selecting the first 29,734 observations so we would have a balanced number of sentences in each class. We used 80 percent of our data to train the models and reserved the other 20 percent for testing.

4 Methods

4.1 Sentence Embeddings

We used DistilBERT to analyze the text of the sentences in our dataset. DistilBERT is a smaller and faster version of the Bidirectional Encoder Representations from Transformers (BERT) language model (Sanh et al., 2020). Using a pretrained model allowed us to create functional tokens without requiring extensive time or data. We tokenized the text of the sentences using the pre-trained DistilBERT tokenizer. We padded and truncated our tensors to a length of 128 to ensure that our inputs were a consistent shape and size. We then used the pre-trained DistilBERT model to generate embeddings from those tokens.

4.2 Dependency Parsing

Dependencies in a sentence refer to the syntactic structure of the sentence; in other words, the relationships between words. Dependency graphs illustrate these relationships, showing how words act on and affect one another. In analyzing the dependencies between verbs, nouns, and other parts

of speech, it becomes possible to identify and extract relationships between tokens.

We used spaCy's medium core English web-trained model for dependency parsing, which automatically generated these graphs (Honnibal et al., 2020). Then, for each sentence, we created a feature dictionary containing different elements of these dependency representations. This included part of speech tags, dependency relationships, and head distances (distances between related words). In dependency parsing, each word in a sentence is considered as a node, and the relationships between the words are represented as directed edges. We set numerical values for part of speech tags and names of dependency relationships to tokenize these graphs so that they could act as inputs for our classification model. We also normalized the head distances to ensure that very large distances in long sentences would not negatively affect the model's performance. We then truncated and padded the resulting tensors to ensure that they could be passed into our classification model.

4.3 AMR Graphs

Abstract meaning representation (AMR) is a graph-based method for representing the semantic meaning of sentences, capturing the core concepts and relationships at play but abstracting away from specific word choices or sentence structure. We generated AMR graphs using AMRLib's parse_xfm_bart_base sentence to graph model, which fine-tuned the transformer-based BART model on the task of "translating" English sentences to AMR graphs (Jascob, 2024). These graphs are automatically represented as strings in PENMAN notation, which represents them in a linear, nested format. We tokenized these strings by splitting on the special characters used to separate words and relationships in PENMAN notation, and created a vocabulary dictionary to assign the remaining strings to numbers. Like we did for the dependency parsing, we truncated and padded the resulting tensors to ensure that they could be passed into our classification model.

4.4 Neural Classification Model

We used a neural classification model to determine whether a sentence was translated, native, or non-native based on our inputs. The input dimensions depended on which type of input we were testing (sentence embeddings, dependency graph features, AMR graphs, or some combination of the three). The same principle applied to the list of features we passed into the classifier. We concatenated these features together for the models with multiple types of input, and then passed them through the hidden layers of our classifier. The hidden layers of our model contained 8 layers: 4 Linear layers, which applied a set of weights to the input parameters, and 4 ReLU layers, one after each linear layer, which output the previous layer's output if it was positive and 0 if it was negative. The first hidden layer output had 384 dimensions (half the dimensions of the DistilBERT model's last hidden state), and each subsequent layer reduced the dimensionality by half. After processing the data, we flattened it to one dimension. We then fed it into a final Linear classification layer to output values representing how likely the sentence was to be in a particular class (translated, native, or non-native). Since these values did not give us probabilities of the sentence being in each class, we applied a softmax layer to amplify the largest input and suppress smaller ones, converting the numbers in the model output to probabilities. These probabilities allowed us to generate class predictions for sentences when testing.

4.5 Model Comparison

To isolate the effects of different types of sentence features on classifier performance, we tested all different possible permutations of the three input types. This resulted in 7 models: BERT only, AMR only, DP only, BERT and AMR, BERT and DP, AMR and DP, and BERT, AMR, and DP. The 8th possible permutation, where no inputs are passed into the model, was presumed to be a random class generator with 33% accuracy, both overall and for each class, because it would take no sentence features as input and therefore would not access any data with which to inform its classification decision. We used the AdamW optimizer with a learning rate of 0.00001 and a batch size of 24 when training our model. Each model was trained separately. We kept the design of the classifier, the sentences used for training, hyperparameters, optimizer, and processing methods constant across all models to better isolate the impact of passing in different sentence representations as model inputs. We trained each model over 10 epochs, reporting the best accuracy over all epochs and accuracy on each of the three classes for each model. We then collected these statistics into a final chart to compare results across models.

5 Model Classification Results

Table 1 presents the overall and within-class accuracy for all seven model configurations. The BERT-only model achieved the highest overall accuracy at 61.95%, with relatively balanced performance across all three classes (native: 60.50%, non-native: 62.10%, translated: 59.95%). This strong baseline shows that embeddings alone capture substantial information about the linguistic differences between these text types, and implicitly encode semantic and potentially structural features.

Models using only structural or semantic features performed significantly worse than BERT. The dependency-parsing-only model achieved 42.65% overall accuracy, while the AMR-only model reached 38.87%. However, both exceeded the 33% random baseline, indicating that syntactic structure and semantic representation do contain signals that distinguish between native, non-native, and translated English. The dependency-only model performed best on native English (50.03%) but struggled with translated text (33.18%), while the AMR-only model showed the opposite pattern, performing best on translated text (46.98%) but worst on non-native text (31.86%).

Combining BERT with structural or semantic features did not seem to improve overall accuracy. The BERT+DP model achieved 61.23% accuracy, and BERT+AMR reached 61.46%, both slightly below the BERT-only baseline. The combined BERT+DP+AMR model similarly achieved 61.40%, showing no advantage from incorporating multiple representation types. While overall accuracy remained stable, the distribution of performance across classes shifted when adding features. For instance, BERT+AMR substantially improved non-native accuracy to 67.40% but decreased translated accuracy to 57.09%, suggesting that different representations may capture different aspects of linguistic variation even if they don't improve aggregate performance.

The DP+AMR model without BERT performed at 42.34% accuracy, comparable to DP alone, indicating that combining structural and semantic representations does not provide much additional benefit.

6 Analysis

Our results reveal several important patterns about the nature of translationese and the role of linguistic representations in its detection. First, the strong performance of BERT alone suggests that contextual embeddings are quite effective in capturing the features that distinguish translated, native, and non-native English. The lack of improvement when adding explicit syntactic or semantic structure indicates that BERT's attention mechanisms may already implicitly encode much of the information present in dependency graphs and AMR representations.

The above-baseline performance of DP-only and AMR-only models confirms that structural and semantic differences exist between the three text types, supporting prior findings that translationese involves changes at multiple linguistic levels. However, the relatively poor performance of these models compared to BERT suggests that lexical cues and patterns captured by sophisticated language models are significantly more telling than structural patterns or semantic clues alone. This aligns with previous research identifying lexical diversity and word choice as key markers of translationese. The weak performance of AMR aligns with the intuition that the underlying meaning of sentences may differ somewhat between translated, native, and non-native text, but likely not very much, especially when sentences are collected in the same context. It also aligns with research indicating that the use of AMR representations as an intermediary step of translation can reduce translationese.

The different class-level performance patterns across models provide insight into the linguistic characteristics of each text type. The dependency-only model's strength on native English but weakness on translated text suggests that translated English may exhibit more varied or non-standard syntactic structures, possibly due to interference from source language syntax.

When BERT is combined with structural features, we observe redistribution of accuracy across classes rather than overall improvement. The BERT+AMR model's increase in non-native accuracy (67.40%) at the cost of translated accuracy suggests that semantic features may help distinguish non-native productions, which may have different semantic patterns due to imperfect language acquisition. This may also be a relic of the data itself: it is possible that representatives of non-English-speaking countries chose to speak about different topics or express different opinions when speaking English than people from English-speaking countries. Meanwhile, BERT+DP's improved performance on translated text (67.48%) may indicate

Configuration	Overall Accuracy	Native Accuracy	Non-Native Accuracy	Translated Accuracy
BERT Only	61.95%	60.50%	62.10%	59.95%
DP Only	42.65%	50.03%	44.76%	33.18%
AMR Only	38.87%	37.75%	31.86%	46.98%
BERT + DP	61.23%	55.49%	60.72%	67.48%
BERT + AMR	61.46%	57.89%	67.40%	57.09%
DP + AMR	42.34%	55.68%	45.13%	25.56%
BERT, DP, + AMR	61.40%	55.74%	65.48%	60.52%

Table 1: Overall and Within-Class Accuracy, By Model

that explicit structural features help identify the anomalies characteristic of translation, but this result is weak considering the DP-only model's bad performance on classifying translated sentences. It may indicate that the combination of structural cues and word choice is impactful in identifying translated text.

These patterns suggest that while translationese manifests across lexical, syntactic, and semantic dimensions, these dimensions may not be independent. The redundancy between BERT embeddings and explicit structural representations implies that modern language models learn to detect syntactic and semantic patterns through exposure to large text corpora, making hand-crafted representations less necessary for classification tasks. This finding suggests that adding explicit linguistic representations to these models is not automatically beneficial and may introduce noise without corresponding gains in accuracy.

7 Conclusion

This study investigated whether incorporating syntactic and semantic structure improves the classification of native, non-native, and translated English. We compared models using BERT embeddings, dependency parsing features, and AMR graphs, both individually and in combination. Our results demonstrate that while BERT alone achieves strong performance, adding explicit structural or semantic representations does not improve overall accuracy and may even make models slightly worse.

The above-baseline performance of dependency and AMR models when used alone confirms that translationese exhibits detectable structural and semantic patterns. However, the lack of complementary benefit when these features are combined with BERT suggests that contextual embeddings already capture much of this information implicitly. This finding supports the hypothesis that modern language models learn complex linguistic patterns, reducing the need for explicit linguistic representations in classification tasks.

Our class-level analysis reveals that different linguistic representations show varying strengths across text types. Dependency features appear particularly useful for identifying translated text, but only in combination with features learned by language models, while AMR has the opposite effect: it helps distinguish translated features alone, but is more useful for detecting non-native text in conjunction with BERT. This suggests that translation and non-native languge effects involve different and complex configurations of divergence from native norms.

These findings have practical implications for both translationese detection and language model training. For detection tasks, our results suggest that BERT-based approaches provide an efficient and effective solution without requiring complex linguistic preprocessing. For curating training data, understanding that translationese manifests mainly through complex patterns that can be captured best by large language models rather than systematic structural deviations may inform strategies for identifying and filtering translated text from training corpora.

Future work could explore whether explicit linguistic representations become more valuable in lower-resource settings where pre-trained models are less effective, or whether finer-grained structural features could provide more targeted information than our aggregated representations. Additionally, investigating whether the source language of translations affects the relative importance of different linguistic levels could shed light on the mechanisms underlying translationese.

Ethical Considerations

This research raises some important ethical considerations. While our work aims to better understand linguistic phenomena, translationese detection systems could potentially be misused to discriminate against translated content or non-native speakers. Language models trained to detect translationese could also potentially be incorporated into content filtering or quality assessment systems, which could have unintended discriminatory effects. If they are used to alter the data used to train language models, this may bias those models away from opinions held primarily by non-native speakers or speakers of other languages who translate their sentences into English. Translationese detection should only be deployed after carefully considering these social implications and should not be used to devalue or exclude contributions from other languages or non-native speakers.

Our finding that language models already implicitly capture many features of translationese suggests that existing NLP systems may be making distinctions based on language background, even when not explicitly designed to do so. This has implications for fairness in NLP applications, as models may show performance differences across native, non-native, and translated text. Researchers developing language technologies should be aware of these potential disparities and try to ensure fair performance across diverse contexts.

Characterizing the properties of non-native and translated text could aid both beneficial and harmful applications. While understanding these patterns can improve machine translation systems, linguistic understanding, and model training, the same knowledge could be used to profile or discriminate against speakers based on their linguistic background.

Limitations

Our study has several limitations that should be considered when interpreting the results. First, our dataset is drawn exclusively from European Parliamentary proceedings, which is a very specific domain. Parliamentary speech may exhibit different patterns of translationese than other genres such as literary translation, technical documentation, or informal text. The formality and subject matter of parliamentary discourse could affect both the prevalence and nature of translationese features, potentially limiting the generalizability of our findings to other contexts. The skill of parliamentary translators and importance of clarity and accuracy when speaking in this setting could also affect what translation artifacts, if any, show up in speech.

Second, we downsampled our dataset to 29,734 sentences per class to achieve balance, discarding a substantial amount of translated and native English data. While this approach prevents class imbalance bias, it may have reduced our models' ability to learn subtle patterns and could affect the stability of our accuracy estimates. A larger dataset, or use of all observations but with higher error weighting for rarer classes, might reveal different patterns or allow us to find more nuanced distinctions.

Third, our dependency parsing and AMR features were generated using automatic parsers, which introduce their own errors and biases. Parser accuracy may vary across the three text types—for instance, the spaCy parser and AMRLib model, both trained primarily on native English, may perform worse on translated or non-native text. Gold AMR annotations would provide cleaner representations but were infeasible given our dataset size and time constraints.

Fourth, our neural architecture is only one method of combining different representation types. We used simple concatenation of embeddings and fixed 16-dimensional learned embeddings for all discrete features (POS tags, dependency relations, AMR tokens). Different methods, like an ensemble of different models trained specifically on embeddings, dependency graphs, or AMR, may have better compared information from different features. The large difference in dimensionality between BART data and dependency or AMR data may also have contributed to the dominance of BART features, which may have made it more difficult for the model to integrate features from the other structures.

Fifth, our training procedure selected the best-performing epoch based on test set accuracy, which could introduce some overfitting to the test data. A validation set may have provided more robust model selection. Additionally, we trained all models for exactly 10 epochs with fixed hyperparameters (learning rate 1e-5, batch size 24), without systematic hyperparameter tuning. A validation set may have also helped determine the ideal hyperparameters for our model.

Finally, our study focuses on classification accuracy as the primary evaluation metric. While accuracy is informative, it does not reveal what

specific linguistic patterns the models learned or which features were most discriminative. More mechanistic interpretability methods and deeper analysis of model predictions, attention weights, or feature importance scores could provide better insight into the nature of translationese and how different representations capture it.

Despite these limitations, our controlled comparison of representation types provides valuable evidence about the role of linguistic structure in translationese detection and the effectiveness of modern large language models for capturing complex linguistic patterns.

References

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation Studies in Scandinavia*, 1:88–95.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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.

Brad Jascob. 2024. amrlib: A python library for amr parsing, generation, and visualization.

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.

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'16)*, pages 4197–4201, Portorož, Slovenia. European Language Resources Association (ELRA).

Daria Pylypenko, Kwabena Amponsah-Kaakyire, Koel Dutta Chowdhury, Josef van Genabith, and Cristina España-Bonet. 2021. Comparing feature-engineering and feature-learning approaches for multilingual translationese classification. *Preprint*, arXiv:2109.07604.

Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. 2016. On the similarities between native, non-native and translated texts. *Preprint*, arXiv:1609.03204.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Shira Wein, Wai Ching Leung, Yifu Mu, and Nathan Schneider. 2022. Effect of source language on AMR structure. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 97–102, Marseille, France. European Language Resources Association.

Shira Wein and Nathan Schneider. 2024. Lost in translationese? reducing translation effect using abstract meaning representation. *Preprint*, arXiv:2304.11501.