Translationese as a Rational Response to Translation Task Difficulty

Anonymous ACL submission

Abstract

Translations systematically diverge from texts originally produced in the target language, a phenomenon widely referred to as *translationese*. Translationese has been attributed to production tendencies (e.g. interference, simplification), socio-cultural variables, and language-pair effects, yet a unified explanatory account is still lacking. We propose that translationese reflects cognitive load inherent in the translation task itself. We test whether observable translationese can be predicted from quantifiable measures of translation task difficulty. Translationese is operationalised as a segment-level *translatedness score* produced by an automatic classifier. Translation task difficulty is conceptualised as comprising source-text and cross-lingual transfer components, operationalised mainly through information-theoretic metrics based on LLM surprisal, complemented by established syntactic and semantic alternatives. We use a bidirectional English-German corpus comprising written and spoken subcorpora. Results indicate that translationese can be partly explained by translation task difficulty, especially in English-to-German. For most experiments, cross-lingual transfer difficulty contributes more than source-text complexity. Information-theoretic indicators match or outperform traditional features in written mode, but offer no advantage in spoken mode. Source-text syntactic complexity and translation-solution entropy emerged as the strongest predictors of translationese across language pairs and modes.

1 Introduction

Previous work in translation studies yielded abundant evidence that any cross-linguistic mediation exhibits systematic statistical deviations from the patterns observed in comparable non-translated, originally-authored texts in the target language. This property of translations is usually referred to as *translationese* (Gellerstam, 1986). Evidence from corpus-based and computational studies confirms that translationese is pervasive, reinforcing its view as an intrinsic feature of cross-linguistic mediation, which produces a translational variant of the target language, often described as a "third code" (Frawley, 1984; Baker, 1993). Crucially, this phenomenon appears to be largely independent of translation quality. In fact, one could argue that texts entirely devoid of translationese may border on transcreation, adaptation, or similar forms of mediated communication that prioritise creative rewriting over fidelity to the source. Such outputs, while possibly effective, diverge from the conventional goals of translation – namely, to render source content faithfully using the most appropriate means of the target language. At the same time, both the strength and the linguistic manifestation of translationese vary substantially. They are conditioned by multiple factors, most prominently translation direction (i.e. source–target pairing) (Evert and Neumann, 2017; Kunilovskaya et al., 2021), register (Redelinghuys, 2016; Kunilovskaya and Corpas Pastor, 2021), translator expertise (De Sutter et al., 2017; Bizzoni and Lapshinova-Koltunski, 2021), and mode (written vs. spoken). Extensive research has also documented recurrent translationese patterns including simplification (Corpas Pastor et al., 2008), shining-through (Teich, 2003), explicitation (Blum-Kulka, 1986; Baker, 1993), and over-normalisation (Hansen-Schirra, 2011).

Despite this wealth of descriptive findings, the field still lacks a principled account that integrates these factors within a unified explanatory framework. We therefore ask whether translationese can be understood as a consequence of cognitive constraints inherent in the translation task, and whether an information-theoretic perspective can provide such a framework.

Building on prior applications of information theory (IT) to language (Shannon, 1948; Levy, 2008), we adopt the view that linguistic encoding is shaped

by cognitive constraints: speakers select forms that balance communicative goals against processing effort (Gibson, 1998; Jaeger, 2010). We extend this perspective to translation, conceptualising it as a task in which production choices are similarly governed by cognitive pressures. Supporting evidence comes from Lim et al. (2024), who show that IT-based difficulty indicators, comparable to those used here, correlate with behavioural measures of translation processing, including production duration and eye-tracking metrics, in an experimental setting. Motivated by this work, we ask whether such indicators can also account for systematic properties of translated texts. Theoretically, more complex source texts should yield more deviant translations, particularly in simultaneous interpreting. Translationese can thus be seen as a rational trade-off between accuracy and fluency, aimed at managing cognitive load under the assumption that producing translationese-marked text may require less effort. This perspective unifies a range of recurrent deviations (explanatory wordiness, simplification, overly literal renderings, or reduced lexical variety) as cognitively rational strategies rather than lapses in translation quality.

Given the differences between monolingual and cross-lingual language processing, any explanation of translated language must consider the source language input and the cognitive demands of cross-linguistic transfer. Much research, especially following Baker (1993), has focused primarily on target-language output without including the source in their analysis. Where the source has been considered, its impact and the cognitive cost of mapping between languages were inferred indirectly, for example, by comparing translations from multiple source languages into a single target language (e.g. Evert and Neumann, 2017; Nikolaev et al., 2020; Hu and Kuebler, 2021)

In this work, we directly connect the presence of translationese in high-quality professional translations and simultaneous interpreting to the difficulty of the translation task, estimated from aligned source–target segments.[1] We investigate two questions: (1) to what extent translationese can be explained by task difficulty, and (2) how much IT measures of difficulty contribute relative to established structural features. By doing so, we address a gap in previous studies, which rarely analyse translationese in the context of aligned source material, and provide a framework for understanding translationese as a cognitive response to the demands of cross-lingual processing.

The remainder of the paper is organised as follows. Section 2 presents the methodology, including the *translatedness score*, a measure of translationese, and representations of *translation task difficulty*. Section 3 describes the experimental setup, covering data, feature handling, and modelling approach. Section 4 reports the results and discusses their implications, and Section 5 offers concluding remarks.

2 Methodology

Investigating the link between translation task difficulty and translationese involves two challenges: quantifying translationese and identifying relevant translation task parameters. The next two sections address these in turn.[2]

2.1 Translatedness Score

Previous work demonstrated that humans, including translation experts, struggle with distinguishing translated and original documents (Baroni and Bernardini, 2006; Wein, 2023), let alone quantifying translationese in a document. To address this limitation, we introduce the *translatedness score* – a measure that captures the extent to which a text exhibits properties of mediated language. This score is defined as the *probability of being a translation*, predicted by a binary classifier trained to predict the text type: translations (targets, tgt) vs. comparable originals (org) in the target language.

A central design requirement is construct validity: the classifier must meaningfully capture translatedness. This entails choosing both an appropriate feature set and an appropriate level of analysis (segment vs. document level). Translationese indicators should be structural (e.g. morphosyntactic patterns, abstract lexical cues) rather than surface features to avoid exploiting domain- or topic-related signals. Character n-grams (Popescu, 2011) or neural embeddings (Borah et al., 2023) were shown to capture the properties of texts that are not directly related to translation-specific linguistic patterns.

Translationese emerges more clearly over larger spans of text. Previous studies employing hand-

[1] "Segment" refers to an aligned unit in the parallel corpus, which may consist of one or more sentences.

[2] To facilitate reproducibility, the code and experimental data will be released upon acceptance.

engineered features (most notably, part-of-speech n-grams) have used chunks ranging from 80 to 2000 words in size (Volansky et al., 2015; Rabinovich and Wintner, 2015). For instance, for translated English and German parliamentary speeches, Kunilovskaya et al. (2024) obtained 80.0% and 88.8% accuracy on full speeches of approximately 700 tokens using a broad set of delexicalised structural features. The best result achieved in Pylypenko et al. (2021) on handcrafted features (PoS-trigrams) is 76.6% and 73.1% for the same target languages.

Although segment-level translationese classification is challenging due to a weaker signal, it presents a practical compromise. First, our translation difficulty indicators are mostly defined at the segment level, and aggregating them to larger text spans risks smoothing over differences between segments. Second, assigning a document-level translatedness score to each segment is suboptimal: it increases the number of observations but violates the assumption of independence, as all segments from the same document inherit the same score regardless of their individual properties. Finally, our focus lies in quantifying the extent of translatedness, rather than maximising classification accuracy. Preliminary experiments confirmed these concerns: both document- and segment-level regression experiments using probability-based scores from document-level classifiers produced noisy results. Based on these considerations, we opted for segment-level classification using the feature set from Kunilovskaya et al. (2024) with minor modifications. Their feature set covers 58 linguistically-motivated phenomena, including grammatical forms, morphological word classes, clause types, syntactic dependencies, word order patterns, discourse elements, and textual measures.

Generally, the feature extraction is straightforward and involves counting the occurrences of Universal Dependencies (UD) tags of the same name. For example, the value for *nn* feature is obtained by counting tokens marked with the PoS tag *NOUN*, and finite and past-tense verb forms *fin, pastv* are counted as the number of tokens marked with the morphological features *VerbForm=Fin* and *Tense=Past*. The features *amod, aux, ccomp, nmod, nsubj* and *obj* are counts of the respective dependency relation tags. Translationese predictors that require additional extraction rules or filtering and emerged as useful in our experiments are described in Table 5 (Appendix A).

Some known translationese predictors are also informative as indicators of translation task complexity. To ensure the validity of the experimental setup, three typical translationese features – lexical density, number of clauses, and mean dependency distance – are deliberately reserved as predictors for subsequent regression experiments. Pairwise collinearity between translationese predictors (measured on targets) and structural source difficulty indicators does not exceed Spearman $\rho = 0.7$. The initial pool of 55 potential translationese indicators was reduced to 37 after filtering out 18 low-variance features (see Section 3.3). For a comprehensive description and a full rationale behind the feature set, we refer the reader to the original study (Kunilovskaya et al., 2024).

2.2 Translation Task Difficulty

Translation task difficulty is conceived as comprising two components: source text comprehension effort and source-target transfer effort. Each component is represented by features that can be categorised as IT or structural (grammatical and lexical), depending on their nature.

We begin with the IT features, which provide a principled way to quantify the processing cost associated with encoding and transferring information. These IT indicators operationalise difficulty by capturing the predictability of linguistic units in context: more predictable items are easier to process and may be reduced or omitted (Delogu et al., 2017; Rutherford et al., 2017), whereas less predictable (more surprising) items incur higher cognitive effort, correlating with longer reading times (Jaeger, 2010; Hao et al., 2020) and neurophysiological responses (Huber et al., 2024; Michaelov et al., 2024). These effects are formalised in surprisal theory (Hale, 2001; Levy, 2008), rooted in Shannon's notion of information (Shannon, 1948). Within this framework, the central measure of information is *surprisal*, defined as the negative log probability of a word given its context (Equation 1).

$$S(w) = -log_2(P(w|context)) \quad (1)$$

Recent work derives these probabilities from large language models (LLMs), primarily of the GPT-2-small family, whose estimates correlate best with experimental measures of processing difficulty (Oh and Schuler, 2023; Shain et al., 2024).

This study uses *average segment surprisal (AvS)*, computed by averaging word-level surprisals

Category	Feature	Description
IT	src_gpt_AvS	Average surprisal of UD tokens in a source segment, derived from dedicated GPT-2 models (English, German).
Grammar	src_tree_depth	Maximum dependency tree depth (longest path from root to any leaf), capturing hierarchical syntactic complexity.
	src_branching	Average branching factor (number of dependents per head), capturing horizontal syntactic complexity.
	src_mdd	Mean dependency distance: linear-to-hierarchical integration effort, reflecting comprehension difficulty (Jing and Liu, 2015).
	src_n_clauses	Number of finite and non-finite clauses per segment, approximated via clausal dependencies *csubj, advcl, acl, xcomp, ccomp*.
Lexis	src_lex_dens	Ratio of unique content words (PoS-disambiguated) to total tokens.
	src_wlen	Average word length in characters, excluding punctuation.
	src_mwe	Number of multiword expressions (*flat*, *fixed*, *compound*), excluding numerals and proper names.
	src_numerals	Number of numerals, including single-word and multiword expressions.
	src_propn	Number of proper names, including single-word and multiword expressions.

Table 1: Source-side features used to model source text comprehension difficulty, grouped by feature type.

across all Stanza-defined words in a segment. AvS derived from monolingual GPT-2 models represents source comprehension difficulty and reflects the predictability of source text words within the source language. AvS derived from translation-direction-specific neural machine translation (NMT) models represents transfer difficulty and captures the uncertainty associated with source-to-target mapping. AvS was used in translation studies to correlate the information content of source and target segments (Kunilovskaya et al., 2023), in clinical linguistics to capture both lexical and structural components of sentence information (Rezaii et al., 2023) and in text readability research (Lowder et al., 2018).

Complementing NMT-based surprisal, *entropy of translation solutions* captures uncertainty over translation choices and thus reflects transfer complexity. It quantifies the number and distribution of alternative target-language realisations available for a given source item, estimated from large parallel corpora. This measure has been used as an indicator of translation difficulty in several studies (Carl et al., 2016; Teich et al., 2020; Kunilovskaya et al., 2025). Paradigmatic entropy of translation solutions is defined in Equation 2.

$$H(T \mid S) = -\sum_{t \in T} P(t \mid S) \log_2 P(t \mid S) \quad (2)$$

Here, $H(T \mid S)$ denotes the entropy of the set of translations T for a source item S. The summation iterates over all possible translation variants $t \in T$, and $P(t \mid S)$ is the probability of observing translation t given the source S. Higher entropy indicates a larger set of similarly probable translation options, implying greater cognitive demand for that source item during source–target transfer.

In addition to IT measures, the study incorporates established structural indicators of translation difficulty. Source-side complexity features, previously used in the *QuEST++* quality estimation framework (Specia et al., 2018), include type–token ratio (TTR), syntactic tree depth and width, number of clauses, and related measures. Cognitive-load-oriented studies in interpreting have further highlighted the role of lexical density, formulaicity (frequency of multiword expressions), subordinations, and numerical expressions (Plevoets and Defrancq, 2020; Kajzer-Wietrzny et al., 2024). Mean dependency distance has been shown to reflect cognitive load in interpreting (Liang et al., 2017) in addition to being a strong translationese indicator (Fan and Jiang, 2019). Source text comprehension difficulty was captured by ten features (*source* subset), presented in Table 1.

Cross-lingual transfer difficulty in this study was represented by four features (*transfer* subset), described in Table 2. All of transfer features are LLM-based and/or IT-inspired.

This basic feature set was extended with eight modified variants. The modifications included (i) AvS versions averaged over content words only (*src_gpt_AvS_content, mt_AvS_content, mean_align_content, tot_entropy_content*), (ii) AvS versions averaged over LLM subwords instead

Feature	Description
mt_AvS	Average target segment surprisal: mean surprisal of UD tokens in a target segment derived from NMT models, conditioned on the full source segment and preceding target context. Higher values indicate less predictable source-target mappings and increased transfer difficulty, hypothesised to trigger translationese.
mean_align	Mean word alignment score: computed using contextualised token representations from *bert-base-multilingual-cased*. Subword alignment is based on dot-product attention (layer 8) with bidirectional softmax filtering (10^{-2}), following *AWESoME* (Dou and Neubig, 2021). Segment-level averages over aligned word pairs; higher scores indicate stronger alignment and lower transfer difficulty.
tot_entropy	Entropy of translation solutions per segment: captures uncertainty in translating source lemmas. Entropy is computed on lemmas (not surface tokens) to account for morphological variation. Translation solution sets use word alignments from both training and test splits of *EPIC–EuroParl–UdS*. Missing values for unaligned words and singletons are replaced with the subcorpus median plus two standard deviations. Higher entropy signals less consistent translation patterns and greater cross-lingual transfer difficulty.
bleu	Teacher-forced pseudo-BLEU: segment-level BLEU between reference targets and token-wise *argmax* predictions under gold prefixes, computed with *sacrebleu*. Higher scores indicate more predictable target realisations and lower transfer difficulty.

Table 2: Transfer features used to model cross-lingual transfer difficulty.

of UD-defined surface tokens (*src_gpt_AvS_subw*, *mt_AvS_subw*) and (iii) versions computed over syntactic subtrees rather than individual words, namely subtree-based alignment (cosine similarity between aligned trees; *mean_cosine*) and entropy over trees (*tot_entropy_trees*). Subtrees were extracted using UD tags (id and head_id), limiting the depths of possible subtrees to 1. Taken together, the 22 indicators capture complementary dimensions of translation task difficulty that may shape translation decisions. The alphabetical list appears in Table 9 (Appendix B).

The general expectation is that greater difficulty increases cognitive load, under which translators prioritise content preservation over target-language naturalness, resulting in higher translatedness of the output. Accordingly, we expect a positive association with translatedness for features signalling structural and lexical complexity of the source (*src_gpt_AvS, src_branching, src_lex_dens, src_mdd, src_wlen, src_n_clauses, src_tree_depth*), translation ambiguity (*tot_entropy*) and cross-lingual predictability as reflected in MT surprisal (*mt_AvS*). Features associated with processing ease or the availability of stereotypical, structurally isomorphic solutions – such as formulaicity (*src_mwe*), alignment strength (*mean_align*), and higher MT performance (*bleu*) – are expected to correlate negatively with translatedness. Two additional indicators, numerals (*src_numerals*) and proper names (*src_propn*), are included as potentially relevant predictors, although their effect direction is less clear.

The experiments used LLM-based features from base and fine-tuned models. Fine-tuning was performed on 120k non-overlapping segments from the *EPIC–EuroParl–UdS* corpus. According to the corpus curators (Kunilovskaya and Pollkläsener, 2026), fine-tuning was thorough and robust, ensuring reliable surprisal estimates. Despite this, downstream regression gains from the fine-tuned models were limited and inconsistent across settings (see Figure 7, Appendix B); therefore, all results reported in this study use the base models.

3 Experimental Setup

3.1 Data

The data come from *EPIC-EuroParl-UdS*, a bidirectional English–German corpora containing written and spoken translations of European Parliament debates (Kunilovskaya and Pollkläsener, 2026). The written data originate from the European Parliament website and their translations. The spoken data contain manual transcriptions of video recordings of parliamentary debates and their simultaneous interpretations.

The corpus is aligned at document and segment levels and balanced across translation directions and modes. It contains UD annotation and word-level surprisal indices from GPT-2[3], and OPUS NMT de-en and en-de models (Tiedemann and Thottingal, 2020), pre-trained and fine-tuned on the written originals to adapt the base models to topical domain.

[3]English (Radford et al., 2019), German (Schweter, 2020)

Table 3 summarises the quantitative parameters of the eight text categories by mode (written vs. spoken), language pair (deen, ende), and text type (original/source vs. target).

mode	lpair	type	words	segs	docs
spoken	deen	src, org	62,900	3,060	165
		tgt	60,408	2,896	
	ende	src, org	69,801	3,260	137
		tgt	60,590	3,092	
written	deen	src, org	77,225	3,211	170
		tgt	85,380	3,216	
	ende	src, org	76,061	3,063	170
		tgt	76,588	3,063	

Table 3: Quantitative parameters of the textual data by mode, language pair (lpair) and text type.

Average segment length ranges from 25.5 to 27.8 words for written data and 19.5 to 21.4 words for spoken data, with standard deviations of approximately 15 and 13 words, respectively. Segments shorter than 4 tokens were filtered out.

3.2 Feature Preprocessing

Classification and regression features went through the same data preprocessing pipeline. Following (Evert and Neumann, 2017), morphological and lexical raw counts and per-word/subtree alignment scores are normalised by segment word count; syntactic and continuous features (`mdd`, `mhd`, `lex_dens`, `ttr`) are averaged across sentences. To enhance comparability and reduce the influence of outliers, highly right-skewed features were transformed using $log(1 + x)$ function, which compresses long tails and makes distributions more symmetric. Skewness was measured using Fisher's moment coefficient (adjusted for sample size), with a threshold of 1.0 used to identify features requiring transformation.

3.3 Machine Learning Models

Translationese classifiers and subsequent regressions were run in the same experimental setup. The machine learning (ML) models were trained separately for each language (de = German, en = English) and mode under a 10-fold cross-validation setup. We used a linear *Support Vector Machine* (SVM) with default *scikit-learn* parameters (kernel='linear', C=1.0). To prevent data leakage, segment-level experiments employed *GroupKFold*, ensuring that all segments from the same document were assigned to the same split, either training or test. Classifiers' performance was evaluated using macro-F1 (precision and recall are well-balanced in all experiments). For regression, we report Spearman correlation coefficient (ρ) between predicted scores and true translatedness, measure of regression fit (R^2) and Mean Absolute Error (MAE).

Many of the count-based features are susceptible to *low variance*, particularly at the segment level, making them noisy and uninformative for ML algorithms. Consequently, we filtered out the 18 lowest-variance features shared across all mode-language datasets. This global decision reduced models' complexity and fixed the input dimensionality at 37 features across classifiers, facilitating direct comparison of their performance. None of the difficulty features had low variance.

Additionally, separately for each mode and target language, we eliminated *collinear features* (Pearson's $r > 0.85$), retaining the feature with the higher univariate F1 score to preserve non-redundant signal. All features were independently scaled to zero mean and unit standard deviation.

Feature selection was performed using cross-validated *Recursive Feature Elimination* (RFECV) from *scikit-learn* (folds=10), which selects features iteratively within each fold. A minimum of two features was enforced. The reported performance after feature selection should be interpreted as conservative, as RFECV favours stable individual predictors and does not capture potential joint effects among weaker features.

To interpret feature contributions in the final classification and regression models, we employed *SHapley Additive exPlanations* (SHAP) (Lundberg and Lee, 2017), a unified framework based on Shapley values from cooperative game theory, which quantifies the average marginal contribution of each feature to the model's predictions across all possible feature coalitions.

4 Results

4.1 Translationese Classification

Table 4 reports macro-F1 scores averaged across folds for the four segment-level classification tasks (with standard deviations), alongside document-level results for comparison.

The performance of the segment-level classifiers is relatively low, with F1 scores around 60%, only 10 percentage points above the random baseline for a balanced binary task. The lower performance compared to document-level classification

mode	support	lang	F1±SD	selected
segment-level				
written	6,274	de	60.85±2.61	13
	6,279	en	55.74±2.94	33
spoken	6,152	de	60.44±1.62	34
	6,156	en	59.50±3.49	12
document-level				
written	340	de	82.27±6.34	21
		en	74.29±5.89	18
spoken	302	de	84.02±5.91	27
		en	76.30±6.91	26

Table 4: Performance of segment- and document-level SVM classifiers. *Selected* reflects the final feature subset after collinearity removal and RFECV (out of 37).

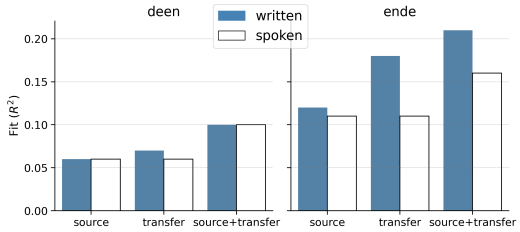
(F1 scores between 74% and 84%) is expected, given the weaker translationese signal in shorter text spans. The signal is further reduced by excluding lexical density and two syntactic complexity indicators from the translationese feature set.

Nevertheless, we argue that the classifier's probabilities for the translated class meaningfully reflect the degree of translatedness: even if a translation is predicted as an original, the lower translatedness score is a useful signal. In fact, a highly confident classifier would produce a left-skewed probability distribution, limiting variability and making the scores less informative as a gradient measure of translatedness. In our experiments, the probabilities are normally distributed, with a subtle shift to the right of the 0.5 threshold, reflecting more salient or uniform patterns in mediated language (see Figure 6, Appendix B).

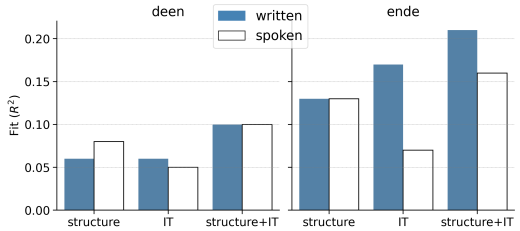
We analyse deviation patterns using feature importances. Feature-level patterns are more reliable at the document level, as segment-level signals are weakened and distorted; this is visible in Figure 3, where SHAP profiles reveal different feature strength and groupings across representations. The results further suggest that translationese is captured primarily by feature combinations rather than strongly deviant individual cues. Only a few features show notable univariate performance, especially in written data (Table 6, Appendix A). As shown in Figure 5, features without statistically significant class differences can still contribute to predictions; for example, lower frequencies of personal pronouns (*ppron*) favour the target class in spoken English (subplot d), despite the non-significant difference in Table 6.

4.2 Regression Results and Discussion

This section tests whether translation task difficulty predicts translatedness score from the classifiers and compares the explanatory power of two difficulty components (source comprehension vs. cross-lingual transfer) and two types of indicators (IT vs. structural).



(a) Source-based and transfer-specific predictors.



(b) Structural and information-theoretic predictors.

Figure 1: Variance in translatedness explained by linear regression models across language pairs and mediation modes. The two panels compare difficulty components and indicator types on the x-axis.

For ease of interpretation, R^2 values are visualised in Figure 1, while the full tabular results are provided in Appendix B (Tables 7 and 8). The regression results show that translation task difficulty explains a modest but consistent share of the variance in translatedness scores, with clear differences by direction and mode. Overall, the models account for between 6% and 21% of the variance. The explanatory power is highest for English→German direction, particularly in the written mode, where the combined model reaches $R^2 = 0.21$. In contrast, the models explain considerably less variance for German→English, where the best-performing models in both written and spoken modes reach only $R^2 = 0.10$, with smaller differences between the modes.

Across all settings, the combined *source+transfer* models consistently outperform models based on a single difficulty type, indicating that comprehension-related and transfer-related factors capture complementary aspects of translation task difficulty. Transfer features tend to be more predictive of translationese

than source complexity in written mode, but this edge disappears in interpreting, where the difficulty components are on par in both translation directions (Figure 1a).

With regard to the type of difficulty indicators (Figure 1b, Table 8), the two feature types perform comparably in German→English written translation ($R^2 = 0.06$); for English→German translation, IT features clearly outperform structural alternatives ($R^2 = 0.17$ vs. 0.13). In spoken mode, however, structural features yield substantially stronger results in both translation directions.

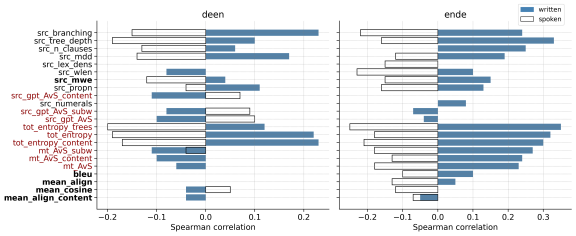


Figure 2: Univariate analysis: Spearman correlation between translation difficulty indicators and translatedness score. Source text difficulty indicators (src_) appear first. Features expected to correlate negatively with translatedness are shown in bold, and IT features are in red on the y-axis. The bars for non-significant ρ values are removed.

An unexpected finding emerged in the analysis of the directionality of individual feature contributions. Figure 2 shows the Spearman correlation between translation task difficulty features and translatedness, separately for written (blue bars) and spoken (empty bars) data in both translation directions. The results reveal striking asymmetries across modes. For many features, the blue and empty bars indicate opposite tendencies. For example, the structural source-complexity features *src_branching* and *src_tree_depth* correlate positively with translatedness in written tasks but negatively in spoken tasks across both language pairs. This suggests that while written translationese tends to mirror source complexity, spoken mode may elicit the opposite response: the more difficult the task, the more fluent and less translationese-marked the output. These findings are further supported by the SHAP importance rankings and granular statistical analysis provided in Appendix B.

The heatmap in Figure 8 shows that IT indicators of transfer difficulty, notably *mt_AvS* and *translation solution entropy*, are the strongest predictors of translationese, particularly in the English→German direction. In contrast, the lighter colour intensity for the German→English direction indicates a weaker association between translation task difficulty and translationese, consistent with the generally lower intensity of translationese in this direction.

5 Conclusion

This study examined translationese as a linguistic marker of cognitive processes involved in human translation and interpreting. By linking measures of translation task difficulty to the degree of translatedness in the target text, we explored whether systematic deviations in mediated language can be explained by the complexity of the translation task.

The results show that translationese can be partly predicted by task difficulty. Regression analyses indicate that difficulty measures explain a non-trivial share of the variance in translatedness scores (R^2 up to 0.21). The relationship is strongest in written translation and particularly pronounced for the English→German direction. Overall, indicators associated with cross-lingual transfer provide stronger explanatory signals than measures of source comprehension, suggesting that the effort involved in mapping structures across languages plays a central role in shaping translationese.

At the same time, the relation between individual features and translatedness differs across mediation modes. Based on univariate analysis, increasing task difficulty tends to correlate with more deviant targets in written mode, whereas the association is reversed in spoken mediation, pointing to different production strategies in translation and interpreting. This contrast suggests that interpreters may rely more strongly on simplification or other compensatory strategies when cognitive load increases.

The comparison of difficulty indicators further shows that IT measures perform at least as well as traditional structural features. In written translation in particular, these measures often provide stronger signals of the relation between task difficulty and translatedness.

Taken together, the findings support the view that translationese can be viewed as a rational communicative mechanism that helps manage cognitive effort while maintaining acceptable output. More broadly, the findings indicate that translationese can serve as an observable signal of cognitive load in cross-lingual processing.

Limitations

Several limitations should be considered when interpreting the results. First, the outcomes depend on the chosen feature set. Linguistic feature extraction is inherently fragile and sensitive to preprocessing decisions, annotation errors, and normalisation choices. Many frequency-based features are also sensitive to sample length, which may affect their stability across segment and document representations.

Second, latent structural dependencies between source and target features may introduce a degree of indirect information leakage, potentially confounding the observed effects of source comprehension on translationese.

Third, the analysis does not explicitly model individual speaker effects. Especially in spoken data, stylistic variation between speakers may influence feature distributions and interact with mediation-related patterns. Future work should therefore account for speaker identity and other individual-level variables.

Fourth, the results are derived from a specific corpus and set of language pairs. While the observed patterns are consistent within this dataset, their generalisability to other domains, genres, or language combinations remains to be tested. Although the analysis suggests systematic differences between written and spoken mediation, these patterns may also reflect differences in topics, communicative conditions and document length across spoken and written subcorpora, although they come from the same global source.

Finally, this study does not attempt a detailed linguistic analysis of the individual predictors and their interactions. Although the modelling results reveal systematic patterns in the feature space, a closer examination of how specific linguistic cues contribute to these patterns would provide additional explanatory depth. Conducting such analysis is left for future work.

We do not foresee direct negative societal impacts from this study. Nevertheless, metrics of translation difficulty or predictability could be misapplied in evaluative settings (e.g., assessing translators or systems) if treated as definitive indicators rather than exploratory analytical tools.

References

Mona Baker. 1993. Corpus Linguistics and Translation Studies: Implications and Applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In Honour of John Sinclair*, pages 233–250. John Benjamins, Amsterdam/Philadelphia.

Marco Baroni and Silvia Bernardini. 2006. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21(3):259–274.

Yuri Bizzoni and Ekaterina Lapshinova-Koltunski. 2021. Measuring Translationese across Levels of Expertise: Are Professionals more Surprising than Students? In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 53–63, Reykjavik. Linköping University Electronic Press.

Shoshana Blum-Kulka. 1986. Shifts of Cohesion and Coherence in Translation. In Lawrence Venuti, editor, *The Translation Studies Reader*, pages 298–313. Routledge, London.

Angana Borah, Daria Pylypenko, Cristina España-Bonet, and Josef van Genabith. 2023. Measuring Spurious Correlation in Classification: 'Clever Hans' in Translationese. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 196–206.

Michael Carl, Moritz Schaeffer, and Srinivas Bangalore. 2016. The CRITT translation process research database. In *New directions in empirical translation process research: Exploring the CRITT TPR-DB*, pages 13–54. Springer.

Gloria Corpas Pastor, Ruslan Mitkov, Naveed Afzal, and Viktor Pekar. 2008. Translation universals: do they exist? A corpus-based NLP study of convergence and simplification. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA'08)*, pages 75–81. Association for Computational Linguistics.

Gert De Sutter, Bert Cappelle, Orphée De Clercq, Rudy Loock, and Koen Plevoets. 2017. Towards a corpus-based, statistical approach to translation quality: Measuring and visualizing linguistic deviance in student translations. *Linguistica Antverpiensia*, 16:25–39.

Francesca Delogu, Matthew W Crocker, and Heiner Drenhaus. 2017. Teasing apart coercion and surprisal: Evidence from eye-movements and ERPs. *Cognition*, 161:46–59.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128.

Stefan Evert and Stella Neumann. 2017. The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German. *Empirical Translation Studies: New Methodological and Theoretical Traditions*, 300:47–80.

Lu Fan and Yue Jiang. 2019. Can dependency distance and direction be used to differentiate translational language from native language? *Lingua*, 224:51–59.

William Frawley. 1984. Prolegomenon to a theory of translation. In William Frawley, editor, *Translation: Literary, Linguistic and Philosophical Perspectives*, pages 159–175. University of Delaware Press, Newark.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. *Translation studies in Scandinavia*, pages 88–95.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 159–166, Stroudsburg, PA. Association for Computational Linguistics.

Silvia Hansen-Schirra. 2011. Between normalization and shining-through. Specific properties of English-German translations and their influence on the target language. In Svenja Kranich, Viktor Becher, Steffen Höder, and Juliane House, editors, *Multilingual discourse production: Diachronic and synchronic perspectives*, volume 12, pages 133–162. John Benjamins Publishing Company, Amsterdam.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic Predictions of People Perusing: Evaluating Metrics of Language Model Performance for Psycholinguistic Modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86.

Hai Hu and Sandra Kuebler. 2021. Investigating Translated Chinese and Its Variants Using Machine Learning. *Natural Language Engineering*, 5:339–372.

Eva Huber, Sebastian Sauppe, Arrate Isasi-Isasmendi, Ina Bornkessel-Schlesewsky, Paola Merlo, and Balthasar Bickel. 2024. Surprisal from language models can predict erps in processing predicate-argument structures only if enriched by an agent preference principle. *Neurobiology of Language*, 5(1):167–200.

Florian T. Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62.

Yingqi Jing and Haitao Liu. 2015. Mean hierarchical distance augmenting mean dependency distance. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 161–170.

Marta Kajzer-Wietrzny, Ilmari Ivaska, and Adriano Ferraresi. 2024. Fluency in rendering numbers in simultaneous interpreting. *Interpreting*, 26(1):1–23.

Maria Kunilovskaya, Koel Dutta Chowdhury, Heike Przybyl, Cristina España-Bonet, and Josef Van Genabith. 2024. Mitigating Translationese with GPT-4: Strategies and Performance. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation*, volume 1, pages 411–430. European Association for Machine Translation.

Maria Kunilovskaya and Gloria Corpas Pastor. 2021. Translationese and register variation in English-to-Russian professional translation. In Vincent X. Wang, Defeng Li, and Lily Lim, editors, *New Frontiers in Translation Studies*, pages 133–180. Springer Nature, Singapore.

Maria Kunilovskaya, Ekaterina Lapshinova-koltunski, and Ruslan Mitkov. 2021. Translationese in Russian literary texts. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 101–112. Association for Computational Linguistics.

Maria Kunilovskaya and Christina Pollkläsener. 2026. EPIC–EuroParl–UdS: Information-Theoretic Perspectives on Translation and Interpreting. In *Proceedings of the 15th Conference on Language Resources and Evaluation (LREC 2026)*. The European Language Resources Association (ELRA).

Maria Kunilovskaya, Heike Przybyl, Elke Teich, and Ekaterina Lapshinova-Koltunski. 2023. Simultaneous Interpreting as a Noisy Channel: How Much Information Gets Through. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 608–618. INCOMA Ltd.

Maria Kunilovskaya, Iuliia Zaitova, Wei Xue, Irina Stenger, and Tania Avgustinova. 2025. Predictability of microsyntactic units across slavic languages: A translation-based study. In *The Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies*, pages 313–322.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Junying Liang, Yuanyuan Fang, Qianxi Lv, and Haitao Liu. 2017. Dependency distance differences across interpreting types: Implications for cognitive demand. *Frontiers in Psychology*, 8(DEC):1–10.

Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and Trevor Cohn. 2024. Predicting human translation difficulty with neural machine translation. *Transactions of the Association for Computational Linguistics*, 12:1479–1496.

Matthew W. Lowder, Wonil Choi, Fernanda Ferreira, and John M. Henderson. 2018. Lexical Predictability During Natural Reading: Effects of Surprisal and Entropy Reduction. *Cognitive Science*, 42(S4):1166–1183.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 4765–4774.

James A. Michaelov, Megan D. Bardolph, Cyma K. Van Petten, Benjamin K. Bergen, and Seana Coulson. 2024. Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of Language*, 5(1):107–135.

Dmitry Nikolaev, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Saeboe, and Omri Abend. 2020. Morphosyntactic predictability of translationese. *Linguistics Vanguard*, 6(1):1–12.

Byung-Doh Oh and William Schuler. 2023. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Koen Plevoets and Bart Defrancq. 2020. Imported load in simultaneous interpreting: An assessment. In Ricardo Muñoz Martín and Sandra L. Halverson, editors, *Multilingual mediated communication and cognition*, The IATIS Yearbook, pages 18–43. Routledge.

Marius Popescu. 2011. Studying translationese at the character level. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2011)*, September, pages 634–639. Association for Computational Linguistics.

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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8596–8611, Online and Punta Cana.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised Identification of Translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Karien Redelinghuys. 2016. Levelling-out and register variation in the translations of experienced and inexperienced translators: A corpus-based study. *Stellenbosch Papers in Linguistics*, 45(1):189–220.

Neguine Rezaii, James Michaelov, Sylvia Josephy-Hernandez, Boyu Ren, Daisy Hochberg, Megan Quimby, and Bradford C. Dickerson. 2023. Measuring Sentence Information via Surprisal: Theoretical and Clinical Implications in Nonfluent Aphasia. *Annals of Neurology*, 94(4):647–657.

Attapol Rutherford, Vera Demberg, and Nianwen Xue. 2017. A Systematic Study of Neural Discourse Models for Implicit Discourse Relation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 281–291.

Stefan Schweter. 2020. German GPT-2 Model. *Zenodo*.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10).

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. *Quality estimation for machine translation*, volume 39 of *Synthesis Lectures On Human Language Technologies*. Morgan & Claypool Publishers. © 2018 by Morgan & Claypool.

Elke Teich. 2003. *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*, volume 5. Walter de Gruyter.

Elke Teich, José Martínez Martínez, and Alina Karakanta. 2020. Translation, Information Theory and Cognition. In Fabio Alves and Arnt Lykke Jakobsen, editors, *The Routledge Handbook of Translation and Cognition*, 1 edition, pages 360–375. Routledge.

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.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

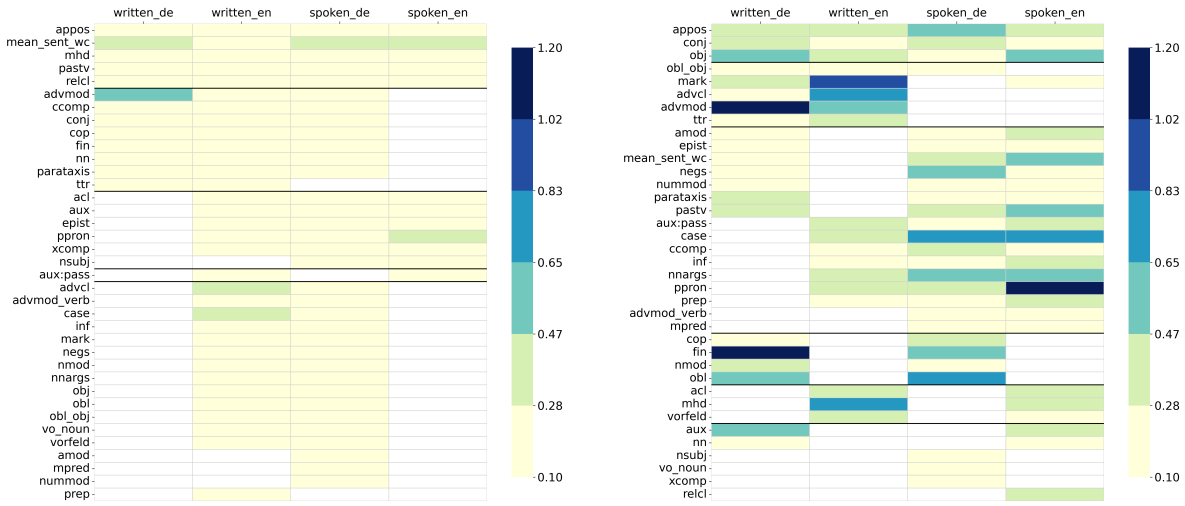
Shira Wein. 2023. Human raters cannot distinguish English translations from original English texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12266–12272, Singapore. Association for Computational Linguistics.

A Appendix. Additional Details on Segment-Level Translationese

This appendix provides details on features and their performance, beginning with a description of some complex predictors. We then look into feature impact through SHAP-based importance rankings and class-prediction contributions, and directional frequency comparisons, complemented with univariate discriminative performance and statistical analysis of differences.

Feature	Description
advmod	Adverbial modifiers per sentence, excluding negatives. Count of *advmod* dependency relations for tokens where morphological features do not contain *=Neg*.
mean_sent_wc	Mean sentence length. Count of UD tokens per sentence.
mhd	Mean hierarchical distance. Average length of paths from the root to all nodes along the dependency edges (Jing and Liu, 2015).
nnargs	Ratio of nouns, inc. proper names, as core verb arguments to the total of arguments. Count of tokens tagged as *NOUN* or *PROPN* to tokens tagged *nsubj*, *obj*, or *iobj*.
relcl	Relative clauses. Count of relative clauses in declarative sentences based on language-specific cues. For English, count of pronouns such as *which, that, whose, whom, what, who* tagged as pronouns with head relation *acl:relcl*. For German, count of pronouns like *der, welch, was, wer*, or pronouns containing *wo* with feature *PronType=Int,Rel* occurring near commas (up to 3 tokens away).
epist	Discourse markers of epistemic stance. Count of total occurrences of items from pre-defined language-specific lists (64 items for English, 74 for German, e.g., *at least, perhaps, angeblich, kein Zweifel*).
mpred	Modal predicates, including modal adjectives in predicative function (e.g., *likely, obvious, supposed, klar, unbedingt, möglich*). For English, count of modal auxiliaries (PoS tag *MD*) excluding *will* and *shall*, including non-*AUX have* followed by nearby infinitive verbs, ignoring causatives, plus modal adjectives with *AUX* dependents. For German, count of core modal verbs (*dürfen, können, mögen, müssen, sollen, wollen*) and modal adjectives in predicative function.
ppron	Personal pronouns. Count of personal pronouns in a segment, filtered by language-specific pronoun lists and morphological features: *PRON* with *Person=* feature, excluding possessive pronouns (*Poss=Yes*).

Table 5: Examples of useful translationese predictors requiring less-trivial extraction rules and/or filtering.



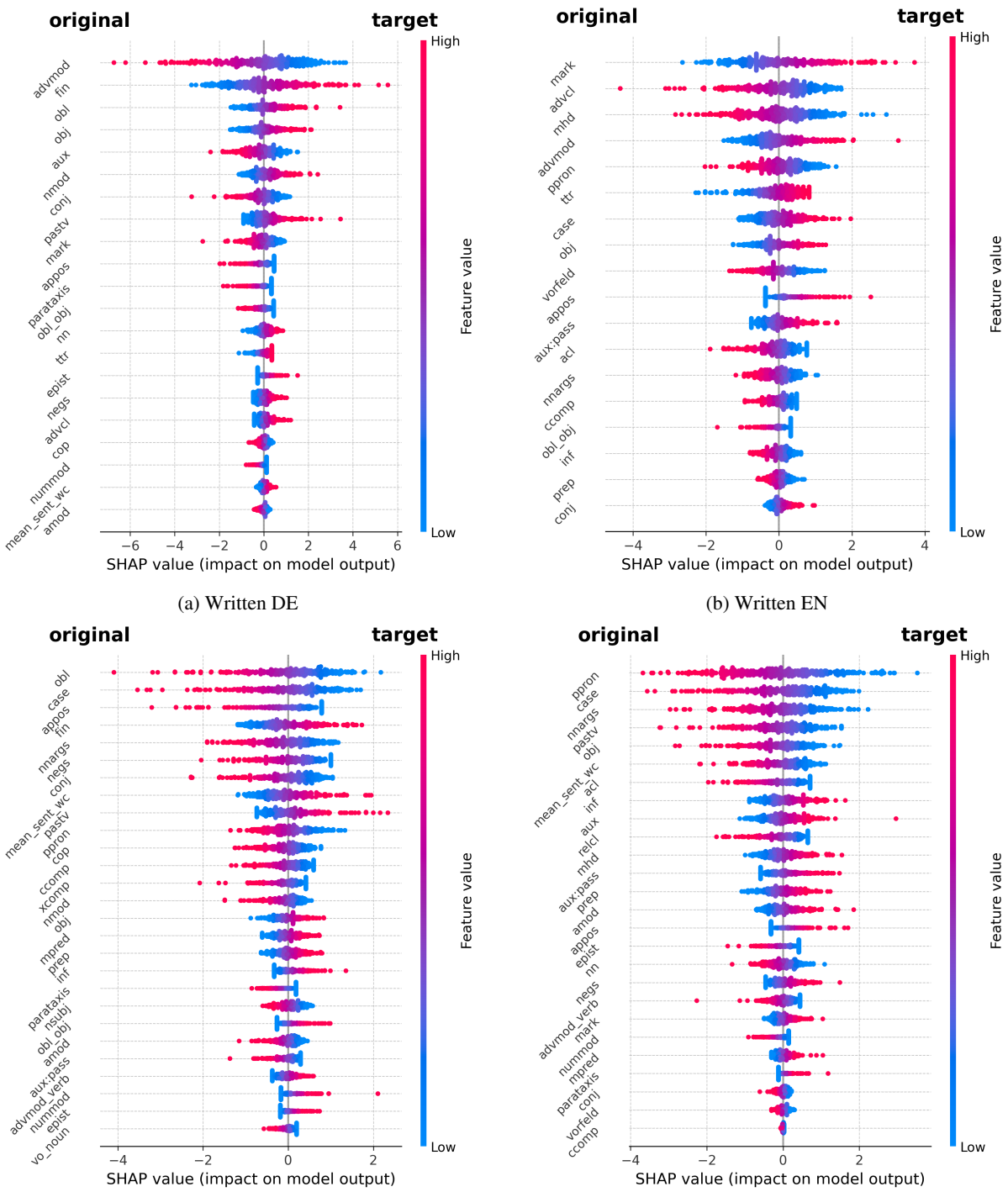
(a) Segment-level features.

(b) Document-level features.

Figure 3: Segment vs. document feature importances: Average absolute SHAP values for translationese predictors by mode and target language on a shared scale. Darker shades indicate higher feature importance; blank cells indicate that a feature was not selected for the given task. Features are grouped by selection overlap: the top block shows predictors selected by all four classifiers, followed by predictors shared by mode (written, spoken) and by target language (German, English). Segment-level analysis dilutes feature importance and obscures patterns that are clearer at the document level. The grouping sizes further suggest that mode yields more consistent deviation patterns than translation direction.

#	feature	written_de		written_en		spoken_de		spoken_en	
		tgt→org	F1_svm	tgt→org	F1_svm	tgt→org	F1_svm	tgt→org	F1_svm
1	acl	↗	58.9	–	45.1	–	51.1	↘	62.2
2	advcl	↗	55.1	–	49.8	–	46.5	↘	61.2
3	advmod	↘	**65.5**	↗	59.1	–	–	↘	56.4
4	advmod_verb	–	52.1	–	39.6	–	44.1	–	56.1
5	amod	↗	56.0	–	48.7	↘	**71.3**	↘	53.5
6	appos	↘	52.7	↗	47.9	↘	**66.5**	–	42.3
7	aux	–	40.6	↘	53.1	↘	57.0	–	–
8	aux:pass	↗	58.2	↘	54.2	–	45.5	–	38.6
9	case	↗	63.7	–	46.5	↘	**68.3**	↘	**66.0**
10	ccomp	↗	55.1	–	43.2	–	44.1	↘	58.0
11	conj	–	50.6	↗	55.7	↘	**70.5**	↘	59.9
12	cop	–	49.7	↗	52.3	–	51.8	–	50.2
13	epist	–	43.9	↗	51.5	↗	56.6	↘	56.2
14	fin	↗	60.5	–	52.9	–	54.0	↘	62.1
15	inf	–	47.7	–	47.7	↘	55.7	↘	58.5
16	mark	–	58.3	–	54.1	–	50.7	↘	57.9
17	mean_sent_wc	↗	59.5	↗	51.6	↘	**68.0**	↘	**66.9**
18	mhd	↗	63.8	–	42.4	↘	**65.8**	↘	63.5
19	mpred	–	39.1	↗	54.4	–	–	–	48.3
20	negs	–	41.6	–	49.2	↘	54.4	–	44.0
21	nmod	↗	61.0	–	–	↘	**67.8**	↘	58.2
22	nn	↗	61.0	↘	54.2	↘	64.8	↘	54.0
23	nnargs	↗	55.6	–	50.7	↘	**67.4**	↘	55.8
24	nsubj	↗	58.9	–	46.9	–	55.1	↘	60.6
25	nummod	–	50.5	–	–	↗	44.7	↘	54.1
26	obj	↗	58.4	–	46.5	–	52.5	↘	**68.7**
27	obl	↗	58.4	↗	54.5	↘	**67.0**	↘	62.8
28	obl_obj	–	50.1	–	48.5	–	50.1	↘	59.6
29	parataxis	↘	55.9	↗	53.3	–	50.0	–	–
30	pastv	↗	63.0	–	49.5	↗	60.6	↘	64.0
31	ppron	–	41.2	↘	52.5	↗	62.3	–	52.8
32	prep	↗	57.8	–	48.7	↘	64.4	↘	54.0
33	relcl	↗	52.1	–	52.4	–	47.0	↘	62.9
34	ttr	↗	60.3	–	53.5	↗	50.0	–	42.5
35	vo_noun	–	50.4	–	44.7	–	50.4	↘	60.9
36	vorfeld	↘	50.4	–	52.1	↘	55.1	↘	54.0
37	xcomp	–	–	–	48.7	↗	51.7	–	42.0

Table 6: Document-level frequency comparison between targets and originals and macro F1 scores for single-feature classifiers. '–' indicates either no statistically significant difference ($p > 0.05$) between classes or cases where a single-feature classifier produced no predictions for one class in at least one fold, rendering the metric undefined. Arrows (↗, ↘) denote significantly higher or lower average feature frequency in target texts. Three features exhibiting statistically significant class differences for all directions and modes are highlighted. Univariate performances with F1 scores > 65 are in bold. Features are sorted alphabetically.



(a) Written DE

(b) Written EN

(c) Spoken DE

(d) Spoken EN

Figure 5: SHAP beeswarm plots for document-level translationese classifications, showing *selected features* contributions to class predictions, sorted by importance for each mode-language combination. Rows correspond to mediation mode (written vs. spoken), columns to target language (German vs. English). Each point corresponds to an individual segment; colour scale from red (high) to blue (low) encodes the feature value, while horizontal position reflects the SHAP value (in the log-odds scale), i.e. the impact of that feature on the model output.

B Appendix. Regression Details

This addentix contains illustrations and supporting evidence for the claims in the main text with regard to regression analysis.

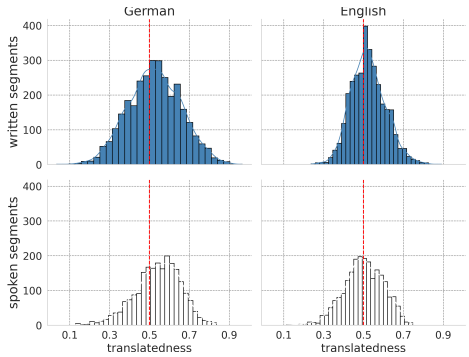


Figure 6: Probability of being a translation (translatedness score) for target segments from the four segment-level translationese classifiers.

lpair	mode	approach	ρ	R^2	MAE	selected/input
deen	written	source	0.27	0.06	0.07	8 / 11
		transfer	0.29	0.07	0.07	6 / 8
		source+transfer	**0.34**	**0.10**	0.06	18 / 19
	spoken	source	0.25	0.06	0.08	6 / 12
		transfer	0.23	0.06	0.08	8 / 9
		source+transfer	**0.30**	**0.10**	0.08	21 / 21
ende	written	source	0.36	0.12	0.10	11 / 11
		transfer	0.44	0.18	0.09	8 / 8
		source+transfer	**0.47**	**0.21**	0.09	19 / 19
	spoken	source	0.34	0.11	0.09	8 / 11
		transfer	0.35	0.11	0.09	6 / 9
		source+transfer	**0.40**	**0.16**	0.08	20 / 20

Table 7: **Difficulty dimension-based analysis:** Results of linear SVM regression for features capturing source comprehension vs. transfer difficulty as well as combined models' results. Metrics include trend, fit, and error, alongside the count of features selected for the final model.

lpair	mode	approach	ρ	R^2	MAE	selected/input
deen	written	structure	0.28	0.06	0.07	11 / 13
		IT	0.28	0.06	0.07	6 / 6
	spoken	structure	**0.28**	**0.08**	0.08	6 / 13
		IT	0.22	0.05	0.08	7 / 8
ende	written	structure	0.37	0.13	0.10	12 / 13
		IT	**0.42**	**0.17**	0.10	6 / 6
	spoken	structure	**0.36**	**0.13**	0.08	13 / 13
		IT	0.28	0.07	0.09	5 / 7

Table 8: **Analysis by feature category:** Results of linear SVM regression for difficulty features grouped by their type (IT vs structural features). In addition to performance metrics, the number of RFECV-selected features, relative to the total post-filter input is reported. The combined models' performance is the same as in Table 7.

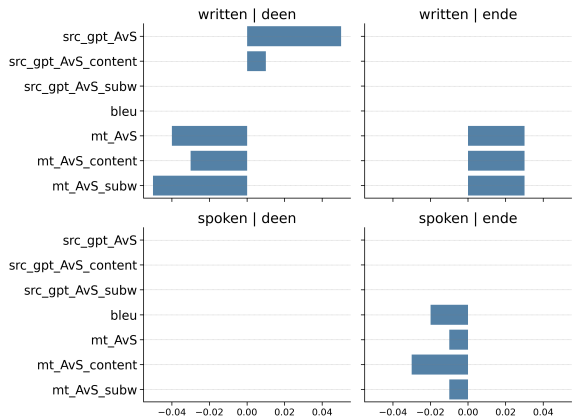


Figure 7: LLM adaptation effect on the regression outcomes, calculated as direction-aligned δ ρ (Spearman, ft - base). The bars represent how much the values of LLM-dependent features in Table 9 (and potentially in Table 7) would increase or decrease, if we used fine-tuned models.

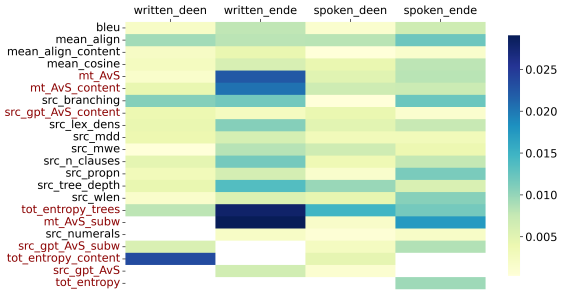
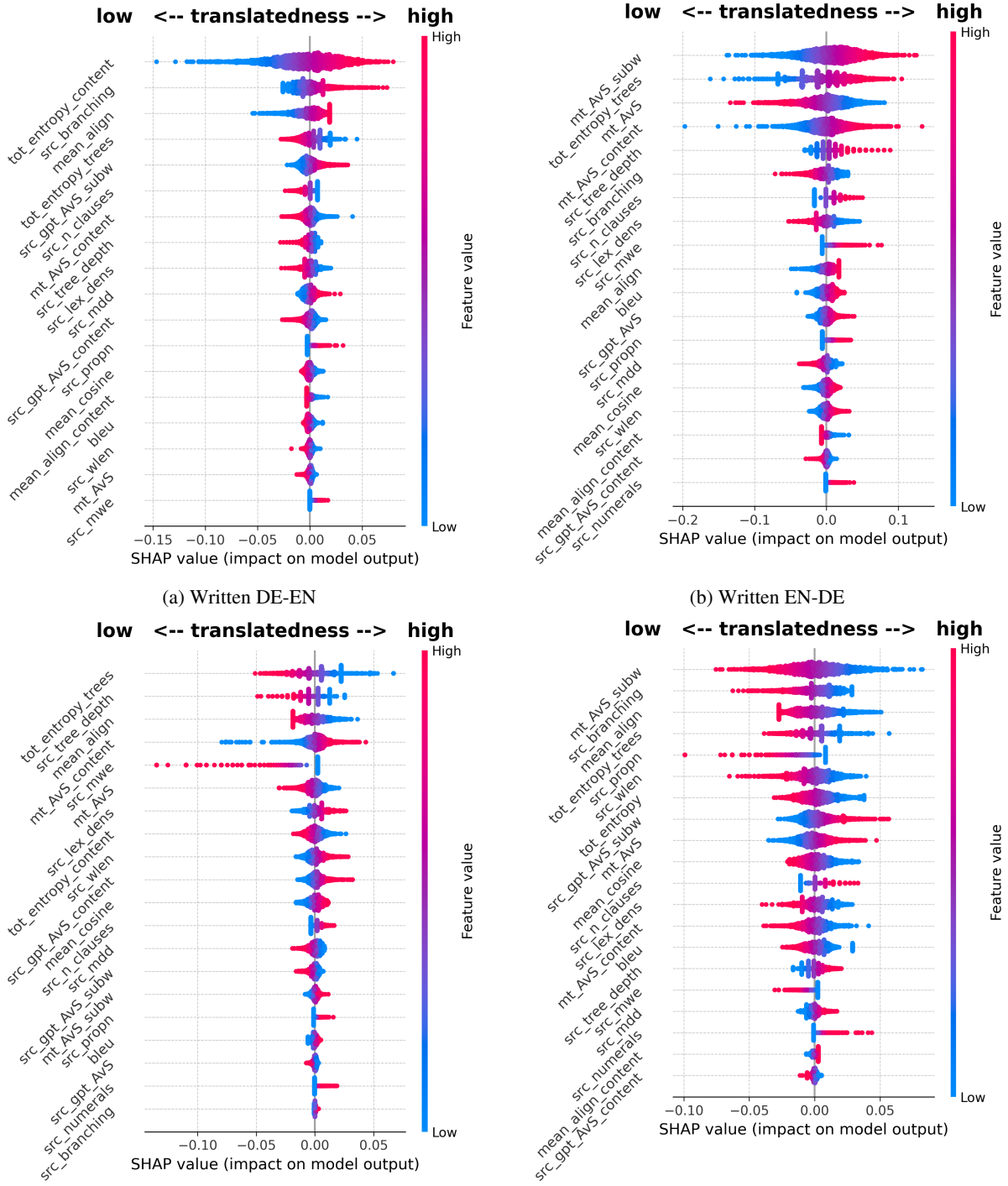


Figure 8: Regression feature importances: Average absolute SHAP values for each translation task difficulty feature across mode-direction combinations. Darker shades indicate higher feature importance, blank cells indicate no contribution in that regression task. Information-theoretical indicators are in red. Features are ordered by the number of experiments in which they were selected, then alphabetically.

#	Feature	written_deen		written_ende		spoken_deen		spoken_ende	
		mean	xy_corr	mean	xy_corr	mean	xy_corr	mean	xy_corr
1	bleu	25.87	–	21.71	0.10	14.47	–	12.85	-0.10
2	mean_align	0.84	–	0.83	0.05	0.72	–	0.72	-0.13
3	mean_align_cont.	0.83	-0.04	0.83	-0.05	0.70	–	0.70	-0.07
4	mean_cosine	0.80	-0.04	0.79	–	0.76	0.05	0.75	-0.12
5	mt_AvS	10.99	-0.06	12.59	**0.23**	12.02	–	12.10	-0.18
6	mt_AvS_content	11.82	-0.10	16.07	**0.24**	12.42	–	15.11	-0.13
7	mt_AvS_subw	10.41	-0.11	11.08	**0.27**	10.94	-0.04	10.89	-0.18
8	src_branching	1.63	**0.23**	2.27	**0.24**	1.55	-0.15	1.89	**-0.22**
9	src_gpt_AvS	6.57	-0.10	6.26	-0.04	7.34	0.10	6.74	–
10	src_gpt_AvS_cont.	8.77	-0.11	8.47	–	9.19	0.07	8.26	–
11	src_gpt_AvS_subw	5.91	-0.08	6.00	-0.07	6.60	0.09	6.36	–
12	src_lex_dens	0.41	–	0.41	–	0.42	–	0.39	-0.15
13	src_mdd	3.87	0.17	3.20	0.19	3.40	-0.14	2.92	-0.12
14	src_mwe	0.00	0.04	0.02	0.15	0.01	-0.12	0.02	-0.15
15	src_n_clauses	2.03	0.06	2.40	**0.25**	1.82	-0.13	2.08	–
16	src_numerals	0.01	–	0.01	0.08	0.00	–	0.01	–
17	src_propn	0.02	0.11	0.03	0.13	0.02	-0.04	0.03	-0.16
18	src_tree_depth	5.24	0.10	5.93	**0.33**	4.59	-0.19	5.18	-0.16
19	src_wlen	5.92	-0.08	4.76	0.10	5.85	–	4.56	**-0.23**
20	tot_entropy	51.95	**0.22**	65.14	**0.32**	41.00	-0.19	50.25	-0.18
21	tot_ent_content	22.99	**0.23**	33.57	**0.30**	18.40	-0.17	25.31	**-0.21**
22	tot_ent_trees	15.32	0.12	19.24	**0.35**	10.67	**-0.20**	12.33	**-0.25**

Table 9: Results of univariate analysis of difficulty indicators, sorted alphabetically. *mean* denotes the average feature value; *xy_corr* represents the Spearman correlation between the feature and the translatedness score. Correlations with an absolute value > 0.20 are bolded to highlight primary predictors. Dash (–) indicate non-significant correlations ($p > 0.05$).



(a) Written DE-EN

(b) Written EN-DE

(c) Spoken DE-EN

(d) Spoken EN-DE

Figure 10: SHAP beeswarm plots for regression, showing the impact of *selected features* on the predicted translatedness scores. Features are sorted by mean absolute SHAP value, indicating their relative contribution to the deviation from the model's base value for each mode-language combination. Rows correspond to mediation mode (written vs. spoken), columns to target language (German vs. English). Each point corresponds to an individual segment; colour scale from red (high) to blue (low) encodes the feature value, while horizontal position reflects the SHAP value (in the log-odds scale), i.e. the impact of that feature on the model output.