Translationese-index: Using Likelihood Ratios for Graded and Generalizable Measurement of Translationese

Yikang Liu[1*], Wanyang Zhang[2], Yiming Wang[1], Jialong Tang[3],
Pei Zhang[3], Baosong Yang[3], Fei Huang[3], Rui Wang[1#], Hai Hu[4#]
[1]Shanghai Jiao Tong University [2]Peking University
[3]Tongyi Lab [4]City University of Hong Kong
Correspondence: yikangliu@sjtu.edu.cn; wangrui12@sjtu.edu.cn; hu.hai@cityu.edu.hk

Abstract

Translationese refers to linguistic properties that usually occur in translated texts. Previous works study translationese by framing it as a binary classification between original texts and translated texts. In this paper, we argue that translationese should be graded instead of binary and propose the first measure for translationese—the translationese-index (T-index), computed from the likelihood ratios of two contrastively fine-tuned language models (LMs). We use synthesized translations and translations in the wild to evaluate T-index's generalizability in cross-domain settings and its validity against human judgments. Our results show that T-index can generalize to unseen genres, authors, and language pairs. Moreover, T-index computed using two 0.5B LMs fine-tuned on only 1-5k pairs of synthetic data can effectively capture translationese, as demonstrated by alignment with human pointwise ratings and pairwise judgments. Additionally, the correlation between T-index and existing machine translation (MT) quality estimation (QE) metrics such as BLEU and COMET is low, suggesting that T-index is not covered by these metrics and can serve as a complementary metric in MT QE.

 https://github.com/yikang0131/TranslationeseIndex

1 Introduction

Translationese refers to linguistic properties that are often introduced in the translation process that are different from those of texts originally written in that language (Gellerstam, 1986). While such properties are not inherently undesirable, they often lead to unnatural and non-native-like language that differs from idiomatic and authentic texts.

Translationese in translations, particularly machine translations (MTs), presents significant challenges in the era of large language models (LLMs).

Many multilingual resources synthesized through MT have been reported as low quality (Kreutzer et al., 2022). Models trained on these "noisy" MT datasets often struggle to generalize effectively in real-world tasks that do not involve translation (Church et al., 2025). This issue extends even to high-resource languages: for Chinese, the second most resource-rich language, LLMs frequently produce unnatural texts resembling translationese in monolingual natural language generation tasks (Guo et al., 2024).

We believe this problem can be alleviated through a quantitative measure of translationese that enables selection of the most authentic and natural translation from multiple MT outputs. However, previous attempts to detect or measure translationese have failed to meet this goal. Existing approaches either develop binary classifiers to distinguish between translated and original texts (Baroni and Bernardini, 2005; Koppel and Ordan, 2011; Volansky et al., 2015; Hu and Kübler, 2021; Pylypenko et al., 2021), or rely on distributional statistics computed across batches of texts rather than individual samples (Freitag et al., 2022; Guo et al., 2024; Li et al., 2025; Flamich et al., 2025), in which original texts are regarded as no-translationese distribution. The former approach lacks continuous measurement capabilities, while the latter cannot score individual translations.

In this paper, we want to find a graded and generalizable measurement of translationese. To this end, we first reframe how translationese should be discovered. We argue that the degree of translationese should be directly compared among translations. This new framework offers two key benefits: first, graded comparisons can be more easily observed, and second, it isolates confounding factors that arise in classification between original and translated texts, thereby enabling better generalization.

Under the new problem formulation, we propose *translationse index* (T-index) and compare it

---
[*]Work done when Yikang Liu was an intern at Tongyi Lab.
[#]Rui Wang and Hai Hu are corresponding authors.

Source text (x)

> The three spectators seemed quite stupefied. They offered no interference, and the boy...

High-translationese translation (ỹ)

> 那三个观众似乎很惊呆了。他们没有进行干涉，那个男孩和男人在地上滚作一团

Low-translationese translation (y)

> 三个旁观者愣住了，他们没敢上前阻拦，只见那男孩和男人在地上滚成一团。

(II) SFT paired scoring models

SFT a base model (θ_0) on translation task

User: Translate the following text: {source}. Assistant: {translation}.

SFT θ_0 with high-translationese → $\tilde{\theta}$

source = x | translation = ỹ | θ_0 → $\tilde{\theta}$

SFT θ_0 with low-translationese → θ

source = x | translation = y | θ_0 → θ

(III) Compute translationese index

translationese ≈ $\log P_{\tilde{\theta}}(y \mid x) - \log P_{\theta}(y \mid x)$

Using log-probability to estimate how much the stylistic feature in y resembles the training distribution of the scoring model.

θ —— translationese-index —— $\tilde{\theta}$

θ and $\tilde{\theta}$ are two reference distributions, resulting in an axis for translationese. Being closer to $\tilde{\theta}$ → more translationese
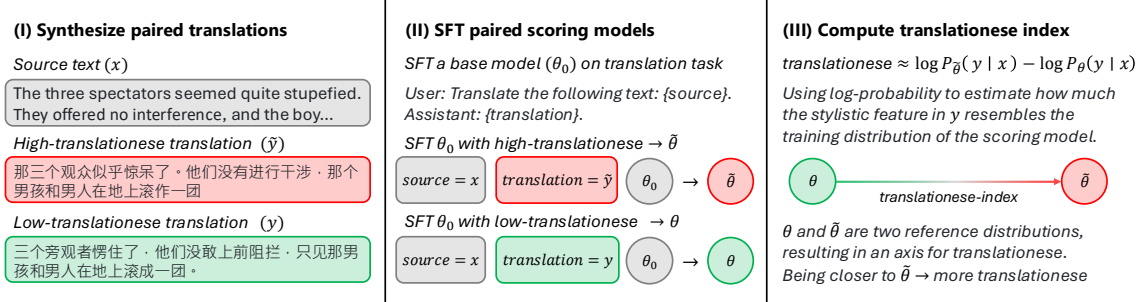
Figure 1: Illustration of the pipeline for translationese measuring using likelihood ratios of a pair of LMs.

with several supervised and unsupervised baselines on synthetic and human-annotated data. We have several major findings: (1) The proposed method is generalizable to multiple genres, authors, and language pairs, even when the backbone LMs are fine-tuned with only 1k synthesized samples. (1) T-index is correlated with both pointwise (Pearson's $r = 0.418$) and pairwise evaluations of translationese by expert human annotators. (3) T-index has very low correlation with existing MT quality estimation (QE) metrics, suggesting that it might be a novel aspect not yet covered by existing metrics.

The paper is organized as follows. We first introduce our new problem formulation and T-index in §2. Next, we describe the multi-genre synthetic benchmark used for translationese measurement in §3. In §4, we evaluate the cross-domain generalization of T-index and several baselines on the synthetic benchmark. Finally, we evaluate T-index against human annotations for MTs in the wild in §5.

2 T-index: using likelihood ratios to measure translationese

Problem formulation. Unlike previous research, which was framed as a binary text-classification task between translated and non-translated text, we compare different translations of the same source text, attempting to provide a continuous measure of translationese. The new framing is motivated by the following rationales:

1. Whether the texts are translated or not does not determine the degree of translationese, i.e., translated texts can also be authentic, and non-translated ones might be unnatural.

2. The binary-classification between translated and non-translated texts is sometimes confounded by features unrelated to translationese, such as the topic of the texts (Amponsah-Kaakyire et al.,

2022; Borah et al., 2023), hindering the discovery of translationese-specific features.

In the new formulation, we do not consider non-translated texts as having "no translationese", but instead aim to measure the degree of translationese of each individual translation.

Specifically, for a source text x and its translation y, we want to estimate the degree of translationese in y. The goal is to find a proper scoring function $f(x, y, \theta)$, parameterized by θ. For each sample (x, y), f will yield a score that directly predicts the degree of translationese.

In order to make the scoring function genuinely generalizable, we consider two possible confounding variables: (1) **genre** and (2) **author**. Ideally, our scoring function should only capture abstract translationese-specific features, rather than genre- or author-related textual features of the translations. Thus, to test whether the scoring function is generalizable and robust, we include test samples from different genres or translated by different authors.

Suppose that we have a set of authors $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$, a set of genres $\mathcal{G} = \{g_1, g_2, \ldots, g_m\}$, we can denote a dataset as $\mathcal{D}_{g_i,a_i}$, where the source text is sampled from the genre g_i, and the translation is produced by the author a_i. $\mathcal{D}$ is a paired dataset, each sample containing a translation with a higher degree of translationese ($\tilde{y}$) and a translation with a lower degree of translationese (y). The goal can be formalized as follows:

$$\max_f \sum_{(x,y,\tilde{y}) \in \mathcal{D}} \mathbb{I}\left[f(x, \tilde{y}, \theta) > f(x, y, \theta)\right].$$

Likelihood ratios as translationese index. Inspired by the success of Likelihood Ratios (LLR) in OOD detection (Ren et al., 2019), we propose to use LLR to measure translationese. Assume that the translation y can be decomposed into three independent parts $\{\mathbf{y}_g, \mathbf{y}_a, \mathbf{y}_t\}$ as

genre component, author component, and translationese component, resulting $\log \mathrm{P}_\theta(y \mid x) = \log [P_\theta(\mathbf{y}_g)P_\theta(\mathbf{y}_a)P_\theta(\mathbf{y}_t)]$. Given a paired dataset $\mathcal{D}$, we can contrastively fine-tune two scoring models θ (on low-translationese samples) and $\tilde{\theta}$ (on high-translationese samples) (see Figure 1). T-index can be formalized as:

$$\begin{aligned} \text{T-index}(y \mid x) &= \log \frac{\mathrm{P}_{\tilde{\theta}}(\mathbf{y}_g)\mathrm{P}_{\tilde{\theta}}(\mathbf{y}_a)\mathrm{P}_{\tilde{\theta}}(\mathbf{y}_t)}{\mathrm{P}_\theta(\mathbf{y}_g)\mathrm{P}_\theta(\mathbf{y}_a)\mathrm{P}_\theta(\mathbf{y}_t)} \\ &\approx \log \mathrm{P}_{\tilde{\theta}}(\mathbf{y}_t) - \log \mathrm{P}_\theta(\mathbf{y}_t). \end{aligned}$$

Intuitively, if a translation y is more likely to be high-translationese, the likelihood of y given by the low-translationese model θ should be lower than that given by the high-translationese model $\tilde{\theta}$. Since the other two components, genre and author, are shared between the two scoring models, we expect them to be canceled out. We also expect LLR to be robust in cross-domain generalization, because the shift in genre and author in testing samples should be captured by both models. We provide empirical confirmation of these assumptions in Appendix A.

Roadmap for the validation of T-index. We show the validity of T-index in two steps. In the first step, we formulate the problem as classification, but instead of having translated and non-translated texts as the classes, we generate two classes of texts with extreme degrees of translationese, using carefully controlled prompts: one with very low translationese, the other with very high translationese, and verify that human annotators agree with the low- vs. high-translationese distinction (§3). We then use T-index to perform text classification to examine the discriminative power and generalizability of T-index for texts on the two ends of translationese (§4). The second step uses real-world translations and abandons the text-classification paradigm (§5). We first ask human annotators to rate the degree of translationese in the sampled translations. We then show that T-index is in line with human ratings using various methods.

3 Constructing a synthetic benchmark of translationese

In this section, we describe how we construct a synthetic dataset containing low-translationese and high-translationese of the same set of source texts (§3.1). We then present corpus-level statistics of the

dataset, demonstrating the difference in linguistic features of the low- and high-translationese texts (§3.1). Finally, we conduct human annotation to demonstrate that humans can indeed capture these differences in degrees of translationese (§3.3).

3.1 Data generation

First, we sample English texts from 7 varied sources (**genre**), including three 19th-century novels written by Charles Dickens and samples of four genres in *The Corpus of Contemporary American English* (COCA; Davies, 2008). The three novels include *Oliver Twist* (1838), *Great Expectations* (1861), and *A Tale of Two Cities* (1859); four genres sampled from COCA include blog, news, magazine, and web texts. We only include paragraph-level samples.

Then, we translate the English texts into Chinese using two different LLMs (**author**), i.e., Qwen2.5-72B-Instruct (Qwen et al., 2025) and LLama3.3-70B-Instruct (Grattafiori et al., 2024). For each source text and translator LLM, we generate two translations using prompts with two different translation strategies: high-translationese (more literal) and low-translationese (more idiomatic). Hence, a total of 14 paired datasets are created. For each dataset, 1,000 triplets are created for training, 100 for validation, and 100 for testing. The mean translation length of low and high-translationese translations is 86.89 ± 40.52 and 83.18 ± 41.87, respectively (in tokens, tokenized by Qwen2.5). Refer to the prompts and examples for two types of translations in the Appendix B.

3.2 Dataset statistics

We compare the statistics of previously studied linguistic features between the low and high-translationese samples, as shown in Table 1. These features are reported to differ between original and translated texts in Chinese. We expect similar differences in our synthetic dataset.

We conduct independent-sample t-tests to test the significance of the difference (see Table 1). Out of the 6 features, type-token ratio, function words, pronouns, and punctuations are in alignment with what has been reported in previous literature on Chinese (Xiao and Hu, 2015; Hu et al., 2018; Hu and Kübler, 2021), suggesting that divergent linguistic patterns between our synthetic low- and high-translationese share similarity with that between the original and translated texts as previously reported and can be used as the starting point for

feature	low	high	p-value
Mean sent. length ↓	24.701	26.465	4e-36
Mean word length ↓	1.713	1.743	3e-34
Type-token ratio ↓	0.739	0.734	7e-06
Freq. of func. words ↑	0.465	0.502	4e-193
Freq. of pron. ↑	0.052	0.059	1e-37
Freq. of punct. ↓	0.156	0.152	1e-09

Table 1: Statistics of the linguistic features for the low- and high-translationese translations. ↓ indicates that the value of that feature is found to be lower in the translated Chinese compared to non-translated, and ↑ suggests the opposite.

studying T-index.

3.3 Human annotation

We conduct human evaluation for two purposes: (1) to validate whether the two types of translations exhibit adequate human-perceivable divergence, and (2) to explore how to collect graded human judgments on translationese. We experiment with two annotation methods: pointwise and pairwise annotations. Pointwise annotation provides direct continuous ratings of the translationese degree, and the pairwise annotation can provide an indirect graded judgment by comparing two translations.

Pointwise annotation. We first experiment on pointwise annotation, each annotator is presented with a source text and a translation. The annotator is asked to rate the translation on a 6-point Likert scale ranging from 0-5, where a higher rating indicates more translationese. We randomly sample 100 translations and their source texts from the synthetic datasets, with equal number of low and high-translationese samples. Three native Chinese speakers who are master students in English/Translation performed the annotation. Low-translationese samples are rated 1.90 ± 1.38 on average, much lower than high-translationese samples, which are rated 3.38 ± 1.42 ($p < 0.001$).

Pairwise annotation. We then conduct pairwise annotation, where the annotator is presented with a pair of translations, with the source text, and forced (without a tie) to choose the one that exhibits more translationese. We sample 50 triplets from the synthetic data, each containing a source text, a low-translationese translation, and a high-translationese translation. The same three annotators are asked to perform the annotation. We take the majority vote of them as the final human judgment. The

agreement between human judgments and the generation strategy is 92.0%, and the inter-rater agreement measured by Fleiss's Kappa is 0.840, which indicates very high agreement.

Both pointwise and pairwise annotations confirm that the two groups of translations exhibit valid divergence at the two ends of the translationese continuum, which are also easily detectable by English-Chinese bilingual speakers.

3.4 More language pairs

We also synthesize French and Germany translations on the basis of the same English source texts with Qwen2.5-72B-Instruct, repeating the generation scheme used for English-Chinese pairs, resulting in 14 en-de and 14 en-fr test sets. Note that we have not validated the quality of generated samples as carefully as what we did to en-zh pairs, and can only provide a preliminary exploration of the cross-linguistic generalization of various measures of translationese with the two language pairs.

4 Classifying synthetic low- and high-translationese

We start with a binary classification task, distinguishing translations of low-translationese from those with high-translationese. Specifically, we evaluate T-index along with several unsupervised (§4.1) and supervised (§4.2) baselines in the cross-domain settings.

4.1 Unsupervised baselines

We first fine-tune an LLM on translations from a specific domain. Then we can use this model as a proxy for its training distribution, and use a scoring function relying on features given by the scoring model to estimate the resemblance between the test sample and the training distribution. Let's say a scoring model is fine-tuned on high-translationese data, then this model is likely to assign higher log-likelihood to high-translationese then low-translationese.

Scoring models. On paired samples from the training domain, we SFT Qwen2.5-0.5B with the translation task to obtain two scoring models. One is fine-tuned on low-translationese translations, and the other on high-translationese translations.

Scoring functions. We include several scoring functions that are tested to be useful in machine-generated text and out-of-distribution detection

Language pair	en-zh_id						en-de_ood	en-fr_ood
Author (LLM as translator)	Qwen2.5-72B-INST_id			LLama3.3-70B-INST_ood			Qwen2.5-72B_id	
Method \ Genre	OT_{id}	$Novel_{ood}$	$COCA_{ood}$	OT_{id}	$Novel_{ood}$	$COCA_{ood}$	All	
supervised baselines: models are trained with two classes of translations with their labels								
DPO (Log-likelihood)	$89.2_{95.8}$	$86.9_{94.5}$	$82.3_{89.5}$	$89.2_{95.5}$	$86.8_{94.4}$	$80.5_{88.2}$	$77.9_{85.6}$	$82.6_{89.7}$
Bradley-Terry RM	$94.7_{98.4}$	$87.6_{94.2}$	$87.0_{93.9}$	$87.3_{94.3}$	$89.9_{95.6}$	$87.2_{93.8}$	$72.9_{79.8}$	$76.5_{84.6}$
XLM-RoBERTa	95.0 __	83.6 __	82.1 __	81.8 __	87.9 __	68.9 __	56.8 __	60.7 __
SVM w/ ling. feats.	71.0 __	73.7 __	65.1 __	72.5 __	69.7 __	63.8 __	— __	— __
scoring model $\tilde{\theta}$ is fine-tuned on high-translationese data: $\tilde{\theta} \approx \min_\theta \mathbb{E}_{(x,y,\tilde{y}) \in \mathcal{D}} [-\tilde{y} \log(f_\theta(x))]$								
Log-likelihood	$80.1_{87.1}$	$79.2_{85.9}$	$79.0_{86.1}$	$80.1_{87.0}$	$77.6_{84.8}$	$76.9_{84.4}$	$71.0_{77.7}$	$75.3_{81.3}$
Entropy	$73.8_{77.0}$	$64.8_{71.0}$	$69.1_{73.7}$	$71.6_{74.4}$	$65.4_{71.5}$	$71.0_{77.9}$	$60.3_{63.9}$	$61.1_{65.4}$
Fast-DetectGPT	$74.3_{80.8}$	$72.1_{78.9}$	$73.1_{80.5}$	$72.0_{79.2}$	$72.0_{79.7}$	$68.5_{74.9}$	$70.2_{76.6}$	$76.0_{83.6}$
Maha. Distance	$55.1_{54.3}$	$52.4_{51.0}$	$50.2_{47.2}$	$54.8_{53.6}$	$52.7_{51.3}$	$50.3_{47.5}$	$51.6_{50.8}$	$51.0_{47.7}$
Relative Maha. Dist.	$78.3_{86.0}$	$70.8_{77.9}$	$76.0_{83.3}$	$72.8_{76.9}$	$73.1_{80.4}$	$70.8_{77.4}$	$50.1_{47.9}$	$50.5_{49.5}$
Trajectory Volatility	$63.6_{69.9}$	$58.5_{59.2}$	$52.1_{49.6}$	$60.1_{60.0}$	$60.1_{62.0}$	$53.1_{52.0}$	$51.1_{49.6}$	$50.7_{47.9}$
scoring model θ is fine-tuned on low-translationese data: $\theta \approx \min_\theta \mathbb{E}_{(x,y,\tilde{y}) \in \mathcal{D}} [-y \log(f_\theta(x))]$								
Log-likelihood	$50.0_{31.5}$	$50.0_{28.6}$	$50.0_{25.6}$	$50.0_{26.9}$	$51.1_{31.0}$	$50.0_{24.4}$	$50.0_{27.4}$	$50.0_{24.9}$
Entropy	$50.3_{29.5}$	$50.0_{32.8}$	$50.0_{28.8}$	$50.0_{29.8}$	$51.1_{32.6}$	$50.0_{25.6}$	$50.0_{38.3}$	$50.0_{37.6}$
Fast-DetectGPT	$53.5_{49.0}$	$50.1_{39.3}$	$50.0_{42.2}$	$51.1_{40.9}$	$50.0_{41.9}$	$50.0_{40.0}$	$50.0_{28.8}$	$50.1_{22.2}$
Maha. Distance	$56.3_{55.0}$	$54.6_{54.3}$	$55.1_{56.1}$	$55.8_{51.5}$	$55.8_{55.2}$	$54.0_{54.4}$	$51.3_{49.6}$	$53.3_{53.2}$
Relative Maha. Dist.	$74.8_{81.9}$	$69.1_{74.9}$	$72.2_{78.5}$	$71.3_{74.5}$	$70.0_{77.0}$	$66.6_{72.3}$	$50.2_{48.9}$	$50.7_{49.9}$
Trajectory Volatility	$58.8_{59.6}$	$57.3_{58.2}$	$61.8_{65.3}$	$57.0_{56.6}$	$57.7_{58.7}$	$60.0_{62.4}$	$51.8_{50.8}$	$51.7_{50.8}$
likelihood ratios of $\tilde{\theta}$ and θ: $\log \mathrm{P}_{\tilde{\theta}}(y \mid x) - \log \mathrm{P}_\theta(y \mid x)$								
Translationese Index	$95.8_{99.2}$	$92.7_{97.8}$	$95.2_{98.8}$	$93.0_{97.9}$	$93.5_{98.3}$	$90.5_{96.2}$	$79.0_{85.5}$	$82.3_{90.0}$

Table 2: Results (accuracy reported with auroc as subscript) for the binary classification between low-translationese and high-translationese. All scoring and classifying models are trained on source-target pairs of *Oliver Twist* translated by Qwen2.5-72B-Instruct. *id* indicates the in-domain test set, and all other columns denote cross-domain test sets (all results are the average across 3 random seeds).

as unsupervised baselines and re-implement them for translationese measurement, including three logits-based functions: Log-likelihood, Entropy, and Fast-DetectGPT (Bao et al., 2024), Mahalanobis Distance (Ren et al., 2023), and three embedding-based functions: Relative Mahalanobis Distance (Ren et al., 2023), and Trajectory Volatility (Wang et al., 2024) (see Appendix C for details).

4.2 Supervised baselines

The methods above do not rely on the labels of the training samples. We also include supervised baselines trained explicitly with labeled samples, including DPO-aligned Reward (Rafailov et al., 2023, 2024), Bradley-Terry Reward Model (Bradley and Terry, 1952; Ouyang et al., 2022), XLM-RoBERTa (Conneau et al., 2020), and SVM with linguistic features (Hu and Kübler, 2021).

4.3 Evaluation metrics

We report ***accuracy*** as our metric for the binary classification task, where the majority baseline is 50%. For methods that yield a continuous score, we compute the threshold that maximizes the accuracy on each test set. We also report ***auroc***, ranging from 0 to 1, where 1 indicates perfect discrimination. For binary classifiers, we only report accuracy.

4.4 Results

We present the results of various measuring (or detecting) methods in Table 2. We mainly evaluate the cross-domain generalization, where models are trained on the single domain of *Oliver Twist* (OT) translated by Qwen2.5-72B-Instruct from English to Chinese, which is denoted as *id* (in-domain). Any other test set from a different genre, translated by another LLM, or from a different language pair, is denoted as *ood* (out-of-domain).

T-index is generalizable. On the in-domain test set, T-index almost perfectly classifies low-translationese and high-translationese. Three supervised methods, DPO with log-likelihood, Bradley-Terry RM and XLM-R, can also achieve an accuracy around 90%. When it comes to the cross-domain test sets, an increasing gap occurs between T-index and the supervised baselines. T-index remains highly discriminative under the influence of genre shift, while a drop in accuracy can be observed in supervised baselines as the genre shift away from the training domain. The limited generalizability of supervised measures is possibly due to learning domain-specific features rather than translationese features (Amponsah-Kaakyire et al., 2022). The same observation applies to the author shift. Yet in this setting, T-index also undergoes a decrease in accuracy.

When evaluated on English-Germany/French translations, the performance drops notably for even T-index. But transfer to these language pairs is non-trivial for DPO, RM, and T-index, which might result from the multilinguality of the pre-trained base model. We leave detailed cross-lingual analysis to future work.

Logits-based features are more discriminative. Performance varies for unsupervised methods depending on the features and models used for scoring. When $\tilde{\theta}$, fine-tuned on high-translationese, is used for scoring, logits-based methods generally outperform embedding-based methods. The simple log-likelihood can already achieve around 80% on most en−zh translations, and the other two methods also perform above random guessing. We suspect that embeddings might encode the semantics of translations, instead of lexical choice and word order that shape translationese, which can be better captured by the logits of LMs.

Logits might encode stylistic features in the training distribution. If the scoring model is θ fine-tuned with low-translationese data, the logits-based functions even yield scores negatively correlated with the degree of translationese (with an AUC under 50). We attribute this observation to the fact that the model fits better to the high-translationese fine-tuning data, with a lower training loss, and θ fine-tuned on low-translationese still assigns higher probability to the high-translationese translations.

The log-likelihood of an LM somehow encodes the stylistic feature that fits its training distribu-

tion[1], making it sensitive to the scoring models. Therefore, T-index can tease apart the stylistic shift caused by confounding factors, by ensembling the likelihoods of θ and $\tilde{\theta}$ that share the same distribution regarding genre and author (see Appendix A).

5 Measuring translationese in the wild

The synthetic translations used in the previous section are elicited with specifically designed prompts, under-representing real-world MTs, where the textual differences might be more nuanced. In this section, we use translations in the wild to investigate whether the continuous scores of T-index can serve as a graded measurement of translationese that aligns with human judgments.

5.1 Collection of human annotations

We sample 50 texts from the same source texts in the previous section and translate them into Chinese with 3 MT systems or LLMs (with the vanilla prompt) randomly selected from a pool of seven MT systems/LLMs (see Appendix B), obtaining 3 translations per source text. Note that we name these translations "in the wild" since we do not instruct the MT-systems or LLMs to produce translations of specific levels of translationese. Thus, all translations contain an arbitrary level of translationese that we need to find out, either with human annotators or T-index.

Following the annotation schemes in the synthetic dataset, we collect pointwise human ratings and pairwise judgments for in-the-wild MTs:

- **Pointwise**: we ask over 30 master students in Translation Studies from prestigious Chinese universities to rate the degree of translationese from 0-5 on 150 (*source*, *translation*) pairs. Each text is rated by more than 10 raters, and then we use the mean rating of each text as the result.

- **Pairwise**: a different set of five annotators is asked to choose the high-translationese translation, given 150 (*source*, *translationA*, *translationB*) triples. We take the majority vote as the ground truth.

Each annotation, for both pointwise and pairwise, takes 1-2 mins. For every 50 annotations, the annotator is compensated with 60 Chinese *yuan*.

In pairwise annotation, the inter-rater agreement (Fleiss' Kappa = 0.287) is substantially lower than

---

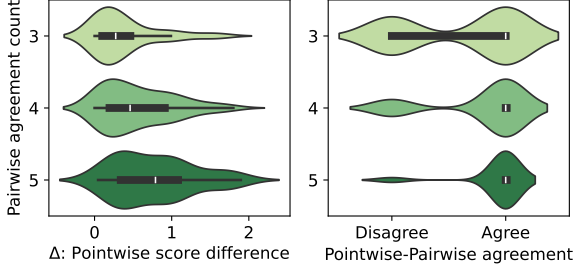[1]The training distribution can be pre-training or post-training distribution

Figure 2: Distribution of human annotation agreement patterns. Left: pointwise score differences by pairwise agreement level. Right: pairwise agreement counts by pointwise-pairwise consistency. Higher annotator consensus corresponds to larger rating differences between translation pairs.

that observed in the synthetic dataset (0.840). Upon observing the annotations closely, we interpret the lower Kappa as the following: for some documents, translation A may demonstrate more translationese in certain positions, while translation B may show translationese in other places, thus making it difficult to make a decision at the document level. However, for other documents, the choice might be easier since one translation clearly demonstrate more translationese as a whole.

Therefore, we categorize pairwise judgments into three groups based on the number of agreement for a given triple: five (unanimous agreement), four, and three out of the five annotations. This categorization captures the difficulty and disagreement in pairwise annotations, which we then compare with pointwise ratings (see Figure 2).

We observe consistent trends between pairwise and pointwise annotations. As pairwise annotation certainty increases (higher agreement among annotators), the differences in pointwise ratings between the chosen and rejected translations become more pronounced. Similarly, agreement between the two annotation methods increases when the paired translations exhibit greater differences in translationese characteristics. We believe that different levels of disagreement among annotators (manifest in the pairwise agreement count) demonstrate that different shades of translationese are indeed observed by human annotators.

5.2 T-index aligns with human judgments

For pairwise evaluation, we compute T-index for each translation and compare the values. The sample in a pair with a greater value is considered to be the prediction given by T-index. We further

include LLM-as-a-judge method for comparison: LLama3.3-Instruct-70B (Grattafiori et al., 2024) and Qwen2.5-Instruct-72B (Qwen et al., 2025) (see prompts in Appendix D).

In §4, we only use data from one single domain of the synthetic dataset for SFT to test the generalization. In this section, we compare scoring models trained with different data: (1) Unpaired samples: SFT data for the two models come from two different domains; (2) Single domain: SFT data for the two models are paired, but only one domain of data is used; (3) Mixed domain: We mix pairs from all domains together with the sample size ranging from 1k to 5k. We also fine-tune BT RM and DPO-aligned models on the same 5k samples as T-index for comparison.

Pairw. agreement count	3	4	5	Pointw.
# samples	N=45	N=66	N=39	N=150
		Agreement↑		Pearson's r↑
T-index				
w/ unpaired (1k)	64.4	66.7	82.1	34.8
w/ single-dom. (1k)	68.9	74.2	**84.6**	34.6
w/ mixed-dom. (1k)	68.9	**77.3**	82.1	32.0
w/ mixed-dom. (3k)	55.6	74.2	74.4	39.2
w/ mixed-dom. (5k)	57.8	74.2	**84.6**	**41.8**
BT RM (5k)	57.8	72.7	76.9	40.7
DPO-aligned (5k)	62.2	66.7	76.9	19.7
LLama3.3-Instruct	57.7	53.0	79.4	—
Qwen2.5-Instruct	62.2	68.1	**84.6**	—
Human Pointwise	60.0	77.3	92.3	—

Table 3: Agreement between automated methods with majority votes in the pairwise annotation (agreement reported) and correlation evaluated against mean ratings in pairwise annotation (Pearson's r reported).

Most automatic methods can predict human pairwise judgments. The results in Table 3 demonstrate that most automated methods achieve above-chance performance in agreement with human pairwise judgments. Among the evaluated methods, T-index and Qwen2.5-Instruct-72B achieve the highest accuracy of 84.6% on pairs with unanimous agreement among all five human annotators. However, a performance gap of approximately 8 percentage points remains compared to the agreement between human pairwise judgments and human pointwise ratings.

T-index correlates moderately with human pointwise ratings. When evaluated against continuous human pointwise ratings, we measure correlation strength using Pearson's r. We find that

T-index and BT RM achieve moderate-to-high correlations (∼0.4) with human mean ratings. The results suggest that training on paired translationese data, as showcased by the synthetic datasets, can help the models capture the gradience of translationese. Notably, higher accuracy on pairwise judgments does not necessarily translate to stronger correlation with pointwise ratings. For example, T-index trained on 1k samples from a single domain achieves similar pairwise accuracy to T-index trained on 5k samples, yet lags 8 percentage points behind in pointwise correlation.

T-index is data-efficient. Ablation experiments on training data in the upper half of Table 3 demonstrate that T-index is robust across different data conditions. Even when scoring models are trained on only 1k samples from different domains (unpaired), T-index can still effectively predict the degree of translationese. While increasing the amount of training data improves correlation with human pointwise ratings, it does not notably impact pairwise agreement performance.

5.3 T-index complements existing automatic QE metrics

We further explore the correlation between T-index and existing automatic QE metrics, including 3 reference-based metrics, xCOMET (Guerreiro et al., 2024), BLEURT-20 (Pu et al., 2021), and BLEU (Papineni et al., 2002), and one reference-free metric, COMET-Kiwi22 (Rei et al., 2022).

Dataset	Source (en)	Translation (zh)
Standard	original English	HT Chinese
	source is original English, and translation is HT reference.	
Reverse	MT English	original Chinese
	translation is original Chinese, and source is MT from Chinese.	
Back-translate	MT English	HT Chinese
	translation is translated Chinese, and source is back-translation.	

Table 4: Three conditions used for the overall MT QE task. The standard MT evaluation is sampled from Flores101 (Goyal et al., 2022); the original Chinese in the reverse test set is from LCMC (McEnery and Xiao, 2004); the HT Chinese in the back-translation test set is from ZCTC (Xiao and Hu, 2015). The MT English is generated by Google Translate.

The correlation is computed under 3 conditions (see Table 4): (1) standard MT evaluation, where the source is originally written and the reference is human translation; (2) reverse test set, where zh-en translations are used for en-zh evaluation, so there is no translationese in references; and

(3) back-translation test set, where the source is back-translated from the human translated references. For each condition, we sample 1000 source-reference pairs, and we use 5 LLMs to generate MTs (see Appendix B).

Results in Figure 3 show only a weak correlation between T-index and existing automatic QE metrics, which indicates that translationese features are not yet covered by them. T-index can therefore be used as a complementary metric for MT QE.

6 Related Work

Translationese identification. Linguistic theories hypothesize that certain features could serve as indicators of translationese (Toury, 1995; Baker, 1996). These features are operationalized through feature engineering in the binary classification task between original and translated texts. For instance, Volansky et al. (2015) and Hu and Kübler (2021) train machine-learning classifiers with linguistic features, such as type-token ratio, POS *n*-grams, and grammar rules. Some information-theoretic features (Lembersky et al., 2012; Rubino et al., 2016; Bizzoni et al., 2020) are also used to identify the translationese. Therefore, these features are often used to estimate the translationese level of a text. For example, Li et al. (2025) uses perplexity, lexical density, and length variance for estimate translationese to filter high-quality training data. However, these methods mostly provide corpus-level statistics, not applicable to MT QE which often requires sample-level information.

Mostly related to our work, Freitag et al. (2022) contrast the likelihood of a natural LM and translationese LM to estimate the naturalness of the MT training data, but only validate it on the classification between original and translated texts without discussing its potential to be used as a finer-grained measurement.

Translationese and translation quality. Though nonnative, translationese it is not necessarily a defect in translation. In translation theory, translators can even purposely foreignize the translations relating to the style and culture in the source language (Venuti, 1994). For HT, translationese is not an obvious indicator of poor translation quality (Kunilovskaya and Lapshinova-Koltunski, 2019). The same goes for MT, the mild translationese is also acceptable for better faithfulness and accuracy (Freitag et al., 2022; Flamich et al., 2025). However, the rigid translationese, more frequently observed
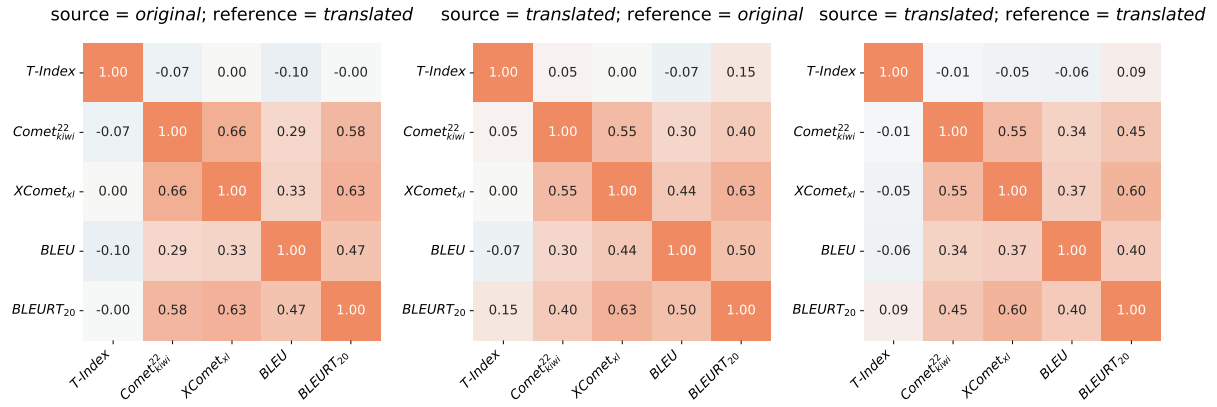
Figure 3: Correlation between the T-index and automatic MT QE metrics. To mitigate the bias that translationese in references or sources leads to, we also use the reverse test set and the back-translation test set as control groups.

in MTs (Freitag et al., 2019; Bizzoni et al., 2020; Luo et al., 2024), is what should be penalized. Translationese is part of the overall translation quality, but it is just one of the many factors that affect the quality.

Unsupervised methods in text classification. Though machine-generated text (MGT) and out-of-distribution (OOD) detection are two text classification tasks, the classification often utilizes unsupervised methods to score two classes of texts. Scores rely on the internal features of the scoring models. Models can be seen as the proxies of training distributions, and the scores quantify how the test samples resemble the training distribution, which is aligned with the objective of translationese measurement. For MGT detection, Gehrmann et al. (2019) and Solaiman et al. (2019) pioneer logits-based features, such as probability, to distinguish between human-written and machine-generated texts. Mitchell et al. (2023) and Bao et al. (2024) further propose perturbed-based methods. For OOD detection, the detection relies on signals, such as different levels of confidence, usually estimated by probabilities and entropy (Hendrycks and Gimpel, 2017; Ren et al., 2019; Hendrycks et al., 2020; Arora et al., 2021) or different geometric properties of hidden states, quantified by distance or between-layer changes (Ren et al., 2023; Jelenić et al., 2024; Wang et al., 2024). These methods are potentially applicable to translationese detection as well.

7 Conclusion and future work

In this paper, we aim to develop a graded and generalizable measure of translationese. To this end,

we reframe translationese measurement as a comparative task between different translations of the same source text, rather than binary classification between translated and non-translated text. Under this new formulation, among evaluated methods, T-index (likelihood ratios of two contrastively fine-tuned LLMs) has the best generalizability and alignment with human ratings and judgments.

We also show that T-index is weakly correlated with several automatic MT QE metrics, suggesting that T-index can be a complementary measure to existing MT QE metrics, which is especially important when existing MT QE metrics focus more on the accuracy and become less reliable in evaluating MTs of higher quality produced by LLMs (Agrawal et al., 2024; Kocmi et al., 2024).

Our work complements previous studies that view translationese as features distinguishing translated from non-translated texts by a new perspective: translationese can also relate to readers' linguistic intuition directly. Building on this, future work could investigate more scalable annotation methods to capture this intuition through comprehensive human experiments. These annotations would then enable the development of finer-grained automated measures for MT system evaluation and post-training.

Beyond MT, this work can be extended to other natural language generation tasks. While binarized features like accuracy or factualness can be more easily automated, there remains a class of features that require graded measurement. Features like translationese or naturalness are more nuanced yet equally essential to the reading experience of LLM-generated texts, opening up important directions for future automated evaluation methods.

Limitations

We primarily verified T-index on the English-Chinese language pair. Further research is needed to see whether the results are generalizable to other language pairs, especially when the two languages are similar and translationese is more difficult to define and detect. We only conduct preliminary human experiments, collecting a limited number of human annotations both in the synthetic benchmark and the in-the-wild dataset. Future research can build upon our work to collect more human annotations to examine human (dis)agreement in greater depth.

Ethics Statements

The source texts used in this paper are sampled from classic novels and well-curated corpora, and the translations are produced by open-sourced LLMs. Therefore, we believe that there is no risk of leakage of personal identifiable information or any ethical issues. The data used in this paper are only intended for research concerning the MT evaluation and should not be interpreted otherwise.

Acknowledgements

References

Sweta Agrawal, António Farinhas, Ricardo Rei, and Andre Martins. 2024. Can automatic metrics assess high-quality translations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14491–14502, Miami, Florida, USA. Association for Computational Linguistics.

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

Udit Arora, William Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10687–10701, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mona Baker. 1996. Corpus-based translation studies: The challenges that lie ahead. In *Terminology, LSP and Translation*, page 175. John Benjamins.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.

Marco Baroni and Silvia Bernardini. 2005. 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, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

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, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Kenneth Church, Boyang Li, Peter Vickers, Shiran Dudy, and Richard Yue. 2025. Emerging trends: translationese. *Natural Language Processing*, 31(3):965–981.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mark Davies. 2008. The corpus of contemporary american english (coca). https://www.english-corpora.org/coca/. Available online, last accessed on February 16, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Gergely Flamich, David Vilar, Jan-Thorsten Peter, and Markus Freitag. 2025. You cannot feed two birds with one score: the accuracy-naturalness tradeoff in translation. *Preprint*, arXiv:2503.24013.

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

Markus Freitag, David Vilar, David Grangier, Colin Cherry, and George Foster. 2022. A natural diet: Towards improving naturalness of machine translation output. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3340–3353, Dublin, Ireland. Association for Computational Linguistics.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. In L. Wollin and H. Lindquist, editors, *Translation Studies in Scandinavia, vol. 1*, pages 88–95. CWK Gleerup.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2024. Do large language models have an english accent? evaluating and improving the naturalness of multilingual llms. *Preprint*, arXiv:2410.15956.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. Association for Computational Linguistics.

Hai Hu and Sandra Kübler. 2021. Investigating translated chinese and its variants using machine learning. *Natural Language Engineering*, 27(3):339–372.

Hai Hu, Wen Li, and Sandra Kübler. 2018. Detecting syntactic features of translated Chinese. In *Proceedings of the Second Workshop on Stylistic Variation*, pages 20–28, New Orleans. Association for Computational Linguistics.

Fran Jelenić, Josip Jukić, Martin Tutek, Mate Puljiz, and Jan Snajder. 2024. Out-of-distribution detection by leveraging between-layer transformation smoothness. In *The Twelfth International Conference on Learning Representations*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, and 1 others. 2024. Findings of the wmt24 general machine translation shared task: the llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46.

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

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Maria Kunilovskaya and Ekaterina Lapshinova-Koltunski. 2019. Translationese features as indicators of quality in English-Russian human translation. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 47–56, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825.

Yafu Li, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025. Lost in literalism: How supervised

training shapes translationese in llms. *Preprint*, arXiv:2503.04369.

Jiaming Luo, Colin Cherry, and George Foster. 2024. To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation. *Transactions of the Association for Computational Linguistics*, 12:355–371.

Anthony McEnery and Zhonghua Xiao. 2004. The Lancaster corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. 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, Dominican Republic. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. 2024. From r to q^*: Your language model is secretly a q-function. In *First Conference on Language Modeling*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. 2019. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2023. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef Van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 960–970.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *Preprint*, arXiv:1908.09203.

Gideon Toury. 1995. *Descriptive Translation Studies–and beyond*. John Benjamins.

Lawrence Venuti. 1994. *The Translator's Invisibility: A History of Translation*, 1st edition edition. Routledge.

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

Yiming Wang, Pei Zhang, Baosong Yang, Derek F. Wong, Zhuosheng Zhang, and Rui Wang. 2024. Embedding trajectory for out-of-distribution detection in mathematical reasoning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Richard Xiao and Xianyao Hu. 2015. *Corpus-Based Studies of Translational Chinese in English-Chinese Translation*. New Frontiers in Translation Studies. Springer Berlin, Heidelberg, Berlin Heidelberg.

A   Empirical confirmation of T-index

To confirm that different components are independent, we quantify the three types of shifts between the training distribution and the testing distribution. First, we use the mean log-likelihood (MLL) as the statistics to measure a distribution represented by the dataset $\mathcal{D}$, which is defined as:

$$\mathrm{MLL}(\mathcal{D};\theta) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \frac{1}{|y|} \log \mathrm{P}_\theta(y \mid x). \quad (1)$$

Then, we obtain the MLL of the training distribution as a reference value, where $\mathrm{ref\_MLL} = \mathrm{MLL}(D_{g_i,a_i,t_i}; \theta_{g_i,a_i,t_i})$. With this reference, we can measure the distribution shift between D_{g_i,a_i,t_i} and testing distribution D_{g_j,a_j,t_j} by the difference between $\mathrm{MLL}(D_{g_j,a_j,t_j}; \theta_{g_i,a_i,t_i})$ and $\mathrm{ref\_MLL}$.

To measure the shift of each independent component, we keep the other two components the same as the training distribution but change the targeted one to the value of the testing distribution. Then the four types of shifts are defined as follows[2]:

$$\mathrm{o\_shift} = \mathrm{MLL}(D_{g_j,a_j,t_j}) - \mathrm{ref\_MLL}$$
$$\mathrm{g\_shift} = \mathrm{MLL}(D_{g_j,a_i,t_i}) - \mathrm{ref\_MLL}$$
$$\mathrm{a\_shift} = \mathrm{MLL}(D_{g_i,a_j,t_i}) - \mathrm{ref\_MLL}$$
$$\mathrm{t\_shift} = \mathrm{MLL}(D_{g_i,a_i,t_j}) - \mathrm{ref\_MLL}$$

We obtain 28 models fine-tuned on the 28 datasets and evaluate them on all 28 tests, resulting in 784 observations in total. We run an OLS regression on the four types of shifts, where overall_shift ∼ genre_shift +

---

[2]Note that o_shift is the overall shift, g_shift is the genre shift, a_shift is the author shift, and t_shift is the translationese shift, and the scoring model here is trained on D_{g_i,a_i,t_i}, for simplicity, we omit the parameter θ_{g_i,a_i,t_i} in the notation.

author_shift + translationese_shift. The linear model explains 97.3% of the variance in the overall_shift; genre_shift ($\beta = 1.0129$, $p < 0.001$), translationese_shift ($\beta = 0.9813$, $p < 0.001$), and author_shift ($\beta = 0.7527$, $p < 0.001$) are all significant predictors. The variance inflation factor (VIF) of the three predictors is 1.001, 1.000, and 1.002, respectively, indicating that the three predictors are independent. The assumption (a) of LLR is empirically confirmed.

With the independence of the three types of shifts, the follow-up question is whether a contrastively fine-tuned model can cancel out the genre and author shifts, the assumption (b). First, we define a model $\theta_{g_i,a_i,t}$ as the model fine-tuned on the dataset $\mathcal{D}_{g_i,a_i,t}$. The contrastively fine-tuned model is $\theta_{g_i,a_i,\tilde{t}}$. We run a paired t-test for the values of genre_shift and author_shift given by θ_t and $\theta_{\tilde{t}}$, which turns out to be insignificant for genre_shift ($p = 0.808$) but significant for author_shift ($p < 0.001$). The results also explain the performance drop of LLR with the author shift in Table 2.

The empirical confirmation suggests that for each model, the log-likelihood can be decomposed into three independent components. However, the unwanted shifts are mostly, but not completely, canceled out by a contrastively fine-tuned model.

B   Details of the datasets

B.1   Synthetic benchmark

Please refer to Table 5 for the prompts used to generate low- and high-translationese in the synthetic benchmark and translation examples.

B.2   Datasets in-the-wild

For the in-the-wild MTs for human annotation, we choose 7 systems of different series. We expect that they can exhibit variance about translationese.

- Google-Translate

- DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025)

- DeepSeek-R1 (DeepSeek-AI et al., 2025)

- Llama3.3-70B-Instruct

- Qwen2.5-3B-Instruct

- Qwen2.5-7B-Instruct

- Qwen2.5-72B-Instruct

Type	Content
Translation Prompts	
Low-translationese	请把以下文本翻译为中文。译文必需符合中文表达，多用小句、流水句以及中文俗语，不一定非要忠实于原文。请直接返回译文。 *[Please translate the following text into Chinese. The translation should be idiomatic Chinese, favoring shorter sentences, run-on sentences, and Chinese colloquialisms. It doesn't have to be strictly faithful to the source text. Please return the translation directly.]*
High-translationese	请把以下文本翻译为中文。译文必需忠实于原文，不要为了中文用语习惯对原文做任何修改。请直接返回译文。 *[Please translate the following text into Chinese. The translation must be faithful to the original text, and no modifications should be made to the text to fit conventions in the Chinese language. Please return the translation directly.]*
Translation Examples	
Source text	The three spectators seemed quite stupefied. They offered no interference, and the boy and man rolled on the ground together; the former, heedless of the blows that showered upon him, wrenching his hands tighter and tighter in the garments about the murderer's breast, and never ceasing to call for help with all his might.
Low-translationese	三个旁观者愣住了，他们没敢上前阻拦，只见那男孩和男人在地上滚成一团。男孩不顾雨点般落在身上的拳打脚踢，死死揪住凶手胸前的衣服，一边拼命地高声呼救，一边越揪越紧。
High-translationese	那三个观众似乎相当惊呆了。他们没有进行干涉，那个男孩和男人在地上滚作一团；前者不顾落在他身上的拳打脚踢，越发力气地抓紧凶手胸前的衣裳，并且一直用尽全力呼救。

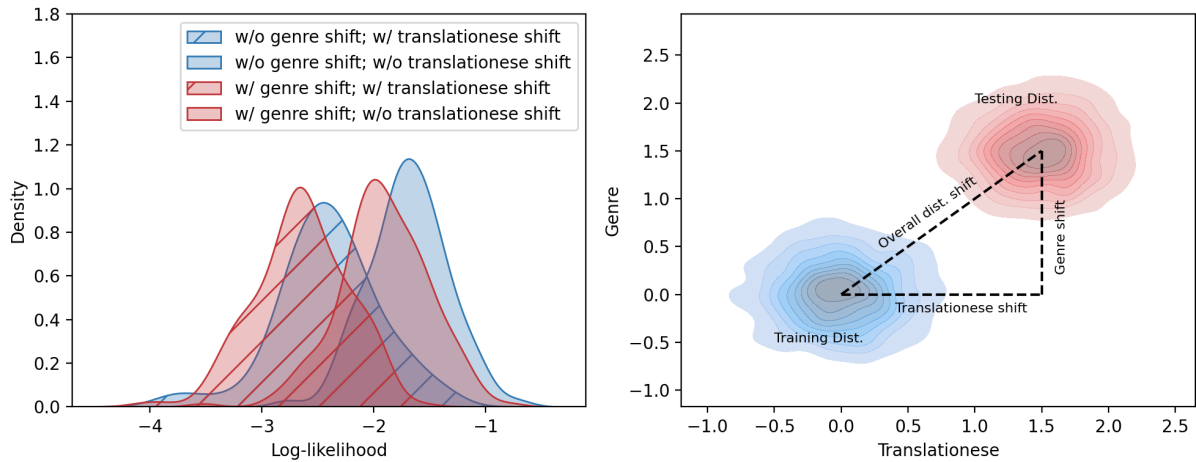Table 5: Translation prompts and examples for low- and high-translationese in the synthetic dataset.

Figure 4: In translationese measurement, the feature wanted is translationese. However, other types of distribution shift might be encoded in log-likelihood as well, illustrated by the left-side figure. The model assigns the highest probabilities to samples from the same distribution as the training data, and assigns lower probabilities when testing samples shift caused by both translationese and genre. The right-side figure intuitively illustrates that the overall distribution shift can be decomposed into independent components, which is the fundamental assumption of LLR.

We use the following five LLMs to produce translations in the section where T-index is compared with existing QE metrics. These models are from the same series. Thus, the results of QE can be more comparable among these models. For each experiment condition, we have 1,000 sources, 1,000 references, and 5,000 translations.

- Qwen2.5-0.5B-Instruct

- Qwen2.5-3B-Instruct

- Qwen2.5-7B-Instruct

- Qwen2.5-32B-Instruct

- QwQ-32B

C More details about baselines

We introduce the high-level intuition of each unsupervised baseline (when the scoring model is trained on high-translationese).

- **Log-likelihood**: LMs assign higher probabilities to samples close to the training distribution. The likelihood of low-translationese samples is expected to be lower than that of high-translationese ones.

- **Entropy**: LMs are less uncertain about in-distribution samples, so the entropy of high-translationese samples are expected to be lower than that of low-translationese ones.

- **Fast-DetectGPT** (Bao et al., 2024): FDG assumes that the likelihood of the original continuation after the context should be higher than that of the alternatives for machine-generated texts, distinguishing them from human-written samples. Similarly, the likelihood of low-translationese translations changes more significantly than that of high-translationese samples under substitution.

- **Mahalanobis Distance** (Ren et al., 2023): MD measures the distance between the last hidden states of the sample and the training distribution. High-translationese samples are closer to the training distribution than low-translationese translations.

- **Relative Mahalanobis Distance** (Ren et al., 2023): RMD provides a background distribution based on MD. OOD samples are expected to be closer to the background distribution, but get away from the training distribution. Here, we measure the relative distance of a sample to the low-translationese distribution and high-translationese distribution.

- **Trajectory Volatility** (Wang et al., 2024): TV measures the changes between adjacent layers of hidden states of model output when the last hidden states of the outputs cluster in a high-density region, which is observed on OOD samples in mathematical reasoning. Here, we expect that low-translationese samples cluster more closely.

Here is the intuition of supervised baselines:

- **DPO-aligned** (Rafailov et al., 2023): Using DPO to align an LLM to prefer high-translationese translation but penalize the low-translationese. The aligned model will assign higher probabilities to high-translationese translations than low-translationese ones.

- **Bradley-Terry RM** (Bradley and Terry, 1952; Ouyang et al., 2022): Training a reward model (RM) with Bradley-Terry loss to assign higher scores to high-translationese samples and lower scores to low-translationese.

- **XLM-RoBERTa** (Conneau et al., 2020): Fine-tuning a pre-trained encoder for classification.

- **SVM with linguistic features** (Hu and Kübler, 2021): Extracting linguistic features and using SVM for classification.

All scoring models are trained on 1 or 2 A100-80G GPUs. It takes around 5 minutes to train a `Qwen2.5-0.5B` base model with the objectives including SFT, DPO, and RM. The implementations and hyperparameters for our model training can be found in our GitHub repository: `https://github.com/yikang0131/TranslationeseIndex`.

D Instruction for annotations and prompts used in LLM-as-a-judge

Table 6: Guidelines and instructions for human annotation

一、任务说明[Task Description]
• 标注人员需要对50条机器翻译数据进行评估[Annotators need to evaluate 50 machine translation examples]
• 每条数据包含一条英文原文和一条中文译文[Each example contains an English source text and a Chinese translation]
• 使用0-5的Likert量表对译文的翻译腔程度进行打分[Use a 0-5 Likert scale to rate the degree of translationese]
• 标注者需要从译文中摘选0-3个翻译腔严重的片段作为评分依据[Annotators should select 0-3 segments with severe translationese as evidence]
• 标注数据将开源但仅用做学术用途，标注者信息会做匿名处理[The annotated data will be open-sourced only for academic purposes, and all information of annotators will be anonymized]

二、翻译腔的定义[Definition of Translationese]
译文用词和语序明显受到翻译过程的影响，导致不符合原生汉语的使用习惯。主要体现在: [Word choice and word order are clearly influenced by the translation process, resulting in expressions that deviate from native Chinese usage. This mainly manifests in:]
1. 用词和语序造成的不自然[Unnatural word choice and word order]
2. 过于直译造成的语义错误（其他类型的翻译错误不计入翻译腔评分）[Semantic errors caused by overly literal translation (other types of translation errors are not counted)]

三、评分标准与摘选要求[Rating Criteria and Selection Requirements]
0分：完全符合中文表达习惯，读起来自然流畅[0 points: Completely conforms to Chinese expression habits, reads naturally and fluently]
1分：稍有不自然，但不影响理解[1 point: Slightly unnatural but does not affect understanding]
2分：个别用词或语序有翻译腔，整体基本可接受[2 points: Individual words or word order show translationese, but overall acceptable]
3分：明显的翻译腔，但基本意思清晰[3 points: Obvious translationese, but basic meaning is clear]
4分：较重的翻译腔，部分表达不符合中文习惯[4 points: Heavy translationese, some expressions deviate from Chinese conventions]
5分：严重的翻译腔，直译痕迹明显或造成理解障碍[5 points: Severe translationese, obvious literal translation or causes comprehension difficulties]

四、示例分析[Example Analysis]
示例1 [Example 1]:
原文[Source]: This issue requires immediate attention.
译文[Translation]: 这个议题需要即刻的关注。
评分[Rating]: 3分
翻译腔片段[Translationese segments]: "即刻的关注"

五、标注注意事项[Annotation Notes]
1. 评分时应重点关注译文的流畅度和自然程度[Focus on fluency and naturalness when rating]
2. 摘选片段时应选择最能体现翻译腔的部分[Select segments that best demonstrate translationese]
3. 摘选数量应与翻译腔程度相匹配[Number of selections should match the degree of translationese]
4. 不要将其他类型的翻译错误记入翻译腔评分[Do not count other types of translation errors in translationese rating]

Table 7: Prompts used in the LLM-as-a-judge method

【LLM-as-a-Judge: 中文翻译质量评估与解析】

任务描述：在以下提供的两个英文句子的中文翻译版本中（A和B），请判断哪一个翻译更符合中文的表达习惯和语境。同时，请简要说明选择的理由。最后，您的回答应包含"A"或"B"，以表明您认为哪一版本更优。

示例：

英文原句：The quick brown fox jumps over the lazy dog.

翻译A：那只敏捷的棕色狐狸跳过了懒惰的狗。

翻译B：快速的棕色狐狸跃过懒散的狗。

评估解析：翻译A使用了"敏捷"和"懒惰"这两个形容词，更加形象生动，更符合中文表达习惯中的具体性和形象性。而"跃过"相较于"跳过"在中文中更具有画面感。因此，翻译A更优。

评估结果：A

实际任务：

英文原文1：{{source}}

翻译A：{{translation_A}}

翻译B：{{translation_B}}