Retranslation at the intersection of style, machines, and plagiarism

Hüseyin Emir Akdağ
Boğaziçi University
Istanbul, Türkiye
huseyin.akdag@boun.edu.tr

Yusuf Mert Aygün
Boğaziçi University
Istanbul, Türkiye
yusuf.aygun2@boun.edu.tr

Mehmet Şahin
Boğaziçi University
Istanbul, Türkiye
mehmet.sahin5@boun.edu.tr

Ena Hodzik
Université de Mons
Mons, Belgium
ena.hodjikj@umons.ac.be

Sabri Gürses
Üsküdar University
Istanbul, Türkiye
sabri.gurses@uskudar.edu.tr

Tunga Güngör
Boğaziçi University
Istanbul, Türkiye
gungort@boun.edu.tr

Abstract

This paper investigates the stylistic distinguishability of literary retranslations within the evolving "translatiosphere". Focusing on the English-Turkish pair, we customized a neural machine translation (NMT) model using three distinct datasets of Alice in Wonderland: original human translations, commercial NMT translations, and generative AI translations. Our methodology employs a sliding-window chunking strategy and 17 morpho-stylistic features tailored for Turkish as an agglutinative language. A Random Forest classifier achieved a Macro F1 score of 83.44%, revealing that human stylistic fingerprints are highly consistent and detectable with just 9 text chunks, whereas LLM-generated outputs exhibit significantly higher variance. Crucially, the model fine-tuned on human data failed to inherit contextual intelligence, instead synthesizing inputs into a homogenized, "statistically safe" signature marked by syntactic rigidity and literalism. We conclude that while human-trained MT internalizes dense linguistic markers, it remains an isolated signature distinct from individual human creativity. These findings provide empirical evidence for identifying machine-generated "translationese" and carry significant implications for plagiarism, authorship, and originality in the digital era.

1 Introduction

The nature of the "translatiosphere"—the semiotic space where translators work and translations exist— (Şahin and Gürses, 2023, p. 1) has changed considerably over the last decades with the adoption of computer-aided translation (CAT) and machine translation (MT) approaches that are powered by corpus-based and natural language processing (NLP) tools (O'Brien, 2023). This affected "the age of retranslation", as it was optimistically called in the beginning of the 21st century (Collombat, 2004). Although retranslations of classics and best-selling books in the literary domain occupy a large share in the publishing industry today, they have been shown to be exploited as a means of unethical practices in the publishing sector, usually in the form of plagiarism or as attempts to increase profits in detriment of the previous translator(s) of the same work (e.g., Turell, 2004; Gürses, 2006, 2008; Şahin et al., 2018; Gürses and Şahin, 2023; Alter, 2026).

With the rise of neural machine translation (NMT), researchers are increasingly experimenting with the use of MT in translation of literary and creative texts, and the style of a human translator has been at the forefront of discussions. The negative impact of the use of MT in literary retranslation has been highlighted in both novice translators (Şahin and Gürses, 2019) and professional translators (e.g., Kenny and Winters, 2020; Guerberof-Arenas and Toral, 2020). Similarly, researchers have provided empirical evidence that only with the help of actual human translators' work can automated systems be stylized in translating literary works (e.g., Yirmibeşoğlu et al., 2023; Gürses et al., 2024; Dallı et al., 2024; Hansen, 2024; Şahin et al., 2025).

The artificial intelligence (AI) boom starting late 2022, on the other hand, has attracted users with its capabilities of customizing the style of a text, and therefore a translation. In this context, retranslations stand as a special case as they have antecedents that can be used for comparison. Based on all of this, the question is raised whether the responsibility of recreating the stylistic features of the antecedent(s) remains or is decreased by the act of reproduction by AI.

In this paper, we analyze style in a customized MT engine trained on a corpus of different translations of a source text to generate a retranslation of the same text. We use existing (human) translations, translations produced by neural MT systems, and translations produced by Gen-AI tools (large language model decoders) to train the MT engine. We aim to answer the following research question:

To what extent can distinctive stylistic features (see Table D1 in Appendix D for the complete taxonomy) be observed in literary retranslations, for the English-Turkish language pair, generated by a customized MT model fine-tuned with

- previous translations of the same source text?
- translations of the same source text created by neural MT systems?
- translations of the same source text created by Gen-AI tools?

To answer our research question, we selected the classical novel Alice in Wonderland (AiW) as a case study. The intuition behind this choice is that this child book is very popular in Turkish and worldwide (Appleton, 2015), and has an interesting history of translation.

We also discuss the implications of translations for plagiarism, creativity, and originality. The main motivation behind our research is the changing nature of the publishing sector, where translations occupy an important place. The ever increasing amount of inorganic, artificial, or synthetic data in the digital world urges us to assess possible repercussions in retranslation.

Our analysis involves a stylistic comparison of different retranslations of the same book. We used a model adapted from Leech and Short (2007) to determine which stylistic features to measure and compare between the outputs produced by different translators or MT models as implemented by Yirmibeşoğlu et al. (2023) and Dallı et al. (2024).

In previous investigations employing the same methods (see Şahin et al., 2025 and Yirmibeşoğlu et al., 2023), it was found that a fine-tuned model trained on literary works (same genre) was closest to the original translator's style (compared to a pretrained model and Google Translate). Our study provides further empirical evidence in current discussions about style, authorship, and originality.

2 Literature Review

2.1 Retranslation

Retranslation has been an established topic since 1990, when Bensimon (1990) and Berman (1990) edited the foundational special issue of *Palimpsestes* (Peeters and van Poucke, 2023, p. 4). Many scholars, such as Gambier (1994), Tahir Gürçağlar (2009), and Paloposki and Koskinen (2010), have continued the discussion on retranslation. While Bensimon (1990) and Berman (1990) viewed retranslation as a process of successive improvement bringing translations closer to the source text, Tahir Gürçağlar adopted a descriptive, sociohistorical approach. She defined retranslation as "either the act of translating a work that has previously been translated into the same language, or the result of such an act, i.e. the retranslated text itself" (Tahir Gürçağlar, 2009). Retranslation deserves special attention in the context of MT and AI as it is not unlikely that previous translation(s) are already part of the digital translatiosphere or could be used in customized MT or LLMs. There are already examples of using translation memory systems, MT, and other tools in the retranslation process (e.g., Rothwell, 2023), which allows the retranslator to see previous solutions in real-time. Although discussed in Gürses and Şahin (2023), there have been no studies investigating the impact of AI on retranslation in an experimental setting.

2.2 Literary Machine Translation

Previous work compares SMT and NMT and investigates the role of literary or in-domain training data, often at document or corpus level (Sluyter-Gäthje et al., 2018; Toral and Way, 2015, 2018). There is an important body of literature that examines literary MT workflows and creativity in literary MT through the use of metaphors (e.g., Ruffo and Macken, 2024; Karakanta et al., 2025). Several works also model user- or translator-level variation, for instance on TED data (Michel and Neubig, 2018; Wang et al., 2021), and examine literary

NMT and stylistic control through literary-domain systems, adaptation challenges, and style prompting, including the MSMT benchmark (Kuzman et al., 2019; Matusov, 2019; Wang et al., 2023). English–Turkish remains sparsely covered: Şahin and Dungan (2014) study the use of Google Translate across literary and other genres in the SMT era, while Şahin and Gürses (2019) examine the use of the same system for English–Turkish literary retranslation after its shift to NMT.

2.3 Style

In the literature on style in human translation, different methodologies have been proposed that focus on both the source and target texts (Saldanha, 2011) or target texts only (Baker, 2000), all foregrounding translator agency. The limitation of MT models is that they model probabilities rather than agency and intention, but corpus and NLP tools have allowed researchers to quantify style. For example, with the use of corpus tools (Dallı et al., 2024; Şahin et al., 2025; Yirmibeşoğlu et al., 2023) and stylometry (Rybicki, 2025), researchers have identified distinct measurable stylistic features in specific literary translators and have provided evidence of style transfer in customized MT models.

3 Methodology

To answer the research question stated in Section 1, we selected the classical novel Alice in Wonderland (AiW) (Carroll, 2008). Using a neural MT model as a base model, we first trained this model using three different types of translations of the book: original translations of AiW, neural machine translations of AiW, and Gen-AI translations of AiW. We obtained three different trained MT models in this way and then retranslated the original English book with each of these models. The output of this process is three retranslations of the book which are then subjected to style analysis.

In this section, we explain the compilation of the training data used in training the base model, the training of the base model, and detailed style analysis of the retranslations.

3.1 Alice in Wonderland (AiW)

Warren Weaver, who published a seminal memorandum for machine translation research in 1949, also researched translations of Alice. From him we learn that in 1866, one year after the novel was published, Lewis Carroll requested from his friends French and German translations of the book for continental sale, but they found it impossible to translate (Weaver, 1964). Today the book has translations in at least 175 languages (Lindseth and Tannenbaum, 2015). There are around 50 translations into Turkish (Tıson, 2015), with more than 20 plagiarised versions. The first Turkish version of AiW was created in 1932 by Ahmet Cevat Emre under the title “Alisin Sergüzeştleri” (Travels of Alice) and was followed by another translation in 1944 by Muzaffer Beşli and Naime Halit Yaşar with the title “Alis Harikalar Diyarında”. The first version is a shortened adaptation of the book and the second was half translated from French. We can call the translation from English by Kısmet Burian that appeared in 1946 with a different title, “Alice Harikalar Ülkesinde”, as the first translation of the book (see Tıson, 2015). Then the first retranslation by Azize Erten appeared in 1953 as "Alis Harikalar Ülkesinde". “Ülkesinde” and “Diyarında” have a similar meaning in Turkish, and retranslations after this date regularly used these titles.

In our research, we identified nine retranslations created until 1995 and we also found that starting with the 1990s several plagiaristic translations, i.e., texts copied from earlier translations, have been published under fake names. Plagiarism in translation has been a topic of research for the last two decades in Turkey (e.g., Gürses, 2006, 2008) and elsewhere (e.g., Turell, 2004). There are at least 10 plagiaristic retranslations of AiW in Turkish. Following the methodologies outlined by Şahin et al. (2018), these plagiaristic versions were identified not only through surface-level similarities but by mapping highly idiosyncratic translator choices, shared omission patterns, and structural anomalies that definitively point to unauthorized textual kidnapping rather than coincidental linguistic convergence. There is a need to separate real and fake, plagiaristic translations in the market, which is not always an easy task. As publishers have also started to use NMT or Gen-AI to produce machine translations under fake human names (Topal and Tanış Polat, 2025), measuring the ability of these tools to create stylistically meaningful retranslations is necessary.

3.2 Compilation of the Training Data

To train the base MT model with original translations of AiW, we selected three human retransla-

tions from three distinct periods — Erten (1953), Dişbudak (1995), and Erzincan Kına (2010). These specific editions were strategically chosen to represent distinct chronological milestones and varying translational approaches within the Turkish literary landscape. Following Azize Erten's inaugural 1953 retranslation, which introduced foundational terms like 'Alis' and 'Ülkesinde', the chosen texts trace the evolution of these lexical precedents. Capturing how certain terms became standard while others faded ensures the neural model is trained on a highly diverse and stylistically balanced foundation. All of the texts have been digitized and used exclusively for non-commercial academic research purposes, in line with the exceptions permitted under Article 35 of the Turkish Law on Intellectual and Artistic Works (FSEK, Law No. 5846)[1]. Training a neural MT model requires a parallel sentence-aligned corpus. We thus aligned the sentences in the English original and the Turkish retranslations for each of the Turkish books as training data. In this way, we obtained a dataset of original translations, Dataset_OT (comprising a total of 57,215 words and 7,782 sentences; see Appendix E for individual breakdowns), consisting of multiple human translations by various professionals in order to capture stylistic diversity.

For training the base MT model with translations of AiW obtained with neural MT models, we translated the English original using four commercial machine translation systems: Google Translate, DeepL, Amazon Translate (AWS), and Microsoft Azure Translator. Since we need sentence-aligned data after the translations to train the base MT model, we had to translate on a sentence basis. However, feeding each source sentence separately to the model causes the context to be lost and may yield an improper translation. Therefore, for each source sentence, we included a context of up to 1000 characters of previous and following sentences. The context size was chosen to balance between cost considerations and the need to reduce noise. In this way, we translated each sentence separately, but provided a context block to enable context-aware translation. Classical MT systems require a target sentence marker `[[TGT_START]]`, which is incorporated into the context block. The target sentence is then extracted from the output with the help of regular expressions, and possible distortions in target markers are addressed accordingly. We collected the English-Turkish sentence pairs in a dataset named as Dataset_NMT.

To train the base model with Gen-AI translations, we translated the original book with four LLMs: Claude Sonnet 4.6 (Anthropic), Gemini 3 Flash Preview (Google), GPT-5.4-mini (OpenAI), and DeepSeek Chat (DeepSeek). LLMs inherently take into account context through structured prompting. Our prompt enforces deterministic output generation and limits the output to only the translated target sentence, which achieves correct sentence alignment and eliminates the need for extraction procedures. We collected the English-Turkish sentence pairs in a dataset named as Dataset_LLM.

Before training the base model, all parallel data was preprocessed to normalize punctuation marks, eliminate unnecessary artifacts, and remove any misaligned sentence pairs.

3.3 Machine Translation Model

The base model used for training is the Helsinki-NLP Opus-MT English-Turkish transformer[2]. To preserve parameter efficiency during fine-tuning, LoRA is applied to `q_proj`, `k_proj`, `v_proj`, and `out_proj` layers with rank $r = 32$, scaling factor $\alpha = 64$, and dropout of 0.1, which results in the adaptation of only 1-2% of model parameters.

The training set (each of Dataset_OT, Dataset_NMT, and Dataset_LLM) is divided into training and validation datasets using a 90/10 split ratio. Sentence pairs are tokenized with the Marian tokenizer with a max sequence length of 256 tokens. Training is conducted using the HuggingFace framework with a 3×10^{-4} learning rate, effective batch size of 8, and mixed precision over 6 epochs. Early stopping based on every 100 step evaluation with a patience of 3 is employed. Inference consists of beam search with beam width 4 and max length of 128 to generate deterministic and high-quality output. We save only LoRA adapters and tokenizer weights.

Ultimately, the training process yielded three distinct fine-tuned models corresponding to our three datasets. We refer to these models as **Human-Model** (base NMT model fine-tuned with original human translations), **NMT-Model** (base

[1]Law No. 5846 on Intellectual and Artistic Works (FSEK), Article 35, Republic of Türkiye.

[2]https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-en-tr

NMT model fine-tuned with NMT translations), and **LLM-Model** (base NMT model fine-tuned with Gen-AI translations). We then retranslated the original English AiW with each of these models to conduct style analysis.

3.4 Morphological Parsing

English and Turkish exhibit vastly different levels of morphological analysis. English is a largely analytic language that relies on distinct words to convey grammatical relations, whereas Turkish is highly agglutinative, often encoding complex syntactic structures within a single word (see Dryer and Haspelmath, 2013). To ensure an accurate comparison of equivalent grammatical and stylistic units across these typologically distinct languages, it is essential to decompose Turkish tokens into their constituent morphemes.

We used the Morphological Parser and Disambiguator developed by Sak et al. (2008) to extract morphological data. The tool initially identifies all feasible morphological parses of a word, followed by selecting the appropriate parse that aligns with the contextual information. It should be acknowledged that morphological parsers are not entirely error-free when processing Turkish language data, although the margin of error is acceptable for obtaining general numerical data instead of fine-grained suffixes.

3.5 Translation Chunking

A fundamental challenge in authorship attribution is the requirement for multiple text samples to train and evaluate machine learning classifiers. Since our experimental setup yielded only one complete, continuous translation for each of **Human-Model**, **NMT-Model**, and **LLM-Model**, traditional document level classification was not feasible.

To overcome this sample size limitation and capture the continuous stylistic features of the texts, we employed a sliding-window segmentation (chunking) methodology (see Akdağ, 2026). Each continuous translated book was divided into overlapping chunks. Specifically, we extracted chunks of 25 sentences in length, using a 5-sentence overlap between consecutive chunks.[3]

This approach transformed single, monolithic translations into robust datasets containing several shorter, stylistically intact text chunks (yielding 69, 53, and 52 chunks for **Human-Model**, **NMT-Model**, and **LLM-Model** translations, respectively). These chunks are not independent texts but localized stylistic snapshots of the translation output, allowing the classification algorithms to identify recurring linguistic patterns.

3.6 Feature Extraction

To capture the distinct stylistic features of human, NMT, and LLM translators, two primary sets of features were extracted from the Turkish text chunks:

- Morpho-Stylistic Features: A predefined set of 17 stylistic and morphological features was extracted to capture the fundamental syntactic and structural tendencies of the translators (see Yirmibeşoğlu et al., 2023). Because Turkish is an agglutinative language, morphological density and word structure are highly indicative of translation style. Extracted features included:
  - Morphological: Average and median morphemes per sentence, average and median morphemes per word.
  - Lexical Richness: Type-Token Ratio (TTR), total number of unique words, number of unique words with more than 10 occurrences.
  - Length Metrics: Mean and standard deviation of word lengths; mean, median, mode, and standard deviation of sentence lengths.
  - Punctuation and Syntax: Normalized frequencies of reduplications, ellipses, questions, and exclamations.
- Keyword Frequencies: Lexical features were extracted based on keyword significance. Using AntConc's Keywords tool (Anthony, 2026), we identified both positive keywords (higher frequency and significance) and negative keywords (lower frequency and significance). For each analysis, the text chunks were compared against a baseline reference corpus consisting of all outputs by the other models or of all other human translations used in training. Keyword extraction was strictly controlled using a 4-term log-likelihood measure, a statistical significance threshold of $p < 0.05$, and the Dice coefficient to measure effect size. Examples of identified lexical markers include specific conjunctions,

[3]Our data, scripts and analysis code are available at: https://github.com/webomurga/stygenai-retranslation

adverbs, and function words (e.g., *aceleyle* 'hastily/in a hurry', *ancak* 'but/however', *elbette* 'of course', *hiçbir* 'nothing', *oldukça* 'quite/rather', *tabii* 'naturally/of course', and *yeniden* 'again/anew').

3.7 Translator Classification

The classification task was framed to determine whether a given text chunk could be accurately attributed to its source (Human, NMT, or LLM). Several machine learning models were evaluated, including Support Vector Machines (Cortes and Vapnik, 1995), K-Nearest Neighbors (Fix and Hodges, 1951), Random Forest (Breiman, 2001), and Burrows' Delta (Burrows, 2002; Mikros et al., 2026). The models were trained on three combinations of the extracted features (stylistic only, keywords only, and all features combined) to determine the most discriminative set of features for translation source attribution.

4 Results

4.1 Multiclass Attribution Performance

To assess the overall distinguishability of the translation sources, a multiclass classification analysis was conducted. The objective is to simultaneously categorize text chunks into their correct translation origin (Human, NMT, or LLM) based on their stylistic and lexical features. We evaluated several classifiers across varying feature sets to determine the most robust approach for handling the high dimensionality of morphological and lexical data.

The results demonstrated that the text chunks generated by different translation methods possess highly distinct stylistic features. When using the combined feature set of all stylistic features and all extracted keywords, the classifiers performed as follows:

- Random Forest: Achieved the highest overall performance with a Macro F1 score of 83.44% (± 2.49%).
- Support Vector Machine (SVM): Achieved a highly competitive Macro F1 score of 82.02% (± 1.88%).
- K-Nearest Neighbors (KNN, k=3): Struggled significantly in the multiclass environment, yielding a Macro F1 score of only 51.57% (± 2.29%).

The significant performance drop in the KNN classifier is largely attributable to the sparse, high-dimensional nature of the lexical keyword space, which tree-based (Random Forest) and margin-based (SVM) classifiers handle much more effectively.

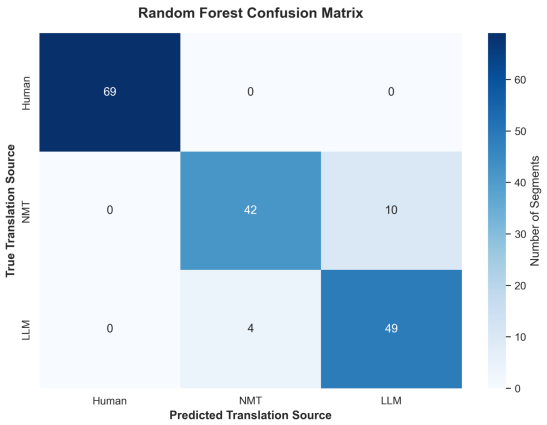

Figure 1: Confusion matrix for the Random Forest multiclass classifier.

As shown in the confusion matrix (Figure 1), the Random Forest classifier demonstrates exceptional precision in isolating Human translations. The vast majority of classification overlap occurs specifically between the NMT and LLM classes. This misclassification pattern is highly indicative of a shared baseline of machine translation artifacts. Despite the architectural differences between encoder-decoder NMT models and autoregressive generative LLMs, both systems appear to share underlying statistical tendencies in their target language generation that distinguish them from human structural choices.

This separability is visually confirmed through a t-distributed Stochastic Neighbor Embedding (t-SNE) projection of the feature space. In the two-dimensional representation (Figure A1), Human translations form a tightly bound and largely isolated cluster, reinforcing the distinctiveness and consistency of human stylistic choices (O'Sullivan, 2025). In contrast, NMT and LLM clusters, while separable, exhibit closer spatial proximity and partial overlap, visually confirming the quantitative findings of the confusion matrix.

4.2 Linguistic Fingerprints

To interpret the decision-making process of the best-performing Random Forest model, we extracted the permutation importance of the combined feature set. This allows us to understand ex-

actly which linguistic phenomena expose the origins of the translated text.

The feature importance analysis (Figure A2) reveals a compelling interplay between morpho-stylistic structures and localized keyword frequencies. The model heavily relies on specific lexical markers alongside structural features like Type-Token Ratio (TTR) and morphological density (average morphemes per word). Specifically, the algorithm identified discourse markers, conjunctions, and function words (such as *ancak* 'but', *elbette* 'of course/sure', *hiçbir* 'nothing', and *tabii* 'naturally/of course') as highly discriminative. In machine translation, these items often suffer from literal translation or over-standardization, whereas human translators employ them with greater pragmatic flexibility.

The distribution of these top features varies significantly across translation sources, demonstrating how certain linguistic tendencies are heavily over-indexed or under-indexed depending on the translation origin.

The violin plots (Figure A3) illustrate these specific probability densities, highlighting both the median values and the broader distribution spread for the most highly discriminative markers across Human, NMT, and LLM outputs.

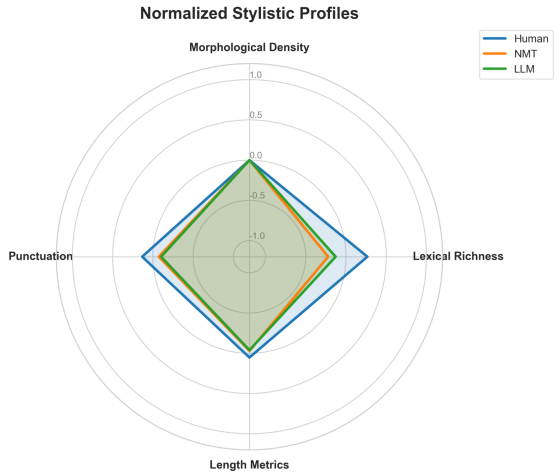

Figure 2: Radar chart mapping the normalized stylistic profiles of Human, NMT, and LLM translations across four macro-categories: Morphological Density, Lexical Richness, Length Metrics, and Punctuation.

Furthermore, when aggregating the morpho-stylistic features into broader linguistic categories, distinct macro-profiles emerge. The radar chart (Figure 2) maps these normalized profiles, illustrating how human translators and machine models diverge sharply in areas such as Lexical Richness, Punctuation, and Morphological Density. For an agglutinative language like Turkish, morphological metrics are exceptionally revealing. The data proves that the attribution models are not merely memorizing vocabulary distributions, but are successfully capturing fundamental differences in syntactic construction and morphological realization.

4.3 Stylistic Consistency

A critical secondary objective of this study is determining the minimum amount of text required to confidently identify the source of a translation. In the context of MT evaluation, understanding the necessary sample size is vital for practical deployment. We conducted a sample size analysis to identify the threshold at which the models reached a robust 85% classification accuracy for each category.

The individual learning curves (Figure A4) reveal significant differences in the stylistic consistency of the three translation sources:

- Human Translation: A robust 85% accuracy threshold was reached with a sample size of only 9 chunks (achieving 92.3% ± 4.7% accuracy).
- NMT Generation: Neural machine translation required 22 chunks to reach the 85% robustness threshold (achieving 89.2% ± 3.1% accuracy).
- LLM Generation: LLM-generated texts exhibited the highest stylistic variance, requiring 32 chunks to achieve the 85% robustness threshold (achieving 88.9% ± 3.8% accuracy).

These findings suggest that original human translations possess highly concentrated and consistent stylistic features that are easily detectable with minimal data. In contrast, LLM-generated translations display much higher intra-textual variance, requiring significantly more text chunks for algorithms to confidently recognize their underlying stylistic features.

When overlaid (Figure 3), the rapid stabilization of the Human curve visually highlights the

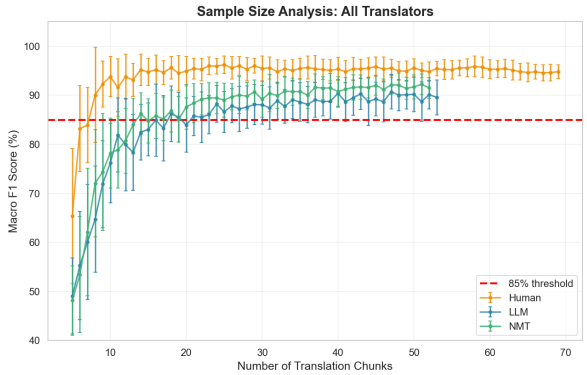

Figure 3: Combined learning curves comparing the stylistic consistency across Human, NMT, and LLM translation sources. The rapid stabilization of the Human curve highlights the dense nature of human stylistic markers compared to machine-generated variance.

highly concentrated and consistent nature of human stylistic markers. In contrast, the outputs of the NMT- and LLM-sourced models display a more gradual incline. The fact that LLM outputs require significantly more text chunks to classify than standard NMT suggests a higher intratextual variance. This is likely a direct artifact of the autoregressive decoding process and temperature sampling inherent to LLMs, which introduces more stylistic fluctuation than the typically deterministic or beam search-driven outputs of standard commercial NMT systems.

4.4 Stylistic Convergence

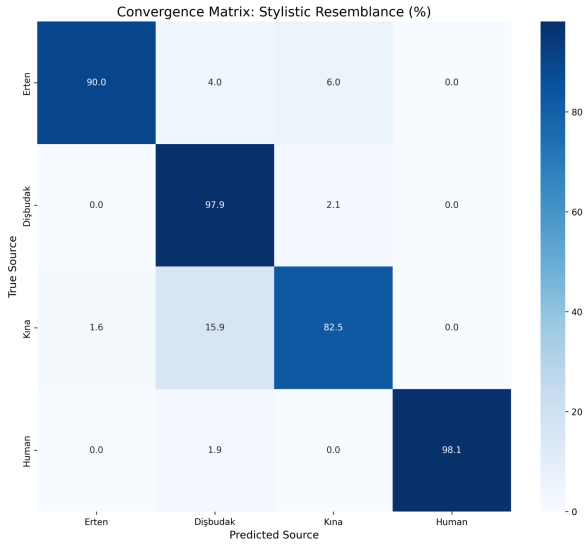

Figure 4: Stylistic convergence matrix detailing the percentage of resemblance among the contributing human translators and the Human-NMT model.

Figure 4 demonstrates that while the neural machine translation model fine-tuned with original human translations successfully internalizes dense stylistic markers, it ultimately synthesizes them into a highly homogenized and isolated signature. As established by its rapid learning curve and the 98.1% self-attribution rate, this human-trained model is remarkably easy for the classifier to detect compared to models trained on commercial MT outputs. This rapid classification occurs because the model efficiently captures and uniformly replicates the highly concentrated morphological and syntactic features found in its human source texts. However, the off-diagonal values in the convergence matrix reveal that the model does not completely resemble any single contributing translator, showing zero percent stylistic overlap with Erten or Kına and a negligible 1.9% with Dişbudak. Rather than mimicking the idiosyncratic variance and unpredictable creative choices of a specific individual, the machine algorithmically averages these human inputs to produce a predictable, statistically safe stylistic fingerprint entirely of its own.

The following verbatim samples illustrate how the Human-Model output differs from the idiomatic and contextual nuances of the human translators, instead opting for a standardized, literalized form. In the opening sentence, the human translators provide three distinct interpretations of "bank" based on the narrative context: a wooden bench, a park bench, or a river bank. The Human-Model output, however, fails to mimic any of these established human interpretations. Instead, it selects the most statistically probable and literal translation, "bankada." In modern Turkish, "banka" primarily refers to a financial institution or a generic seat, making the model choice feel out of place in a 19th century literary context. This failure of mimicry demonstrates that the Human-NMT model prioritizes literal equivalence over the contextual intelligence present in its training data.

The samples given in Appendix B highlight the distinctive signature of the Human-Model as a force of lexical normalization. The human translators employ vocabulary that conveys surprise or extraordinariness, such as "şaşacak" or "olağanüstülük." The Human-Model output defaults to "dikkat çekici," which is the standard dictionary equivalent for "remarkable." While the model had access to the more creative lexical choices of Erten and Kına, it essentially ignored them in favor of a "safe" statistical mean. This

confirms that the model does not inherit the specific lexical fingerprints of its trainers but instead creates a neutralized version of the target language.

The divergence in the sample given in item Appendix B is primarily syntactic. The human translators reconfigure the English comparative "not much larger than" into more natural Turkish phrasings like "farksız görünen" or "pek de büyük sayılmazdı." Conversely, the Human-Model output mirrors the English structure almost exactly with "çok daha büyük olmayan." This adherence to the source text's syntax, despite the availability of more fluid human alternatives, creates a rigid "translationese." This rigidity is a key component of the Human-Model's isolated signature, making it easily distinguishable from the more adaptable human styles it was meant to emulate.

By failing to adopt the contextual interpretations of the human data and instead defaulting to literalism and syntactic rigidity, the Human-Model establishes its own unique, isolated identity. This identity is defined by a lack of the idiomatic flexibility found in the human training set, which explains why the model remains highly identifiable to the classifiers despite its human training data origin.

4.5 Lexical Creativity

As a final example, we examine the translation of neologisms, which serves as a concrete indicator of lexical creativity (Kumcu, 2008). In Chapter 2 in AiW, "The Pool of Tears," Lewis Carroll famously introduces the word *curiouser*. This is a deliberate grammatical error: Carroll applies the inflectional comparative suffix "-er" to a polysyllabic adjective (which rigidly requires the analytic "more curious" in standard English). He explicitly flags this deviation in the text, noting that Alice was so startled that she "quite forgot how to speak good English." Translating this intentional morphological violation poses a significant challenge, particularly for machine translation models.

The examples in Appendix C demonstrate a clear divergence in the way the different translation paradigms handle lexical anomalies. The human translator successfully captures the situational meaning of *tuhaf* ("strange/bizarre"). However, likely due to the difficulty of replicating this specific English morphological error in Turkish, the text is normalized using standard comparative structures *daha tuhaf* ("more strange/bizarre"), thereby losing the neologism itself.

The generative AI model mirrors this human strategy closely, opting for semantically accurate *daha da tuhaf* ("even more strange/bizarre") while similarly normalizing the grammatical structure. This aligns with the LLM's tendency to produce highly fluent, context-aware target texts, even if it smooths over deliberate source text anomalies.

Conversely, the neural machine translation model demonstrates a complete inability to handle the creative deviation. It processes the anomalous token rigidly and probabilistically: not only does it normalize the grammar, but it also completely misinterprets the semantic context, rendering "curious" in its literal, primary sense as *meraklı* ("inquisitive"). This underscores a fundamental limitation in NMT systems: they fail to recognize deliberate grammatical deviations as creative choices, resulting in outputs that miss both the neologism and the contextual meaning that justifies it.

5 Conclusions

This study measured the level of distinguishability of stylistic features in the retranslations of a literary work produced by three different MT models. The MT models were built through fine-tuning the Helsinki-NLP Opus-MT English-Turkish transformer model with three distinct sets of texts for the literary work Alice in Wonderland: (1) three original translations; (2) translations generated by four major MT providers, and (3) translations generated by four major Gen-AI tool providers. The analysis of the outputs was based on the style vector analysis methodology employed in Yirmibeşoğlu et al. (2023). We used morphological parsing, translation chunking methods, and corpus analysis methods for translator classification.

The Human-Model (OPUS-MT base model fine-tuned with three original translations) produced a retranslation with more consistent and distinct stylistic features than the retranslations generated by the models fine-tuned with MT or Gen-AI outputs, which were close to one another and showed overlap in stylistic features. The use of discourse markers, conjunctions, and function words was the major discriminative feature of the Human-Model output. Lexical richness and the use of ellipsis were also significantly higher. In short, this output displayed the most homogenized and isolated stylistic profile among the three mod-

els as it successfully internalized dense stylistic markers from the MT model fine-tuned on original human translations. However, when examined closely in comparison to the three human translations used for fine-tuning the base MT model, the Human-Model output failed to present the level of contextual awareness of its fine-tuning data and tended to apply literal interpretations of the source text and follow syntactic rigidity. This finding suggests that using original human translations of a literary work for fine-tuning MT models to create a retranslation of the same work is still far from producing contextually sensitive and syntactically flexible retranslations. This underscores the value of original human translations, particularly retranslations in our case.

Style transfer in literary machine translation has proven to be possible through fine-tuning with aligned corpora of a specific translator, which was recommended for building systems "endowed with a personalized imprint, ensuring consistency in linguistic features throughout the translated content, while potentially facilitating reader engagement, fostering a sense of familiarity, and contributing to the branding of the translator's reputation and cultural heritage." (Dallı et al., 2024, p. 17). Regardless of the variety of source texts involved in such corpora, a consolidated stylistic imprint of a translator would allow such a transfer. However, in the case of the current study, blending the highly individual styles of multiple translators during fine-tuning resulted in an averaging effect. Consequently, rather than inheriting the creative idiosyncrasies of its human training data, the Human-Model produced a retranslation that was lexically less rich, syntactically less flexible, and ultimately synthesized into a rigid, homogenized signature.

Finally, our study has implications for issues such as plagiarism, creativity, and originality. Some of the Turkish translations of Alice in Wonderland might already be available on the Internet, but, if so, whether they are processed by NMT or Gen-AI tools is not a fact that can be easily verified. In our study, we have not found direct evidence of copying from original translations by the models fine-tuned with NMT or Gen-AI outputs. However, it would be worth comparing them with all existing translations in future studies. Our methodology can contribute to the efforts towards identifying original translation solutions, in other words, the level of creativity, in translations of the same source text. This in turn would provide evidence of possible plagiarism and protect translator rights in an era where data are at high risk of being exploited without author's or translator's consent.

6 Limitations

While the classification results demonstrate a high degree of separability between human and machine-generated texts, several methodological constraints must be contextualized.

First, to overcome the limitation of having a single continuous translation per model configuration, this study utilized a sliding-window chunking strategy. While this successfully captured localized stylistic snapshots and generated sufficient data points for training, it inherently introduces partial collinearity between overlapping chunks. The algorithms are classifying highly localized stylistic decisions rather than document-level independent variables.

Secondly, the feature importance metrics (particularly the high discriminative power of morphological density and specific conjunctions) are deeply tied to the English-to-Turkish translation direction. Because Turkish is a highly agglutinative language, morphological choices carry significant stylistic weight. It remains an open question whether the specific lexical markers that so clearly separate LLM, NMT, and Human outputs in this study would manifest with the same discriminative power in analytic target languages, or whether the models' "machine-like" signatures would shift entirely to syntactic ordering. These constraints highlight the need for cross-linguistic validation, although they do not diminish the localized efficacy of the established morpho-stylistic identification method (see also Guerberof-Arenas and Toral, 2022).

Finally, even though it is already a known fact that Gen-AI tools are getting better at generating highly stylized outputs, this study did not investigate the impact of prompting (Castaldo and Monti, 2024; Yamada, 2023; van Egdom et al., 2024) on quality or literary style or using detailed translation briefs to improve quality through post-editing completed by LLMs (Macken, 2024). Instead, we used a simple prompt to allow for a better examination of the impact of fine-tuning data on style transfer.

Carbon Impact Statement

This research involved computational experiments using an NVIDIA A100 GPU hosted on the Google Colab platform. We estimate the total computational time for training and evaluation to be approximately 4 hours (including initial configuration trials). Using a standard GPU power consumption of 400W, a data center PUE of 1.1, and an average grid carbon intensity of 445 g CO_2e/kWh, the total energy consumption is estimated at 1.76 kWh, resulting in a carbon footprint of approximately 0.78 kg CO_2e. Emissions were calculated following the methodology of the ML CO_2 Impact Calculator (Lacoste et al., 2019).

Acknowledgements

As cited in the paper as well, we adapted the baseline code created within the framework of a scientific research project funded by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) to implement the sliding-window chunking mechanism utilized in our methodology. The project was titled *"Literary Machine Translation to Produce Translations that Reflect Translators' Style and Generate Retranslations"* (Grant No: 121K221). We would like to thank all team members for their valuable contributions:

- Principal Investigator: Mehmet Şahin
- Researchers: Ena Hodzik, Tunga Güngör
- Post-doctoral Researcher: Sabri Gürses
- Graduate Students: Zeynep Yirmibeşoğlu, Harun Dallı, Olgun Dursun

References

Akdağ, H. E. (2026). Directed Attention is All You Need: Profiling Style from Limited Text Data. Proceedings of the Second Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2026), 14–27. https://aclanthology.org/2026.sigturk-1.2/

Alter, A. (2026, April 10). Where Does Publishing's A.I. Problem Leave Authors and Readers. The New York Times. https://www.nytimes.com/2026/04/10/books/shy-girl-ai-publishing.html

Anthony, L. (2026). AntConc (Version 4.4.0) [Computer software]. Waseda University, Tokyo, Japan. https://www.laurenceanthony.net/software/AntConc

Appleton, A. (2015, July 23). The mad challenge of translating "Alice's Adventures in Wonderland". Smithsonian Magazine. https://www.smithsonianmag.com/arts-culture/mad-challenge-translating-alices-adventures-wonderland-180956017/

Baker, M. (2000). Towards a methodology for investigating the style of a literary translator. Target. International Journal of Translation Studies, 12(2): 241–266.

Bensimon, P. (1990). Présentation [Presentation]. Palimpsestes, 4, ix-xiii.

Berman, A. (1990). La retraduction comme espace de la traduction [Retranslation as a space for translation]. Palimpsestes, 4, 1–7. https://doi.org/10.4000/palimpsestes.596

Breiman, L. (2001). Random forests. Machine Learning.

Burrows, J. F. (2002). "Delta": a measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing.

Carroll, L. (2008). Alice's adventures in Wonderland (The Millennium Fulcrum Edition 3.0). Project Gutenberg. https://www.cs.cmu.edu/~rgs/alice-table.html

Castaldo, A., and J. Monti (2024). Prompting Large Language Models for Idiomatic Translation. In Proceedings of the 1st Workshop on Creative-text Translation and Technology, 32–39. European Association for Machine Translation. https://aclanthology.org/2024.ctt-1.4/

Collombat, I. (2004). Le XXIe siècle : l'âge de la retraduction [The 21st century: the age of retranslation]. Translation Studies in the new Millennium. 2.https://hal.science/hal-01452331

Cortes, C., and V. Vapnik (1995). Support-vector networks. Machine Learning.

Dallı, H., O. Dursun, E. Hodzik, S. Gürses, T. Güngör, and M. Şahin (2024). Giving a translator's touch to the machine. Palimpsestes, 38: 15–36. https://doi.org/10.4000/12sp6

Dişbudak, B. (1995). Alice Harikalar Ülkesinde. İstanbul: Görsel Yayınları.

Dryer, M. S., and M. Haspelmath, eds. (2013). The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology, Leipzig. https://wals.info/

Erten, A. (1953). Alis Harikalar Ülkesinde. İstanbul: Varlık Yayınları.

Erzincan Kına, K. (2010). Alice Harikalar Diyarında. İstanbul: İthaki Yayınevi.

Fix, E., and J. L. Hodges (1951). Discriminatory analysis: Nonparametric discrimination.

Gambier, Y. (1994). La retraduction, retour et détour [About retranslation]. Meta, 39(3), 413–417. https://doi.org/10.7202/002799ar

Guerberof-Arenas, A., and A. Toral (2020). The impact of post-editing and machine translation on creativity and reading experience. Translation Spaces, 9(2): 255–282. https://doi.org/10.1075/ts.20035.gue

Guerberof-Arenas, A., and A. Toral (2022). Creativity in translation: Machine translation as a constraint for literary texts. Translation Spaces, 11(2): 184–212. https://doi.org/10.1075/ts.21025.gue

Gürses, S. (2006). İntihal: Nasıl Olağanlaştırılıyor [Plagiarism: How is It Being Normalized]. Presentation at the Translation Ethics Symposium, Istanbul University, 8 December 2006.

Gürses, S. (2008). Çeviri İntihalleri: Kültür Ekosisteminin Paraziti [Translation Plagiarisms: The Parasite of the Cultural Ecosystem]. Çeviri Etiği Toplantısı Bildiri Kitabı, İstanbul Üniversitesi Yayınları, 133–146.

Gürses, S. and Şahin, M. (2023). The shifting value of retranslations and the devaluing effect of plagiarism: The complex history of Dostoevsky (re)translations in Turkish. Parallèles – numéro 35(1), April, 47–67. https://doi.org/10.17462/para.2023.01.03

Gürses, S., M. Şahin, E. Hodzik, T. Güngör, H. Dallı, and O. Dursun (2024). Çeviribilim Çalışmalarında Çevirmenin Üslubu ve Makinenin Üslubu [The Style of the Translator and the Style of the Machine in Translation Studies]. Çeviribilim ve Uygulamaları Dergisi, 36: 100–124. https://doi.org/10.37599/ceviri.1468718

Hansen, D. (2024). Évaluation experte d'un prototype d'aide à la traduction créative: la traduction littéraire automatique individualisée au regard de ses enjeux traductologiques, éthiques et sociétaux [Expert evaluation of a prototype for creative translation assistance: Individualized literary machine translation with regard to its translatological, ethical, and societal implications]. Doctoral dissertation, Université de Liège.

Karakanta, A., Nas, M., and Dorst, A. G. (2025, June). Metaphors in Literary Machine Translation: Close but no cigar?. In Proceedings of Machine Translation Summit XX: Volume 1, 276–286. https://aclanthology.org/2025.mtsummit-1.21/

Kenny, D., and M. Winters (2020). Machine translation, ethics and the literary translator's voice. Translation Spaces, 9(1): 123–149. https://doi.org/10.1075/ts.00024.ken

Kumcu, A. (2008). "Alice Çeviri Diyarında: Kültürel Motifler Açısından Karşılaştırmalı Bir Çeviri Eleştirisi" [Alice in Translationland: A Comparative Translation Analysis with Regard to Cultural Motifs], Hacettepe Çeviribilim ve Uygulamaları Dergisi, No: 18, 123–144.

Kuzman, T., Š. Vintar, and M. Arčan (2019). Neural machine translation of literary texts from English to Slovene. In *Proceedings of the Qualities of Literary Machine Translation*, 1–9, Dublin, Ireland, August. European Association for Machine Translation. https://aclanthology.org/W19-7301/

Lacoste, A., A. Luccioni, V. Schmidt, and T. Dandres (2019). Quantifying the Carbon Emissions of Machine Learning. arXiv preprint arXiv:1910.09700. https://arxiv.org/abs/1910.09700

Leech, G. N., and M. Short (2007). *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. Pearson Education.

Lindseth, Jon A., Tannenbaum, Alan (2015) (eds.): *Alice in a World of Wonderlands: The Translations of Lewis Carroll's Masterpiece*, vol. I, 739-–740. New Castle: Oak Knoll Press.

Macken, L. (2024). Machine translation meets large language models: Evaluating ChatGPT's ability to automatically post-edit literary texts. In Proceedings of the 1st Workshop on Creative-

text Translation and Technology. European Association for Machine Translation. https://aclanthology.org/2024.ctt-1.7/

Matusov, E. (2019). The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, 10–19, Dublin, Ireland, August. European Association for Machine Translation. https://aclanthology.org/W19-7302/

Michel, P., and G. Neubig (2018). Extreme adaptation for personalized neural machine translation. *arXiv:1805.01817*. https://arxiv.org/abs/1805.01817

Mikros, G., V. Sosoni, V. Manousakis, and K. Polychroniou (2026). Burrows' Delta as a Convergent Validator: Stylometric Analysis for Complementary Machine Translation Evaluation. Journal of Quantitative Linguistics, 1–29. Taylor & Francis. https://doi.org/10.1080/09296174.2026.2612931

O'Brien, S. (2023). Human-centered augmented translation: against antagonistic dualisms. Perspectives, 32(3): 391–406. https://doi.org/10.1080/0907676X.2023.2247423

O'Sullivan, J. (2025). Stylometric comparisons of human versus AI-generated creative writing. Humanities and Social Sciences Communications, 12(1): 1–6. Springer Nature. https://doi.org/10.1057/s41599-025-05986-3

Paloposki, O., and Koskinen, K. (2010). Reprocessing texts. The fine line between retranslating and revising. Across Languages and Cultures, 11(1), 29–49. https://doi.org/10.1556/Acr.11.2010.1.2

Peeters, K., and van Poucke, P. (2023). Retranslation, thirty-odd years after Berman. Parallèles, 35(1), 3–27. https://doi.org/10.17462/PARA.2023.01.01

Rothwell, A. (2023). Retranslating Proust Using CAT, MT, and Other Tools. In Computer-Assisted Literary Translation, 106–125. Routledge.

Ruffo, P., Daems, J., and Macken, L. (2024). Measured and perceived effort: assessing three literary translation workflows. Tradumàtica, (22), 238–257. https://doi.org/10.5565/rev/tradumatica.378

Rybicki, J. (2025). Can machine translation of literary texts fool stylometry? Digital Scholarship in the Humanities, 40(1): 268–276. https://doi.org/10.1093/llc/fqaf010

Sak, H., T. Güngör, and M. Saraçlar (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In Proceedings of the International Conference on Natural Language Processing, 417–427. Springer. https://doi.org/10.1007/978-3-540-85287-2_40

Saldanha, G. (2011). Translator style: Methodological considerations. The Translator, 17(1): 25–50. https://doi.org/10.1080/13556509.2011.10799478

Sluyter-Gäthje, H., F. Barteld, and H. Zinsmeister (2018). Neural machine translation for literary texts. Unpublished manuscript.

Şahin, M., and N. Dungan (2014). Translation testing and evaluation: A study on methods and needs. *Translation & Interpreting*, 6(2): 67–90.

Şahin, M., D. Duman, S. Gürses, D. Kaleş, and D. Wools (2018). Towards an empirical methodology for identifying plagiarism in translation. In Perspectives on Retranslation: Ideology, Paratexts, Methods, 161–196. Routledge.

Şahin, M., and S. Gürses (2019). Would MT kill creativity in literary retranslation? In Proceedings of the Qualities of Literary Machine Translation, 26–34. European Association for Machine Translation. https://aclanthology.org/W19-7304/

Şahin, M., and S. Gürses (2023). A call for a fair translatiosphere in the post-digital era. Parallèles: 1–17. http://dx.doi.org/10.17462/para.2022.02.01

Şahin, M., E. Hodzik, S. Gürses, T. Güngör, Z. Yirmibeşoğlu, O. Dursun, and H. Dallı (2025). Stylistic Features and Creativity in Machine Translation of Literary Texts. Yearbook of Translational Hermeneutics, 5(1): 165–196. https://doi.org/10.52116/yth.vi1.105

Tahir Gürçağlar, Ş. (2009). Retranslation. In M. Baker and G. Saldanha (eds.), *Routledge Encyclopedia of Translation Studies*, 233–236. Routledge.

Tıson, A. (2015) The History of Turkish Translations of Alice, *Alice in a World of Wonderlands: The Translations of Lewis Carroll's Masterpiece*, Ed. Jon A. Lindseth, Oak Knoll, Delaware, 601–604.

Topal, M. Ş., and N. Tanış Polat (2025). Yazın Çevirisinde Yapay Zekâ: Etik Değerlendirmeler ve Geleceğe Yönelik Çıkarımlar. Çeviribilim ve Uygulamaları Dergisi, 38: 108–126. https://doi.org/10.37599/ceviri.1662979

Toral, A., and A. Way (2015). Machine-assisted translation of literary text: A case study. *Translation Spaces*, 4(2): 240–267. https://doi.org/10.1075/ts.4.2.04tor

Toral, A., and A. Way (2018). What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment*, 263–287. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-91241-7_12

Turell, M. T. (2004). Textual kidnapping revisited: The case of plagiarism in literary translation. International Journal of Speech, Language and the Law, 11(1): 1–26. https://doi.org/10.1558/ijsll.v11i1.1

van Egdom, G.-W., C. Declercq, and O. Kosters (2024). "Can make mistakes": Prompting ChatGPT to Enhance Literary MT Output. In Proceedings of the 1st Workshop on Creative-text Translation and Technology, 10–20. European Association for Machine Translation. https://aclanthology.org/2024.ctt-1.2/

Wang, Y., C. Hoang, and M. Federico (2021). Towards modeling the style of translators in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1193–1199, Online, June. Association for Computational Linguistics. https://aclanthology.org/2021.naacl-main.94/

Wang, Y., Z. Sun, S. Cheng, W. Zheng, and M. Wang (2023). Controlling styles in neural machine translation with activation prompt. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2606–2620, Toronto, Canada, July. (Preprint arXiv:2212.08909, Dec. 2022.) Association for Computational Linguistics. https://aclanthology.org/2023.findings-acl.163/

Weaver, W. (1964). *Alice in many tongues: the translations of Alice in wonderland*. The University of Wisconsin Press, Madison.

Yamada, M. (2023). Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. In Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track, 195–204. Asia-Pacific Association for Machine Translation. https://aclanthology.org/2023.mtsummit-users.19/

Yirmibeşoğlu, Z., O. Dursun, H. Dallı, M. Şahin, E. Hodzik, S. Gürses, and T. Güngör (2023). Incorporating human translator style into English-Turkish literary machine translation. Proceedings of the 24th Annual Conference of the European Association for Machine Translation, 419–428. https://aclanthology.org/2023.eamt-1.40/

Appendix A Visualizations of Quantitative Analyses

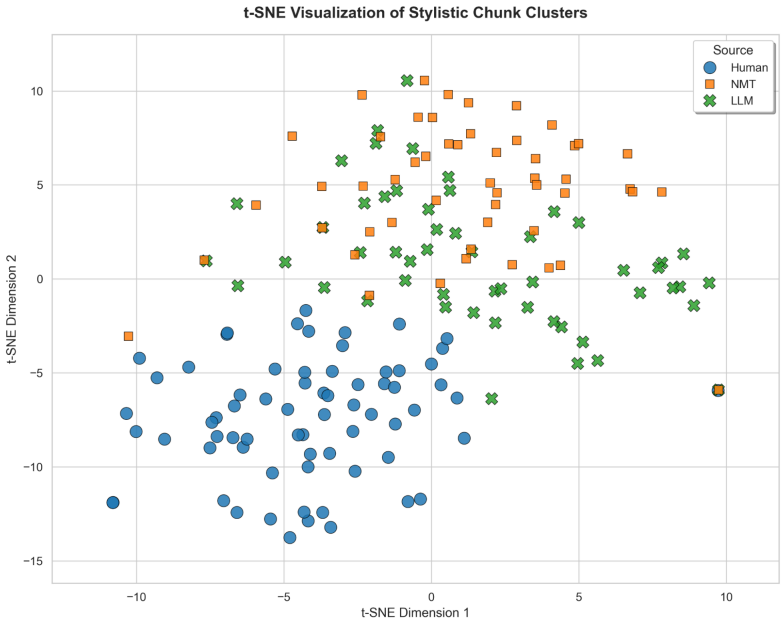

Figure A1: Two-dimensional t-SNE projection of the multidimensional stylistic feature space. Each point represents a 25-sentence text chunk. Human chunks form a tightly isolated grouping, while NMT and LLM chunks occupy adjacent and partially overlapping spaces.

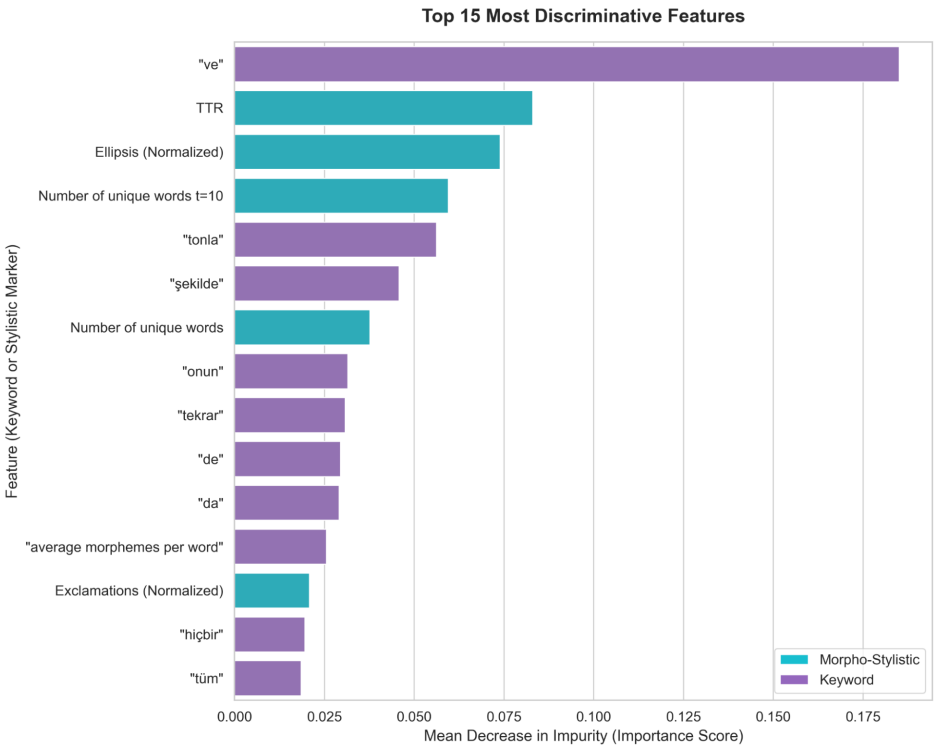

Figure A2: Top most discriminative features identified by the Random Forest classifier, ranked by Mean Decrease in Impurity. Keyword markers and morpho-stylistic features are color-coded to highlight the blend of structural and lexical identifiers used by the model.

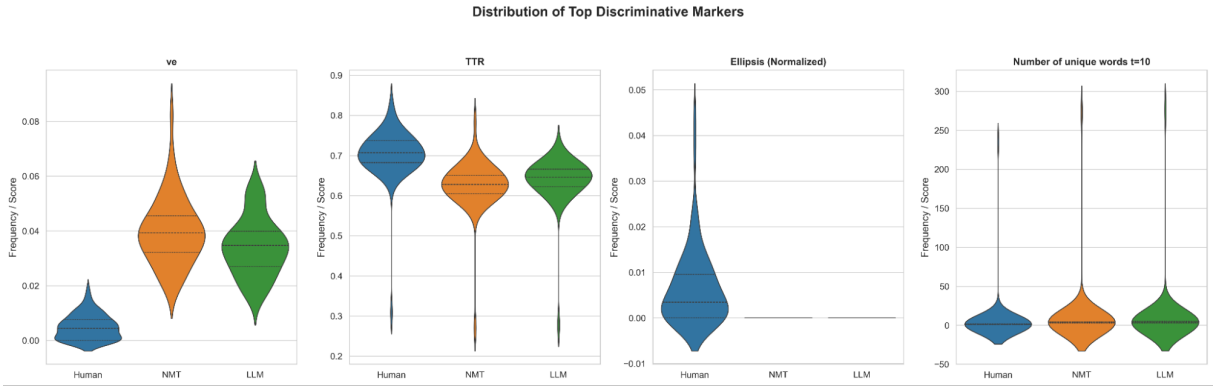

Figure A3: Violin plots illustrating the probability density and distribution variance of top discriminative markers across the three translation sources.

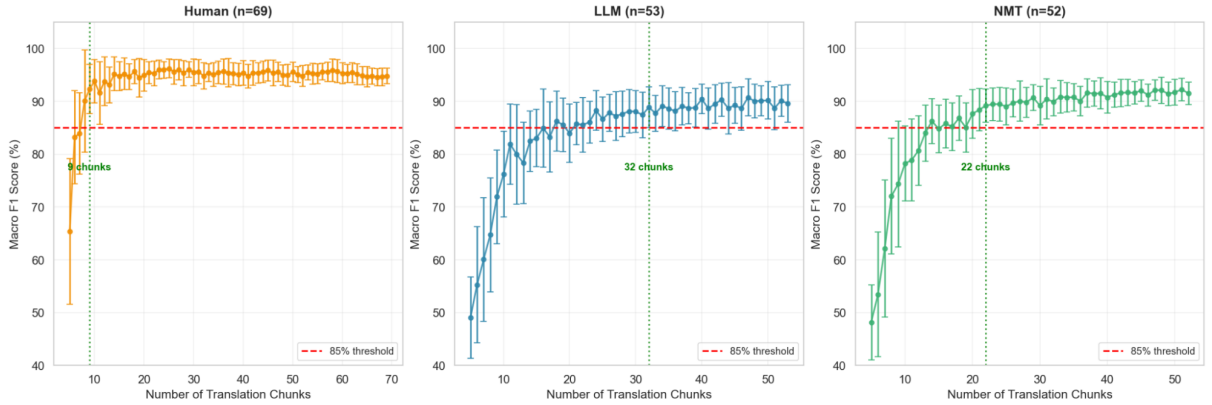

Figure A4: Sample size analysis depicting classification accuracy as a function of the number of text chunks for each translation category. The dashed lines represent the 85% robustness threshold.

Appendix B Additional Translation Examples

Source Text: “There was nothing so very remarkable in that.”

(1) Erten:

Bunda şaşacak bir şey yoktu ;
this-LOC be.amazed-FUT.PART one thing non.existent-PST ;

(2) Dişbudak:

Bunda o kadar şaşacak bir şey yoktu elbette .
this-LOC that much be.amazed-FUT.PART one thing non.existent-PST of.course .

(3) Kına:

Bunda öyle çok fazla bir olağanüstülük yoktu ;
this-LOC so very much one extraordinary-NMLZ non.existent-PST ;

(4) Human-Model:

Bunda çok dikkat çekici bir şey yoktu ;
this-LOC very attention pull-ADJ one thing non.existent-PST ;

Excerpt B1: Comparison of translations for the phrase “nothing so very remarkable,” demonstrating the machine model’s tendency toward lexical normalization compared to human creative choices.

Source Text: “...a small passage, not much larger than a rat-hole:”

(5) Erten:

...fare deliğinden farksız görünen bir geçide girildiğini
...rat hole-3SG.POSS-ABL different-PRIV appear-PRES.PART a passage-DAT enter-PASS-VN-ACC
anladı ;
understand-PST ;

(6) Dişbudak:

...fare deliğinden pek de büyük sayılmazdı .
...rat hole-3SG.POSS-ABL much-INTENS large count-PASS-NEG-PST.HAB ;

(7) Kına:

...fare deliğinden çok fazla geniş olmayan küçük bir geçide
...rat hole-3SG.POSS-ABL very much wide be-NEG-PRES.PART small a passage-DAT
açılıyordu :
open-PROG-PST ;

(8) Human-Model:

...fare deliğinden çok daha büyük olmayan küçük bir geçite
...rat hole-3SG.POSS-ABL much more large be-NEG-PRES.PART small a passage-DAT
gittiğini fark etti :
go-VN-ACC notice-PST ;

Excerpt B2: Comparison of translations for the phrase “not much larger than a rat-hole,” highlighting the machine model’s rigid adherence to source-text syntax versus human syntactic restructuring.

Appendix C Idiomatic Translation Excerpts

Source Text: “ ‘Curiouser and curiouser!’ cried Alice (she was so much surprised, that for the moment she quite forgot how to speak good English);”

(9) Human-Model:

Alis: “Galiba daha tuhaf,” diye bağırdı (o kadar şaşırmıştı ki, o
Alice: "I.guess more strange," say-ADV shout-PST (that degree get.surprised-INFER-PST COMP, that
anda nasıl düzgün konuşulacağını tamamen unutmuştu);
moment-LOC how properly speak-PASS-FUT.PART-3SG.POSS-ACC completely forget-INFER-PST);

(10) LLM-Model:

“Daha da tuhaf!" diye bağırdı Alice (o kadar şaşırmıştı ki, o
"More FOC strange!" say-ADV shout-PST Alice (that degree get.surprised-INFER-PST COMP, that
an için nasıl iyi İngilizce konuşulacağını tamamen
moment for how good English speak-PASS-FUT.PART-3SG.POSS-ACC completely
unutmuştu);
forget-INFER-PST);

(11) NMT-Model:

“Daha meraklı ve meraklı!" diye bağırdı Alice (o kadar şaşırdı ki, bir
"More curious and curious!" say-ADV shout-PST Alice (that degree get.surprised-PST COMP, one
an için iyi İngilizce konuşmayı unuttu)..."
moment for good English speak-VN-ACC forget-PST)..."

Excerpt C1: Comparison of machine translations for the phrase “Curiouser and curiouser!”, highlighting the models’ differing approaches to idiomatic expressions and complex sentence structures.

Appendix D Stylistic Features Taxonomy

Word Level Features	Sentence Level Features	Morphological Data	Keywords
Type-token ratio	Ellipsis sentences	Average morphemes per sentence	Positive and negative keywords extracted with AntConc for each dataset
Number of unique words	Question sentences	Median morphemes per sentence	
Number of unique words, threshold = 10	Exclamation sentences	Average morphemes per word	
Mean word length (characters)	Mean sentence length	Median morphemes per word	
Standard deviation of word lengths	Standard deviation of sentence lengths		
Reduplications	Median of sentence lengths		
	Mode of sentence lengths		

Table D1: Stylistic features used in translator style analysis (adapted from Yirmibeşoğlu et al., 2023).

Appendix E Corpus Statistics

Human Translation Training Data (Dataset_OT)	Word Count	Sentence Count
Erten (1953)	19,055	2,335
Dişbudak (1995)	18,388	3,106
Erzincan Kına (2010)	19,772	2,341
Total Training Corpus	**57,215**	**7,782**

Table E1: Word and sentence counts for the original human translations used to fine-tune the Human-Model.

Generated Retranslations (Evaluation Outputs)	Word Count	Sentence Count
Human-Model Output	17,593	2,461
NMT-Model Output	18,362	2,208
LLM-Model Output	18,944	2,177

Table E2: Word and sentence counts for the complete generated texts analyzed in the study.