The Devil is in the Details: Assessing the Effects of Machine-Translation on LLM Performance in Domain-Specific Texts

Javier Osorio[1], Afraa Alshammari[2], Naif Alatrush[2], Dagmar Heintze[2], Amber Converse[1], Sultan Alsarra[3], Latifur Khan[2], Patrick T. Brandt[2], Vito D'Orazio[4]

[1]University of Arizona [2]University of Texas at Dallas
[3]King Saud University [4]West Virginia University

Correspondence: josorio1@arizona.edu

Abstract

Conflict scholars increasingly use computational tools to track violence and cooperation at a global scale. To study foreign locations, researchers often use machine translation (MT) tools, but rarely evaluate the quality of the MT output or its effects on Large Language Model (LLM) performance. Using a domain-specific multilingual parallel corpus, this study evaluates the quality of several MT tools for text in English, Arabic, and Spanish. Using ConfliBERT, a domain-specific LLM, the study evaluates the effect of MT texts on model performance and finds that MT texts tend to yield better results than native-speaker written texts. The MT quality assessment reveals considerable translationese effects in vocabulary reduction, loss of text specialization, and syntactical changes. Regression analysis at the sentence level reveals that such distortions, particularly reductions in general and domain vocabulary rarity, artificially boost LLM performance by simplifying the MT output. This finding cautions researchers about uncritically relying on MT without considering MT-induced data loss.

1 Introduction

Political scientists, like many other domain-specific users, often rely on computational tools to make sense of large volumes of data. In particular, conflict scholars increasingly use computational methods to analyze global dynamics of political conflict and cooperation in foreign locations. To do so, researchers frequently rely on machine translations (MT) to translate political text from different languages (Boschee et al., 2018; Halterman et al., 2023). Despite the growing research on MT quality (Liu and Zhu, 2023; Kahlon and Singh, 2023; Lee, 2023; Ahrenberg, 2017), social scientists seldom evaluate the quality of the MT output nor its consequences on model performance. Careful researchers may be concerned about MT quality due to data loss or incorrect translations, particularly for low-resource languages (De Vries et al., 2018; Licht et al., 2024; Bartaškevičius, 2024) or specialized domains requiring precise terminology (Cambedda et al., 2021). MT-induced changes to the source text, known as translationese effect (Gellerstam, 1986), may result in considerable alterations of the output text, making translationese especially crucial to investigate in domain-specific translations where seemingly minor distortions of the output text may lead to incorrect inference. Moreover, there is little work analyzing the impact of MT quality on Large Language Model (LLM) performance (Huang and Liu, 2024). Consequently, the quality of the MT text often gets overlooked, and its effects on LLM performance remain ignored. For researchers tracking conflict around the world, disregarding translationese or its effects on LLM performance may lead to missing important signals about security threats or cooperation.

By using a multilingual parallel corpus from the United Nations (Ziemski et al., 2016), this study analyzes the quality of various MT tools for English, Arabic, and Spanish and evaluates the effects of MT distortions on LLM performance on tasks related to political conflict and cooperation. In particular, the study evaluates four MT tools, Google Translate (GT) (Google Cloud, 2024), DeepL Translate (DeepL) (DeepL, 2024), Google Translate within the Deep Learning Translator (Deep) (Deep Translator, 2020), and OPUS Machine Translation (OPUS) (Tiedemann and Thottingal, 2020), and evaluates the performance of their MT outputs using ConfliBERT (Hu et al., 2022), a domain-specific LLM specialized on political conflict.

This research offers several contributions. The study carefully evaluates the quality of various MT tools using a domain-specific parallel corpus in English, Arabic, and Spanish. Contrary to the ex-

Proceedings of Machine Translation Summit XX Volume 1, pages 315–332, June 23-27, 2025

pectation that LLMs work better when processing native-speaker written/translated texts (NST), experiments in this study indicate that ConfliBERT performs better using MT output. By disentangling sentence-level characteristics, the analysis reveals MT distortions related to nouns, verbs, lemmas, vocabulary complexity, and sentence structure. The study further explores the effects of MT distortions on LLM performance at the sentence level using regression analysis. The results reveal that MT distortions, primarily the vocabulary loss of general and domain-specific rarity, generate simplified text representations that artificially boost model performance, particularly for translated text from Arabic and Spanish to English. Such simplification favors model performance on MT output over NST. These results represent a double-edged sword for researchers using MT tools who may face a trade-off between achieving higher LLM performance at the expense of domain-specific words that may be relevant to their subject of study.

2 Related Works

Pre-trained Language Models (PLM) such as BERT (Devlin et al., 2018a) achieve great results by using continued pre-training on domain-specific data to capture its unique vocabulary, semantics, and language style (Gururangan et al., 2020). Taking advantage of this capability, political scientists use different models to study party manifestos (Mens and Gallego, 2024), voter partisanship (Potter et al., 2024), social movements (Caselli et al., 2021; Hürriyetoğlu et al., 2022; Radford, 2020), dictionary development (Radford, 2021; Osorio et al., 2019), codebook-based classification (Halterman and Keith, 2024) and annotations (Ziems et al., 2024), among other tasks. Similarly, conflict scholars use specialized language to study political violence and cooperation. For that purpose, ConfliBERT (Hu et al., 2022) is a domain-specific model specialized on political conflict that yields superior performance compared to generic LLMs.

Researchers using non-English text generally rely on MT tools to pre-process the original text. However, MTs can heterogeneously distort the data, thus affecting the model performance (Osorio et al., 2024). Assessing MTs from Spanish to English, previous research shows a net summarization of the original text, which reduces the verboseness of the original text and results in an adaptation of the target text to linguistic characteristics of English. This adaptation to English linguistic standards yields high quality-metric results for MT text (Osorio et al., 2024). The summarization effect can further artificially enhance ConfliBERT EN's performance, as the English language generally favors more concise text (Yang et al., 2023).

When evaluating the translation quality from Arabic to English, Osorio et al. (2024) found that MTs artificially extend the source text. This data increase in MTs from Arabic is penalized in English, as metrics show a notable decrease in translation quality relative to the original Arabic text. However, this data increase appears to introduce more linguistic elements that artificially boost ConfliBERT EN's performance on the MT corpus.

The predominant body of research is in favor of languages with abundant resources; thus, more recent studies use translation tools to mitigate the scarcity of training data, including (De Vries et al., 2018), which used Google Cloud (2024) (GT) to translate official transcripts of European Parliament debates written in the official majority of the EU's languages into English. To improve inference in prompting multilingual LLMs, Etxaniz et al. (2023) translated from languages that are comparatively less represented in available LLMs, like Spanish, into English to leverage the fact that English makes up the majority of training data in multilingual LLMs. Other recent studies further compare translation tools (Ibrahim, 2021; Akki and Larouz, 2021; Behr and Braun, 2023) and quality translation metrics (Mathur et al., 2020; Sabtan et al., 2021; He et al., 2021; Lee et al., 2023).

3 Data and Annotations

This research uses the United Nations Parallel Corpus (UNPC) (Ziemski et al., 2016), containing 86,307 official United Nations (UN) Security Council documents translated by professional UN translators. Since the UN operates in six official languages, these translations are considered the Gold Standard Record (GSR). Out of the official UN languages, this study uses NST texts written in English (EN), Spanish (ES), and Arabic (AR). In total, the UNPC contains 11,365,709 fully aligned sentences across languages. This study uses a random sample of 11,326 sentences from UN Security Council documents related to human rights, the protection of civilians, and terrorism. The resulting sample provides a uniquely valuable multilingual parallel corpus in the domain of political conflict and co-

operation. Having the same GSR content across multiple NST sentences within the UNPC provides a *ceteris paribus* leveled field to compare the effects of different MT tools on model performance.

This study uses the UNPC sentences previously annotated by Osorio et al. (2024),[1] which classify the content of the sentences according to PLOVER (Open Event Data Alliance, 2018), an ontology often used in political science to categorize different types of material and verbal interactions based on the cooperative or conflictive conduct of the parties involved. Annotators classified the full sample of sentences according to three tasks. *Relevance* is a binary classification identifying whether a sentence is relevant for political conflict or cooperation or not. *QuadClass* is a multi-class classification task categorizing whether the sentences indicate verbal conflict, verbal cooperation, material conflict, material cooperation, or non-relevant sentences. Finally, *BinQuad* is a binary classification for each QuadClass category identified above, indicating whether the sentence can be categorized as the respective PLOVER category or not, thereby representing one of the other three categories.

The annotations have the following distributions[2]. In the Relevance binary task, coders identified 52% sentences as not relevant and the rest 48% as relevant. For the multi-class QuadClass task, coders identified 14% sentences as Material Conflict, 13% as Material Cooperation, 8% as Verbal Conflict, 11% as Verbal Cooperation, and 53% as not relevant. Finally, the BinQuad binary task of QuadClass categories produced the following distribution for Material Conflict (yes 14%, no 86%), Material Cooperation (yes 13%, no 87%), Verbal Conflict (yes 8%, no 92%), and Verbal Cooperation (yes 11%, no 89%). All experiments used balanced datasets, with the number of randomly selected sentences capped to match the smallest category size in each task.

4 Translation Quality Assessment

Using UNPC text in English (EN), Spanish (ES), and Arabic (AR), we conduct a series of MTs using different tools. Our analysis uses bidirectional MT to convert the entire sample of Spanish and Arabic texts into English (ES to EN, AR to EN) and vice versa (EN to ES, EN to AR). To conduct the translations, we use four commonly used MT tools: Google API Translate (GT) (Google Cloud, 2024), DeepL Translate (DeepL) (DeepL, 2024), Deep Learning Translator (Deep) (Deep Translator, 2020), and OPUS Machine Translation (OPUS) (Tiedemann and Thottingal, 2020).[3] Google translate employs subword tokenizers optimized on extensive multilingual corpora (Kudo and Richardson, 2018). This approach addresses out-of-vocabulary challenges by merging frequent character sequences into subwords and transforming tokenized subwords into dense embeddings within the Google-managed Transformer architecture. Positional encodings enable the self-attention mechanism to align and predict tokens in the encoder-decoder pipeline. These vectors are continually refined through large-scale training on vast datasets, a process referred to as dynamic tuning (Google Cloud, 2024; Vaswani et al., 2017). DeepL provides free and subscription-based translating services between a variety of languages via its website or an API (DeepL, 2024). Deep is a lightweight Python package that invokes the public Google Translate service. It accesses a standard and universal shared model that lacks dynamic tuning capabilities. This causes lower accuracy or a failure to accurately capture complex and domain-specific words (Deep Translator, 2020; Google Cloud, 2024). Google Translate was selected via the deep translation package to establish a baseline comparison between the paid and free versions of the most used MT tool in the literature (Wu et al., 2016). OPUS, a Hugging Face Transformers library, presents a suite of state-of-the-art pretrained translation models (Tiedemann and Thottingal, 2020). In particular, its Helsinki-NLP/opus-mt-ar-en and Helsinki-NLP/opus-mt-es-en models are specifically trained to translate from Arabic to English and from Spanish to English, respectively (OPUS, 2016).

Building on (Han et al., 2022), we use four quality assessment metrics: SacreBLEU (Post, 2018), METEOR (Banerjee and Lavie, 2005), BERTScore (Devlin et al., 2018b), and COMET (Crosslingual Optimized Metric for Evaluation of Translation), using the wmt20-comet-da model (Rei et al., 2020). SacreBLEU, and METEOR are lexical-based metrics measuring the similarity between NST and MT text using mathematical or heuristic methods, COMET is a neural-based metric (Rei et al., 2020),

[1] See appendix D for details on the annotation process.
[2] Details in Appendix E
[3] See Appendix F on MT tools development, training data relevance, and Appendix G for quality metric evaluation.

whereas BERTScore uses an embedding-based metric that relies on deep learning methods (Lee et al., 2023). Quality scores range from 0 to 1, with high values indicating greater NST-MT similarity (Chatzikoumi, 2020; Zhang et al., 2020). The metrics can be ranked according to their degree of flexibility. SacreBLEU employs the Moses tokenizer, an advanced preprocessing tool that facilitates score comparability and was first created for the Moses statistical machine translation system (Post, 2018). The Moses tokenizer uses heuristics and rules unique to a given language to normalize text and to handle punctuation or special characters. METEOR is more flexible and calculates the similarity of word alignments. COMET is a state-of-the-art neural-based MT evaluation metric. BERTScore is the most flexible metric as it considers contextual correctness and synonyms.

Using each UNPC NST text as a reference, Figure 1 presents the quality scores from the different metrics applied to each MT output. Results show that different tools generate varying degrees of quality across languages. While for AR to EN and ES to EN MTs, SacreBLEU, METEOR, and BERTScore indicate that DeepL provides the best quality output, COMET considers OPUS to be the most accurate MT tool for these language combinations. For EN to ES, OPUS yields the best MT quality based on all metrics, while for EN to AR, COMET disagrees with all other metrics and considers DeepL the best-performing MT tool. While these metrics offer a first assessment of the MT quality, they do not permit an in-depth understanding of MT-induced translationese effects on the source text or assess more subtle changes in meaning and nuance. Consequently, these metrics do not fully capture whether MT-induced changes influence LLM performance. The following section evaluates LLM performance across MT texts to see if the results align with quality assessment suggestions.

5 Model Performance Across MT Tools

Following the quality assessment, we test the effect of MTs on LLM performance. To do so, we use ConfliBERT (Hu et al., 2022), a domain-specific pre-trained language model specifically designed to analyze political texts, to evaluate the UNPC NST and MT texts for three classification tasks: Relevant (binary) classification, QuadClass (multiclass) classification, and BinQuad (binary) classification of each QuadClass category. For each task, the fine-tuning uses three versions of the ConfliBERT family, namely ConfliBERT Arabic (Alsarra et al., 2023), ConfliBERT Spanish (Yang et al., 2023), and ConfliBERT English (Hu et al., 2022), with cased and uncased variations, resulting in a total of 6 different models. All models use balanced datasets for each task. By keeping the UNPC content and the use of ConfliBERT constant, we analyze variations in performance derived from different MT tools, including Deep, DeepL, GT, OPUS, and the NST texts. First, we split the data into training, testing, and developing using 70-15-15 rule. Second, for each model, we perform the evaluation using 10 seeds and 5 epochs. Finally, we run a total of 114 fine-tuning tasks on those models and their corresponding datasets. We used a HPC system with a single A100 GPU 20GB and a single V100 GPU 32GB, and a learning rate of 4e-05, with a training batch size of 8 and a maximum sequence length of 512 for both binary and multiclass classifications. Figures 2-4 present the F1 scores highlighting the top-performing models in red. Overall, results show that processing MT text yields better results than analyzing NST text. This is puzzling since domain-specific models would be expected to perform better with NST texts.

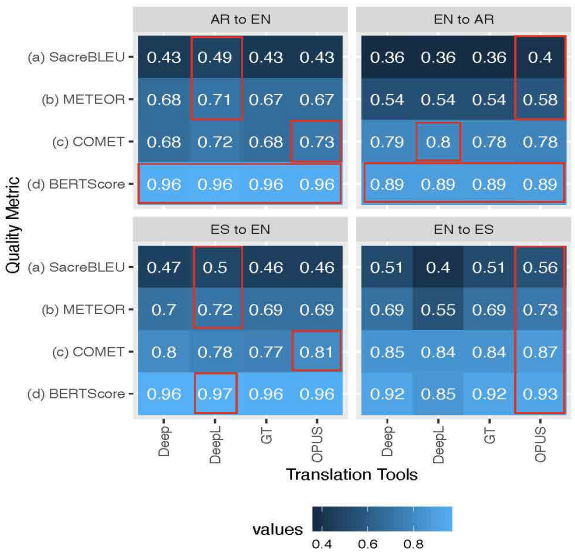


Figure 1: Quality Assessment Metrics

5.1 Relevant Binary Classification

Figure 2 reports the F1 performance of the ConfliBERT models for the relevant binary task on the NST and MT texts across languages. Red squares indicate best models with p-values at $p<0.01$ or lower. Overall, the results show high performance

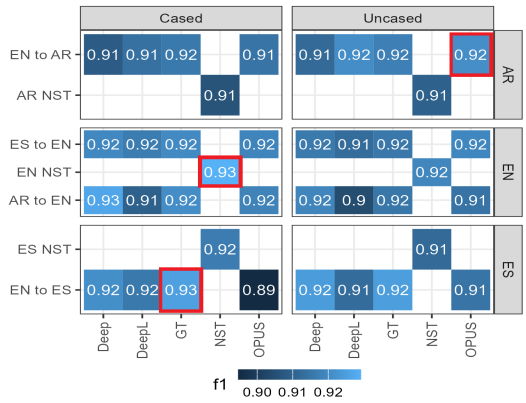


Figure 2: Binary Relevance Classification

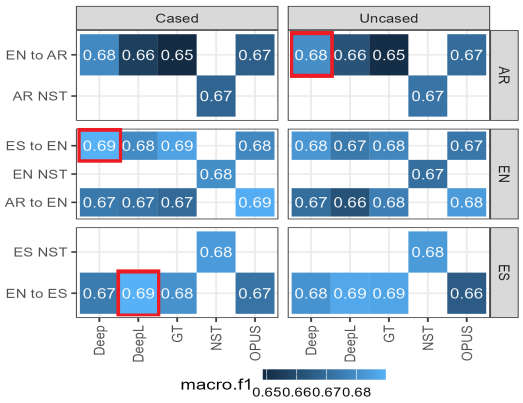


Figure 3: QuadClass Classification

levels and little variations across MT outputs. Most importantly, the results show that MT texts yield marginally better results in Arabic and Spanish than models processing the NST texts. Yet, these results do not hold for English, where models using NST text outperform those analyzing MT text. For Arabic, ConfliBERT uncased performs best (F1 0.921) on the EN to AR corpus translated with OPUS. This result is statistically significantly better than NST text in Arabic (F1 0.91). For English, ConfliBERT cased on English NST text shows the highest performance (F1 0.929) and performs statistically significantly better than AR to EN Deep (F1 0.927). Finally, the results for Spanish indicate that ConfliBERT cased performs best using the EN to ES GT translation (F1 0.925) and significantly better than NST text in Spanish (F1 0.917).

5.2 QuadClass Multi-Class Classification

Figure 3 presents the results of the QuadClass classification task. Overall, the results of the QuadClass classification report lower performance than the Relevant binary task. This is understandable as a five-categories multi-class classification is more difficult than a dichotomous task, and the latter has more training examples than the former. In general, these results also indicate that analyzing MT text performs marginally better than processing NST texts across languages. For Arabic, ConfliBERT uncased reports the highest F1 (0.680) for the QuadClass on the EN to AR Deep translated text. This result is statistically significantly better than the NST text Arabic model (F1 0.672). For English, ConfliBERT cased processing the ES to EN Deep translation generates the best QuadClass performance (F1 0.69), while the NST model in English (F1 0.68) has a statistically significantly lower performance. Finally, the results for Spanish show that ConfliBERT cased performs best using EN to ES DeepL translation (F1 0.69). This result is barely better than Spanish NST text (F1 0.684).

5.3 BinQuad Binary Classification

Figure 4 shows binary classification results for Material Conflict (panel 4.a), Material Cooperation (panel 4.b), Verbal Conflict (panel 4.c), and Verbal Cooperation (panel 4.d). The analysis shows heterogeneous results. For specific QuadClass instances in Arabic, MT generally performs better than Arabic NST text. However, the results for NST text versus MT text in English and Spanish are mixed.

Panel 4.a shows the Material Conflict scores. In general, the results indicate that MT text performs better than NST text in Arabic and English, but the Spanish models show a comparable performance in NST and MT texts. For Arabic, ConfliBERT uncased has the best performance (F1 0.885) when processing the Deep EN to AR translation. In contrast, NST text Arabic has a statistically lower performance (F1 0.863). For English, ConfliBERT cased performs the best (F1 0.908) using the ES to EN Deep text. This result is statistically significantly better than using English NST text (F1 0.890). For Spanish, ConfliBERT cased reports the best performance with the EN to ES OPUS text (F1 0.876). However, this result is not statistically different from the NST text in Spanish (F1 0.876).

Material Cooperation results (panel 4.b) indicate that MT works as well as NST Arabic and English texts and sometimes works better than Spanish NST text. For Arabic, ConfliBERT cased with NST Arabic text yields the best performance (F1 0.819). Yet, it is not different from ConfliBERT uncased with EN to AR OPUS translation (F1 0.818). For English, the top performing model is ConfliBERT

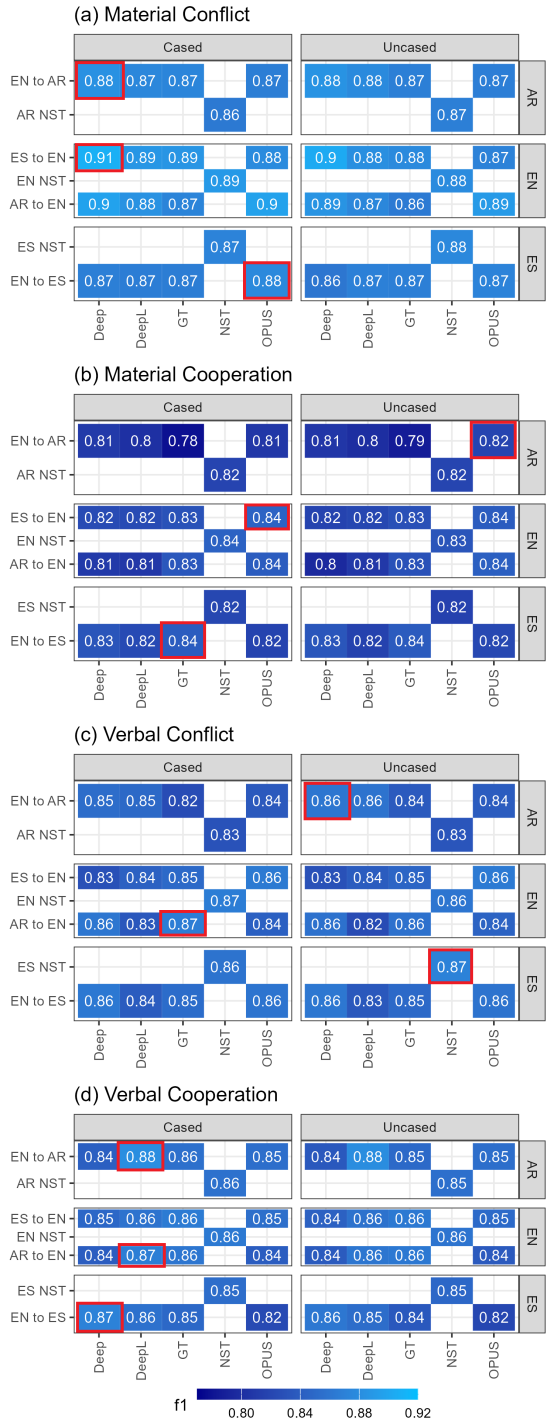


Figure 4: Binary QuadClass Classification

cased using the ES to EN OPUS translation (F1 0.844). However, this result is not statistically superior to English NST text (F1 0.843). Finally, the results for Spanish show that the EN to ES GT text yields the best results with ConfliBERT cased (F1 0.838), while the Spanish NST text model has lower performance (F1 0.82).

Panel 4.c reports the Verbal Conflict F1 scores. In general, the results show that MT texts perform better in Arabic, but these findings do not hold for English and Spanish. The results for Arabic indicate that ConfliBERT uncased has the best performance with EN to AR Deep text (F1 0.86). This score is better ($p<0.001$) than the Arabic NST text performance (F1 0.833). For English, the AR to EN GT translation using ConfliBERT cased has the best result (F1 0.867). However, this score is not different ($p=0.805$) from the English NST text model (F1 0.866). For Spanish, ConfliBERT uncased works the best when using Spanish NST text (F1 0.87). However, this result is not different ($p=0.197$) from its closest competitor, the EN to ES Deep text with ConfliBERT cased (F1 0.862).

Finally, Verbal Cooperation results in panel 4.d indicate that MT texts yield better results than NST texts in Arabic and Spanish, but models using sentences in English perform as well as those using MT texts. For Arabic, ConfliBERT uncased processing EN to AR DeepL translations has the best performance (F1 0.88). This result is statistically superior to the Arabic native model (F1 0.856). For English, the top performing model is ConfliBERT cased processing AR to EN DeepL translation (F1 0.867). Although this model performs better than the English NST text model (F1 0.863), the difference is not statistically significant. Finally, the results for Spanish indicate that processing the EN to ES Deep translation with ConfliBERT cased yields the best performance (F1 0.87). In contrast, the Spanish NST text model reports a lower performance (F 0.853) at statistically significant levels.

Overall, this section shows that LLM performance does not necessarily align with the MT quality suggestions. The following sections try to identify the determinants of model performance.

6 Corpus Rarity and Vocabulary Loss

To disentangle the characteristics of MT outputs that yield marginally superior ConfliBERT performance compared to processing NST texts, this section analyzes MT-induced distortions to the original corpora. First, we measure the total vocabulary size after preprocessing using spaCy's `en_core_web_trf` transformer pipeline for English (Honnibal et al., 2020), spaCy's `es_dep_news_trf` transformer pipeline for Spanish (Honnibal et al., 2020), and the Farasa segmenter (Al-shaibani, 2021) for Arabic. The vocabulary size for each language represents the total number of unique words included in the MT corpora. Figure 5 shows the vocabulary sizes. We

find that the MT corpora consistently have a lower vocabulary size than the respective NST corpus. This finding aligns with the characteristics of translationese, where translated text tends to show reduced lexical diversity compared to original text (Riley et al., 2020). Therefore, there may be a convergence of MT on similar phrasing, reducing the need to learn diverse patterns as in the native text.

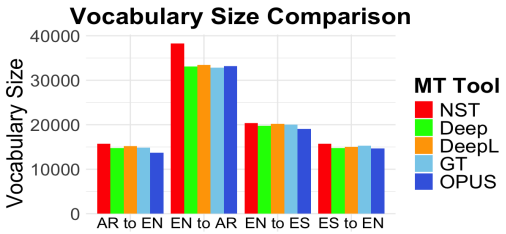


Figure 5: Native-MT Vocabulary Size Comparison

To further explore this reduction effect, we measure lexical rarity per sentence. Following (Proisl, 2022), we define lexical rarity as the proportion of tokens in a text that does not appear in the 5,000 most common tokens for a domain. We consider two types of rarity: general and domain. General rarity relies on the 5,000 most common tokens for a language, regardless of subject. Domain rarity uses the 5,000 most common tokens in the sentences from the UNPC to measure rarity as it relates to a political corpus. We use rarity as a proxy for lexical complexity and consider it the prime indicator for the reduction of text complexity and loss of context in the MT texts. A reduction in rarity for MT text represents a decline in the number of unique tokens compared to the NST text. Consequently, a reduction in the mean rarity of the MT corpus represents a loss of language complexity compared to the NST corpus. The loss of rarity may be particularly relevant for domain-specific researchers where key terms or technical words may carry particular substantive value. In addition to rarity, we measure the number of unique lemmas, nouns, and verbs in each sentence as additional measures of linguistic features (see Appendix H).

Figure 6 shows the mean general and domain rarity scores. Using a pairwise Wilcox test from the `stats` R package (R Core Team, 2023), we compare the MT text to the NST corpora in terms of rarity and find that the English and Arabic MTs all have statistically significantly lower general and domain rarity scores than the respective NST corpus. Spanish, however, does not display the same effect. In contrast, MTs using Deep, DeepL, and GT all have statistically significant higher general rarity scores than the Spanish NST corpus. However, regarding domain rarity, the difference between the Spanish NST text and the MT output using Deep, DeepL, and GT is not significant. OPUS translations are not significantly different from Spanish NST text in general rarity, but show a statistically significant reduction in domain rarity.

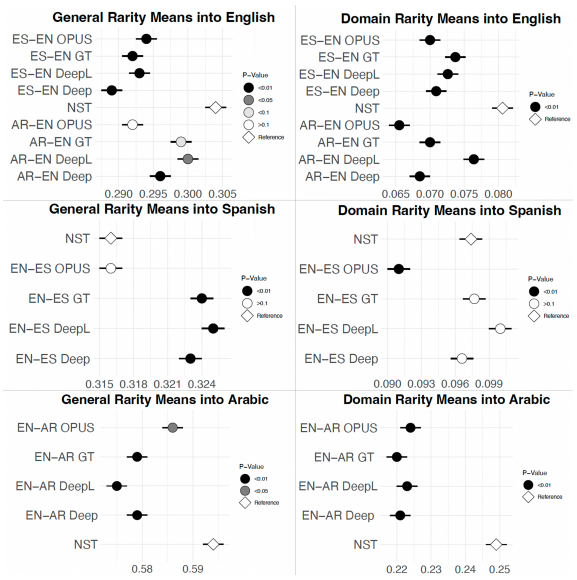


Figure 6: General and Domain Rarity Means

These results and the reduced vocabulary size show that MTs into English and Arabic generate in a significant loss of rare tokens. While this loss may simplify the text and facilitate model classification, these translationese-induced simplifications may lead to the loss of critical context where specific words and their substantive meaning are essential.

The significant increase in rarity for some MT tools into Spanish is likely due to a bias toward brevity using more complex words. While native speakers may opt for longer but simpler phrasing in NST text, the MT text may result in a brief but vocabulary-heavy phrasing. This may be due to the training data for translation into Spanish favoring these characteristics. The reduction in text complexity further indicates that there could be overfitting to MT text in fine-tuning. Models trained on the MT text, which has lower text complexity, may not perform well when tasked with classifying NST text or even text from another MT tool that introduces higher or different types of complexity. This finding, consequently, warrants additional consideration when fine-tuning using MT text.

These findings resonate with major challenges in Neural Machine Translation (NMT). First, NMT

systems perform poorly in specialized domains for which the system has not been trained for. Second, NMT systems are weak at translating low-frequency (rare) words, especially in cases where there are many inflections (as in verb conjugations in Spanish). Third, NMT systems struggle with long sentences, which are disproportionately underrepresented in the UNPC (Koehn and Knowles, 2017). Finally, Vanmassenhove et al. (2019) similarly find that MT systems fail to reach the diversity of phrasing and vocabulary of natural human language. Therefore, there is a loss of information, context, and the unique voice of the speaker/writer of the source text in this process. While event classification may not suffer from this loss, other tasks, such as Named Entity Recognition (NER), may experience poorer performance using MT.

7 Dependency Distance

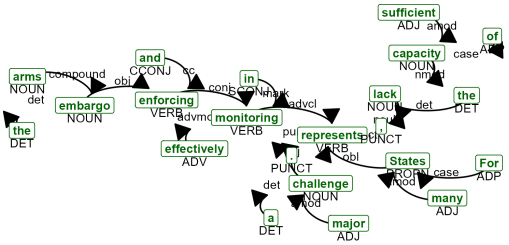
To explore the MT effects on model performance, the study also analyzes the dependency distance mean (DDM) of each sentence across languages and the dependency distance mean difference (DDM^d) of each MT output to their corresponding NST text. DDM is the average syntactical distance between the root of a sentence to other parts of speech and is generally regarded as an indicator of sentence complexity (Liu et al., 2017, 2022). A high DDM refers to highly complex sentences. Relatedly, DDM^d is interpreted here as the distortion caused by the MT tool when compared to its target NST text, such that a negative DDM^d indicates syntactical simplification and a positive DDM^d shows increasing syntactical complexity by the MT tool.

To get the DDM, we use `textdescriptives`, `spacy`, `spacy_transformers`, and libraries with `en_core_web_sm`, `bert-large-arabertv02`, and `es_core_news_sm` models for English, Arabic, and Spanish, respectively. For example, Table 1 presents the same English NST sentence in comparison to its English MT using DeepL from Arabic and Spanish texts. As Table 1 shows, small differences in the MT output are consequential for the dependency tree, the DDM, and DDM^d of the MTs into English. See Appendix I for details.

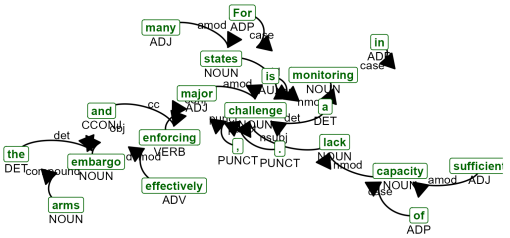
8 Sentence-Level Prediction Confidence

To better understand ConfliBERT's performance across NST and MT texts, we analyze the effects of different sentence-level characteristics on model performance. We first estimate the degree of con-

(a) EN NST: "For many States, the lack of sufficient capacity represents a major challenge in effectively monitoring and enforcing the arms embargo." DDM = 2.63.



(b) AR to EN using DeepL: "For many states, lack of sufficient capacity is a major challenge in monitoring and effectively enforcing the arms embargo." DDM = 3.09, DDM^d compared to EN NST = 0.46.



(c) ES to EN using DeepL: "The lack of sufficient capacity is a major challenge for many States in the effective monitoring and enforcement of arms embargoes." DDM = 2.45, DDM^d compared to EN NST = -0.18.

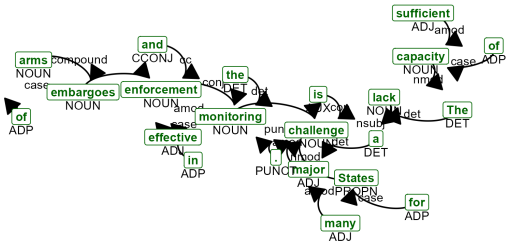


Table 1: Dependency Distance Example

fidence to which ConfliBERT correctly classifies each sentence. Then, we use regression analysis to explain the levels of prediction confidence based on various sentence-level characteristics.

To calculate prediction confidence at the sentence level, we use the ConfliBERT-uncased model in the Binary Relevant classification task. The methodology generates label predictions and confidence scores by applying the softmax function to its output logits (Devlin et al., 2018b). The methodology processes the logits of each sentence through softmax, converting them into probabilities ranging from (0,1), thus indicating the model's confidence in correctly assigning the chunk to a specific class. In this way, the prediction reflects the probability of ConfliBERT's correct classification.

For sentences longer than 512 tokens, we apply a chunking strategy, splitting them into segments

of 512 tokens each (Pappagari et al., 2019; Park et al., 2022), and independently classify and generate a predicted label and confidence score for each segment. To ensure an accurate sentence-level prediction, we apply majority voting to determine the final label and average the confidence scores across all chunks in a sentence. This method ensures that the final confidence prediction reflects the model's certainty across the entire sentence, allowing us to handle longer texts without losing context or compromising accuracy. Averaging the confidence scores provides a robust measure of the model's overall confidence.

9 Explaining Model Performance

To explore the determinants of model performance, we further analyze the sentence-level prediction confidence for the binary classification task using a linear regression model as indicated in equation 1.

$$y_i = \alpha + \beta_1 V_i + \beta_2 N_i + \beta_3 L_i + \beta_4 R_i^g + \beta_5 R_i^d + \beta_6 DDM_i + \beta_7 DDM_i^d + \epsilon_i \quad (1)$$

where y_i is the predicted confidence of ConfliBERT correctly identifying the binary classification for sentence i. The independent variables refer to sentence characteristics that could affect model performance, including the number of verbs (V_i), the noun count (N_i), unique lemmas (L_i), general rarity (R_i^g), domain rarity (R_i^d), the dependency distance mean (DDM_i), and the DDM difference (DDM_i^d) caused by MT, the latter is only included in MT texts. α and ϵ represent the intercept and the errors, respectively. To facilitate the comparison of coefficients, we standardize V_i, N_i, L_i, and DDM_i to range from 0,1 for the count measures, and a [-1,1] range for DDM_i^d. Using equation 1, we regress these sentence-level characteristics to explain ConfliBERT's performance for the binary classification task across NST and MT outputs. Appendix J reports the regression results.

Following Ward et al. (2010) and Brandt et al. (2022), we evaluate the contribution of each variable on the probability of correct classification by comparing the contribution of each sentence-level characteristic to the regression Root Mean Standard Error (RMSE) using stepwise elimination. RMSE is the standard deviation of the residuals away from the regression line. A low RMSE indicates that the observations closely revolve around the regression line, which suggests a good model fit. The stepwise elimination approach consists of first running the full regression and calculating the RMSE, then dropping one variable at a time from equation 1 and comparing the change in the RMSE from each subsequent model. A large RMSE increase after eliminating a certain variable indicates a greater model fit loss, suggesting that this variable largely contributes to the model performance. Since each regression has its own RMSE (see Appendix K), we favor the comparability of results by calculating the Model Fit Loss as a percentage using as baseline the full model's RMSE. This provides a standardized measure for cross-model comparison in which lower Model Fit Loss values indicate worse model performance after each variable elimination.

Figure 7 presents the Model Fit Loss by stepwise elimination across native and MT texts. The baseline in each panel is the full model RMSE from equation 1. The Model Fit gradually decreases after subsequently dropping each independent variable in each elimination step; the magnitude of the drop indicates the contribution of each eliminated variable. In general, Figure 7 shows that all models experience substantial performance loss after eliminating the general and domain rarity variables. This shows that general rarity and domain-specific rarity have considerable leverage in explaining ConfliBERT performance for binary classification. Text translated using OPUS seems particularly sensitive to the contributions of general and domain rarity in any translation direction.

These results offer an important finding suggesting that highly specialized words are crucial for explaining model performance. As MT tools generally reduce the vocabulary richness (see Figure 5) and decrease the number of specialized or domain-specific terms in the MT text (see Figure 6), the resulting translation output is a simplified representation of the native text containing sentences with fewer tokens and simpler words. This MT text generally makes it easier for LLMs to process. However, this performance gain comes at the cost of lower vocabulary richness in key terms.

10 Discussion and Conclusion

Table 2 presents a general summary of the main results. Based on these findings, we derive the following main conclusions: First, MT quality assessment scores provide limited insight about which MT tool performs best across classification tasks. Most quality scores recommend DeepL and OPUS as the best tool for Arabic and English, and OPUS

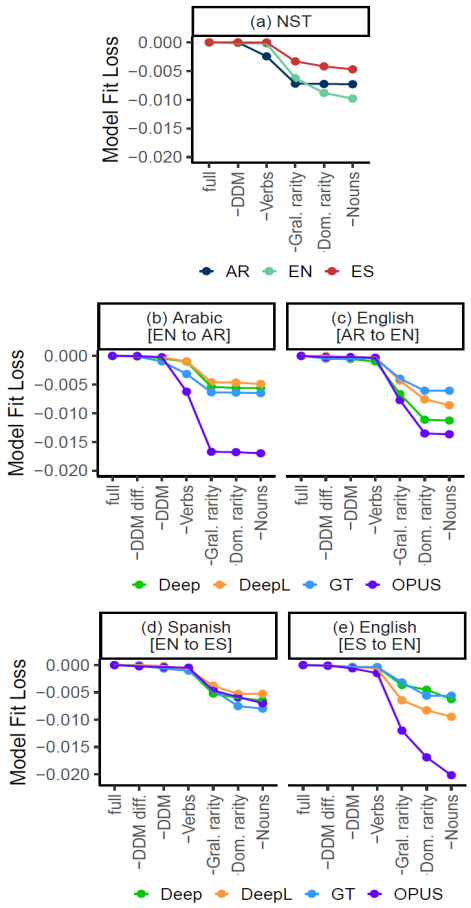


Figure 7: Model Fit Loss by Stepwise Elimination

for Spanish. However, these tools rarely outperform other MT texts across classification tasks.

Second, the sentence-level analysis reveals that all MT tools induce a reduction in vocabulary complexity. In addition, Arabic and English translations suffer from a reduction in both general and domain-specific lexical rarity. This suggests an important simplification of key terms that may be of particular relevance to domain experts. However, we detect an increase in general rarity in Spanish, and no general changes in domain rarity.

Third, although LLMs are expected to perform better with native documents than with MT texts, results across downstream tasks indicate that LLMs generally perform better with MT texts than with native corpora. Yet, no single MT tool consistently reports the top performance across languages.

Finally, using regression analysis and a stepwise deletion approach to assess the contributions of different sentence-level characteristics on model performance, the analysis indicates that highly specialized words—represented by general and domain-specific rarity—have the most leverage in explaining model performance for binary tasks.

Based on a specific application in the field of political science, this study suggests an important trade-off for the use of MT tools that could be extended to other domains. On the one hand, results indicate that MT tools may substantially reduce the time and effort for human analysis to process large volumes of text, and such MT texts tend to yield better results when using specialized LLMs for a variety of tasks. In simple terms, it seems that machines talking to machines tend to generate better results. On the other hand, the use of MT tools tends to produce translationese outputs that reduce vocabulary richness, particularly for rare terms that may be of high substantive value to domain experts. For human translators operating in highly technical fields, such vocabulary loss may prove unacceptable despite the artificially superior machine-to-machine performance.

Finding	Arabic	English	Spanish
Best MT Quality	OPUS	DeepL	OPUS
MT Voc. size	Decrease	Decrease	Decrease
Gral. rarity	Decrease	Decrease	Increase
Domain rarity	Decrease	Decrease	Not signif.
Best Binary	OPUS	Native	GT
Best QuadClass	Deep	Deep	DeepL
Best Mat. Conf.	Deep	Deep	OPUS
Best Mat. Coop.	OPUS	OPUS	GT
Best Verb. Conf.	Deep	GT	Native
Best Verb. Coop.	DeepL	DeepL	Deep
Main performance contributors	R_i^g	R_i^g and R_i^d	R_i^g and R_i^d

Table 2: Summary of Results

11 Limitations

The study has several limitations. First, this analysis is circumscribed to the political domain. Therefore, results may not be generalizable to other domains. Second, the conclusions derived from the regression analysis are based on a relatively simple binary classification task. Future research should evaluate if these findings hold in more sophisticated downstream tasks such as multi-class and multi-label classification, or named entity recognition. Third, MT tools were trained on either the UNPC itself (OPUS) or similar multilingual UN text (GT, Deep), or can be expected to have been trained on it (DeepL). MT tools can, therefore, be expected to achieve a higher translation accuracy due to their familiarity with UN text. Furthermore, some MT tools are not free, limiting their acceptability. Additionally, the study does not include other MT tools such as ChatGPT (OpenAI, 2022) or Gemini (Gimine, 2023). However, the selection of MT tools focuses on the most used tools in the chosen languages. Fourth, fine-tuning ConfliBERT on multiple languages with large datasets consumes significant time and computational resources.

References

Lars Ahrenberg. 2017. Comparing machine translation and human translation: A case study. In *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*, pages 21–28, Varna, Bulgaria. Association for Computational Linguistics.

Fouad Akki and Mohammed Larouz. 2021. A comparative study of english-arabic-english translation constraints among efl students. *International Journal of Linguistics and Translation Studies*, 2(3):33–45.

Maged Saeed Al-shaibani. 2021. Magedsaeed/farasapy: A python implementation of farasa toolkit. https://github.com/MagedSaeed/farasapy. (Accessed on 04/16/2025).

Sultan Alsarra, Luay Abdeljaber, Wooseong Yang, Niamat Zawad, Latifur Khan, Patrick Brandt, Javier Osorio, and Vito D'Orazio. 2023. ConfliBERT-Arabic: A pre-trained Arabic language model for politics, conflicts and violence. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 98–108, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Osvaldas Bartaškevičius. 2024. *Light machine translation post-editing effort and quality evaluation of news texts translated from English to Lithuanian*. Ph.D. thesis, Kauno technologijos universitetas.

Dorothée Behr and Michael Braun. 2023. How does back translation fare against team translation? an experimental case study in the language combination english–german. *Journal of survey statistics and methodology*, 11(2):285–315.

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, and James Starz. 2018. ICEWS Weekly Event Data.

Patrick T. Brandt, Vito D'Orazio, Latifur Khan, Yi-Fan Li, Javier Osorio, and Marcus Sianan. 2022. Conflict forecasting with event data and spatio-temporal graph convolutional networks. *International Interactions*, 48(4):800–822. Accessed: 2024-08-27.

Giulia Cambedda, Giorgio Maria Di Nunzio, and Viviana Nosilia. 2021. A study on automatic machine translation tools: A comparative error analysis between deepl and yandex for russian-italian medical translation. *Umanistica Digitale*, (10):139–163.

Tommaso Caselli, Osman Mutlu, Angelo Basile, and Ali Hürriyetoğlu. 2021. Protest-er: Retraining bert for protest event extraction. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 12–19, Online. Association for Computational Linguistics. Accessed: 2023-07-10.

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Erik De Vries, Martijn Schoonvelde, and Gijs Schumacher. 2018. No longer lost in translation: Evidence that google translate works for comparative bag-of-words text applications. *Political Analysis*, 26(4):417–430.

Deep Translator. 2020. deep-translator: A flexible free and unlimited python tool to translate between different languages in a simple way using multiple translators. https://github.com/nidhaloff/deep-translator. (Accessed on 04/16/2025).

DeepL. 2024. DeepL Translator. https://www.deepl.com/translator. (Accessed on 04/16/2025).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english? *Preprint*, arXiv:2308.01223.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. In L. Wollin and H. Lindquist, editors, *Translation studies in Scandinavia: Proceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II*, number 75 in Lund Studies in English, pages 88–95. CWK Gleerup, Lund.

Gimine. 2023. Gimine: Open-source data mining platform. https://gimine.com. Accessed: 2024-09-12.

Google Cloud. 2024. Google cloud translation api. Accessed: 2025-04-16.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Andrew Halterman and Katherine A. Keith. 2024. Codebook llms: Adapting political science codebooks for llm use and adapting llms to follow codebooks. *Preprint*, arXiv:2407.10747.

Andrew Halterman, Philip A. Schrodt, Andreas Beger, Benjamin E. Bagozzi, and Grace I. Scarborough. 2023. Creating custom event data without dictionaries: A bag-of-tricks. *arXiv preprint arXiv:2304.01331*.

Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, and Goran Nenadic. 2022. Examining large pre-trained language models for machine translation: What you don't know about it. *Preprint*, arXiv:2209.07417.

Pinjia He, Clara Meister, and Zhendong Su. 2021. Testing machine translation via referential transparency. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 410–422.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. Accessed: 2025-04-16.

Yibo Hu, MohammadSaleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick Brandt, and Vito D'Orazio. 2022. Conflibert: A pre-trained language model for political conflict and violence. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5469–5482.

Yan Huang and Wei Liu. 2024. Evaluating the Translation Performance of Large Language Models Based on Euas-20. *arXiv preprint*. ArXiv:2408.03119 [cs].

Ali Hürriyetoğlu, Osman Mutlu, Fırat Duruşan, Onur Uca, Alaeddin Gürel, Benjamin J. Radford, Yaoyao Dai, Hansi Hettiarachchi, Niklas Stoehr, Tadashi Nomoto, Milena Slavcheva, Francielle Vargas, Aaqib Javid, Fatih Beyhan, and Erdem Yörük. 2022. Extended Multilingual Protest News Detection - Shared Task 1, CASE 2021 and 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 223–228, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dalia Ibrahim. 2021. Google translation and the question of ideology in political news headlines. *SAHIFATUL-ALSUN*, 37(37):57–80.

Navroz Kaur Kahlon and Williamjeet Singh. 2023. Machine translation from text to sign language: a systematic review. *Universal Access in the Information Society*, 22(1):1–35.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. ArXiv preprint arXiv:1706.03872.

T. Kudo and J. Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proc. 2018 Conf. Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 66–71.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint*. ArXiv:1910.09700 [cs].

Sangmin-Michelle Lee. 2023. The effectiveness of machine translation in foreign language education: a systematic review and meta-analysis. *Computer Assisted Language Learning*, 36(1-2):103–125.

Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4):1006.

Hauke Licht, Ronja Sczepanski, Moritz Laurer, and Ayjeren Bekmuratovna. 2024. No more cost in translation: Validating open-source machine translation for quantitative text analysis. Discussion Paper.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.

Shanshan Liu and Wenxiao Zhu. 2023. An analysis of the evaluation of the translation quality of neural machine translation application systems. *Applied Artificial Intelligence*, 37(1):2214460.

Xueying Liu, Haoran Zhu, and Lei Lei. 2022. Dependency distance minimization: A diachronic exploration of the effects of sentence length and dependency types. *Humanities and Social Sciences Communications*, 9(1):1–9.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *CoRR*, abs/2006.07264.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. *Preprint*, arXiv:2006.06264.

Gaël Le Mens and Aina Gallego. 2024. Positioning political texts with large language models by asking and averaging. ArXiv preprint arXiv:2311.16639.

Open Event Data Alliance. 2018. Political language ontology for verifiable event records. `https://github.com/openeventdata/PLOVER`. Accessed: 2022-10-01.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`. Accessed: 2024-09-12.

OPUS. 2016. An overview of the OPUS collection. `https://opus.nlpl.eu/`. Accessed: 2025-04-05.

Javier Osorio, Sultan Alsarra, Amber Converse, Afraa Alshammari, Dagmar Heintze, Naif Alatrush, Latifur Khan, Patrick T. Brandt, Vito D'Orazio, Niamat Zawad, and Mahrusa Billah. 2024. Keep it local:

Comparing domain-specific llms in native and machine translated text using parallel corpora on political conflict. In *The 2nd International Conference on Foundation and Large Language Models*.

Javier Osorio, Viveca Pavon, Sayeed Salam, Jennifer Holmes, Patrick T. Brandt, and Latifur Khan. 2019. Translating CAMEO verbs for automated coding of event data. *International Interactions*, 45(6):1049–1064.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844.

Hyunji Hayley Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient classification of long documents using transformers. *Preprint*, arXiv:2203.11258.

Matt Post. 2018. A call for clarity in reporting bleu scores. ArXiv preprint arXiv:1804.08771.

Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4244–4275, Miami, Florida, USA. Association for Computational Linguistics.

Thomas Proisl. 2022. textcomplexity: Linguistic and stylistic complexity. Accessed: 2025-04-16.

R Core Team. 2023. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Accessed: 2025-04-16.

Benjamin J Radford. 2020. Multitask Models for Supervised Protests Detection in Texts. *arXiv preprint arXiv:2005.02954*.

Benjamin J. Radford. 2021. Automated dictionary generation for political eventcoding. *Political Science Research and Methods*, 9(1):157–171.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

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.

Yasser Sabtan, Mohamed Hussein, Hamza Ethelb, and Abdulfattah Omar. 2021. An evaluation of the accuracy of the machine translation systems of social media language. *International Journal of Advanced Computer Science and Applications*, 12.

Nick Schäferhoff. 2024. The history of google translate (2004–today): A detailed analysis. Accessed: 2025-04-16.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

United States Environmental Protection Agency. 2015. Greenhouse gas equivalencies calculator. Accessed: 2025-04-16.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Michael D Ward, Brian D Greenhill, and Kristin M Bakke. 2010. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375. Publisher: SAGE Publications Ltd.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, and others. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wooseong Yang, Sultan Alsarra, Luay Abdeljaber, Niamat Zawad, Zeinab Delaram, Javier Osorio, Latifur Khan, Patrick T. Brandt, and Vito D'Orazio. 2023. Conflibert-spanish: A pre-trained spanish language model for political conflict and violence. In *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, pages 287–292.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *Preprint*, arXiv:1904.09675.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational linguistics - Association for Computational Linguistics*, 50(1):237–291.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Replication Files

The data and replication files are available in GitHub at https://github.com/javierosorio/devil_in_the_details_mtsummit25.

B Ethical Considerations

This research utilizes United Nations Parallel Corpus as a source of information but does not involve human research subjects. By evaluating MT tools on political conflict and cooperation, we aim to help low-resource languages expand their high-quality data on such scarce contents (Magueresse et al., 2020). Creating a gold standard content of UN data aligned per sentence across multiple native languages sets up the foundation for researchers to use as labeled data sets for the specified languages of English, Arabic, and Spanish. And even extrapolate the work into the other three UN official languages, such as French, Chinese, and Russian, which are also considered lower-resource languages when compared to English.

C Sustainability Statement

Following Lacoste et al. (2019), this section presents the estimated energy cost and its corresponding carbon impact statement. The experiments reported in this study were conducted using the National Center for Supercomputing Applications (NCSA) in Illinois, and the University of Arizona offers High Performance Computing (HPC). The study used 418 hours of computation on type gpuA100*4 (TDP of W) hardware. The total estimated emissions are 45.14 kgCO_2eq. According to the United States Environmental Protection Agency United States Environmental Protection Agency (2015), this amount of emissions is equivalent to driving 115 miles in an average gasoline-powered passenger vehicle.

D Annotation Process

As indicated in Osorio et al. (2024), the annotation process involved eight steps:

- First, 12 human coders with domain-specific knowledge in political science and international relations received extensive training on the codebook. These annotators possessed bilingual skills in either English and Spanish, or English and Arabic.
- Second, the coders worked on various sets of randomly sampled 300 aligned sentences. For each set, we had three or four coders. Each human coder processed each individual sentence.
- Third, coders performed a first round of sentence classification blindly. Preventing coders from seeing the annotations conducted by other coders prevents artificial inter-coder correlation. Coders classified each sentence into any of the QuadClass categories or marked them as non-relevant.
- Fourth, after finishing the first round of blind annotations, coders compared their annotations in a non-blind revision round. This helps to rectify discrepancies between coders and strengthen their mastery of the codebook.
- Fifth, sentences with unanimous agreement are considered GSR annotations.
- Sixth, for those sentences in which there was no initial unanimous agreement, coders resolved disagreements in a third round of reviews to enhance inter-coder reliability.
- Seventh, for unresolved sentences, a final coder made the ultimate classification decision.
- Finally, sentences with unresolved classifications or multiple QuadClass labels were excluded from the final dataset.

E Annotation Result

Figures 8 and 9 present the distribution of annotations for the binary classification (relevant or not) and the QuadClass classification, indicating whether a sentence can be categorized as Material Conflict (Mat Conf), Material Cooperation (Mat Coop), Verbal Conflict (Verb Conf), Verbal Cooperation (Verb Coop), or not relevant. For the Binary QuadClass task, the study uses each of the QuadClass as a binary classification.

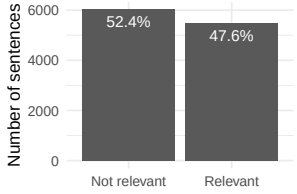


Figure 8: Binary Annotations

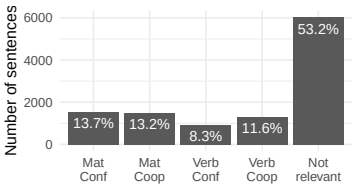


Figure 9: QuadClass Annotations

F MT Tools Training

Each of the MT tools used in this analysis was trained on different corpora of multilingual text, some of which may have included or are known to include multilingual UN documents. While GT and, consequently, Deep, as GT variant, were originally trained on UN documents, they do not specifically include the UNPC as training data (Schäferhoff, 2024). OPUS, in contrast, specifically includes the UNPC as part of its multilingual training corpora, which may result in OPUS showing exceptionally high accuracy in MT the NST text into the target language (OPUS, 2016). DeepL does not specify which training data was used to train the model but emphasizes that DeepL uses a web crawler to find and validate translations on the internet (DeepL, 2024). Consequently, it is possible that DeepL also included UN multilingual documents as training data. While all MT tools can, therefore, be assumed to have been trained on some variant of multilingual UN data, this can be expected to affect the MT output insofar as it is likely to show lower translationese effects than could be expected from an MT model that has not 'seen' the data before when compared to the original UNPC corpus. This limitation notwithstanding, we expect MT tools to differ in terms of their quality and expect their training on UN data not to favor one tool over the others.

G MT Quality Evaluation Metric Configurations

This appendix provides additional technical details related to the configuration used for the MT quality evaluation metrics implemented in the study.

- **SacreBLEU**: Implemented with default settings (tokenizer='13a', force=False, lowercase=False) to ensure reproducibility across multi-lingual settings.
- **METEOR**: Implemented via NLTK library with default settings.
- **BERTScore**: Calculated using bert-base-multilingual-cased model.
- **COMET**: Computed using wmt20-comet-da model with default configurations.

H Nouns, Verbs, and Lemmas

Figure 10 presents the noun count comparison, Figure 11 the verb count comparison, and Figure 12 the lemma count comparison across the native language corpora and MT corpora across languages.

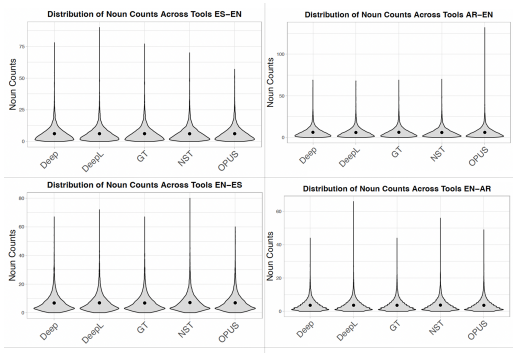


Figure 10: Noun Count Difference

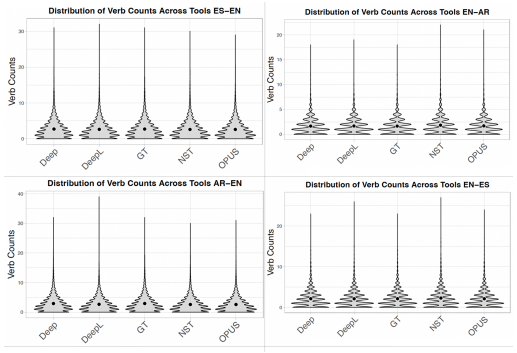


Figure 11: Verb Count Difference

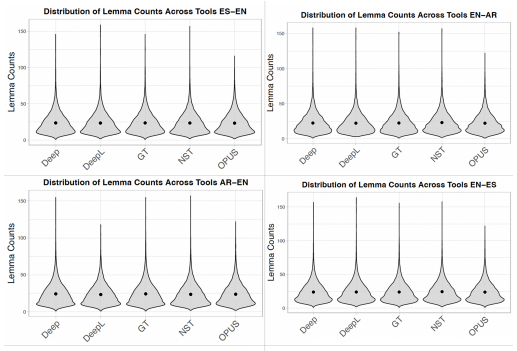


Figure 12: Lemma Count Difference

I Dependency Distance

Figure 13 presents the distribution of the dependency distance mean (DDM), and Figure 14 shows the DDM difference (DDM^d) between the MT texts and their corresponding native language.

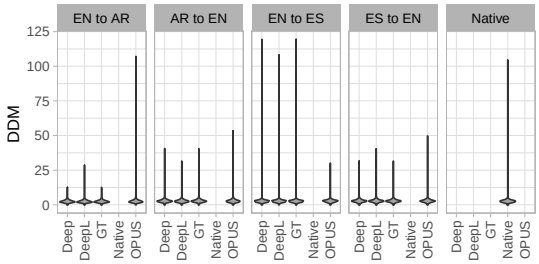


Figure 13: Dependency Distance Mean

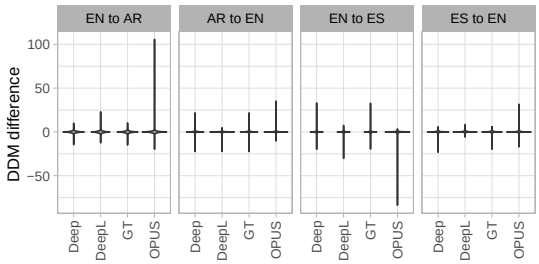


Figure 14: Dependency Distance Mean Difference

J Regression Results

Figure 15 presents the results of the regression analysis indicated in equation 1, where the dependent variable in the probability of ConfliBERT correctly categorizing each sentence in the binary classification task. Coefficients present the point estimate with confidence intervals at 95% of statistical significance. Estimates to the right of the 0 threshold indicate that such sentence characteristic increases the model performance. In contrast, estimates to the left of the threshold indicate a reduction in model performance.

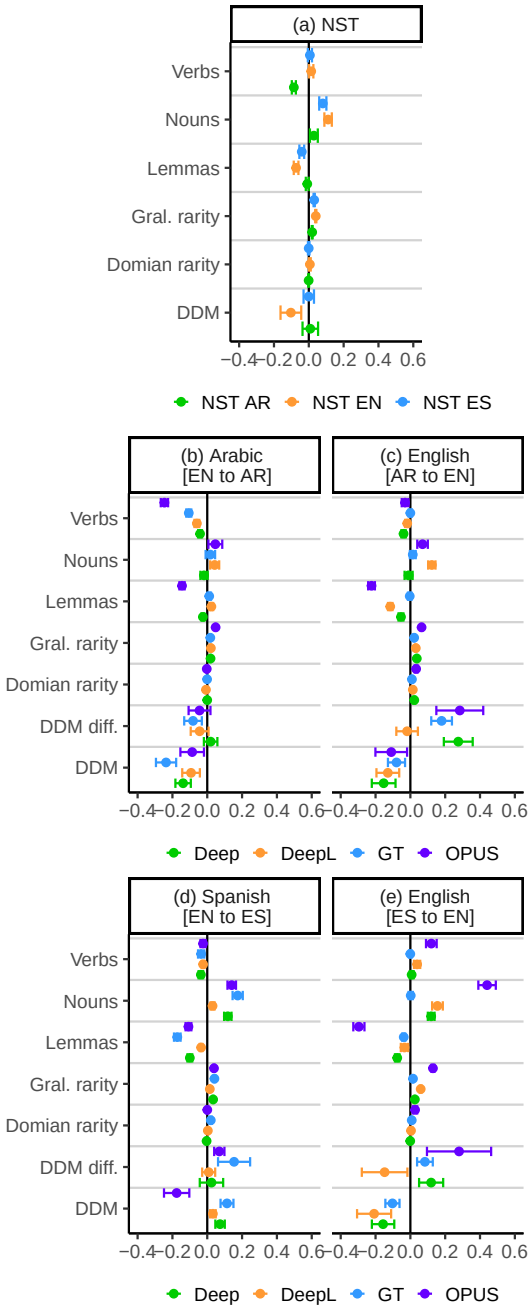
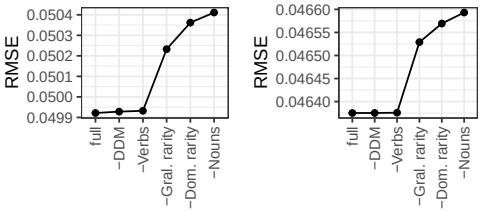


Figure 15: Determinants of Model Performance

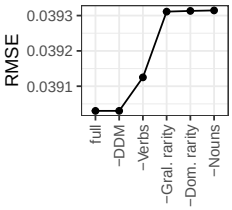
K Individual RMSE plots

The Figures in this Appendix present the original Root Mean Standard Errors (RMSE) generated by each regression. In these plots, the higher RMSE value indicates broader disturbances and, consequently, a lower model fit for ConfliBERT correctly predicting the binary classification task. Figure 16 reports the RMSE from the regressions using the native languages. Figure 17 reports the RMSE from the regressions using the Arabic to English MT output. Figure 18 reports the RMSE from the regressions using the Spanish to English MT text. Figure 19 reports the RMSE from the regressions using the English to Arabic MT documents. Figure 20 reports the RMSE from the regressions using the English to Spanish MT sentences.



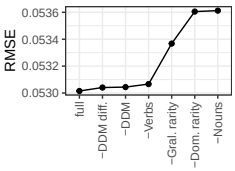
(a) EN NST

(b) ES NST

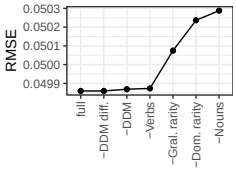


(c) AR NST

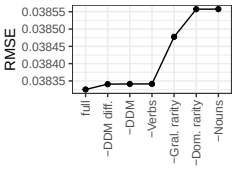
Figure 16: RMSE from NST Text



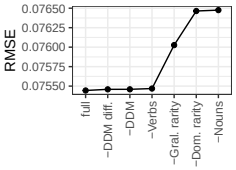
(a) AR to EN Deep



(b) AR to EN DeepL

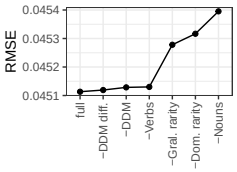


(c) AR to EN GT

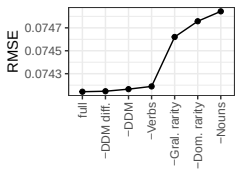


(d) AR to EN OPUS

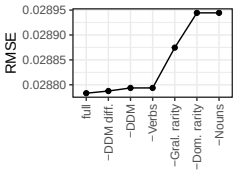
Figure 17: RMSE from Arabic (AR) to English (EN)



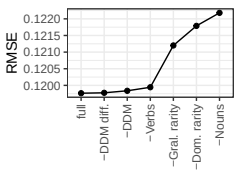
(a) ES to EN Deep



(b) ES to EN DeepL

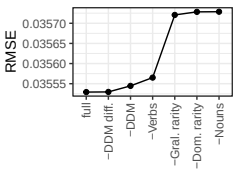


(c) ES to EN GT

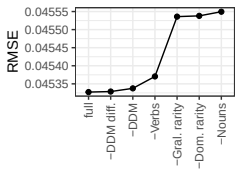


(d) ES to EN OPUS

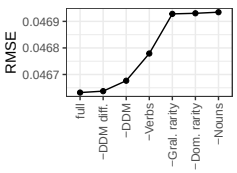
Figure 18: RMSE from Spanish (ES) to English (EN)



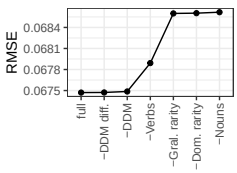
(a) EN to AR Deep



(b) EN to AR DeepL

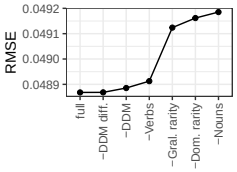


(c) EN to AR GT

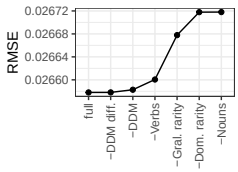


(d) EN to AR OPUS

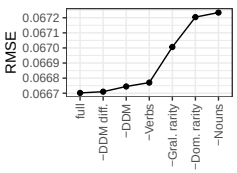
Figure 19: RMSE from English (EN) to Arabic (AR)



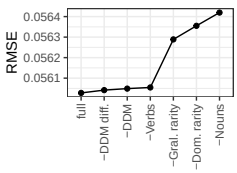
(a) EN to ES Deep



(b) EN to ES DeepL



(c) EN to ES GT



(d) EN to ES OPUS

Figure 20: RMSE from English (EN) to Spanish (ES)

L Acknowledement

This research was supported in part by NSF awards DMS-1737978, DGE-2039542, OAC-1828467, OAC-2311142, and DGE-1906630, ONR awards N00014-17-1-2995 and N00014-20-1-2738, Army Research Office Contract No. W911NF2110032. This work used Delta at NCSA / University of Illinois through allocation CIS220162 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants 2138259, 2138286, 2138307, 2137603, and 2138296. Part of the materials presented in this study were processed using the High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII).