



Proceedings of the
**First Workshop on
Style in GenAI-Translated Content
(StyGenAI)**

15 June 2026

Edited by

Natalia Resende, Sheila Castilho, Sami Ul Haq and Maria Clara Menezes



The papers published in these proceedings are—unless indicated otherwise—covered by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence. You may copy, distribute, and transmit the work, provided that you attribute it (authorship, proceedings, publisher) in the manner specified by the author(s) or licensor(s), and that you do not use it for commercial purposes. The full text of the licence may be found at: <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>

© 2026 The authors

ISBN: 9789403901381

DOI: 10.26116/9789403901381

URL: <https://doi.org/10.26116/9789403901381>

Contents

Preface by the Workshop Organisers	i
StyGenAI 2026 Committees	iii
Workshop Programme	v
Papers	1
Nooredeen Awwad, Ebtihal Enfes and Kolawole John Adebayo. <i>Tracing Style in English–Arabic Translation: A Stylometric Comparison of Human and Machine Outputs</i>	3
Antonio Castaldo and Johanna Monti and Sheila Castilho. <i>Emotion Profiling in LLM-Based Literary Translation: Systematic Shifts Across MT and Post-Editing</i>	17
Hüseyin Emir Akdağ, Yusuf Mert Aygün, Mehmet Şahin, Ena Hodzik, Sabri Gürses and Tunga Güngör. <i>Retranslation at the intersection of style, machines, and plagiarism</i>	27
Marthe Lamote, Ewoenam Tokpo, Tom Vanallemeersch, Sara Szoc and Koen Van Winckel. <i>Style and Terminology in Commercial Flows: Mixing APE with Iterative Feedback</i>	47
Natalia Resende and Sheila Castilho. <i>The Style of Machines: A Stylometric Study of LLM Generation and Translation</i>	57
Beniamin Sopot and Dorothy Kenny. <i>Toying with Style: Can GenAI Mimic a Literary Translator’s Voice?</i>	67
Aurora Trapella, Lieve Macken and Alessandra Molino. <i>Lexical Variation in English–Italian News Translation: A Comparative Study of Google Translate and ChatGPT</i>	79
Lise Volkart and Pierrette Bouillon. <i>Lexical and Syntactic Diversity: Still Lost in Machine Translation?</i>	89

Preface by the Workshop Organisers

This volume contains the proceedings of the 1st Workshop on Style in Generative AI-Translated Content (StyGenAI 2026), co-located with the 27th Annual Conference of the European Association for Machine Translation (EAMT 2026). StyGenAI is the first workshop dedicated specifically to the study of style in Generative AI-translated content.

The advent of Generative Artificial Intelligence has profoundly reshaped the landscape of translation. Large language models are increasingly used as translation agents capable of producing linguistically fluent and stylistically sophisticated texts. As a consequence, translation can no longer be evaluated solely in terms of adequacy and accuracy, but also in terms of stylistic choices and effects. While the stylistic characteristics of human translation and conventional machine translation have been widely studied, the stylistic footprint of Generative AI remains comparatively underexplored. StyGenAI was established to address this gap by bringing together expertise from translation studies, computational linguistics, stylistics, natural language processing, and professional translation practice.

The workshop welcomed empirical, methodological, and conceptual contributions examining stylistic regularities, deviations, and innovations in AI-generated translation. Particular attention was given to questions of genre, language pair, prompting strategies, evaluation methodologies, and human-AI interaction in translation workflows. The workshop provided a forum for researchers and practitioners interested in understanding how large language models shape translated texts, how their outputs differ from both human translation and traditional machine translation systems, and how stylistic patterns emerge, stabilise, or drift across genres, languages, and translation settings. The accepted papers reflect the diversity of approaches currently being developed to investigate style in GenAI-mediated translation.

Submissions and programme The technical programme included eight papers covering a broad range of topics. Contributions explore emotion profiling in literary translation, lexical and syntactic diversity in machine translation, stylometric analyses of LLM-generated and translated texts, interactions between style and terminology in commercial translation workflows, retranslation at the intersection of style and plagiarism, lexical variation in news translation, stylometric comparisons of human and machine translation, and the capacity of Generative AI systems to emulate the voice of literary translators. Together, these papers demonstrated the growing interest in understanding style as a central dimension of AI-assisted and AI-generated

translation. In addition to the technical programme, we were honoured to host two distinguished keynote speakers. Dr Marzena Karpinska (Simon Fraser University) who delivered the first keynote of the workshop. Her research focuses on the evaluation and machine translation of creative texts, with particular emphasis on the stylistic footprint of Generative AI and the ability of language models to generate and translate creative content across different styles. Professor Carl Vogel (Trinity College Dublin) who presented the second keynote on Theory of Style in Language. His longstanding work on computational approaches to style analysis, authorship, discourse, and interpretation provides valuable perspectives on the identification and analysis of stylistic signals in both human and AI-generated texts. As the first workshop devoted exclusively to style in Generative AI translation, StyGenAI represents an important step towards establishing a research agenda for this emerging area. The contributions collected in this volume demonstrate both the complexity of stylistic phenomena in AI-generated translation and the need for interdisciplinary approaches capable of addressing them.

Acknowledgments We sincerely thank all those who contributed to the success of the workshop: the authors for their submissions and enthusiasm for this emerging field; the members of the Programme Committee for their thoughtful reviews and valuable feedback; the keynote speakers for sharing their expertise; and the EAMT 2026 organisers for their support and collaboration.

We hope that these proceedings stimulate further research into the stylistics of Generative AI translation and foster new collaborations across disciplinary boundaries. We also hope that StyGenAI serves as the beginning of an ongoing conversation about style, creativity, authorship, and evaluation in an era increasingly shaped by Generative AI.

We hope you enjoy reading the papers as much as we did!

June 2026,

Natália Resende, Sheila Castilho, Sami Ul Haq and Maria Clara Menezes.

StyGenAI 2026 Committees

Organising Committee & Workshop Chairs

Natalia Resende, School of Languages, Literatures and Cultural Studies, Trinity Centre for Literary and Cultural Translation/ADAPT Centre, Trinity College Dublin, Ireland

Sheila Castilho, School of Applied Language & Intercultural Studies (SALIS)/ADAPT Centre, Dublin City University, Ireland

Sami Ul Haq, School of Applied Language & Intercultural Studies (SALIS)/ADAPT Centre, Dublin City University, Ireland

Maria Clara Menezes, School of Applied Language & Intercultural Studies (SALIS), Dublin City University, Ireland

Programme Committee

Paşa Abdullah Bayramoğlu — Üsküdar University

Maria Zafar — South East Technological University

Miquel Esplà-Gomis — Universitat d'Alacant

Benjamin Sopot — Dublin City University

Lama Alharthi — Dublin City University

Antonio Castaldo — University of Naples L'Orientale and University of Pisa

Maria Fernandez-Parra — Swansea University

Maria del Mar Sánchez Ramos — Universidad de Alcala

Lieve Macken — Ghent University

Manuel Lardelli — University of Padua

David Orrego-Carmona — University of Warwick

Gokhan Dogru — Pompeu Fabra University

Christophe Declercq — Utrecht University

Dimitris Asimakoulas — University of Surrey

Michael Carl — Kent State University

Martin Popel — UFAL, Faculty of Mathematics and Physics, Charles University

Miguel Ángel Jiménez Crespo — Rutgers University

Akiko Sakamoto — Kansai University

Abigail Walsh — ADAPT Centre, Dublin City University

Constantin Orasan — University of Surrey

Nora Aranberri — University of the Basque Country

Aletta G. Dorst — Leiden University

Wichaya Pidchamook — Thammasat University

Federico Gaspari — ADAPT Centre, Dublin City University

Vicent Briva-Iglesias — SALIS, CTTS, ADAPT Centre, Dublin City University

Antonio Toral — Universitat d’Alacant
Goedele De Sterck — Universidad de Salamanca
Antoni Oliver — Universitat Oberta de Catalunya
Chinonso Osuji — ADAPT Research Centre, Dublin City University
Patrick Cadwell — Dublin City University
Monja Jannet — Dublin City University
Ana Guerberof Arenas — University of Groningen
Rudy Loock — Université de Lille
Callum Walker — University of Leeds
Vincent Vandeghinste — Instituut voor de Nederlandse Taal, Leiden // Centre for Computational Linguistics, KU Leuven
Susana Valdez — Leiden University Centre for Linguistics
Sanjun Sun — Beijing Foreign Studies University
Mihaela Vela — Universität des Saarlandes
James Hadley — Trinity College Dublin
Sari Hokkanen — Tampere University
Sébastien Le Maguer — Trinity College Dublin / ADAPT Centre
Mustafa Keshkeia — Dublin City University
Kirti Vashee — Translated

Workshop Programme

Time	Activity
09:00 – 09:05	Opening notes
09:05 – 10:00	Keynote Speech: Dr Marzena Karpinska
10:00 – 10:30	Oral Presentations: Style in GenAI-Generated Literary Translation <i>Toying with Style: Can GenAI Mimic a Literary Translator's Voice?</i> - Beniamin Sopot and Dorothy Kenny
10:30 – 11:00	Coffee break
11:00 – 12:30	Oral Presentations: Style in GenAI-Generated Literary Translation (cont.) <i>Emotion Profiling in LLM-Based Literary Translation: Systematic Shifts Across MT and Post-Editing</i> - Antonio Castaldo, Johanna Monti and Sheila Castilho <i>Retranslation at the Intersection of Style, Machines, and Plagiarism</i> - Hüseyin Emir Akdağ, Yusuf Mert Aygün, Mehmet Şahin, Ena Hodzik, Sabri Gürses and Tunga Güngör
12:00 – 12:30	Oral Presentations: Comparative Studies on Style in NMT, LLMs and HT <i>Lexical and Syntactic Diversity: Still Lost in Machine Translation?</i> - Lise Volkart and Pierrette Bouillon
12:30 – 13:30	Lunch break
13:30 – 14:30	Keynote Speech: Prof Carl Vogel - <i>Theory of Style in Language</i>
14:30 – 15:00	Oral Presentations: Comparative Studies on Style in NMT, LLMs and HT (cont.) <i>Tracing Style in English–Arabic Translation: A Stylometric Comparison of Human and Machine Outputs</i> - Nooredeen Awwad, Ebtihal Enfes and Kolawole John Adebayo
15:00 – 15:30	Coffee break
15:30 – 17:00	Oral Presentations: Comparative Studies on Style in NMT, LLMs and HT (cont.) <i>Style and Terminology in Commercial Flows: Mixing APE with Iterative Feedback</i> - Marthe Lamote, Ewoenam Tokpo, Tom Vanallemeersch, Sara Szoc and Koen Van Winckel <i>Lexical Variation in English–Italian News Translation: A Comparative Study of Google Translate and ChatGPT</i> - Aurora Trapella, Lieve Macken and Alessandra Molino <i>The Style of Machines: A Stylometric Study of LLM Generation and Translation</i> - Natália Resende and Sheila Castilho
17:00 – 17:15	Closing remarks

Papers

Tracing Style in English–Arabic Translation: A Stylometric Comparison of Human, Neural, and LLM-Generated Outputs

Nooredeen Awwad

School of Computing,
University of Kent, UK
nawwad65@gmail.com

Ebtihal Enfes

Computer Engineering Department,
Eastern Mediterranean University, Turkey
ebtihal.enfes@emu.edu.tr

Kolawole J. Adebayo

International Engineering College,
Maynooth University, Ireland
kolawole.adebayo@mu.ie

Abstract

This paper investigates stylistic variation in English–Arabic news translation across human and machine-generated outputs. We compile a balanced dataset of 250 source texts and generate 1,000 translations using human translators, Google Cloud Translation, GPT-4O, and a Qwen3-8B-based open-source model. We perform a stylometric analysis measuring structural and lexical properties, including entropy, lexical diversity, repetition, and length ratios. Results reveal consistent, system-specific stylistic tendencies. Machine-generated translations exhibit normalization effects, with differences in compression, repetition, and lexical diversity across systems and target languages. These findings demonstrate that modern translation systems produce distinct stylistic signatures, motivating evaluation beyond adequacy and fluency toward fine-grained stylistic and discourse-level measures.

1 Introduction

Modern machine translation (MT) systems increasingly produce fluent and semantically accurate outputs. Advances in neural machine translation (NMT), particularly Transformer-based architectures (Vaswani et al., 2017), and instruction-following large language models (LLMs) have improved translation quality across many language pairs (Zhang et al., 2023; He et al., 2024; Huang

and Liu, 2024; Hendy et al., 2023). As a result, machine-generated translations often perform well on conventional evaluation metrics such as BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020). However, adequacy and fluency do not fully capture the linguistic properties of translation. Systems that preserve meaning may still differ in lexical richness, structural variation, repetition, compression, and discourse organisation.

Research in translation studies and computational linguistics has long shown that translated texts exhibit systematic regularities, including simplification, normalisation, and explicitation (Baker, 1993; Volansky et al., 2015; Rabinovich et al., 2017). These features, commonly associated with translationese, may appear through reduced lexical diversity, increased repetition, and more regular structural patterns. Prior work suggests that such effects can be amplified in machine-generated translation, particularly in neural MT systems (Vanmassenhove et al., 2019). More recent studies further indicate that LLM-based translation may introduce stylistic homogenisation because alignment and instruction-tuning objectives favour fluent, predictable outputs (Qian et al., 2024; Kong and Macken, 2025). Entropy- and distribution-based methods now make it possible to examine these tendencies through interpretable stylometric measures (Yao and Fan, 2025).

Despite these developments, systematic comparisons across human translation, commercial neural MT, closed-source LLMs, and open-source LLMs remain limited. Existing research has often focused on adequacy-oriented evaluation or isolated human–machine comparisons. Less attention has been paid to whether modern translation systems exhibit consistent and distinguishable stylistic signatures across translation paradigms

and directions. This gap is especially relevant for English–Arabic translation, where stylistic variation beyond conventional quality assessment remains underexplored.

English–Arabic translation provides a useful setting for this analysis because the two languages differ substantially in morphology, syntax, and discourse organisation (Hatim and Mason, 1990; Dickins et al., 2016; Ryding, 2005). Arabic allows greater flexibility in sentence structure, lexical elaboration, and cohesion strategies, whereas English generally favours more constrained syntactic patterns (Enfes et al., 2026). Consequently, semantically equivalent translations may still differ in readability, redundancy, lexical density, and discourse flow (House, 2015; Vanmassenhove et al., 2019).

News translation is also an appropriate domain because it combines stable communicative goals with meaningful variation in discourse structure and journalistic style (van Dijk, 1988). Contemporary MT and LLM-based translation research continues to rely heavily on news and public-affairs domains for evaluation and benchmarking (Kocmi et al., 2024). However, comprehensive stylistic comparisons spanning commercial MT systems, proprietary LLMs, open-source LLMs, and human translation remain underexplored.

To address this gap, this paper investigates whether modern MT and LLM-based translation systems exhibit identifiable stylistic signatures relative to human translation in English–Arabic news translation. We compile a balanced bidirectional dataset of news and public-affairs texts and generate translations using human translators, Google Translate, GPT-4o, and a Qwen3-8B-based open-source model. Using a multilingual stylometric framework, we analyse lexical and structural variation through entropy-based, repetition-sensitive, and distributional measures, complemented by qualitative examination of discourse-level phenomena such as simplification, redundancy, and compression.

Our contributions are threefold:

- We introduce a controlled bidirectional English–Arabic news translation dataset covering sentence- and paragraph-level outputs across human, MT, and LLM-based conditions.
- We develop a multilingual stylometric framework combining entropy-based, lexical,

structural, and repetition-sensitive measures for comparing translation outputs.

- We show that contemporary MT and LLM-based systems exhibit consistent, system-specific stylistic signatures, motivating style-aware approaches to MT evaluation beyond adequacy alone.

2 Related Work

2.1 Translationese, simplification, and stylometric variation

Research in translation studies has long argued that translated texts exhibit systematic linguistic regularities that distinguish them from non-translated language, a phenomenon commonly referred to as translationese. Foundational work identifies recurring tendencies such as simplification, explicitation, normalization, and levelling out (Baker, 1993). Subsequent corpus-based studies demonstrated that these phenomena are observable through measurable linguistic properties, including reduced lexical diversity, lower syntactic complexity, increased repetition, and more regular structural patterns (Volansky et al., 2015; Rabinovich et al., 2017; Kong and Macken, 2025). Simplification, in particular, has been extensively examined using features such as sentence length, clause density, and lexical richness. For example, Liu and Afzaal (2021) report lower syntactic complexity in translated English texts, supporting the view that translation often favours structurally simpler constructions.

Within machine translation research, these questions have increasingly been reframed in terms of machine translationese, referring to systematic regularities introduced by machine-generated translation. Prior studies show that neural MT outputs often exhibit stronger normalization effects than human translation, including reduced lexical richness and increased structural regularity (Toral et al., 2018; Vanmassenhove et al., 2019). Such findings suggest that stylistic compression in MT output is not merely anecdotal but computationally measurable. Recent work has further extended this line of inquiry using entropy- and distribution-based approaches. Yao and Fan (2025), for example, demonstrate that lexical and syntactic variation across human, neural MT, and LLM-generated translations can be effectively quantified using entropy-based measures.

Stylometric analysis has played an important role in operationalising these phenomena computationally. Rabinovich et al. (2017) show that translated texts encode detectable statistical patterns that can even reveal typological relationships between languages, while Volansky et al. (2015) provide a systematic characterisation of translationese features across multiple linguistic dimensions. Together, these studies establish that translationese and machine translationese are not only theoretically observable but also computationally separable through lexical, structural, and distributional signals.

Despite these advances, existing work has largely focused on isolated comparisons between human and machine translation or on individual MT paradigms. Comparatively little research has examined whether commercial neural MT systems, proprietary LLMs, and open-source LLMs exhibit distinct and internally consistent stylistic signatures within a unified comparative framework, particularly in English–Arabic translation.

2.2 LLM-based translation and prompting

Large language models have recently become competitive alternatives to conventional neural MT systems, with strong performance across language pairs and domains. Prior work shows that prompt formulation, example selection, and multi-step prompting can substantially affect translation quality, robustness, and error reduction (Zhang et al., 2023; He et al., 2024).

Unlike conventional NMT systems, LLM-based translation is shaped by prompting conditions, decoding behaviour, alignment objectives, and interaction design. These factors may affect not only semantic correctness, but also lexical variation, sentence structure, repetition, and discourse organisation. Recent studies suggest that instruction tuning and alignment can encourage stylistic homogenisation by favouring fluent, predictable, and regular outputs (Qian et al., 2024; Kong and Macken, 2025).

However, most research on LLM-based translation still focuses on adequacy, fluency, hallucination reduction, or prompting efficiency. Less attention has been paid to whether different LLM-based systems produce distinguishable stylistic behaviours relative to human translation and conventional neural MT.

2.3 Evaluating translation beyond adequacy and fluency

Machine translation evaluation has traditionally centred on adequacy and fluency, often measured using automatic metrics such as BLEU (Papineni et al., 2002) and neural metrics such as COMET (Rei et al., 2020). Although these metrics are useful for assessing semantic correctness, they are less sensitive to stylistic and discourse-level variation (Freitag et al., 2022). Translations with similar adequacy scores may still differ in readability, lexical richness, compression, repetition, and structural organisation.

Recent work has therefore called for broader and more context-sensitive evaluation frameworks. Liu et al. (2024) argue for incorporating translation-studies perspectives into MT evaluation, especially communicative purpose and contextual appropriateness. Qian et al. (2024) show that LLM-based evaluation is sensitive to prompt design, annotation guidelines, and reference information, while Song et al. (2025) show that comparative judgment frameworks can improve consistency in human evaluation.

These developments show that translation quality cannot be reduced to semantic equivalence alone. However, evaluation remains largely accuracy-oriented and often overlooks stylistic and structural variation. Stylometric analysis offers a complementary approach by capturing lexical, structural, and discourse-level properties that conventional MT metrics may miss.

2.4 Machine translationese in news translation and the Arabic–English setting

News translation is a useful domain for studying stylistic variation because it combines stable communicative goals with variation in discourse structure, lexical choice, and rhetorical organisation (van Dijk, 1988). It also remains central to MT and LLM-based evaluation and benchmarking (Kocmi et al., 2024). Prior work shows that machine translationese is detectable in news translation: Kong and Macken (2025), for example, find measurable differences in lexical diversity, sentence structure, and discourse cohesion between neural MT, LLM-generated, and human English–Chinese news translations.

The Arabic–English setting is especially relevant because the two languages differ substan-

tially in morphology, syntax, lexical formation, and discourse organisation. Arabic’s rich morphology, flexible sentence structure, and lexical variation create challenges for MT systems and may make stylistic regularities more visible in translated output. However, prior work on Arabic–English MT has focused mainly on adequacy, error analysis, and system-level quality assessment. Abdelaal and Alazzawie (2020), for instance, identify omissions, insertions, and lexical inaccuracies in Google Translate news outputs.

To our knowledge, no prior work has jointly compared human translation, commercial neural MT, proprietary LLMs, and open-source LLMs within a unified English–Arabic stylistic framework. The present study addresses this gap by combining entropy-based, lexical, structural, and repetition-sensitive measures with qualitative discourse analysis.

3 Methodology

3.1 Study design and dataset

The study uses a controlled within-item comparative design to examine stylistic variation across human and machine-generated translations. Each source text was translated under all experimental conditions, enabling direct system comparison while controlling for source-content variation. This design helps minimise confounds related to topic, genre, or source-text differences (Song et al., 2025).

The dataset contains 250 contemporary news and public-affairs texts: 125 English-source items translated into Arabic and 125 Arabic-source items translated into English. It includes short sentences, long sentences, and short paragraphs to examine whether stylistic tendencies remain stable across different discourse lengths and structural conditions.

Table 1 summarises the distribution of text types, translation directions, and source-token statistics.

Text type	Items	En→Ar	Ar→En	Tokens	Mean
Short sent.	100	50	50	1,219	12.2
Long sent.	100	50	50	2,668	26.7
Paragraphs	50	25	25	2,028	40.6
Total	250	125	125	5,915	23.7

Table 1: Source-text dataset statistics by translation direction, including item counts and mean source-token length.

The source texts were manually selected from

approximately 600 candidate items drawn from BBC News, Sky News, and Al Jazeera. The screening process excluded fragments, ticker-style content, title-only strings, incomplete passages, and texts dependent on missing context. The final dataset preserves natural variation in news discourse while maintaining comparability across translation conditions.

Existing English–Arabic parallel corpora were not used because the study requires controlled within-item comparison across contemporary human and machine translation systems. Many public corpora also provide limited information about translation provenance, post-editing status, or generation conditions, which is problematic for stylistic analysis.

Each source item was translated under four conditions, resulting in 1,000 translated outputs.

3.2 Translation conditions

Each source text was translated under four conditions: human translation, Google Translate, GPT-4o, and Qwen3-8B. Google Translate served as a commercial neural MT baseline, while GPT-4o and Qwen3-8B represented closed-source and open-source LLM-based translation settings, respectively.

Google Translate outputs were generated using the Google Cloud Translation API in April 2026 with the default neural MT configuration.¹ GPT-4o translations were generated through the OpenAI API using deterministic decoding, and Qwen3-8B was run with the same prompting configuration. No post-editing was applied to any machine-generated outputs.

Human translations were produced independently by two bilingual graduate translators who are native speakers of Arabic. Translators worked only from the source text and did not have access to machine-generated outputs. Instructions emphasised preserving meaning, tone, and completeness without summarisation, omission, or stylistic simplification. These translations served as the human reference condition.

For the two LLM-based conditions, the same translation prompt was used across all items to minimise prompt-induced variability. The prompt instructed the models to preserve meaning, tone, named entities, dates, numbers, and paragraph

¹See <https://docs.cloud.google.com/translate/docs/advanced/nmt-model>.

structure while returning only the translated text. Temperature was set to 0 for all LLM outputs to ensure deterministic behaviour and improve comparability across systems.

The prompt used for GPT-4o and Qwen3-8B is provided in Appendix A.1.

3.3 Data preparation and alignment

All translations were merged into a unified comparison table using shared item identifiers. Each record contains the source text, source language, target language, text type, and the four aligned translations. Outputs were stored in UTF-8 format and manually checked for encoding problems, formatting corruption, empty outputs, and incorrect-language generations.

Preprocessing was intentionally minimal in order to preserve stylistic characteristics. Processing was limited to whitespace cleanup and Unicode normalisation where necessary. No grammatical correction, spelling correction, or stylistic editing was applied after translation generation. Moreover, tokenisation was performed separately for English and Arabic. English tokenisation used standard whitespace and punctuation segmentation. Arabic tokenisation employed rule-based clitic segmentation and orthographic normalisation because tokenisation substantially affects distributional and length-sensitive measures in morphologically rich languages.

Because English–Arabic and Arabic–English outputs are not directly comparable at the surface level, all analyses were conducted separately by translation direction. Paragraph-level items were additionally analysed independently from sentence-level items for measures sensitive to discourse structure and sentence segmentation.

3.4 Stylometric analysis

The quantitative analysis focuses on lexical, structural, repetition-based, and discourse-level features associated with translationese and machine-generated language (Vanmassenhove et al., 2019; Liu and Afzaal, 2021; Yao and Fan, 2025). These measures capture multiple dimensions of stylistic behaviour rather than treating translation quality or simplification as a single property.

Structural measures. Output length was measured in tokens and compared with source length using the following length-ratio metric:

$$\text{LengthRatio}_i = \frac{|T_i|}{|S_i|}$$

where $|T_i|$ denotes the number of tokens in translation i and $|S_i|$ denotes the number of tokens in the corresponding source text. Values above 1 indicate expansion, while values below 1 indicate compression. Mean sentence length was also computed for each system and translation direction. For paragraph-level items, sentence counts were used to identify splitting, merging, or restructuring.

Lexical distribution and entropy. Lexical predictability was estimated using token-level Shannon entropy:

$$H(T) = - \sum_{w \in V} p(w) \log_2 p(w)$$

where V represents the output vocabulary and $p(w)$ the relative frequency of token w . Entropy was calculated over concatenated outputs grouped by system, translation direction, and text type to improve stability for shorter texts. Higher entropy indicates greater lexical dispersion, while lower entropy indicates greater predictability.

Lexical richness was measured using moving-average type-token ratio (MATTR) (Covington and McFall, 2010), which is less sensitive to text length than raw type-token ratio. MATTR was calculated separately for each system-by-direction subset.

Repetition and simplification. Repetition was measured using repeated-token, repeated-bigram, repeated-trigram, and duplicate-sentence rates. Repeated bigram rate was used as the main repetition measure because it captures local phrasal recurrence while remaining less sparse than higher-order n -grams.

Simplification was treated as a multidimensional construct, reflected in lower lexical richness, reduced entropy, shorter output length relative to the source, and lower mean sentence length. This follows prior work showing that simplification emerges through interacting lexical and structural tendencies (Liu and Afzaal, 2021; Yao and Fan, 2025).

Discourse-level behaviour. Paragraph-level outputs were analysed to assess discourse-level behaviour beyond sentence-level translation,

including boundary preservation, sentence restructuring, expansion or compression, and added discourse framing or connectives.

3.5 Qualitative analysis

Quantitative analysis was complemented by close qualitative examination of representative translation examples. The qualitative component was designed as metric-guided close reading rather than a separate annotation study. Its purpose was to explain how stylistic tendencies observed quantitatively manifested in concrete translational choices.

The qualitative categories were defined prior to analysis based on the stylistic dimensions investigated quantitatively and on prior research on translationese and machine translationese. The analysis focused on repetition, redundancy, explicitation, compression, omission, register shift, literalness, awkward phrasing, and discourse restructuring.

Examples were selected using a purposive sampling procedure guided by quantitative results. Outputs exhibiting strong compression, expansion, entropy change, repetition increase, lexical-richness reduction, or sentence-length variation were first identified at the item level. Candidate cases were then grouped by translation direction, text type, and system, after which representative examples illustrating recurring system-level tendencies were selected for close comparison against the source text and competing translations.

3.6 Statistical analysis

Because each source item was translated under all experimental conditions, the study employs a repeated-measures design. Statistical analyses were conducted on aligned outputs sharing the same source item and were performed separately for each translation direction.

Because stylistic measures were not assumed to follow normal distributions, non-parametric repeated-measures tests were used throughout. Omnibus system differences were assessed using the Friedman test, followed by pairwise Wilcoxon signed-rank tests with Holm correction for multiple comparisons. Effect sizes were reported alongside significance tests using Kendall’s W for Friedman analyses and rank-biserial correlation for Wilcoxon comparisons as shown in Table 2.

Sentence-level and paragraph-level analyses were additionally reported separately in order to avoid masking discourse-sensitive effects through aggregation.

To ensure reproducibility, we make available all prompts, preprocessing scripts, stylistic-analysis code, and aggregated experimental outputs on this link <https://github.com/NRAwwad/stylometric-english-arabic-translation>.

Metric	χ^2	p	W
Sent. length	29.489	< .001	.098
Token count	32.328	< .001	.108
Token entropy	28.997	< .001	.097
Token/source	32.328	< .001	.108

Table 2: Significant Friedman tests for English-to-Arabic sentence-level outputs.

4 Findings from Quantitative Analysis

4.1 Overall stylistic profiles

Table 3 summarises the principal stylistic features across translation conditions and target languages. Although many numerical differences are modest in magnitude, the repeated-measures analysis indicates that several patterns are statistically consistent across aligned source items, particularly for Arabic-target outputs (see Section 3.6). The findings therefore suggest stable system-level stylistic tendencies rather than random variation.

A clear asymmetry emerges across translation directions. Arabic-target outputs exhibit stronger divergence across systems than English-target outputs, especially with respect to compression, lexical dispersion, and sentence expansion. This pattern is consistent with the greater morphological and syntactic flexibility of Arabic, where stylistic variation may become more visible at the lexical and structural levels (Enfes et al., 2026).

For Arabic-target texts, Google Translate produces the most expansive outputs overall, with the highest mean token count (22.71), sentence length (19.83), entropy (4.25), and MATTR-50 (0.959). These results suggest a tendency toward lexical elaboration and structural expansion relative to the other systems. By contrast, Qwen3-8B consistently produces the most compressed Arabic outputs, with the lowest token count (21.80), sentence length (19.14), entropy (4.18), and source-length ratio (0.917). The lower entropy and reduced lex-

Target	Condition	Tokens	Sent. len.	Entropy	MATTR	Bigram rep.	Ratio
Arabic	Google	22.71	19.83	4.25	0.959	0.0032	0.965
Arabic	Human	22.01	19.38	4.20	0.957	0.0020	0.930
Arabic	GPT-4o	22.09	19.26	4.20	0.956	0.0029	0.933
Arabic	Qwen3-8B	21.80	19.14	4.18	0.956	0.0023	0.917
English	Google	29.29	25.17	4.41	0.883	0.0131	1.256
English	Human	29.78	25.27	4.42	0.881	0.0108	1.274
English	GPT-4o	29.98	25.17	4.42	0.874	0.0127	1.282
English	Qwen3-8B	29.43	24.47	4.43	0.879	0.0107	1.273

Table 3: Main stylistic measures by target language and translation condition. Values represent item-level means.

ical dispersion suggest stronger normalization and concentration effects in the open-source condition.

GPT-4o and the human translation condition occupy an intermediate position and remain closely aligned across most Arabic-target measures. This convergence is notable because GPT-4o was generated under deterministic decoding conditions, suggesting that alignment objectives and instruction-following behaviour may nevertheless produce outputs structurally close to human journalistic translation norms.

Differences among English-target outputs are comparatively smaller but remain systematic. GPT-4o produces the highest mean token count (29.98) and source-length ratio (1.282), indicating a mild tendency toward expansion relative to the source. Qwen3-8B exhibits the highest entropy (4.43), although the difference is extremely small and should therefore be interpreted cautiously as only marginally greater lexical dispersion. Google Translate produces the highest MATTR-50 (0.883) together with the highest repeated-bigram rate (0.0131). The co-occurrence of lexical variation and local phrase repetition suggests a more uneven lexical distribution in which varied vocabulary co-exists with recurrent discourse framing and formulaic phrasal patterns.

Overall, the findings indicate that contemporary translation systems exhibit stable but distinct stylistic profiles. Google Translate shows stronger expansion and repetition tendencies, Qwen3-8B exhibits greater compression and lower lexical concentration in Arabic-target outputs, while GPT-4o remains consistently closest to the human translation condition across several lexical and structural dimensions.

4.2 Machine outputs relative to the human translation condition

Table 4 reports paired deviations between each machine condition and the corresponding human translation for the same source item.

Target	Cond.	Tok. ratio	Len. ratio	Ent. Δ	MATTR Δ	Bi-rep. Δ
Arabic	Google	1.039	1.029	0.052	0.0019	0.0012
Arabic	GPT-4o	1.003	0.995	0.002	-0.0008	0.0009
Arabic	Qwen3	0.992	0.989	-0.021	-0.0010	0.0002
English	Google	0.990	1.013	-0.009	0.0021	0.0023
English	GPT-4o	1.013	1.011	0.002	-0.0065	0.0019
English	Qwen3	1.005	0.995	0.003	-0.0019	-0.0001

Table 4: Machine outputs relative to human translations. Ratios above 1 indicate higher values than humans; deltas show absolute differences.

For Arabic-target texts, Google Translate is on average 3.9% longer than the human translation and shows higher sentence length and entropy, reinforcing its tendency toward expansion and lexical elaboration. GPT-4o remains very close to the human condition across most measures, while Qwen3-8B is slightly shorter and shows the largest negative entropy shift, suggesting stronger compression and lexical concentration.

English-target differences are smaller but still consistent. Google Translate produces slightly shorter outputs than the human condition, but with longer sentence length, higher MATTR-50, and higher repeated-bigram rates, suggesting recurrent local phrasing despite relatively diverse vocabulary. GPT-4o remains close to the human baseline, while Qwen3-8B produces shorter sentences and slightly lower lexical richness.

Overall, the differences are modest and should be interpreted as tendencies rather than categorical distinctions. Their consistency across aligned items nevertheless suggests identifiable system-level translational preferences under controlled prompting conditions.

4.3 Paragraph-level behaviour

Paragraph-level analysis provides a useful setting for observing stylistic restructuring beyond sentence-level translation, including segmentation, clause coordination, and discourse continuity.

For Arabic-target paragraphs, the differences between systems are relatively small. Google Translate produces slightly longer and higher-entropy outputs than the human condition, while Qwen3-8B is marginally more compressed and has the lowest MATTR-50 value. This pattern is consistent with the sentence-level results, where Qwen3-8B showed reduced lexical dispersion in Arabic-target translation.

The English-target paragraphs show clearer variation. As shown in Table 5, the human condition has the highest mean token count, while

Target	Cond.	Tokens	Sents.	Sent. len.	Entropy	MATTR	Bi-rep.
Arabic	Google	37.16	1.88	23.82	5.05	0.938	0.0041
Arabic	Human	36.92	1.84	23.80	5.01	0.930	0.0061
Arabic	GPT-4o	36.92	1.84	23.80	5.01	0.930	0.0061
Arabic	Qwen3	36.68	1.88	23.36	5.00	0.925	0.0044
English	Google	51.56	1.80	32.28	5.14	0.809	0.0286
English	Human	52.92	1.88	30.87	5.16	0.806	0.0266
English	GPT-4o	52.60	1.92	30.18	5.18	0.809	0.0257
English	Qwen3	50.68	2.00	28.23	5.16	0.817	0.0221

Table 5: Paragraph-level stylistic properties of human and machine translations by target language and system.

GPT-4o remains close in both length and entropy. Qwen3-8B produces shorter outputs but has the highest mean number of sentences per paragraph, suggesting a tendency toward sentence splitting and structural simplification.

Google Translate also produces shorter English-target paragraphs than the human condition, but with higher repeated-bigram rates and longer mean sentence length. This suggests a smoothing tendency, where clauses are combined into longer units while relying more on recurrent phrasal patterns.

Overall, the paragraph-level results show that stylistic divergence is not limited to lexical choice. Machine translation systems also differ in sentence restructuring, paragraph rhythm, and discourse organisation, especially in English-target connected discourse.

4.4 Surface lexical similarity to the human translation condition

Table 6 reports a surface-level lexical similarity analysis between the human translation condition and the machine-generated outputs. For each item, the nearest machine condition was identified using token-set Jaccard similarity against the corresponding human translation. The analysis was included to examine lexical alignment patterns rather than translation quality itself.

GPT-4o is the nearest condition to the human translation in 179 out of 250 items, substantially exceeding both Google Translate and Qwen3-8B. Mean Jaccard similarity is also highest for GPT-4o (0.903), indicating considerably stronger lexical overlap with the human translation condition.

Because token-set Jaccard similarity ignores token order, sequence-sensitive similarity was additionally examined using word error rate (WER), where the human translation served as the reference and machine outputs as hypotheses. Lower WER values indicate closer token-sequence align-

ment with the human translation. The WER results confirm the same overall pattern: GPT-4o exhibits substantially lower mean WER (0.0603) than Google Translate (0.3196) and Qwen3-8B (0.4269).

Importantly, surface lexical similarity should not be interpreted as a direct measure of translation quality or stylistic adequacy. High lexical overlap may arise from similar lexical choices while still differing in discourse organisation, sentence rhythm, or stylistic naturalness. Nevertheless, the findings indicate that GPT-4o consistently produces outputs that are lexically and sequentially more aligned with the human translation condition under the controlled prompting setup used in this study.

Condition	Nearest	Rate	Jac.	Mean WER	Corp. WER
Google	39/250	0.156	0.673	0.3196	0.3128
GPT-4o	179/250	0.716	0.903	0.0603	0.0654
Qwen3-8B	32/250	0.128	0.556	0.4269	0.4058

Table 6: Surface similarity to human translations. “Nearest” gives the machine condition with the highest token-set Jaccard similarity; lower WER indicates closer sequence-level similarity.

Overall, the quantitative findings demonstrate that contemporary translation systems exhibit distinct but internally consistent stylistic tendencies across lexical, structural, and discourse-level measures. Google Translate displays stronger expansion and discourse-smoothing behaviour, Qwen3-8B exhibits greater compression and reduced lexical dispersion, while GPT-4o remains systematically closest to the human translation condition across multiple surface and structural dimensions.

5 Findings from Qualitative Analysis

The qualitative analysis examined 84 representative examples selected from the quantitative outputs. The examples were chosen because they showed strong compression, expansion, entropy change, repetition increase, lexical-richness decrease, or sentence-length change relative to the human translation. Table 8 presents representative examples of the main qualitative patterns observed in the close reading, including simplification, redundancy, and paragraph restructuring.

Three main patterns emerged. First, compression did not always reduce adequacy: shorter formulations were sometimes concise and journalistic, but stronger compression occasionally weakened nuance, emphasis, or discourse signals, es-

pecially in Qwen3-8B outputs. Second, Google Translate frequently produced fluent but formulaic English-target outputs, including repeated reporting frames and recurrent phrasal templates, consistent with its higher repeated-bigram rates. Third, stylistic divergence extended beyond lexical choice to discourse organisation, with some systems restructuring sentence boundaries, smoothing transitions, or merging clauses into longer syntactic units.

GPT-4o requires a slightly different interpretation. Although it was the machine condition most lexically aligned with the human translations, lexical similarity did not imply full stylistic equivalence. Qualitative inspection still showed differences in discourse pacing, syntactic preference, and lexical selection, although these were generally less pronounced than for Google Translate and Qwen3-8B.

Overall, the qualitative findings reinforce the quantitative results: contemporary MT and LLM-based systems exhibit stable translational preferences affecting compression, repetition, lexical concentration, and discourse organisation.

6 Discussion

The findings show that contemporary translation systems exhibit stable and measurable stylometric tendencies even when their outputs are broadly fluent and semantically adequate. Across both English–Arabic and Arabic–English directions, the systems differed in lexical dispersion, compression, repetition, sentence segmentation, and discourse organisation. These differences were captured through complementary measures, including entropy, MATTR, length ratios, and repetition statistics.

A central result is that machine translationese is not a uniform phenomenon across systems. Google Translate generally produced more expansive Arabic-target outputs with higher entropy and lexical dispersion, whereas Qwen3-8B showed stronger compression tendencies and lower lexical dispersion. GPT-4o often remained closer to the human translation condition across several measures, but it still displayed detectable system-specific regularities. This suggests that stylometric behaviour emerges from the interaction of translationese effects, model architecture, decoding behaviour, alignment objectives, and system-specific generation constraints.

These findings are especially relevant for instruction-tuned LLMs. Prior work suggests that alignment and preference optimisation can increase linguistic homogenisation and reduce distributional diversity in generated text (O’Mahony et al., 2024; Guo et al., 2025). The stable stylometric profiles observed here may therefore reflect not only translational simplification, but also broader regularisation effects associated with instruction following, deterministic decoding, and RLHF-style optimisation. In this sense, stylistic patterns in LLM translation should be interpreted as outcomes of constrained generation behaviour rather than as direct evidence of linguistic competence or deficiency.

The paragraph-level results further show that stylistic divergence becomes more visible in connected discourse. While sentence-level averages capture local lexical and structural tendencies, paragraph-level analysis reveals differences in discourse rhythm, segmentation, and information packaging. Some systems tended to smooth discourse by merging clauses into longer syntactic units, whereas others produced shorter and more segmented structures. This supports previous work on document-level and discourse-aware machine translation, which argues that sentence-level evaluation can obscure coherence and discourse phenomena (Maruf et al., 2021).

The similarity audit raises an additional methodological issue for GenAI translation research. GPT-4o showed substantially closer lexical-overlap and sequence-level similarity to the human translation condition than the other machine systems. However, such similarity should not be interpreted as direct evidence of human equivalence. Surface-level measures such as Jaccard similarity and WER capture token and sequence correspondence, but they do not fully account for pragmatic, rhetorical, or discourse-level behaviour. Moreover, the human translation condition represents a specific controlled baseline rather than a single universal model of human translation style.

More broadly, the results support evaluation approaches that extend beyond adequacy and fluency alone. Current MT evaluation remains strongly oriented toward semantic correctness, yet translations that preserve meaning may still differ in redundancy, lexical concentration, discourse rhythm, and rhetorical packaging (Fuentes et al., 2022; Freitag et al., 2022; Huang and Liu, 2024). These dif-

ferences are particularly important in news translation, where style contributes to readability, perceived neutrality, and journalistic register.

Overall, the study shows that stylometric analysis can complement conventional MT evaluation by identifying forms of translational regularisation, stylistic homogenisation, and discourse restructuring that may remain invisible to adequacy-based metrics alone.

7 Limitations

Several limitations should be considered when interpreting the findings.

First, the study examines stylometric regularities rather than style in a fully comprehensive linguistic or literary sense. The quantitative measures used here primarily capture surface- and distribution-level properties such as lexical dispersion, repetition, segmentation, and compression. While these measures are informative for identifying translational tendencies, they do not directly model deeper pragmatic, rhetorical, or discourse-functional dimensions of style (Lagutina et al., 2019).

Second, the dataset remains relatively modest for stylometric inference, consisting of 250 source items and 1,000 translated outputs distributed across translation directions, text types, and systems. Although the repeated-measures design improves comparability across conditions, larger corpora would provide more robust estimates of stylistic variation and improve generalisability across domains.

Third, the study focuses exclusively on contemporary news discourse. News translation provides a useful controlled setting because it combines relatively stable communicative goals with meaningful stylistic variation, but the findings may not generalise directly to literary, conversational, technical, or social-media translation.

Fourth, the LLM-based outputs were generated using deterministic decoding (temperature 0) in order to maximise reproducibility and reduce stochastic variation across aligned items. While this setting is methodologically useful for controlled comparison, it likely reduces stylistic variability and may amplify regularisation effects associated with constrained decoding.

Finally, the study compares only a limited number of translation systems. The findings therefore should not be interpreted as representative of all

neural MT systems or all LLM-based translation models.

8 Conclusion

This paper presented a controlled stylometric comparison of human, neural MT, and LLM-based translation in English–Arabic news translation. Using a balanced bidirectional dataset and a repeated-measures design, the study examined stylistic variation across sentence- and paragraph-level translation through quantitative stylometric analysis and qualitative close reading.

The findings show that contemporary translation systems exhibit stable, system-specific stylometric tendencies even when outputs remain broadly fluent and semantically adequate. Differences emerged in lexical dispersion, compression, repetition, sentence segmentation, and discourse organisation across both translation directions. Paragraph-level analysis further revealed discourse restructuring and stylistic smoothing that are less visible in sentence-level evaluation alone.

The similarity audit highlights the importance of methodological transparency in GenAI translation research. GPT-4o showed substantially closer lexical and sequence-level similarity to the human translation condition than the other machine systems, but such overlap should not be interpreted as evidence of human equivalence. Instead, it reinforces the need to document translation workflows, decoding settings, and comparison procedures when analysing stylistic behaviour in LLM-based systems.

More broadly, the study suggests that stylometric analysis can complement adequacy-oriented evaluation by capturing translational regularisation, discourse restructuring, and stylistic homogenisation in contemporary MT and LLM outputs.

References

- [Abdelaal and Alazzawie2020] Abdelaal, Noureldin Mohamed and Abdulkhaliq Alazzawie. 2020. Machine translation: The case of arabic-english translation of news texts. *Theory and Practice in Language Studies*, 10(4):408–418.
- [Baker1993] Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Baker, Mona, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In Honour of John Sinclair*, pages 233–250. John Benjamins, Amsterdam and Philadelphia.

- [Covington and McFall2010] Covington, Michael A and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.
- [Dickins et al.2016] Dickins, James, Sándor Hervey, and Ian Higgins. 2016. *Thinking Arabic Translation: A Course in Translation Method: Arabic to English*. Routledge, London and New York, 2 edition.
- [Enfes et al.2026] Enfes, Ebtihal, Nooredeen Awwad, and Kolawole John Adebayo. 2026. Cross-dialectal transfer learning for offensive language detection in arabic. *IEEE Access*, 14:48182–48197.
- [Freitag et al.2022] Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.
- [Fuentes et al.2022] Fuentes, Belén Saldías, George Foster, Markus Freitag, and Qijun Tan. 2022. Toward more effective human evaluation for machine translation. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 76–89.
- [Guo et al.2025] Guo, Yanzhu, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526.
- [Hatim and Mason1990] Hatim, Basil and Ian Mason. 1990. *Discourse and the Translator*. Longman, London and New York.
- [He et al.2024] He, Zhiwei, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.
- [Hendy et al.2023] Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.
- [House2015] House, Juliane. 2015. *Translation Quality Assessment: Past and Present*. Routledge, London and New York.
- [Huang and Liu2024] Huang, Yan and Wei Liu. 2024. Evaluating the translation performance of large language models based on eas-20.
- [Kocmi et al.2024] Kocmi, Tom, Eleftherios Avramidis, and Rachel Bawden. 2024. Findings of the WMT24 general machine translation task. In *Proceedings of the Ninth Conference on Machine Translation*. Association for Computational Linguistics.
- [Kong and Macken2025] Kong, Delu and Lieve Macken. 2025. Decoding machine translationese in English-Chinese news: LLMs vs. NMTs. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 99–112, Geneva, Switzerland. European Association for Machine Translation.
- [Lagutina et al.2019] Lagutina, Ksenia, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and PG Demidov. 2019. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195. IEEE.
- [Liu and Afzaal2021] Liu, Kanglong and Muhammad Afzaal. 2021. Syntactic complexity in translated and non-translated texts: A corpus-based study of simplification. *PLOS ONE*, 16(6):e0253454.
- [Maruf et al.2021] Maruf, Sameen, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys (CSUR)*, 54(2):1–36.
- [O’Mahony et al.2024] O’Mahony, Laura, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. 2024. Attributing mode collapse in the fine-tuning of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, volume 2, page 2.
- [Papineni et al.2002] Papineni, Kishore, 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.
- [Qian et al.2024] Qian, Shenbin, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Fred Blain. 2024. What do large language models need for machine translation evaluation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3674, Miami, Florida, USA. Association for Computational Linguistics.
- [Rabinovich et al.2017] Rabinovich, Ella, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540.
- [Rei et al.2020] Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural

- framework for mt evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, pages 2685–2702.
- [Ryding2005] Ryding, Karin C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press, Cambridge.
- [Song et al.2025] Song, Yixiao, Parker Riley, Daniel Deutsch, and Markus Freitag. 2025. Enhancing human evaluation in machine translation with comparative judgement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20536–20551, Vienna, Austria. Association for Computational Linguistics.
- [Toral et al.2018] Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the third conference on machine translation: Research papers*, pages 113–123.
- [van Dijk1988] van Dijk, Teun A. 1988. *News as Discourse*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- [Vanmassenhove et al.2019] Vanmassenhove, Eva, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.
- [Vaswani et al.2017] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- [Volansky et al.2015] Volansky, Vered, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.
- [Yao and Fan2025] Yao, Guangyuan and Lingxi Fan. 2025. An entropy-based study of simplification in chatgpt translations compared to neural machine translation and human translation across genres. *PLOS ONE*, 20(12):e0339762.
- [Zhang et al.2023] Zhang, Biao, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In Krause, Andreas, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

A Appendix

A.1 LLM Translation Prompt

LLM translation prompt

Translate the following news text into {target_language}.

Requirements: preserve meaning accurately; preserve the original tone and register as much as possible; keep named entities, dates, and numbers correct; keep paragraph breaks if present; do not add explanations; do not summarize; do not omit content; output only the translation.

Text: {source_text}

Table 7: Prompt used for GPT-4o and Qwen3-8B translation conditions.

A.2 Representative Qualitative Examples

Pattern	Example and interpretation
Simplification / compression	Source: كما أنه لا يزال عضواً في اللجنة المركزية لحركة فتح. Human: He is also still a member of the Central Committee of the Fatah Movement. Qwen3-8B: He is still a member of the Central Committee of Fatah. Interpretation: Qwen3-8B preserves the main meaning but reduces lexical specificity by omitting “also” and shortening “the Fatah Movement” to “Fatah”.
Redundancy / formulaic repetition	Source: القيادة المركزية الأمريكية: لم تتمكن أي سفينة من اختراق الحصار الأمريكي المفروض على مضيق هرمز منذ دخوله حيز التنفيذ يوم الاثنين. Human: The US military said that no vessel has managed to break through the American blockade imposed on the Strait of Hormuz since it came into effect on Monday. Google: US Central Command: No ship has been able to breach the blockade. The US military announced that no ship has been able to breach the US blockade imposed on the Strait of Hormuz since it went into effect on Monday. Interpretation: Google produces a fluent but repetitive rendering, repeating both the reporting frame and the phrase “no ship has been able to breach”.
Paragraph restructuring	Source: يبلغ ارتفاع بنت جبيل عن سطح البحر ٧٧٠ متراً، ما يجعل منها موقعاً مشرفاً على نقاط عديدة في الجنوب وخاصة في القطاع الأوسط، وهذا ما قد يفسر الإصرار الإسرائيلي للسيطرة عليها. Human: The height of Bint Jbeil is 770 meters above sea level, which makes it a commanding location over many points in the south, especially in the central sector. This may explain the Israeli insistence on controlling it. Google: Bint Jbeil is 770 meters above sea level, which makes it a location overlooking many points in the south, especially in the central sector, and this may explain the Israeli insistence on controlling it. Interpretation: Google preserves the factual content but merges two human sentences into one smoother syntactic unit.

Table 8: Representative qualitative examples illustrating simplification, redundancy, and paragraph restructuring.

Emotion Profiling in LLM-Based Literary Translation: Systematic Shifts Across MT and Post-Editing

Antonio Castaldo^{1,2}, Johanna Monti¹, Sheila Castilho³

¹University of Naples “L’Orientale” ²University of Pisa ³Dublin City University

antonio.castaldo@phd.unipi.it, jmonti@unior.it,

sheila.castilho@adaptcentre.ie

Abstract

This paper investigates whether LLM translations exhibit identifiable emotional profiles and how post-editing reshapes them toward human-like norms. We compare LLM translations of Margaret Atwood’s *Oryx and Crake* with their post-edited versions and a human translation, using a large-scale corpus of contemporary Italian science-fiction as a baseline. We examine emotion through lexicon-based and multilingual modeling, conducting a fine-grained analysis of emotional variation across systems. We find that MT systems introduce model-specific and statistically significant emotional fingerprints across translations, leading to a limited preservation of an author’s voice.

1 Introduction

Large Language Models (LLMs) have quickly become the predominant approach to machine translation (MT), particularly in literary translation, where the scope extends beyond mere transfer of meaning, encompassing the preservation of an author’s voice, style and their aesthetic intent (Zhang et al., 2025).

Recent research in the field of LLMs has demonstrated significant improvements in their ability to handle figurative language (Castaldo, 2024; Raunak et al., 2023) and context-related issues (Castilho et al., 2023). These advancements make them especially promising for literary translation, where such features are fundamental. However, despite these improvements, the extent to which they may be able to preserve an author’s voice remains underexplored. This gap becomes even

more relevant when considering post-editing (PE). While it is generally assumed that PE brings MT outputs closer to HT in terms of quality, it remains unclear to what extent PE preserves the author’s voice present in the source text, or whether it is influenced by stylistic interference introduced by the MT system.

In this work, we study how authorial voice is rendered through the lens of emotional profiling, investigating whether different MT paradigms introduce distinct emotional patterns and how these are transformed during post-editing. The author’s emotional voice, therefore, refers to the way the author intended emotions to be expressed in the original text. We conduct a comparative study on a set of translations of the literary novel “*Oryx and Crake*” by Canadian author Margaret Atwood (Atwood, 2004), from English to Italian. We analyze outputs from three LLM systems, alongside their post-edited versions and a published human translation (HT).

To complement this analysis and evaluate the robustness of our findings, we use a large-scale corpus of contemporary Italian science fiction from Wattpad. From this corpus, we sample 1M tokens and generate parallel translations using an LLM (TowerPlus-2B) and an NMT system (NLLB-200-600M). This large-scale setup allows us to move beyond individual examples and test for statistically significant patterns in emotional variation. Using lexicon-based and multilingual modeling approaches, we examine emotion shifts between source texts and translations, i.e. the differences in how emotionally charged language is rendered, finding that both systems introduce systematic and statistically significant changes in emotional properties.

Our findings highlight that MT systems exhibit distinctive emotional fingerprints, reflected in systematic differences in how emotional content is distributed, intensified, or attenuated through-

© 2026 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

out the text. The effectiveness of post-editing in restoring the author’s voice appears to be strongly constrained by the quality of the underlying MT model. We further find these emotional fingerprints to be statistically significant and particularly concentrated at emotionally salient peaks, where the author’s emotional voice would ideally be most faithfully preserved.

2 Related Work

In the field of translation studies, a large amount of scientific literature has focused on whether translated texts exhibit features that are statistically different from naturally occurring human writing (Koppel and Ordan, 2011; Volansky et al., 2015; Mauranen and Kujamäki, 2004). The research approach based on corpus linguistics is grounded in the pioneering work done by Baker (1993), who introduced the notion of “translation universals”. According to the author, translation universals are hypotheses of linguistic features common to translated texts regardless of the source and target languages, a phenomenon often referred to as *translationese*. These findings laid the methodological foundation for stylometric analysis of translated texts, leading to the hypothesis that if human translators leave identifiable fingerprints, MT systems should too potentially in more systematic ways.

To this end, subsequent studies have confirmed the hypothesis, finding that MT systems display traces of translationese (Bizzoni et al., 2020). In relation to NMT, Vanmassenhove et al. (2019) specifically documented a loss of lexical diversity in NMT outputs across several language pairs, potentially representing a universal fingerprint. This was confirmed by a later study, where the authors identified the tendency of NMT systems to exacerbate frequent patterns (Vanmassenhove et al., 2021). Despite the growing body of literature in the field, the stylistic effects of post-editing on LLM outputs are less understood.

Studies on post-editing have typically focused on NMT systems where the concept of *post-edited* has been introduced as a distinct register that characterizes post-edited translations, and differs from both MT outputs and human translations. Post-editing has been found to introduce intermediate stylometric profiles and stylistic normalisation (Toral, 2019). More recent work, grounded in Baker’s theory, revealed evidence of post-edited in NMT outputs of literary texts, finding that post-

edited translations come closer to the MT outputs than to human translations in terms of stylistic features (Castilho and Resende, 2022).

Most studies on translationese and post-editing have focused on surface-level features, such as lexical diversity and frequency distributions. While these metrics provide valuable insights, in the context of literary translation, they provide only a partial view of style. In literary texts, an author’s voice is not only reflected in structural and lexical choices, but also in the distribution and intensity of emotions conveyed throughout the narrative. In this regard, researchers have developed resources to identify and analyze emotions in text (Mohammad and Turney, 2010; Passaro et al., 2017), in both monolingual and multilingual settings. Lexicon-based methods have been demonstrated to produce accuracies comparable to machine learning and transformer-based methods, with the additional benefit of transparency and interpretability (Teodorescu and Mohammad, 2023; Öhman, 2021).

In this work, we make use of the available resources to conduct an emotion-based analysis of MT outputs. Specifically, we model the emotional profile of texts and investigate how it is transformed through the translation process. We extend this analysis to post-edited outputs, exploring whether post-editing mitigates or reinforces these emotional shifts, and whether post-edited exhibits distinctive emotional fingerprints.

3 Experimental Setup

We conduct two main experiments: one is a controlled study focused on a literary excerpt from Margaret Atwood’s *Oryx and Crake*, of which we possess a human professional translation, three translations generated by different LLMs and three post-edited MT outputs, as described in Section 3.1. On this corpus, we rely on the Italian Emotive Lexicon (Passaro et al., 2017) to compare the emotional density and intensity across the translation versions. Through this experiment, we aim to investigate whether MT systems introduce model-specific behavior in regards to the preservation, amplification or reduction of emotional voice in the text, as well as how post-editing affects these aspects.

The second study is a large-scale analysis on a corpus of naturally occurring literary texts, that we scrape from a fan-fiction platform called Wattpad.

From Wattpad, we scrape one thousand stories consisting of 25M tokens in total. The stories were extracted from the relevant science-fiction category and have all been published in the last five years. For this second experiment, we adopt a more scalable approach using the NRC Emotion Lexicon (Mohammad and Turney, 2010) in its multilingual version.

3.1 Data Collection

The translations analyzed in this study were obtained from a previous study, in which 121 segments of Margaret Atwood’s *Oryx and Crake* were translated using different MT systems and subsequently post-edited by professional literary translators (Castaldo et al., 2025). We selected *Oryx and Crake* because the novel contains stylistically rich passages, figurative language, and neologisms, making it particularly suitable for studying emotional variation. Such characteristics allow for different renderings and creative deviations across translation versions. The diverse composition of the corpus, with professional translations and post-edited versions of different state-of-the-art MT systems, makes it an ideal resource for our study. We refer to that work for full details on the translation and post-editing pipeline.

To empirically validate our emotion analysis, we compiled a large-scale reference corpus of contemporary Italian science-fiction. The corpus was constructed by scraping Wattpad, a major user-generated fiction platform, using the Python library Scrapy¹. We collected every available chapter from 1,000 stories belonging to the science-fiction genre, as identified by the platform’s category system. Language identification was performed using langdetect², retaining only texts classified as **Italian**, resulting in a corpus of approximately 25 million tokens.

The Wattpad corpus allows us to assess whether the emotional profiles observed in MT, LLM, PE, and HT outputs reflect systematic tendencies of the models. By comparing these outputs against a broad sample of naturally occurring Italian science-fiction prose, we can more reliably identify which emotional shifts are consistently introduced, amplified, or suppressed by the different translation processes.

¹<https://scrapy.org>

²<https://pypi.org/project/langdetect/>

3.2 Lexical Resources

As mentioned in the previous section, two main lexical resources are used in this study: the Italian Emotive Lexicon (Passaro et al., 2017) and the NRC Emotion Lexicon (Mohammad and Turney, 2010). Both resources adopt the same eight basic emotions, modeled after Plutchik (Plutchik, 1985), but diverge in the way they are composed and in the data they expose.

Italian Emotive Lexicon. The Italian Emotive Lexicon was built from data collected from native Italian speakers via an online survey, where participants were asked to produce five words associated with each emotion. This list of seed words was then used to populate the lexicon. This lexicon provides a list of words connected to each of the eight emotions, with part-of-speech tagging and frequency data. Interestingly, it also exposes emotion centroids and the cosine similarity of each lexical item to the centroid.

NRC Emotion Lexicon. The NRC Emotion Lexicon, on the other hand, was built for English through crowdsourcing where annotators labeled predefined words with emotions. The NRC Emotion Lexicon exposes a list of words with binary associations to each emotion. In addition to emotions, each word is also labeled for polarity (positive, negative). In the following versions, it has also been adapted to multilingual tasks, supporting over 108 languages, including Italian.

4 Results

In this section, we investigate whether different translation versions (HT, LLM, and PE) exhibit distinct emotional profiles, and to what extent these profiles converge toward HT. We focus on two main dimensions: the amount of emotion-related tokens present in the text compared to content words, and how strongly associated are the words with the emotion categories in the lexicon.

To this end, we first examine the number of tokens matched in the emotion lexicon relative to the total number of words. This provides an indication of how frequently emotion lexical items are used.

Table 1 provides an overview of the raw counts of emotion-related tokens across systems. If MT systems preserve the emotional profile of the source text, we would expect their values to be aligned with the human translation. On the op-

MT			PE		
System	Match	Words	System	Match	Words
GPT-3.5	718	2175	GPT-3.5	745	2279
GPT-4	717	2190	GPT-4	728	2232
Mistral	636	2026	Mistral	731	2253
HT: 738 / 2218					

Table 1: Emotion-related matched tokens from the lexicon and word counts, per translation version.

posite, systematic deviations may indicate model-specific tendencies.

As shown in Table 1, MT outputs differ in their degree of lexical alignment with the emotion lexicon. In particular, some systems produce fewer matched tokens, suggesting a tendency to under-represent emotion-related vocabulary, while others approach the levels observed in HT. Post-edited versions generally increase the number of matched tokens, indicating that human intervention tends to reintroduce or reinforce emotional lexical choices in line with HT. The lack of preservation of emotional lexicon could be attributable to accuracy errors in the translation or hallucinated outputs. We verify the alignment between the different MT versions, using translation quality metrics³ and the HT as reference. We provide quality metrics results in Table 2, divided per system and translation version.

System	BLEU	chrF	COMET
MT			
GPT-4	28.63	53.55	0.79
GPT-3.5	27.11	54.20	0.82
Mistral	15.86	43.24	0.77
PE			
GPT-4	25.70	52.63	0.78
GPT-3.5	25.29	53.05	0.78
Mistral	21.24	49.02	0.79

Table 2: Automatic evaluation scores across MT outputs and their post-edited (PE) versions.

We observe from Table 2 that overall MT systems achieve comparable quality across all metrics, and even the systems receiving high scores exhibit deviations in terms of emotion lexicon from HT. The only observable exception is Mistral which is the worst performing system and also displays most divergence in terms of matched tokens,

³The COMET model we use throughout our analysis is wmt-22-comet-da (Rei et al., 2022).

as displayed in Table 1.

4.1 Emotion Analysis

To provide a more fine-grained investigation of how emotion is distributed and how strongly it is expressed, we complement our analysis with calculations on emotional density and intensity. While the two metrics are certainly related, they measure different phenomena. *Density* captures the proportion of emotion-lexicon matches among content words, whereas *intensity* captures the average strength of those matches based on the cosine scores provided by the lexicon. Therefore, a sentence may have few words strongly associated with emotions, resulting in high intensity and low density, or many weakly associated words resulting in the opposite direction.

4.2 Emotional Density

We calculate *emotional density* as the proportion of content words in a sentence with matched tokens from the lexicon. This measure serves as a proxy for how much emotion-related vocabulary is present in the sentence, independently of the specific emotion category involved. Emotional density shows HT to have the highest density, with the PE output from GPT-4 approaching it. While the PE outputs from Mistral and GPT-3.5 also show an increase in the number of matched tokens, their density increase is diluted by the higher number of content words. By aggregating the results per group (HT, MT, PE), PE improves the density of MT, making it approach HT. The post-edited version of GPT-4 is the translation that most approaches HT in terms of content words, emotion-related tokens and density. This may be attributed to the high performance of the LLM which in previous studies has proved to approach HT in terms of both quality (Hendy et al., 2023) and creativity (Castaldo et al., 2025).

4.3 Emotional Intensity

Emotional intensity measures how strongly the lexical choices in a sentence are associated with emotional content overall. To compute this score, first we match each lemmatized content word in the sentence to the lexicon, and retrieve its cosine similarity scores for the eight basic emotions. Then, we average these scores across all matched words for each emotion, obtaining sentence intensity as the mean of the resulting eight emotion

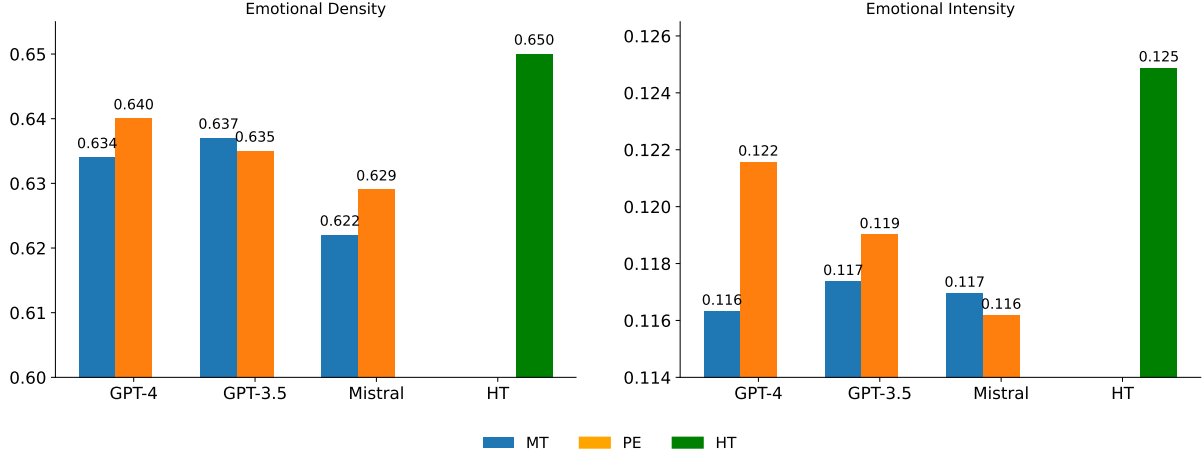


Figure 1: Emotional density and intensity across MT systems, their post-edited (PE) versions, and human translation (HT).

scores. We show the results of both metrics side-by-side in Figure 1 to allow easier comparison.

The results of emotional intensity agree with emotional density in terms of overall tendency. PE seems to increase the emotional density of the corpora, making the translations approach HT. Interestingly, we find a drastic increase in intensity after GPT-4 has been post-edited, suggesting that given the higher quality of the model’s outputs it may be more easily adapted by the translators and lead to better overall preservation of the original author’s voice. While the other models display only a marginal increase in intensity, the only exception is Mistral which exhibits a contraction in terms of intensity. Given that Mistral is the smallest and worst performing model according to quality metrics, the finding is consistent with our preliminary assumption that model performance affects the ability of the translators to adapt its content to maximize the preservation of emotions, as per the source.

4.4 Emotion Preservation

By computing sentence-level emotional intensity over the corpora, we are able to trace an intensity trajectory across the text. To do so, we apply a 5-line rolling mean, then aggregate the resulting values per translation version. The per-sentence standard deviation, shown at the bottom of the figure, captures the degree of disagreement between versions at each point in the text. We present the resulting trajectories of HT, PE and MT in Figure 2, where the horizontal axis represents the line index of the literary excerpt and the vertical axis the emotional intensity, allowing us to directly compare where the versions track closely and where

they diverge the most.

Interestingly, we find that the majority of the disagreements is found in high-intensity peaks, which leads us to investigate how each model behaves in this regard. Therefore, to further investigate the relationship between model quality and emotional fidelity, we identify the twenty sentences with the highest emotional intensity in HT and measure how often each version preserves them. We define a peak as preserved when the translation intensity falls within one standard deviation of HT. Table 3 reports peak preservation rates across MT and PE systems out of the twenty identified peaks.

Across MT outputs, we find preservation rates to be low, ranging from 38% to 46%, with no substantial difference between models. Post-editing consistently improves preservation, but the effect appears highly dependent on the model and does not mirror the trends observed for overall intensity. The post-edited GPT-4 achieves a modest gain over its MT output, but is outperformed by GPT-3.5 which reaches 64%. We argue that this behavior could reflect a tendency in translation behavior induced by the MT quality. GPT-4’s higher fluency and coherence may lead translators to introduce several small local adjustments across the text, that increase the overall intensity of the translation, while not directly addressing high-intensity peaks. This would be consistent with Castaldo et al. (2025) that showed in their experiments how GPT-4 received more edits in post-editing, that required much fewer editing time. By contrast, GPT-3.5 may lead the translators to introduce more substantial revisions, recovering most of the emotional

peaks found in HT.

A further noteworthy result comes from the post-edited version of Mistral, which not only fails to improve over its MT output but drops to 31%, the lowest rate across all systems.

This finding reinforces the earlier assumption that low-quality outputs may hinder translators’ ability to recover emotional intensity, as their attention must first be directed toward resolving issues of accuracy and fluency before they can effectively address stylistic aspects.

Table 3: Preservation of emotional intensity peaks across MT and PE versions. A peak is considered preserved if the translation intensity falls within one standard deviation of HT.

System	Type	Preserved	Rate
GPT-3.5	MT	5/20	38%
GPT-4	MT	6/20	46%
Mistral	MT	6/20	46%
GPT-3.5	PE	8/20	62%
GPT-4	PE	7/20	54%
Mistral	PE	4/20	31%

5 Large-Scale Emotion Analysis

To assess whether the emotional patterns observed in Section 4.1 generalize beyond a single literary text, we conduct a large-scale validation experiment on a separate corpus of contemporary Italian science fiction drawn from Wattpad. We sample 1M tokens from this corpus and generate parallel translations from Italian into English using two systems: TowerPlus-2B (Alves et al., 2024), an LLM-based MT system, and NLLB-200-600M (NLLB Team et al., 2022), a neural MT system. Emotion detection is performed using the NRC multilingual lexicon (cf. §3.2) and then confronted with a valence-arousal-dominance (VAD) regression model, and a multilingual emotion classifier⁴ to ensure that the observed differences are not driven by the specific choice of resource. Emotion shift is calculated as the difference between the target and source emotion scores, normalized by text length.

Both systems introduce statistically significant shifts in emotion-related properties relative to the source ($p < 0.05$), with a consistent tendency to

overproduce emotion-related tokens. TowerPlus generally amplifies emotion more than NLLB, though the direction of the shift depends on the representational framework used. At the level of emotion categories, TowerPlus systematically overproduces positive emotions relative to NLLB, with statistically significant differences for trust ($p = 0.0013$), surprise ($p = 0.0032$), and anticipation ($p = 0.0436$). These results confirm that different MT paradigms do not preserve emotion-related aspects of the source in the same way: while the magnitude of the differences is modest, they are systematic enough to be reliably detected across a corpus of this scale. This is consistent with the model-specific emotional profiles identified in the Oryx and Crake analysis, and suggests that the tendency of LLM-based systems to consistently introduce shifts in emotional content is not an artifact of a single text, but a more general property of the system.

The large-scale data further allows us to examine whether model differences are distributed across the text or concentrated at specific peaks, mirroring the pattern observed in Section 4.3 where the majority of disagreement was found in high-intensity peaks. Our results show that TowerPlus and NLLB tend to converge on low-emotion passages while diverging substantially in high-emotion source segments.

6 Discussion

Taken together, the results reveal a set of consistent patterns. Overall, HT is the upper bound and PE consistently moves MT closer to HT, in terms of both emotional density and intensity, and in line with previous research on post-editing (Castilho and Resende, 2022). However, the degree to which PE recovers emotional properties is not uniform across models, and the results suggest that MT output quality affects the translator’s ability to intervene effectively on emotional content.

GPT-4 post-editing produces the largest increase in overall emotional intensity, with translators not only reintroducing more emotion-related vocabulary but selecting words more strongly associated with emotional content. We interpret this as evidence that higher-quality outputs are more easily adapted in terms of stylistic and emotional content, as translators may attend to more nuanced interventions rather than focusing on correcting fluency and accuracy errors.

⁴<https://huggingface.co/Lsthf/multilingual-emotion-classification>

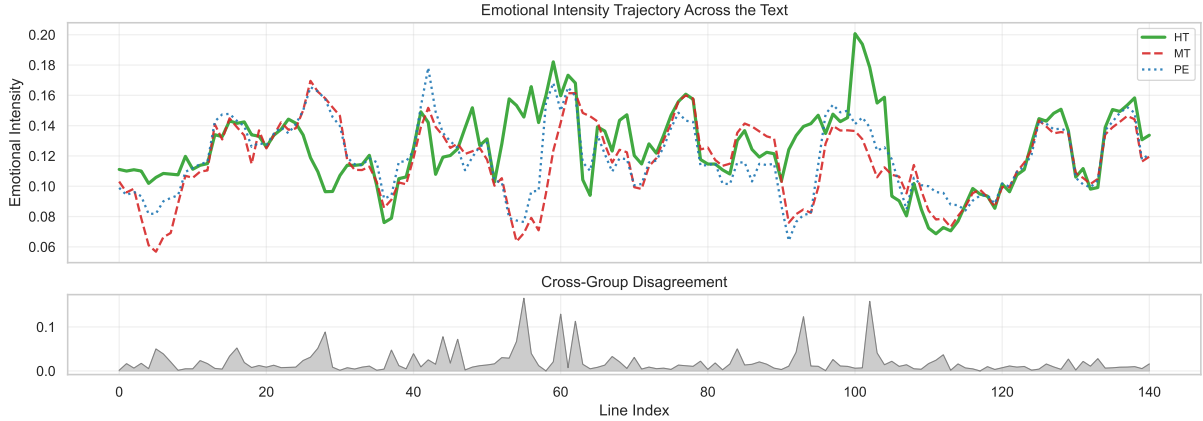


Figure 2: Emotional trajectory across MT systems, their post-edited (PE) versions, and human translation (HT). Trajectory is smoothed using a 5-line rolling mean. The bottom of the figure shows per-sentence standard deviation, indicating disagreements across versions.

While post-editing increases intensity and density across all models, this behavior does not always translate to better preservation of emotional peaks, as we see in the difference between the post-edited versions of GPT-4 and GPT-3.5 (cf. §4.4). GPT-4 is outperformed in peak preservation by GPT-3.5, and we argue that this divergence reflects a difference in how translators engage with the two outputs. The higher fluency of GPT-4 may encourage a more evenly distributed revision that raises the overall emotional density and intensity, without specifically addressing high-intensity peaks. GPT-3.5, on the other hand, may lead translators to introduce more substantial revisions, recovering emotional peaks in the process. This interpretation is consistent with Castaldo et al. (2025), who found that GPT-4 outputs attracted more numerous but less time-consuming edits, suggesting a pattern of small refinements across the translation, rather than substantial revisions.

This finding is consistent with the results obtained for Mistral. Although post-editing increases the model’s emotional density, it results in a contraction of emotional intensity, and a peak preservation rate equal to 31%, which is the lowest rate across all systems. This suggests that beyond a certain threshold of quality, translators may not be able to recover emotional content, while being confronted with more pervasive fluency and accuracy errors.

The large-scale analysis conducted on the Wattpad corpus situates these findings empirically, finding that the emotion shifts introduced by both TowerPlus and NLLB across 1M tokens are statistically significant and not specific to the document

taken into analysis. The results confirm that the emotional shifts introduced by model-specific behavior are a systematic property of MT systems, leading to a partial preservation of an author’s voice. Finally, our analysis confirms that model differences concentrate in high-emotion segments, suggesting that the emotional fingerprint of the model may be actually found in segments where an author’s voice may need to be more faithfully preserved.

7 Conclusion

In this study, we investigated how emotions are distributed and expressed across different translation versions, starting from a fine-grained analysis on a literary excerpt from Margaret Atwood’s *Oryx and Crake* and validating our findings with a large-scale analysis on a corpus of contemporary Italian science fiction drawn from Wattpad.

Our results consistently show that MT systems introduce model-specific emotional fingerprints that are detectable both at the level of individual translations and across large corpora. In the controlled study, we found that LLM and NMT outputs differ from HT in terms of emotional density and intensity, and that these differences are not uniformly distributed across the text but concentrate at emotionally salient peaks, where the preservation of an author’s voice is most critical.

Post-editing consistently moves MT outputs closer to HT, but its effectiveness depends on the quality of the underlying system. Higher-quality outputs appear to facilitate stylistic adaptation leading to a higher preservation of emotional content, whereas lower-quality outputs redi-

rect the translator’s attention toward the correction of errors, leaving less room for emotional recovery. The large-scale analysis confirms our findings through an independent study, showing that the tendency of MT models to introduce model-specific emotional fingerprints, especially in high-intensity peaks, may be a systematic property of these models.

Taken together, these findings have direct implications for literary translations, suggesting that MT systems have distinctive emotional profiles and post-editing alone may not be sufficient to fully recover the emotional patterns observed in HT, especially when the underlying MT quality is low.

8 Limitations

Although we complemented the analysis with both neural and lexicon-based methods, the approaches adopted in this study remain approximations of emotional expression and may not fully capture its complexity.

We also acknowledge that our analysis focuses exclusively on translation versions without directly modeling the source text itself. As a result, our investigation of authorial voice is necessarily constrained by how that voice is rendered in translation, rather than by a direct comparison with the emotional properties of the original text.

In relation to the large-scale experiment, we cannot fully exclude the possibility that some of the stories extracted from Wattpad had already undergone machine translation or machine-assisted editing prior to publication on the platform. Although this risk is difficult to eliminate at scale, it may introduce noise into the comparison between naturally occurring texts and MT outputs.

Finally, while our findings suggest that MT systems exhibit distinctive emotional fingerprints, we do not evaluate reader perception. The observed emotional shifts are computationally measurable, but it remains unclear to what extent they are perceptible to readers or whether they significantly affect literary reception. Future work could therefore integrate reception studies and human evaluations to assess emotional fidelity in literary translation.

Acknowledgments

This work has been funded by the Italian National PhD programme in Artificial Intelligence, partnered by University of Pisa and University

of Naples “L’Orientale”, through a doctoral grant (ID 39-411-24-DOT23A27WJ-6603) established by Ex DM 318, of type 4.1, co-financed by the National Recovery and Resilience Plan.

The third author benefits from being member of the ADAPT SFI Research Centre at Dublin City University, funded by the Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2.

References

- Alves, Duarte M., José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An Open Multilingual Large Language Model for Translation-Related Tasks, February. arXiv:2402.17733.
- Atwood, Margaret. 2004. *Oryx and Crake*. Number v.1 in The MaddAddam Trilogy Ser. Knopf Doubleday Publishing Group, New York.
- Baker, Mona. 1993. Corpus Linguistics and Translation Studies — Implications and Applications. In *Text and Technology*, page 233. John Benjamins, June.
- Bizzoni, Yuri, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How Human is Machine Translation? Comparing Human and Machine Translations of Text and Speech. In Federico, Marcello, Alex Waibel, Kevin Knight, Satoshi Nakamura, Hermann Ney, Jan Niehues, Sebastian Stüker, Dekai Wu, Joseph Mariani, and Francois Yvon, editors, *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online, July. Association for Computational Linguistics.
- Castaldo, Antonio, Sheila Castilho, Joss Moorkens, and Johanna Monti. 2025. Extending CREAMT: Leveraging large language models for literary translation post-editing. In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 506–515, Geneva, Switzerland, June. European Association for Machine Translation.
- Castaldo, Antonio. 2024. Prompting Large Language Models for Idiomatic Translation. In *CTT, WS*.
- Castilho, Sheila and Natália Resende. 2022. Post-Editese in Literary Translations. *Information*, 13(2), January.
- Castilho, Sheila, Clodagh Quinn Mallon, Rahel Meister, and Shengya Yue. 2023. Do online Machine Translation Systems Care for Context? What

- About a GPT Model? In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 393–417, Tampere, Finland, June. European Association for Machine Translation.
- Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. February. arXiv: 2302.09210.
- Koppel, Moshe and Noam Ordan. 2011. Translationese and Its Dialects. In Lin, Dekang, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA, June. Association for Computational Linguistics.
- Mauranen, Anna and Pekka Kujamäki. 2004. *Translation Universals: Do they exist?* John Benjamins Publishing, January. Google-Books-ID: 4bQ9AAAAQBAJ.
- Mohammad, Saif and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA, June. Association for Computational Linguistics.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.
- Passaro, Lucia, Alessandro Bondielli, and Alessandro Lenci. 2017. Learning affect with distributional semantic models. *Italian Journal of Computational Linguistics*, 3:23–36, 12.
- Plutchik, Robert. 1985. On emotion: The chicken-and-egg problem revisited. *Motivation and Emotion*, 9(2):197–200, June.
- Raunak, Vikas, Arul Menezes, Matt Post, and Hany Hassan Awadalla. 2023. Do GPTs Produce Less Literal Translations? May. arXiv: 2305.16806.
- Rei, Ricardo, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névél, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.
- Teodorescu, Daniela and Saif Mohammad. 2023. Evaluating Emotion Arcs Across Languages: Bridging the Global Divide in Sentiment Analysis. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4124–4137, Singapore, December. Association for Computational Linguistics.
- Toral, Antonio. 2019. Post-edited: an Exacerbated Translationese. In Forcada, Mikel, Andy Way, Barry Haddow, and Rico Sennrich, editors, *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland, August. European Association for Machine Translation.
- Vanmassenhove, Eva, Dimitar Shterionov, and Andy Way. 2019. Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In Forcada, Mikel, Andy Way, Barry Haddow, and Rico Sennrich, editors, *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland, August. European Association for Machine Translation.
- Vanmassenhove, Eva, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In Merlo, Paola, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online, April. Association for Computational Linguistics.
- Volansky, Vered, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118, April.
- Zhang, Ran, Wei Zhao, and Steffen Eger. 2025. How Good Are LLMs for Literary Translation, Really?

Literary Translation Evaluation with Humans and LLMs. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10961–10988, Albuquerque, New Mexico, April. Association for Computational Linguistics.

Öhman, Emily. 2021. The Validity of Lexicon-based Sentiment Analysis in Interdisciplinary Research. In Hämäläinen, Mika, Khalid Alnajjar, Niko Partanen, and Jack Rueter, editors, *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 7–12, NIT Silchar, India, December. NLP Association of India (NLPAI).

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 semi-otic 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 (Colombat, 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 pre-trained 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, socio-historical 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 translationsphere 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 (Tison, 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 Tison, 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 ex-

tracted 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 transformer². 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

¹Law 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,

³Our 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% ($\pm 2.49\%$).
- Support Vector Machine (SVM): Achieved a highly competitive Macro F1 score of 82.02% ($\pm 1.88\%$).
- K-Nearest Neighbors (KNN, $k=3$): Struggled significantly in the multiclass environment, yielding a Macro F1 score of only 51.57% ($\pm 2.29\%$).

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.

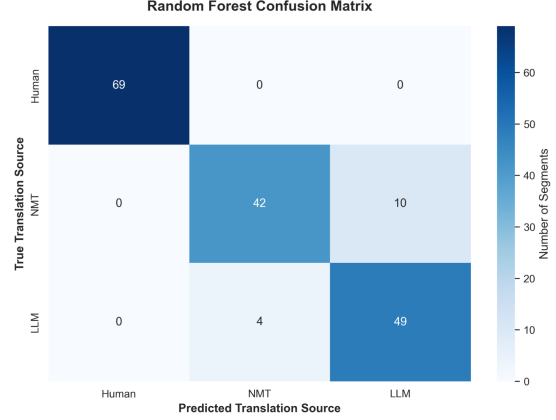


Figure 1: Confusion matrix for the Random Forest multiclass classifier.

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.

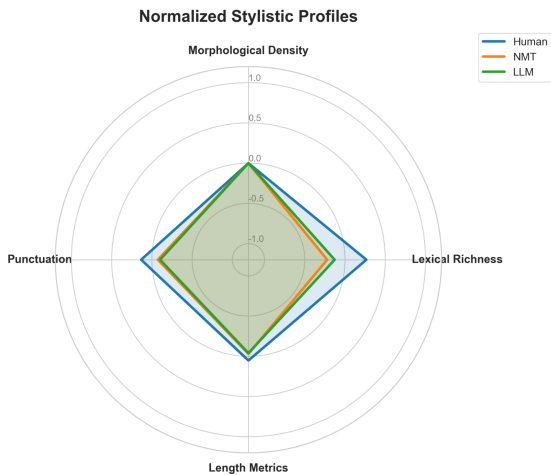


Figure 2: Radar chart mapping the normalized stylistic profiles of Human, NMT, and LLM translations across four macro-categories: Morphological Density, Lexical Richness, Length Metrics, and Punctuation.

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 $92.3\% \pm 4.7\%$ accuracy).
- **NMT Generation:** Neural machine translation required 22 chunks to reach the 85% robustness threshold (achieving $89.2\% \pm 3.1\%$ accuracy).
- **LLM Generation:** LLM-generated texts exhibited the highest stylistic variance, requiring 32 chunks to achieve the 85% robustness threshold (achieving $88.9\% \pm 3.8\%$ 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

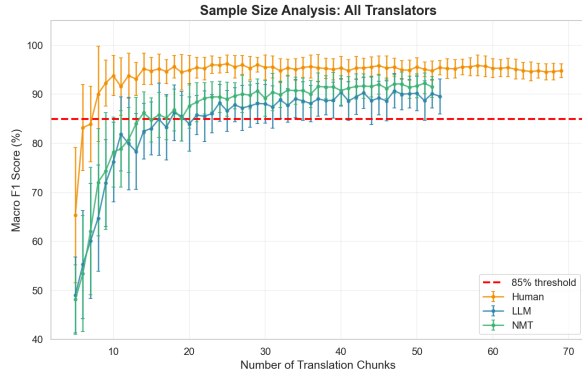


Figure 3: Combined learning curves comparing the stylistic consistency across Human, NMT, and LLM translation sources. The rapid stabilization of the Human curve highlights the dense nature of human stylistic markers compared to machine-generated variance.

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 intra-textual 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

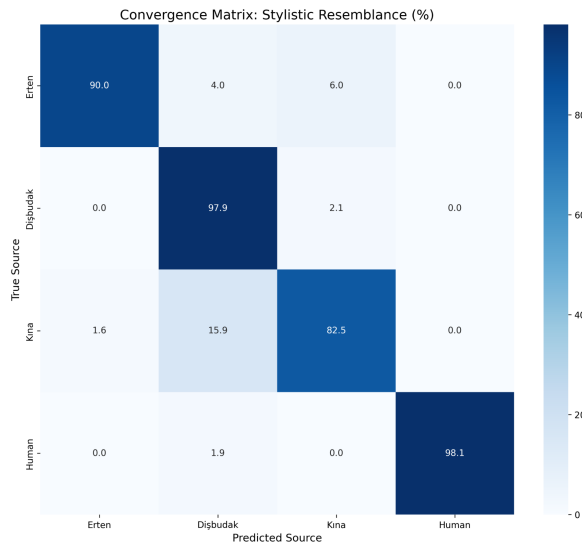


Figure 4: Stylistic convergence matrix detailing the percentage of resemblance among the contributing human translators and the Human-NMT model.

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 retractions. This underscores the value of original human translations, particularly retractions 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 CO₂e/kWh, the total energy consumption is estimated at 1.76 kWh, resulting in a carbon footprint of approximately 0.78 kg CO₂e. Emissions were calculated following the methodology of the ML CO₂ 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 XXI^e 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 détournement [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, İstanbul 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. Dalli, 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/ceviriri.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 translational, 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 translationsphere 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.v11.105>
- Tahir Gürçağlar, Ş. (2009). Retranslation. In M. Baker and G. Saldanha (eds.), *Routledge Encyclopedia of Translation Studies*, 233–236. Routledge.

- Tison, 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/ijsl1.v11i1.1>
- van Egdom, G.-W., C. Declercq, and O. Kusters (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. Dalli, 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

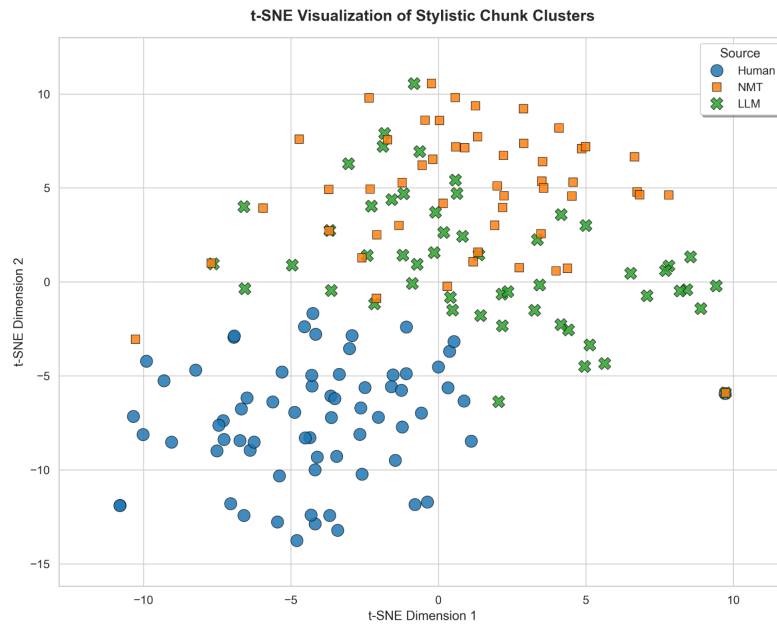


Figure A1: Two-dimensional t-SNE projection of the multidimensional stylistic feature space. Each point represents a 25-sentence text chunk. Human chunks form a tightly isolated grouping, while NMT and LLM chunks occupy adjacent and partially overlapping spaces.

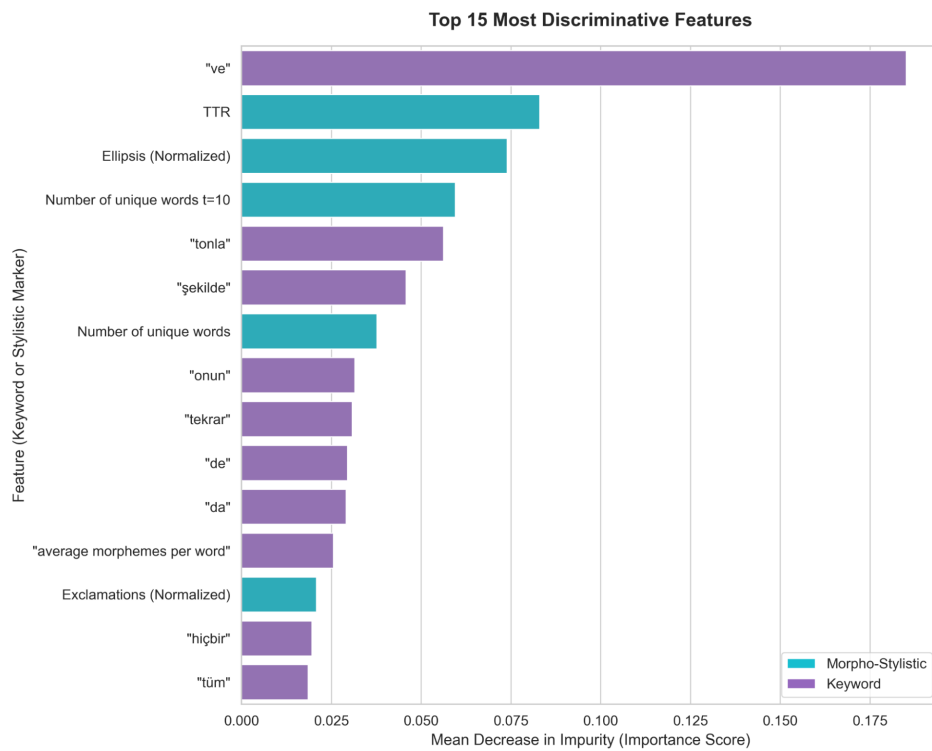


Figure A2: Top most discriminative features identified by the Random Forest classifier, ranked by Mean Decrease in Impurity. Keyword markers and morpho-stylistic features are color-coded to highlight the blend of structural and lexical identifiers used by the model.

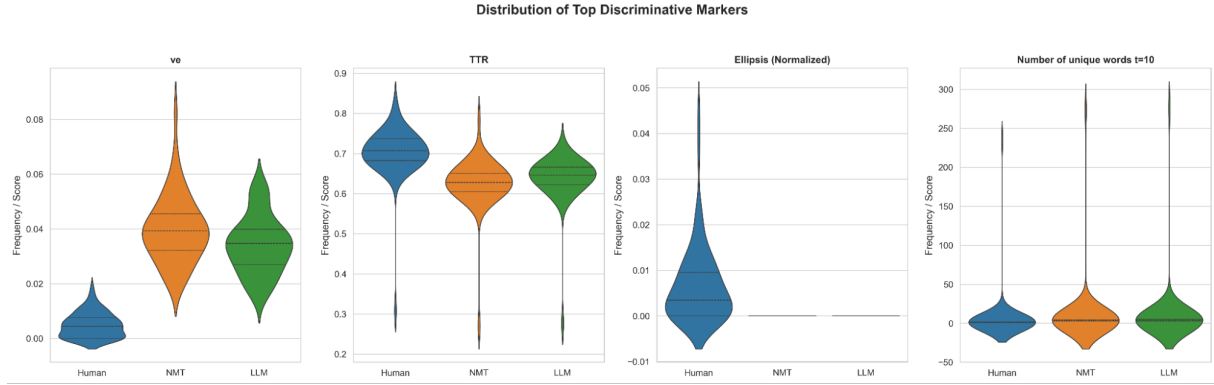


Figure A3: Violin plots illustrating the probability density and distribution variance of top discriminative markers across the three translation sources.

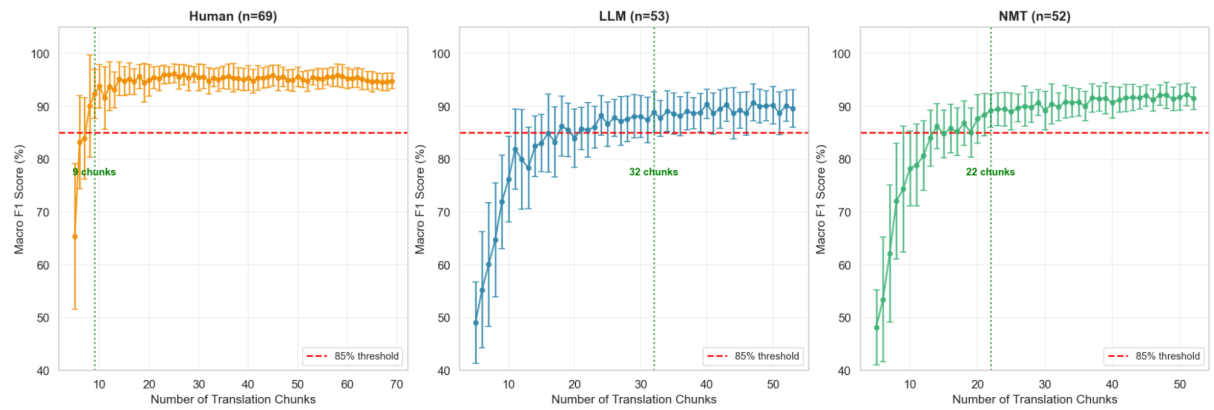


Figure A4: Sample size analysis depicting classification accuracy as a function of the number of text chunks for each translation category. The dashed lines represent the 85% robustness threshold.

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.

Style and Terminology in Commercial Flows: Mixing APE with Iterative Feedback

Marthe Lamote Ewoenam Tokpo Tom Vanallemeersch

Sara Szoc Koen Van Winckel

CrossLang

{firstname.lastname}@crosslang.com

Abstract

MT systems have demonstrated strong capabilities in natural language translation, especially in the era of multilingual large language models, which exhibit impressive text understanding and generation capabilities across multiple languages. However, meeting specific stylistic and lexical requirements of end users remains challenging. In this paper, we explore the effectiveness of an automated post-editing approach (APE) in a real-world MT production environment using LLMs and iterative human feedback. We further examine current limitations and potential areas for improvement. To this end, we propose an end-to-end methodology for APE and provide analytical and empirical insights into challenges of APE in commercial settings, primarily relying on client and expert feedback and analysis. We analyze the performance of LLM-based APE across five main stylistic dimensions, assess the terminology correction component of the pipeline, and demonstrate how iterative feedback, through prompt refinement and glossary enhancement, contributes to improving both stylistic and terminological consistency.

1 Introduction

Machine Translation (MT) systems have shown considerable adeptness in translating natural language text (Zhu et al., 2024b; Alves et al., 2024),

especially in the era of multilingual large language models (MLLMs), which are increasingly exhibiting impressive capabilities in understanding and generating texts in multiple languages. In spite of these advances, it remains challenging to generate translations that meet specific stylistic and lexical requirements of end users (Raunak et al., 2023). This limitation is even more prominent in out-of-the-box pretrained translation systems.

To address the above limitation, MT systems are often finetuned on user-specific data to adapt the model to stylistic and lexical preferences of users (Finkelstein et al., 2026). Finetuning a model, however, imposes significant resource demands that are often not available in run-time environments. Additionally, finetuning, if not properly carried out, could cause catastrophic forgetting of useful knowledge previously learned by the model (Zhai et al., 2024; Luo et al., 2023).

Another potential solution to this issue is in-context learning (ICL). This generally involves selecting a set of translation examples to demonstrate the task (Dong et al., 2024). These task exemplars are appended to the input prompt to form part of the context for text generation. Although ICL does not impose significant resource demands, it is sensitive to the choice of exemplars, so that poor exemplars could have a detrimental effect on the output quality (Zhu et al., 2024a). This makes ICL challenging to deploy in situations where a suitable translation memory is not available.

In practice, some type of post-editing often needs to be carried out to provide fine-grained stylistic and lexical adaptations to suit user preferences. The use of LLMs for automatic post-editing (APE) has recently been adopted in localization frameworks (Uguet et al., 2024).

In this paper, we explore the effectiveness of an

© 2026 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

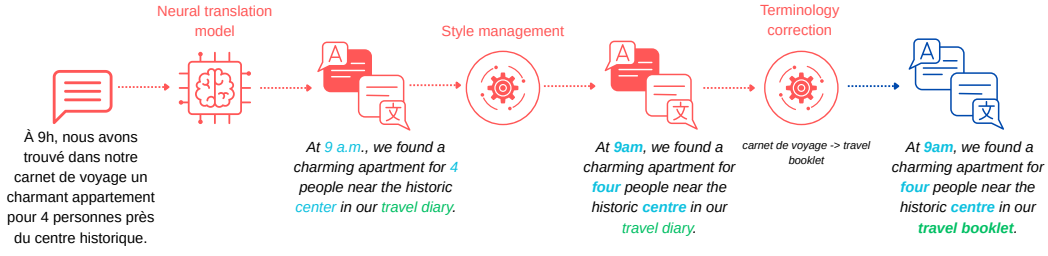


Figure 1: A depiction of our end-to-end APE pipeline with an example translation. Words that undergo style adaptations are shown in **blue**, whereas words that undergo terminology correction are shown in **green**. Correctly resolved words according to user preferences are in **bold**.

APE approach using LLMs and iterative human feedback in a real-world production setting, its current limitations, and possible improvements. To this effect, we (1) propose an end-to-end methodology for APE, showing how different phases of post-editing can be effectively automated using LLMs (Figure 1) and (2) provide some analytical and empirical insights on challenges of APE in commercial settings, primarily relying on client and expert feedback in the process. We used GPT-4.1 for all LLM implementations in this project. This choice is motivated by the fact that this model is a highly optimized general-purpose model designed for speed and responsiveness, which was an important consideration for this project.

2 Related works

Recent works have explored the use of LLMs for APE. Raunak et al. (2023) explore the use of GPT-4 to post-edit outputs from an NMT system. Their work indicates the capacity of GPT-4 to produce meaningful edits that align with human judgement. They report an improvement on WMT-22 in four language directions based on COMET scores. In spite of these score improvements, they note a tendency for the GPT-4 model to hallucinate by proposing edits in its chain of thought that it does not incorporate into the final post-edited translation. Uguet et al. (2024) also show in their work that LLMs demonstrate great potential in APE but are not fully mature in their current state for this purpose; instead of improving the raw MT translations, they consistently deteriorated the unedited “raw” translation outputs in evaluation metrics such as BLEU (Papineni et al., 2002).

Ki et al. (2024) investigate LLaMA-2 7B for APE. They show reliable improvements for results

in Chinese-English, English-German and English-Russian. However, they show that the LLM struggles to incorporate fine-grained instructions; rather, it focuses on broad and generic instructions for post-editing. Chen et al. (2024) use an iterative approach to refine translations. An LLM, over multiple rounds, refines the translation to improve the naturalness of the translation. Their approach is based on the observation that humans typically modify translations repeatedly to improve them. They report that this approach improves fluency and naturalness, especially for high-resource languages.

The above works primarily explore the use of APE with standard benchmark datasets, which often do not reflect and reveal frequently encountered real-world challenges. Hence, our aim is to investigate the effectiveness of APE with LLMs in an MT production environment, particularly in commercial settings. Nunziatini and Diego (2024) explore post-editing translations in a production environment but limit their scope to gender-inclusive post-edits. They show GPT-4’s capability of generating post-edits that mitigate gender stereotypes, but point out the tendency of the model to make unnecessary and inconsistent changes. Closest to our line of work is Vidal et al. (2022), who give a brief presentation on the possibilities of incorporating LLMs in an APE workflow in a real-word production environment. Our aim is to take this further by exploring and proposing a more detailed APE workflow in production environments, focusing on stylistic and terminological adaptations. We also provide a fine-grained and in-depth analysis of APE based on human feedback and analysis.

3 APE Approach

In this section, we describe the end-to-end pipeline that makes up our automated approach to post-editing. Our APE pipeline consists of style management, terminology correction, and iterative human feedback.

3.1 Style Management

The style management component of our APE, as depicted in Figure 2, begins with the conversion of the client style guide containing the language preferences and requirements into concise instructions that can be integrated into our APE architecture. The rules of the style guide, initially contained in a PDF style guide, are extracted and converted to instructions using the LLM.

After this initial extraction stage, a manual validation process is conducted by our internal prompt engineer and linguist. This process is essential in ensuring that all necessary rules are properly captured and that the instructions are clear, concise, non-ambiguous, and understandable. The output is engineered into a prompt for the LLM using various prompt engineering strategies in a sandbox environment, where a sample of the client’s data is used for prompt testing. An example of the automatic conversion and manual validation of a style guide section can be found in the appendix.

3.2 Terminology Correction

The terminology correction step begins with the conversion of the glossary into a predefined format, which also allows for the addition of notes or metadata (such as part of speech) to guide the LLM on how to apply specific term translations, especially in instances where the usage of the term can be ambiguous.

The formatted glossary is used to identify segments in the source corpus containing relevant keywords. Segments selected through this glossary lookup are sent to the term correction component. For this step, the source sentence and the translation are added to the prompt with the identified glossary terms and, when available, the terms’ metadata. The prompt is set up to direct the term correction model to integrate the terms in the sentence without breaking already correct translations and to ensure final grammatical correctness.

3.3 Iterative human feedback

The APE pipeline supports flexible adaptation of both the prompts and glossary resources throughout the project lifecycle, guided by ongoing linguistic feedback. Recurring issues identified in the APE output are systematically analyzed by the linguist and prompt engineer, and the system configuration is adjusted accordingly. When errors originate from the style management step, prompt refinements can be applied to better steer the model’s behavior. Issues arising from the term correction step can be mitigated by enriching the glossary with additional metadata to more effectively guide the model’s application of terminology. This iterative feedback loop enables continuous improvement of the system and ensures sustained alignment with evolving client requirements.

4 Experimental setup

As mentioned above, the main objective of this work is to investigate the efficacy of an APE pipeline with LLMs as the core component in a *real-world industry setting*. To this end, we use an existing dataset from a large-scale French-to-English project carried out for a client in the travel industry. The original dataset consists of a total of almost 49,738 segments from 118 different input files, consisting of website content and travel descriptions.

To allow for a thorough and qualitative analysis of how client style rules are applied across different stylistic dimensions, we downsampled the dataset to a sample of 250 instances. This sample was extracted after randomly shuffling the dataset using Excel’s `rand()` function. Upon inspection, it showed that the extracted segments covered most of the input files, thus creating a representative sample of the full project. After a preprocessing step to remove extraneous instances consisting solely of non-translatable content, a sample of 242 lines remained.

The source texts are used as inputs to our proposed APE pipeline, saving the outputs from each intermediate step in the pipeline for qualitative analysis, as depicted in Figure 1.

The qualitative analysis was performed by the linguist assigned to the project, who has extensive knowledge of the client style guide and glossary. In addition, we conducted a quantitative evaluation using five established MT metrics: BLEU, TER (Snover et al., 2006), ChrF (Popović,

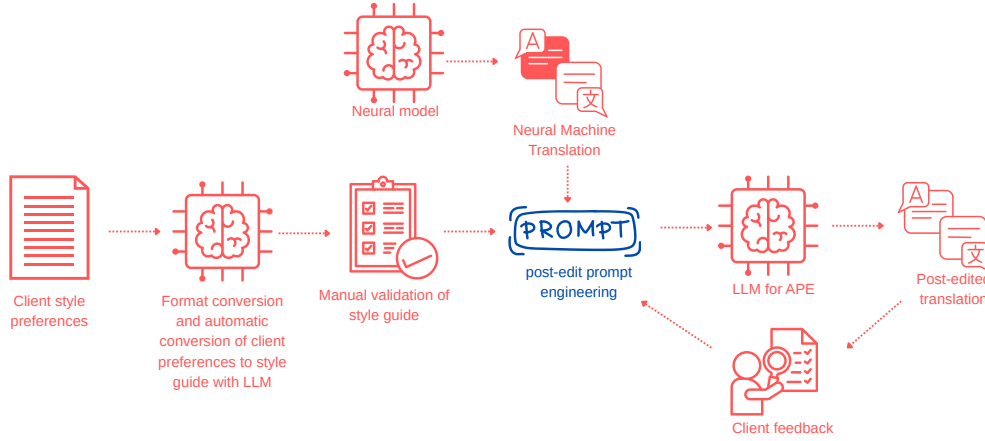


Figure 2: An illustration of the style management pipeline, showing the connections between the various steps.

2015), COMET (Webber et al., 2020), and BertScore (Zhang et al., 2019).

We use *gpt-4.1-2025-04-14* for all LLM implementations in this work, with a temperature value of 0.2 and a maximum output length of 200 tokens. All other configurations are set to OpenAI’s default values. We use *gpt-4.1-2025-04-14* for the initial “raw” translations as well.

5 APE Analysis

We analyze the quality of APE outputs generated by GPT-4.1 across five main stylistic dimensions. These dimensions, discussed below, are derived from the client style guide and serve as the framework for our analysis. Statistics of the APE outputs relating to these dimensions are provided in Table 1. We will further discuss this in Section 5.2 In addition to these dimensions, we also assess the term correction component of the APE pipeline, which is likewise implemented using GPT-4.1. Lastly, we will examine the benefits of iteratively integrating human feedback.

5.1 Stylistic dimensions

Language and form This stylistic dimension primarily covers the required language variant and general instructions aimed at ensuring structurally sound and grammatically correct translations, with an emphasis on clarity and simplicity. In this use case, the translation is required to be in natural clear British English with accessible wording and well-formed sentence structures, adhering to conventions outlined in the Guardian Style Guide¹.

¹<https://www.theguardian.com/info/series/the-guardian-style-guide>

Formatting and punctuation This dimension addresses instructions related to formatting and punctuation, encompassing a wide range of conventions derived from both general language standards and client-specific preferences. These include the use of colons, quotation marks, dashes, and commas², as well as formatting rules for numbers, dates, times, units of measurement, and currencies.

Content conventions This dimension covers guidelines governing the translation of foreign place names, currencies, and abbreviations.

Inclusive and sensitive language This dimension comprises guidelines requiring inclusive, non-stereotyping, and bias-aware language, with particular attention to gender-neutral and accessibility-conscious formulations.

Tone and register This last dimension relates to tone and register, defining the desired voice of the translation. This category differs from the previous ones in that its instructions—such as “Use an immersive, authentic, and engaging travel writing style” and “Preserve local color”—are inherently less explicit and therefore more difficult to operationalize and evaluate systematically.

5.2 Stylistic analysis

Language and form In general, the baseline translation exhibits a high level of grammatical accuracy with relatively few issues identified. Nevertheless, we identify eight segments that are modified by the APE system with respect to this stylistic

²This also includes the Oxford comma, which is the last comma before the coordinating conjunction (such as *and* or *or*) in lists. Its use is a matter of style preference and is generally not preferred by the client.

Stylistic dimensions	Modified instances	Improvement	Deterioration	Non-functional modification	Unapplied cases
Language & Form	8	7	1	0	0
Formatting & Punctuation	45	45	0	0	1
Content Conventions	4	1	3	0	0
Inclusive & Sensitive Language	1	1	0	0	0
Tone & Register	117	14	3	100	0

Table 1: Statistics of modifications made by the APE pipeline regarding the five stylistic dimensions.

dimension. Six segments feature minor edits that improve grammatical correctness and enhance the fluency of the translation, as is the case in example (1). One of the remaining segments involves the correction of a word to its British spelling. The last segment, however, is a less desirable instance in which a less common and archaic British variant is introduced.

- (1) **Source:** *Vous prenez la direction de l’ouest, avec derrière vous les vagues de l’Atlantique et leur grondement solennel.*

Raw MT: *You head west, leaving behind you the waves of the Atlantic and their solemn rumble.*

APE: *You head west, leaving behind the waves of the Atlantic and their solemn rumble.*

Formatting and punctuation The most frequently observed interventions in the analyzed sample include the following:

- Removal of redundant Oxford comma: the Oxford comma was removed when not required for clarity.
- Conversion of units of measurement: as the French source text employs the metric system, the corresponding imperial units were systematically added in parentheses in the translation.
- Parenthetical clauses: en dashes with surrounding spaces were used to replace em dashes without spaces in parenthetical constructions.
- Number formatting: numbers from one to ten were spelled out, while numbers above ten were expressed in numerals.

These guidelines are implemented during the style management step, resulting in the correction of formatting and punctuation in 46 segments, one

of which is shown in example (2). One instance is identified in which a formatting guideline is not applied as expected. In this segment, numbers below ten are rendered as numerals instead of in their spelled-out form. This represents the only false negative observed. Notably, the segment contains both HTML tags and template variables, suggesting that the failure to apply the style rule may be an anomaly caused by the presence of such elements.

- (2) **Source:** *Ruelles escarpées, galeries et balcons de bois ouvragé, le vieux centre a été réhabilité : on se croirait au XIXe siècle.*

Raw MT: *Steep alleyways, galleries, and crafted wooden balconies—the old town centre has been revitalized: it feels like stepping back into the 18th century.*

APE: *Steep alleyways, galleries and crafted wooden balconies – the old town centre has been revitalized: it feels like stepping back into the 18th century.*

Content conventions In the analyzed sample, four segments are identified in which these instructions affect the translation output. Of these, one segment shows an improvement over the baseline translation, whereas three segments reflect a deterioration.

All four instances could be traced back to the instruction “Verify and use official English names for places/companies where they exist; otherwise, preserve the local name”. The positive example involves a French brand name that was initially translated into English by the baseline MT system, which the APE step correctly restores to the French name, as no official English equivalent exists for the brand in question. In the three negative examples, proper nouns were first rendered in their correct English forms by the baseline MT, after which the APE step reverts them to their French spellings. An example is shown in (3).

- (3) **Source:** *Excursion sur les pas des pharaons, à Guizéh et Saqqarah.*

Raw MT: *An excursion following in the footsteps of the pharaohs, at Giza and Saqqara*

APE: *An excursion following in the footsteps of the pharaohs, at Giza and Saqqarah.*

Inclusive and sensitive analysis In this dimension, there is one instance in the analyzed sample where the inclusive language guidelines prompt a modification; the example is shown in (4). The style guide specifies the use of gender-neutral language where appropriate, recommending singular *they* in cases of unknown gender and the avoidance of unnecessarily gendered nouns (e.g., using *postal workers* instead of *postmen*). In the observed example, the French term *cavalier* is translated as *horseman* by the baseline MT system. This is subsequently revised to the gender-neutral variant *horse rider* during the style management step. In addition, the associated possessive pronouns are adjusted from *his* to *their*, ensuring consistency with the prescribed inclusive language conventions.

- (4) **Source:** *Le cavalier mongol avec son aigle au bras est à la fois un symbole national et une figure de l’imaginaire collectif.*

Raw MT: *The Mongolian horseman with his eagle on his arm is both a national symbol and a figure in the collective imagination.*

APE: *The Mongolian horse rider with their eagle on their arm is both a national symbol and a figure in the collective imagination.*

Tone and register The analysis of the selected sample identifies 117 segments in which the modifications cannot be clearly attributed to any specific instruction from the defined stylistic dimensions. While some of these changes may stem from tone-related instructions, this cannot be established with certainty.

Among these segments, 17 demonstrates meaningful changes to the baseline, of which 14 constitute improvements and 3 result in a deterioration of translation quality. The remaining 100 segments involve changes that neither correspond to identifiable stylistic instructions nor produce an objective difference in meaning or quality.

This tendency is illustrated by cases involving purely preferential lexical substitutions, such as replacing *prefers* with *favours* and subsequently reverting *favour* back to *prefer*, as is shown in example (5).

- (5) **Source:** *Votre guide sait quel acacia le guépard préfère ces jours-ci, qu’une hyène a récemment mis bas à deux petits adorables, que les éléphants favorisent ce point d’eau plutôt qu’un autre ...*

Raw MT: *Your guide knows which acacia the cheetah prefers these days, that a hyena has recently given birth to two adorable cubs, and that the elephants favour this waterhole over another ...*

APE: *Your guide knows which acacia the cheetah favours these days, that a hyena has recently given birth to two adorable cubs, and that the elephants prefer this waterhole to another ...*

As also observed by Nunziatini and Diego (2024), the style management step appears prone to introducing unnecessary preferential modifications, despite explicit prompt constraints such as “Do not change phrasing or word choices unless required by these guidelines.”

A closer examination of these instances reveals a notable pattern: in 95 out of the 100 cases involving unnecessary and non-functional changes, such modifications occur in segments that do not require any stylistic revision. Only five cases exhibit co-occurrence with genuinely required, guide-line-driven revisions. This distribution suggests that the model predominantly introduces preferential changes in segments where no rule-based intervention is required, indicating a tendency to “over-edit” in the absence of clear stylistic triggers.

5.3 Term correction

Out of the total of 242 segments in the sample, 88 contain at least one term listed in the glossary. Not all of these instances require intervention, however, as the baseline MT output often already aligned with the prescribed terminology. This is particularly common in the travel domain, which is relatively non-technical: in 40 out of the 88 segments, no correction is necessary, and the translations are therefore left unchanged. In the remaining 48 segments, the term correction step modifies the output to ensure compliance with the glossary.

The majority of the term correction interventions are successful, but seven instances are identified in which glossary entries are applied incorrectly. In all cases, the issue involves relatively common terms for which the glossary prescribes

a more specific translation that is not appropriate in every context. In example (6), the French noun *salon* is associated in the glossary with the translations *trade fair* and *show*. While valid in certain contexts, these translations do not cover the full semantic range of the noun in French. As a result, the glossary is applied outside its intended scope, leading to incorrect translations—such as rendering *salon* as *trade fair* or *show* in contexts where it instead refers to a lounge.

(6) **Source:** *Bassin à l'ombre de la cour intérieure, Honesty Bar au **salon**, allées éclairées à la bougie, tout ici évoque l'élégance.*

Raw MT: *A pool in the shade of the courtyard, an honesty bar in the **lounge**, candlelit paths—everything here exudes elegance.*

APE: *A pool in the shade of the courtyard, an honesty bar in the **trade fair**, candlelit paths – everything here exudes elegance.*

5.4 Iterative human feedback

Section 3.3 conceptually described how recurring issues identified in the APE output can be addressed through targeted modifications to the prompt and/or by enriching the glossary entries with additional metadata. To evaluate the effectiveness of this feedback loop in practice, the sample was reprocessed using an updated prompt and revised glossary designed to mitigate the issues outlined above. This allows for a comparative assessment of whether the proposed adjustments lead to measurable improvements in translation quality and system behavior.

The analysis of the initial APE outputs reveals 13 segments in which translation quality deteriorates following the style management step, as well as 6 instances in which the term correction results in mistranslations.

The update to the style prompt followed involved the addition of explicit guidelines targeting specific stylistic issues, such as clusters of Latinate nouns and zeugmatic constructions in lists. The most significant modification, however, is a shift in overall prioritization: the revised prompt places greater emphasis on natural, idiomatic target-language expression rather than strict adherence to source-text structure.

The translations of all 13 segments that previously exhibited a decline in quality following the style management step were improved with the

updated prompt, resulting in more fluent and idiomatic renderings across all cases. One such improvement is visualized in example (7).

(7) **Source:** *Aux alentours, le Piémont multiplie les séductions, de l'élégante capitale Turin au charme provincial d'Alba ou d'Asti, le tout porté par une gastronomie de haute volée.*

Previous APE: *All around, Piedmont **multiplies its allure**, from the elegant capital Turin to the provincial charm of Alba or Asti, all carried by **high-flying gastronomy**.*

Improved APE: *All around, Piedmont **offers countless attractions**, from the elegant capital Turin to the provincial charm of Alba or Asti – all carried by **outstanding cuisine**.*

With regards to the glossary-related issues, five out of the seven previously identified errors were resolved through the addition of enriched metadata designed to better guide the model's selection and application of terminology. In this way, example (6) is improved, as the terminology step preserves the appropriate translation suggested by the baseline MT system. Two instances remained incorrect, in which the model selected the less idiomatic option out of two available glossary translation options despite the instruction to “choose the best translation according to the context”.

Apart from the two remaining errors, all previously identified issues were successfully addressed through prompt refinement and glossary enhancement, demonstrating the effectiveness of the feedback loop. These results further highlight the value of iterative collaboration between linguists and prompt engineers in the development and refinement of robust automatic post-editing pipelines.

6 Quantitative Evaluation

In Table 2, we show the results of the quantitative evaluation of our APE approach³. We evaluate the outputs of the style management and terminology correction components to examine how much improvement each corresponding step contributes.

We first observe a strong correlation between the five adopted metrics: COMET, BLEU, TER, ChrF, and BertScore. The latter four particularly showed perfect correlation, further validating the process.

³We draw the readers' attention to the fact that the target references used in the quantitative evaluation, while being human-edited, were adapted from outputs of a previous run of the APE pipeline. See discussion in the Limitations section.

system	BLEU	TER	ChrF	COMET	BertScore
Raw MT	60.044	34.410	75.119	0.873	0.775
Style Management	62.489	32.630	76.404	0.873	0.777
Style & Terminology	62.892	32.345	76.786	0.873	0.780

Table 2: Comparison of raw MT and APE outputs.

Generally, the various metrics show a consistent improvement at each step in our pipeline. We see a relatively big improvement in score after the style management step, and a further, although less significant, improvement after the terminology correction step.

We observe no considerable change in COMET scores across the three outputs. This is likely due to the fact that COMET, being a semantic metric, is relatively insensitive to certain stylistic and terminological changes that do not affect the semantics of the text.

7 Conclusion

In this work, we investigate the effectiveness of an APE approach based on GPT-4.1, enhanced through an iterative feedback loop, in a real-world commercial setting. We propose an end-to-end APE approach to this effect and carry out a detailed qualitative analysis across five stylistic dimensions, as well as the terminology correction component.

Our analysis of the five stylistic dimensions indicates that using GPT-4.1 in an APE pipeline considerably improves translations in the language and form, formatting and punctuation, and inclusive and sensitive language dimensions. However, it struggles to improve translations with respect to the content conventions. It is difficult to ascertain whether the improvements with regard to tone and register can be directly attributed to the style guide or whether these modifications are influenced by the intrinsic style of the LLM.

Our analysis of the term correction component shows that it reliably preserves terms that are already correctly translated and effectively corrects instances where the baseline MT output does not comply with the glossary. At the same time, it exhibits a tendency to over-apply glossary entries for relatively common terms whose prescribed translations are context-dependent. In such cases, the system may enforce a specific glossary equivalent even when it is not appropriate for the given context.

The above-mentioned issues can be mitigated by prompt refinement and glossary enhancement based on linguistic feedback. Our findings indicate that the human feedback loop is pivotal in an APE pipeline, allowing the system to be systematically improved to address reported issues and reflect the exact preferences of the client, as well as to adapt to the dynamic business environment of clients.

Finally, we show, through quantitative evaluation using five industry-established metrics, how the various components of our APE architecture consistently improve translations.

Limitations

A real-world test case introduces several challenges, reflecting the operational realities of commercial localization workflows. Chief among them in this work is obtaining target references for quantitative evaluation. As earlier described, in this use case, references were obtained through human post-edits on outputs from a previous run of the APE system, which may introduce a bias in favor of our APE system. Additionally, although the use of standard MT quality metrics such as BLEU are effective at capturing overall improvements and terminological corrections, they may not adequately capture fine-grained stylistic improvements by the setup. Nevertheless, the detailed qualitative analysis carried out in this work gives an outlook that is consistent with the quantitative results, thus further confirming our conclusions.

Secondly, due to the proprietary and client-specific nature of the project, the full prompts used for our experiments and analysis are not disclosed. Consequently the replicability of this work is affected.

Ethics statement

To ensure ethical standards and protect client anonymity, all examples in this work have been modified. Care has however been taken to ensure that these modifications do not alter the core content that underpins the analysis and conclusions.

References

- Alves, Duarte M, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.
- Chen, Pinzhen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190.
- Dong, Qingxiu, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA, November. Association for Computational Linguistics.
- Finkelstein, Mara, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, et al. 2026. Translategemma technical report. *arXiv preprint arXiv:2601.09012*.
- Ki, Dayeon and Marine Carpuat. 2024. Guiding large language models to post-edit machine translation with error annotations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4253–4273.
- Luo, Yun, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. corr abs/2308.08747 (2023). *arXiv preprint arXiv:2308.08747*.
- Nunziatini, Mara and Sara Diego. 2024. Implementing gender-inclusivity in mt output using automatic post-editing with llms. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 580–589.
- Papineni, Kishore, 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.
- Popović, Maja. 2015. chrF: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.
- Raunak, Vikas, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024.
- Snoover, Matthew, Bonnie Dorr, Richard Schwartz, Linea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.
- Uguet, Celia, Fred Bane, Mahmoud Aymo, João Torres, Anna Zaretskaya, and Tània Blanch Miró Blanch Miró. 2024. Llms in post-translation workflows: comparing performance in post-editing and error analysis. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 373–386.
- Vidal, Blanca, Albert Llorens, and Juan Alonso. 2022. Automatic post-editing of mt output using large language models. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 84–106.
- Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu. 2020. Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Zhai, Yuexiang, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2024. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*, pages 202–227. PMLR.
- Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Zhu, Shaolin, Menglong Cui, and Deyi Xiong. 2024a. Towards robust in-context learning for machine translation with large language models. In Calzolari, Nicoletta, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629, Torino, Italia, May. ELRA and ICCL.
- Zhu, Wenhao, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024b. Multilingual machine translation with large language models: Empirical results and analysis. In Duh, Kevin, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico, June. Association for Computational Linguistics.

A Appendix

This shows an example of a section of the style guide that was automatically converted to an APE instruction and then manually validated.

1. Style guide section: Inclusive writing

- Exercise caution when referring to cities, countries or regions that are politically contested. When giving examples involving multiple regions, use comparable references. For example, avoid mixing countries with states or continents.
- Do not generalise or stereotype people based on their region, culture, age or gender, not even with positive stereotypes.
- Do not use rude, derogatory or racially biased terms.
- Accessibility: Avoid using words that imply pity or illness. Instead of ‘victim of’, ‘suffering from’, ‘afflicted by’ or ‘crippled by’, use ‘person who has’ or ‘person with’. Instead of ‘the disabled’, ‘the handicapped’, ‘the blind’ or ‘the deaf’, use ‘disabled people’, ‘blind people’, ‘deaf people’ or ‘people with hearing loss’.
- Gender-neutral language: Avoid gendered terms (postal workers instead of postmen, firefighter instead of fireman). Use ‘they’ when a person’s gender is unknown or irrelevant (Ask your concierge if you need help. They are there to support you.)

2. LLM conversion:

- Use inclusive, gender-neutral language and avoid stereotypes.

3. Manually validated instructions:

- Avoid stereotypes, generalizations, and offensive or pitying terminology.
- Use gender-neutral language (postal workers not postmen, they for unknown gender).
- Use person-first language (person with... not victim of...).

The Style of Machines: A Stylometric Study of LLM Generation and Translation

Natalia Resende
Trinity College Dublin
ADAPT Centre
resenden@tcd.ie

Sheila Castilho
Dublin City University
ADAPT Centre
sheila.castilho@dcu.ie

Abstract

This paper investigates whether large language models (LLMs) develop a recognisable stylistic signature and whether such a signature persists across a text generation task and a translation task in different languages. Drawing on stylometric approaches traditionally used in authorship and translator attribution studies, we analyse outputs from ChatGPT-4o in English, Brazilian Portuguese, and Spanish, across original literary composition and translation conditions. The analysis focuses on five feature categories: lexical density, lexical richness, vocabulary sophistication, sentence length, and punctuation style. Our results show a LLM cross-linguistic profile characterised by higher lexical richness and density, rarer vocabulary, and a preference for the em-dash punctuation mark. This pattern appears in both generation and translation, though it is stronger in original text production.

1 Introduction

The notion that individual authors leave a distinctive imprint on their writing is foundational to stylometry and computational literary studies. It has long been argued that each author possesses a set of stylistic idiosyncrasies (i.e., preferences in vocabulary, sentence construction, rhythm, and punctuation) that function as a kind of linguistic fingerprint, consistent enough across texts to be identified and distinguished from those of other authors (Holmes, 1994; Holmes, 1998).

© 2026 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

In the translation domain where the translators, who have long been rendered invisible by prevailing norms of equivalence and transparency (Venuti, 1995), have gained more recognition as secondary authors, whose own stylistic features shape the target text (Saldanha, 2014). Baker’s (Baker, 2000) proposal for a corpus-based methodology to investigate translator style established that translators exhibit consistent, recoverable patterns of lexical and syntactic choice. These patterns persist across different source texts and source languages, and cannot be attributed entirely to the influence of the original author or the structural constraints of the target language. Corpus stylistic work has confirmed (Mauranen and Kujamäki, 2004; Volansky et al., 2015; Baroni and Bernardini, 2006) and refined these findings, identifying translator fingerprints across a range of linguistic features including reporting verb constructions, the use of emphatic italics, lexical density, and type-token ratio (Saldanha, 2011; Munday, 2008).

In this context, the emergence of large language models (LLMs) as versatile text generators and translators raises a question that the field has only recently begun to address: do these systems develop a writing style of their own?

Early evidence suggests that LLMs might have a “style”. Research has shown that ChatGPT tend to systematically overuses lexical items. Juzek and Ward (2025) identified 21 focal words whose usage had spiked anomalously in scientific abstracts following the widespread adoption of ChatGPT, most notoriously the verb *delve* at rates far exceeding their frequency in human writing. They confirmed that these words are disproportionately represented in GPT-generated text relative to human-authored baselines. Kobak and colleagues (2025) found an abrupt increase in a cluster of LLM-

preferred style words when analysing over fourteen million PubMed abstracts between 2010 and 2024. This suggested that at least 10% of 2024 abstracts was processed by an LLM. These findings point not only to occasional word choice differences but to something that could be more systematic: a recognisable and persistent lexical signature that distinguishes LLM output from human texts.

Whether this signature found in LLM-generated texts extends beyond lexical choice to the broader stylistic dimensions studied in authorship and translator attribution research (such as sentence length, lexical richness, lexical density, punctuation preferences, and vocabulary sophistication) remains an open question. Unlike human authors or translators, LLMs are not individuals shaped by biography, mother tongue, or conscious aesthetic choices. They are trained on vast corpora of human-produced texts and refined through reinforcement learning from human feedback; their outputs are thus the product of statistical patterns learned across a broad range of domains, registers, and genres (OpenAI, 2023). Whether this training process produces something that can be called a *style* is an open and consequential question.

Building on the established finding that both human authors and human translators leave distinctive stylistic traces in their texts, a key question is whether a comparable stylistic profile emerges for LLMs, and whether such a profile persists across both generative and translation conditions — that is, whether the fingerprint of an LLM is visible not only when the system is composing original text but also when it is operating as a translator. Furthermore, drawing on Baker’s (1996) proposal that translated texts tend to exhibit universal features regardless of the source language, we ask whether any such AI stylistic traces persist across the different languages in which the text is generated or translated. Therefore, this paper investigates the following questions:

1. *Does ChatGPT-4o exhibit outputs that are stylistically distinguishable from human-authored texts?*
2. *If such a stylistic tendency exists, is it consistent across generation, translation and language conditions in the texts examined?*

This paper is organised as follows. The Methodology (section 2) outlines the corpus design (section 2.1) and data collection procedures (section

Language	Author	Translator	# words
EN	Human	—	1831
		Human	1235
		AI	1164
	AI	—	1467
PT-Br	Human	—	1257
		Human	2035
		AI	1754
	AI	—	2006
ES	Human	—	1036
		Human	1898
		AI	1357
	AI	—	1450

Table 1: Corpus structure: twelve text categories across language, authorship, and translation conditions.

2.3). The results are then presented (section 3) followed a discussion of the findings in relation to stylistics and their implications for translation studies (section 4). Finally, we present our concluding remarks and directions for future work (section 5).

2 Methodology

2.1 Corpus Design

To investigate whether ChatGPT exhibits a distinctive writing style and whether that style persists across authorship and translation roles, we constructed a small but balanced parallel corpus comprising twelve text categories across three languages: English (EN), Brazilian-Portuguese (PT-Br), and Spanish (ES). The corpus covers two conditions, original authorship and translation, and two agent types: human and AI (ChatGPT-4o).

Three human authors, one writing in EN, one in PT-Br, and one in ES, produced original short stories. The same AI system (ChatGPT-4o) also produced one original story in each language. Each human-authored story then served as a source text for two translations: one produced by a human translator and one by ChatGPT-4o. This yields twelve text categories, as summarised in Table 1.

The source text for translations (both human translation and AI translation) into PT-Br and ES was the short story originally produced in EN by the human author. Conversely, the source text for the human and AI translations into EN was the short story written in PT-Br by a human author. This asymmetry reflects a methodological choice grounded in the global dynamics of literary trans-

lation. EN is consistently ranked as the world’s dominant source language for translation: according to the UNESCO Index Translationum dataset (UNESCO,), it has historically occupied the first position among source languages, and (CSA Research, 2020) found that English is involved in 19 of the top 20 most translated language pairs worldwide. This dominance is even more pronounced in literary translation, where the flow of texts tends to run from English into other languages far more often than the reverse (Venuti, 1995; Heilbron, 1999). The PT-Br-authored story was chosen as the source for the EN translation because it was the first human-written text produced in a language other than EN as part of the data collection process, making it the natural starting point for exploring translation into the globally dominant language.

2.2 Rationale

Before presenting the results, it is worth making the comparative logic of the study explicit. The two conditions, translation and original authorship, are analysed here as two sources of evidence. In the translation condition, the human and AI texts share the same source text, meaning that source language structure and author style are held constant across both agents. Any systematic difference observed between human and AI translations across three language pairs cannot be attributed to source text properties, and we interpret this as suggestive evidence of a cross-linguistic AI stylistic tendency. The authorship condition is necessarily less controlled: with one author per language, individual stylistic variation cannot be fully disentangled from language-level differences. However, the authorship findings are not used here as independent evidence for an AI signature; rather, if the same patterns detected in the more controlled translation comparison also appear in the authorship condition generated by chatGPT-4o, we interpret this convergence as further support for the claim. The translation condition provides the cleaner test; the authorship condition shows whether the tendency generalises beyond the translation task.

2.3 Data Collection Procedures

Authorship condition. Both human authors and ChatGPT-4o received identical written instructions for text production. The only exception was the prohibition on the use of technological assistance, which applied exclusively to human participants

and was not included in the AI prompt.

- a third-person narrator,
- a target audience of adults (18+), a target length of 1,000–1,200 words,
- a fixed narrative brief (a character unearths a mysterious buried object that triggers reflections on family history),
- and an explicit prohibition on the use of any technological assistance including AI tools or writing enhancement software (only humans received this instruction).

Human authors were recruited on the basis of their prior experience in creative writing and their language background (one L1 EN speaker, one L1 PT-Br speaker, and one L1 ES speaker).

Translation condition. Human translators received written guidelines instructing them to translate each source story without recourse to any machine translation or AI-assisted tool (including Google Translate, DeepL, or ChatGPT), while allowing the use of print or online dictionaries. ChatGPT-4o received a minimal prompt: “*Translate this short story into [target language]*”. Human translators were additionally instructed to preserve the names of characters as given in the source text. This constraint was introduced to ensure comparability with ChatGPT-4o’s behaviour, which, based on preliminary tests, tends not to alter character names under minimal prompting conditions. No further stylistic constraints were imposed on the AI or on the human translators in the translation condition, in order to reflect the zero-shot setting that characterises typical LLM-based translation practice.

2.4 Feature Extraction

Five categories of stylistic features were extracted computationally from each text:

1. **Lexical density.** Computed as the ratio of content words (nouns, verbs, adjectives, and adverbs) to the total number of tokens in each text. Higher values indicate a greater proportion of information-carrying words and hence denser, more compact phrasing.
2. **Lexical richness (MATTR).** Measured using the Moving-Average Type-Token Ratio

(Covington and McFall, 2010), which computes the mean TTR over a sliding window (50 words in our experiment) of fixed length across the text. Unlike simple TTR, MATTR is robust to differences in text length, making it suitable for comparing documents of varying sizes.

3. **Vocabulary sophistication.** Operationalised as the mean Zipf frequency (van Heuven et al., 2014) of the content words (nouns, verbs, adjectives, and adverbs) in each text, estimated using the `wordfreq` Python library (v3.1.1) (Speer, 2022) for EN, PT-Br, and ES. Zipf frequency is defined as $\log_{10}(\text{frequency per billion tokens})$; lower values therefore correspond to rarer, more sophisticated vocabulary.
4. **Mean sentence length.** Calculated as the average number of tokens per sentence across each document.
5. **Punctuation style.** Computed as per-token ratios for a set of typographically salient marks including em-dashes (–) and excluding the hyphens, single quotation marks (excluding apostrophes), and ellipses (...). These marks were selected on the basis of their known sensitivity to authorial and translator style (Baker, 2000).

3 Results

We present results for the five feature categories introduced in Section 2.4, organised around the two research questions: (i) whether GPT-4o output differ from human-authored texts, and (ii) whether this stylistic signature (if any) persists across both the original writing, translation and language.

3.1 Lexical Features

Lexical richness (MTTR). Figure 1 shows the results for lexical richness. We note that, ChatGPT originals consistently display higher MTTR values than their human counterparts across all three languages. The gap is most pronounced in EN, where the GPT original (MTTR = 0.808) exceeds the human original (0.742) and both translation texts. In PT-Br and ES the difference is smaller but in the same direction ($\Delta = +3\%$ and $+1\%$, respectively). In the translation condition, GPT translations are marginally richer than human translations in all

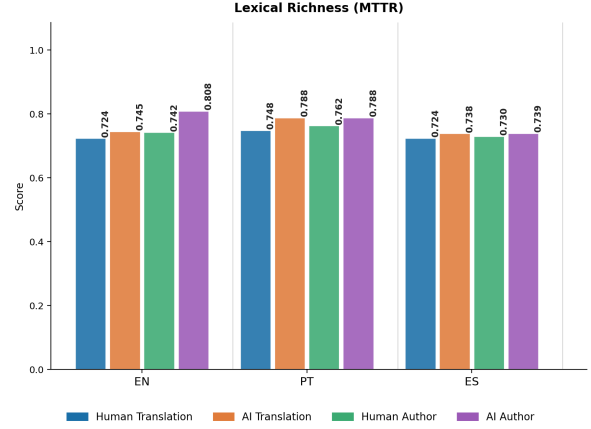


Figure 1: Lexical richness (MTTR) across all conditions and languages.

languages (e.g. EN: 0.745 vs. 0.724; PT: 0.788 vs. 0.748), suggesting that the lexical-richness signature of the system is present, albeit attenuated, even when GPT operates as a translator rather than an author.

Lexical density. Figure 2 shows the results for lexical density. We can see that, GPT originals are denser than human originals in all three languages (EN: 0.549 vs. 0.474; PT: 0.540 vs. 0.485; ES: 0.495 vs. 0.466), indicating a higher proportion of content words relative to total tokens. In the translation condition, the human-GPT difference is smaller and less consistent, though GPT translations show slightly higher density in all languages.

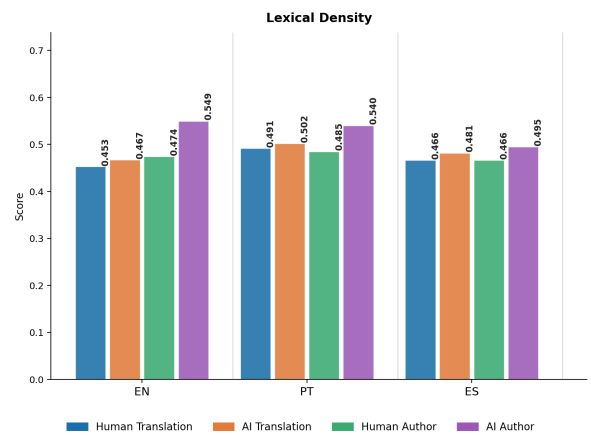


Figure 2: Lexical density across all conditions and languages.

Vocabulary sophistication. In figure 3, we note that the lower Zipf frequency values indicate rarer, more sophisticated vocabulary. GPT originals use markedly rarer words than human originals in EN

(mean Zipf: 3.20 vs. 3.56) and PT-Br (3.51 vs. 3.70), while the difference in ES is negligible (5.65 vs. 5.78). In the translation condition the pattern reverses slightly: human translations tend to use words of similar or slightly lower frequency than GPT translations, suggesting that the lexical sophistication gap is a feature of GPT’s generative mode rather than its translation mode.

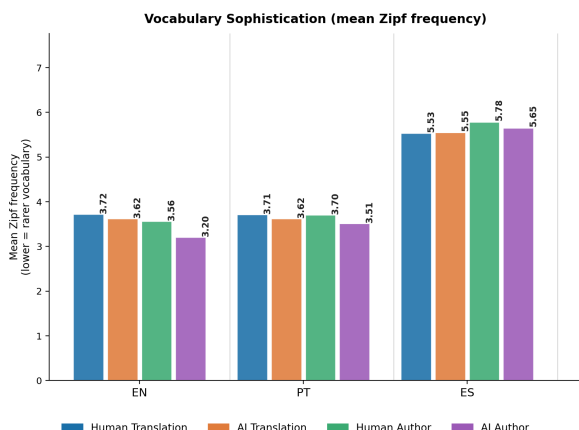


Figure 3: Vocabulary sophistication (mean Zipf frequency) across all conditions and languages. Lower values indicate rarer, more sophisticated vocabulary.

3.2 Sentence Length

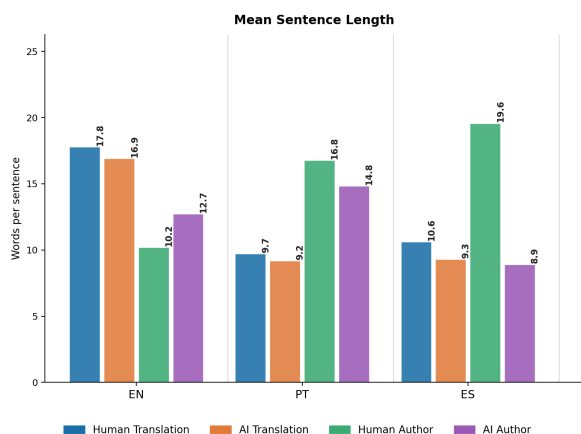


Figure 4: Mean sentence length (words per sentence) across all conditions and languages.

Mean sentence length (Figure 4) shows the largest absolute differences in the dataset. In the original writing condition, GPT originals are systematically shorter-sentenced than human originals in two of three languages: ES and PT-Br. The ES contrast is particularly striking: the human original has a mean sentence length nearly 2.2 times that of the GPT original (Human: 19.6 vs. GPT: 8.9 words). In EN, however, the GPT

original is slightly longer-sentenced than the human original (GPT: 12.7 vs.; Human: 10.2 words). In the translation condition, the pattern diverges markedly between languages. EN human translations are longer-sentenced than GPT translations (Human Tr.: 17.8; GPT Tr.: 16.9 words). A possible explanation is that the PT-Br source text already featured longer sentences on average (16.8 words), which may have influenced both translations independently of any agent-specific tendency. In PT-Br and ES, by contrast, both human and GPT translations produce shorter sentences than the human originals written in those languages, which may also reflect the influence of the shorter-sentenced EN source text on both translations rather than any intrinsic property of the target language or the translating agent.

3.3 Punctuation Style

Figure 5 shows the per-1,000-token rates for em-dashes, commas, and periods across all conditions.

Em-dash. The em-dash is the single most discriminating feature in the dataset. Human translators and human authors use it rarely or not at all: every human text scores 0.0 em-dashes per 1,000 tokens in the translation condition across all three languages, and the human originals score 3.2 (EN), 0.0 (PT-Br), and 0.0 (ES). GPT, by contrast, uses em-dashes extensively: GPT translations score 10.1 (EN), 27.3 (PT), and 25.3 (ES) per 1,000 tokens, and GPT originals score 12.5 (EN), 9.5 (PT-Br), and 2.9 (ES). Within this corpus, this constitutes a near-categorical stylistic distinction between human and GPT-4o output across both conditions and all three languages. We acknowledge, however, that em-dash frequency varies considerably among human authors and that GPT-4o’s tendency to favour it reflects a default behaviour under zero-shot prompting conditions, one that could be suppressed through explicit stylistic instructions. Nevertheless, the consistency of this pattern across three languages and both authorship and translation conditions, without any interference in the prompting, suggests that em-dash overuse is a feature of GPT-4o’s default stylistic output, even if it does not constitute a universal or immutable distinction between human and AI writing.

Comma and period. Comma usage is high in the EN human translation (94.3/1k) relative to both

the GPT EN translation (70.1/1k) and the EN human original (44.6/1k). GPT EN translation uses commas at an intermediate rate. In PT-Br and ES, the pattern is reversed: human translations use fewer commas than GPT translations, and the originals use more commas than the translations. A possible explanation is that the human translation into EN was influenced by the amount of comma introduced in the source text in PT-Br but ChatGPT not so much. Similarly, the low frequency of comma introduced in the EN source influenced the human translation into ES and PT-Br to be reduced compared to their original counterparts in ES and PT-Br in which the authors tend to introduce much more comma. Period frequency broadly mirrors the inverse of comma usage, with GPT translations in PT-Br and ES using more periods than their human counterparts (PT: 108.2 vs. 86.7/1k; ES: 105.1 vs. 95.8/1k), consistent with GPT’s shorter mean sentence lengths in those languages.

3.4 Summary: GPT vs. Human Relative Differences

Figure 6 presents a heatmap of the percentage difference between GPT and human values for each feature across the six conditions (three languages $\times$ two tasks). The pattern that emerges is consistent: GPT shows a robust positive divergence from human output on em-dash usage in the translation task ($+\infty\%$ over a human baseline of zero in PT and ES; $+24\%$ in EN original), and positive divergence on MTTR and lexical density across both tasks. Vocabulary sophistication divergence is most pronounced in the original writing task (EN: -10% , indicating rarer words in GPT), while sentence length divergence is largest in the ES original condition (-55% , indicating shorter GPT sentences). Crucially, the direction of divergence on em-dash, lexical richness, and density is consistent across both tasks, suggesting that GPT stylistic signature is present in both the authorship and translation conditions at the lexical level and punctuation level.

4 Discussion

4.1 A Recognisable GPT Stylistic Signature

The results suggest that, in the texts examined, GPT-4o exhibits a stylistic tendency that sets it apart from the human authors and translators across linguistic dimensions. This signature is not reducible to a single feature but is instead dis-

tributed across lexical and punctuation domains. Three components are particularly robust.

First, and most strikingly, the em-dash functions as a near-categorical marker of GPT authorship. Its complete absence from all human translation texts and from human original writing in PT-Br and ES, contrasted with its consistent and substantial use by GPT in both tasks and all three languages, suggests that it is not a translational borrowing from any particular source language or a genre convention but a idiosyncratic preference of the default model under zero shot prompting. This finding extends earlier work on GPT lexical preferences (Juzek and Ward, 2025; Kobak et al., 2025) from the vocabulary domain to punctuation: just as GPT overuses certain words, it overuses certain typographic conventions.

Second, GPT originals are consistently richer in lexical terms than human originals. Higher MTTR and lexical density values, alongside rarer average word frequencies, paint a picture of GPT as an over-lexicalised author that deploys a broader vocabulary and a higher proportion of content words than human literary authors writing under equivalent conditions. This aligns with the observation that LLM outputs tend toward a particular kind of formal, informationally dense register (OpenAI, 2023), though the mechanism in a literary prompting context deserves further investigation.

Third, the GPT stylistic lexical tendencies are present in both the original writing and translation conditions, though it is more pronounced in original authorship. This is consistent with the hypothesis that when GPT operates as a translator, the source text exerts a partial normalising constraint on its output, pulling certain surface features closer to the source author’s style and this is visible when we analyse the influence of the sentence length of the source texts influence the length of the sentences in the target text. Even so, the em-dash signature is not suppressed by the translation task, suggesting that some stylistic preferences are robust enough to persist despite source text influence.

4.2 Implications for Translation Studies

From a translation studies perspective, these findings reinforce the view that the translator, whether human or artificial, is not a stylistically neutral medium but an agent whose own preferences leave traceable marks on the target text (Baker, 2000;

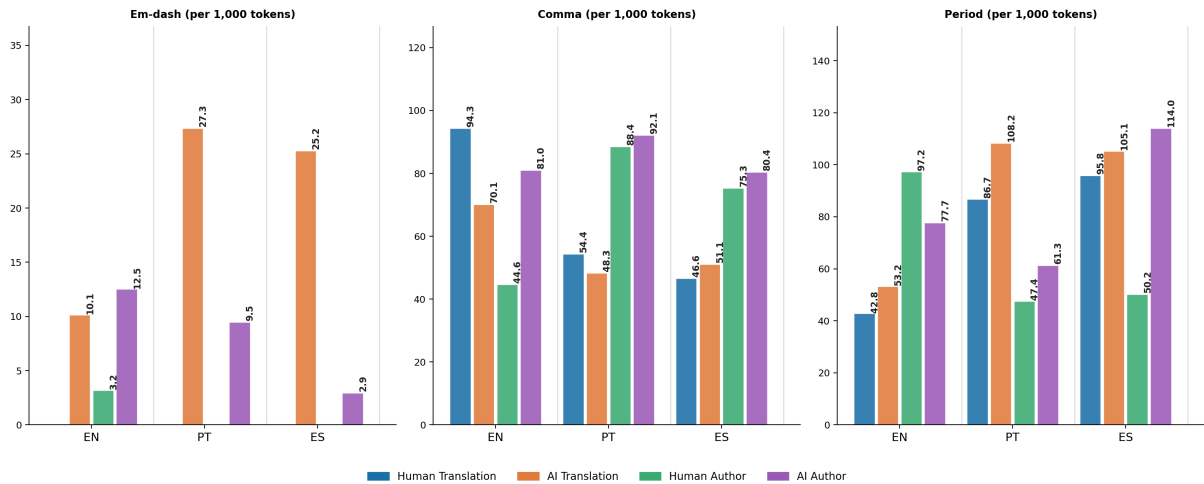


Figure 5: Punctuation signatures: em-dash, comma, and period rates (per 1,000 tokens) across all conditions and languages. Note the near-zero em-dash usage in all human-authored texts versus the consistently elevated rates in GPT outputs.

Baker, 1996; Baker, 1993). The finding that GPT’s punctuation preferences persist even in the translation condition implies that researchers and practitioners evaluating MT output should be alert not only to semantic accuracy and fluency but to the stylistic imprint that the system imposes on target texts.

The sentence length results also raise questions about source language interference in AI-mediated translation. The pattern observed in the EN translations, where both human and GPT translations exhibit longer sentences than their respective target language norms, is consistent with the classical phenomenon of interference (Toury, 1995), whereby source text syntactic patterns leave traceable marks on the target text. GPT-4o proves no less susceptible to this effect than human translators, suggesting that source text influence operates similarly across both human and AI-mediated translation. The PT-Br and ES translation conditions point in the same direction: in both languages, translations produce shorter sentences than the human and AI originals written in those languages, possibly as a result of interference from the shorter-sentenced EN source text. This cross-linguistic consistency suggests that source language interference in sentence length is not an isolated phenomenon but a broader tendency in translation, regardless of the target language or the agent involved.

4.3 Limitations

Several limitations should be noted. The corpus is small, twelve texts, each of approximately 1,000-

2,000 words, and results cannot be taken as statistically representative. The human participants, though carefully selected, each represent a single authorial voice, meaning that observed human-GPT differences could partly reflect individual idiosyncrasy rather than a categorical human-AI contrast. The analysis is based on surface-level stylistic features and does not address deeper syntactic or narrative dimensions of style. Finally, the study covers a single AI system (ChatGPT-4o) at a single point in time; the stylistic properties of LLMs are known to shift across model versions and fine-tuning iterations. Future work should address these limitations through larger, more diverse corpora, multiple AI systems, and the inclusion of inferential statistics. Additionally, all AI outputs were generated using a minimal zero-shot prompt, which may not reflect the stylistic behaviour of ChatGPT-4o under the more directive prompting strategies commonly used in practice. Users can readily instruct LLMs to suppress specific stylistic features, such as em-dash usage, or to preserve source text punctuation conventions; however, whether more explicit, i.e., detailed prompting would also attenuate the lexical patterns observed here remains an open question for future research. Finally, no replication runs were conducted for the AI-generated texts; it cannot be excluded that different runs would yield different stylistic profiles. We also acknowledge that the absence of qualitative analysis reflects a deliberate methodological choice consistent with the corpus stylistic tradition within which this study is situated (Stamatatos, 2009; Burrows, 2002; Baker, 1996), in

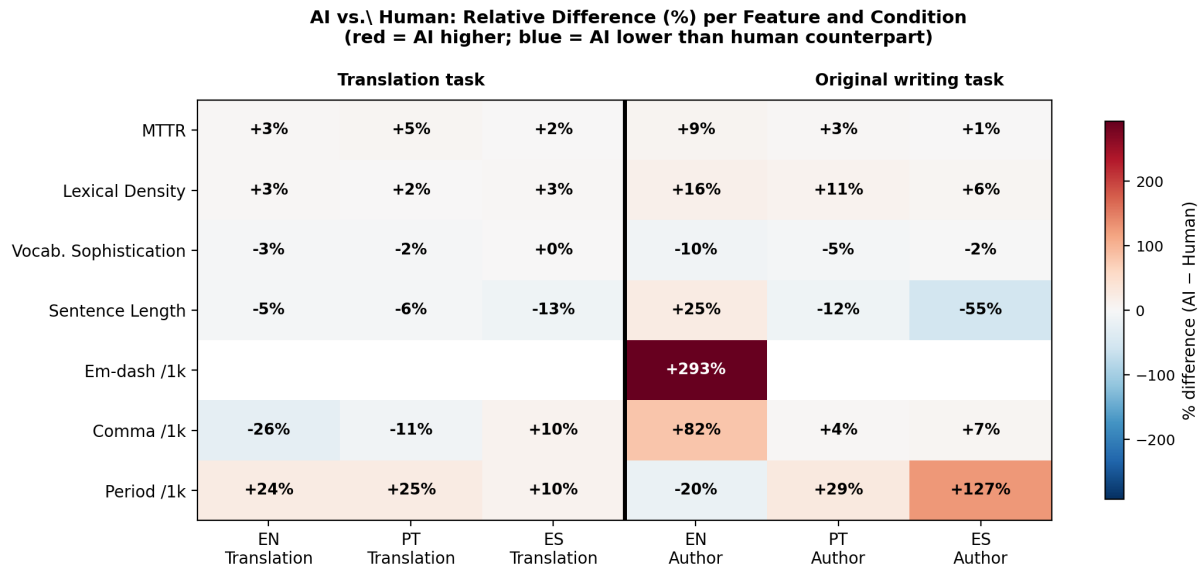


Figure 6: Percentage difference between GPT and human values for each stylometric feature, across translation and original writing conditions in three languages. The vertical black line separates the two task types. Red indicates GPT-higher values; blue indicates GPT-lower values.

which quantitative, corpus-based methods are considered the most appropriate means of detecting stylistic patterns that operate below the threshold of conscious perception and would not be recoverable through qualitative close reading alone.

5 Conclusion

This paper has investigated whether ChatGPT-4o presents a recognisable stylistic signature when producing literary text in original and translation conditions, across EN, PT-Br, and ES. The analysis of five stylometric feature categories (lexical density, lexical richness, vocabulary sophistication, sentence length, and punctuation style) reveals tendencies in the texts examined whereby GPT outputs show higher lexical richness, greater lexical density, rarer vocabulary, and a marked preference for the em-dash. This profile is present in both original authorship and translation conditions, though it varies across languages. In addition, results reveal that these stylistic tendencies tend to be more pronounced when GPT functions as an author than when it operates as a translator, suggesting that source text influence partially constrains the model’s stylistic autonomy.

In computational stylometry, they provide preliminary evidence that ChatGPT-4o may develop stylistic tendencies detectable by the same methods used to identify human authorial style, warranting investigation at larger scale. In translation studies, they suggest that the AI translator

may function as an active stylistic agent whose tendencies are not fully neutralised by the translation task. And in AI language research, they extend the emerging literature on LLMs lexical markers into the domain of punctuation and literary genre, suggesting that the stylistic tendencies of ChatGPT-4o may be richer and more systematic than isolated lexical observations have so far revealed.

Future work will expand the corpus to include multiple human authors and translators per language, additional AI systems, and more fine-grained syntactic features, while also developing inferential methods appropriate for the small-corpus regime characteristic of literary stylometric research. Furthermore, replication across different story selections and literary traditions will be necessary to assess the generalisability of the findings reported here.

6 Acknowledgements

We would like to express our sincere gratitude to the authors Adriana Martins, Lisa Ronan, Ana Olivares and to the translators Thais Gianmarco, Catarina Freitas, and the students of the Masters in Translation Technologies at DCU who contributed to the translation from English into Spanish. All participants generously volunteered their time and creativity in writing and translating the short stories that form the basis of this study. Their collaboration and enthusiasm made this research possible. The authors benefit from being member of the

ADAPT SFI Research Centre at Dublin City University, funded by the Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2.

Data and Materials Availability

The files and materials used in the experiments are available in the following repository.

References

- [Baker1993] Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Baker, Mona, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In Honour of John Sinclair*, pages 233–250. John Benjamins, Amsterdam and Philadelphia.
- [Baker1996] Baker, Mona. 1996. Corpus-based translation studies: The challenges that lie ahead. In *Terminology, LSP and Translation*, page 175. John Benjamins.
- [Baker2000] Baker, Mona. 2000. Towards a methodology for investigating the style of a literary translator. *Target: International Journal of Translation Studies*, 12(2):241–266.
- [Baroni and Bernardini2006] Baroni, Marco and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.
- [Burrows2002] Burrows, John F. 2002. ‘delta’: a measure of stylistic difference and a guide to likely authorship. *Lit. Linguistic Comput.*, 17:267–287.
- [Covington and McFall2010] Covington, Michael A. and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100.
- [CSA Research2020] CSA Research. 2020. The state of the linguist supply chain: Translators and interpreters in 2020. Technical report, CSA Research.
- [Heilbron1999] Heilbron, Johan. 1999. Towards a sociology of translation: Book translations as a cultural world-system. *European Journal of Social Theory*, 2(4):429–444.
- [Holmes1994] Holmes, David I. 1994. Authorship attribution. *Computers and the Humanities*, 28(2):87–106.
- [Holmes1998] Holmes, David I. 1998. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117.
- [Juzek and Ward2025] Juzek, Tom S and Zina B. Ward. 2025. Why does ChatGPT “delve” so much? exploring the sources of lexical overrepresentation in large language models. In Rambow, Owen, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6397–6411, Abu Dhabi, UAE, January. Association for Computational Linguistics.
- [Kobak et al.2025] Kobak, Dmitry, Rita González-Márquez, Emőke Ágnes Horvát, and Jan Lause. 2025. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813.
- [Mauranen and Kujamäki2004] Mauranen, Anna and Pekka Kujamäki. 2004. *Translation universals: Do they exist?*, volume 48. John Benjamins Publishing.
- [Munday2008] Munday, Jeremy. 2008. *Style and Ideology in Translation: Latin American Writing in English*. Routledge Studies in Linguistics. Routledge, New York and London.
- [OpenAI2023] OpenAI, R. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1.
- [Saldanha2011] Saldanha, Gabriela. 2011. Translator style: Methodological considerations. *The Translator*, 17(1):25–50.
- [Saldanha2014] Saldanha, Gabriela, 2014. *Style in, and of, Translation*, chapter 7, pages 95–106. John Wiley Sons, Ltd.
- [Speer2022] Speer, Robyn. 2022. rspeer/wordfreq: v3.0, September.
- [Stamatatos2009] Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.
- [Toury1995] Toury, Gideon. 1995. *Descriptive Translation Studies and Beyond*. John Benjamins, Amsterdam and Philadelphia.
- [UNESCO] UNESCO. Index translationum: World bibliography of translation. <https://data.unesco.org/explore/dataset/tran001/information/>. Accessed: 05/2026.
- [van Heuven et al.2014] van Heuven, Walter J. B., Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67(6):1176–1190.
- [Venuti1995] Venuti, Lawrence. 1995. *The Translator’s Invisibility: A History of Translation*. Routledge, London and New York.
- [Volansky et al.2015] Volansky, Vered, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Toying with Style: Can GenAI Mimic a Literary Translator’s Voice?

Benjamin Sopot
SALIS/CTTS
Dublin City University
Ireland

benjamin.sopot2@mail.dcu.ie

Dorothy Kenny
SALIS/CTTS
Dublin City University
Ireland

dorothy.kenny@dcu.ie

Abstract

In this paper we explore the potential of generative AI to produce stylistically-aware literary translations. We use three prompting strategies (zero-shot, one-shot and few-shot) to attempt to elicit translations in the style of Polish literary translator Jan Rybicki from ChatGPT 5.4 and Claude Sonnet 4.6. We test these strategies on a 1,000-word extract from a novel for which we have a previously unpublished translation by Rybicki. Stylo-metric analyses (based on average sentence length, sentence length distribution, and most frequent words) shows that none of the prompts we tested, regardless of the strategy, brought either LLM significantly closer to Rybicki’s style.

1 Introduction

The use of artificial intelligence (AI) systems in the process of literary translation is now a commercial reality, albeit a controversial one (Kenny, 2025). While academic research in the area has tended to focus on technical challenges related to training and evaluating the neural machine translation (NMT) models and—since 2022—large language models (LLMs) used in AI-assisted literary translation, a growing body of work is also concerned with the impact of automation on the style of translated literary texts. Such work includes studies that explore how AI systems can emulate the style of individual human translators. The current research contributes to this body of work by investigating the potential of two LLMs to produce translations in the style of well-known Polish literary translator, Jan Rybicki.

This paper serves as a pilot study for a broader research project that seeks to explore the potential of personalized AI-assisted translation to augment the human translator’s performance and experience of translation. One of the research questions asks whether personalized AI translation models can capture a human translator’s literary style. As analysts, we learn about Rybicki’s style from a large corpus of his previous translations, which he has shared with us for the purposes of this project. This corpus can also be used to supply prompting data and human reference translations for our experiments with stylistically-aware AI-assisted translation, as illustrated in this paper. Ultimately this research hopes to contribute to the currently intense professional and media debate about the use of AI in literary translation and other creative practices. It is conducted in a context in which others active in the creative industries routinely move to protect their (literal) voice and likeness (see, for example, Equity 2026), but where conversational LLMs already claim to be able to produce translations in the (metaphorical) voice of individual human translators (Kenny, in preparation). We proceed from the vantage point of Computer-Assisted Literary Translation or ‘CALT’, which is “founded on the idea that the possibilities held out by translation technologies...should be explored by both scholars and practitioners of the art” (Rothwell et al., 2024: 6; emphasis in the original); and, in keeping with the tenets of CALT, we are open to the idea that such technologies have the potential to augment rather than displace human translation. Whether LLMs really can emulate an individual translator’s style, however, and whether such LLMs might be useful to the literary translator in question – for example by producing a draft translation that consistently uses his/her style – can only be ascertained through targeted research of the kind we conduct here.

2 Previous Research

2.1 NMT, literary translation and translator style

Previous work in AI-assisted literary translation investigates how ‘generic’ (often free online) machine translation (MT) can be applied to literary text, and on how bespoke ‘literary-adapted’ MT systems can be created by training or fine-tuning engines using (large quantities of) literary texts. Kenny (2025) provides an overview of the latter research. Literary-adapted NMT systems that aim to emulate the style of an individual human translator are rare, but not unheard of. Hansen and Esperança-Rodier (2022), for example, showed that, notwithstanding the flaws in its output, a bespoke English-French NMT system fine-tuned on translations by Nathalie Serval produced texts bearing some lexical and syntactic features that were consistent with her style. Likewise, Yirmibeşoğlu et al. (2023) fine-tuned a pre-trained generic English-Turkish NMT model using data from a corpus of literary translations by Nihal Yeginobalı (1927-2020) and concluded that their fine-tuned model captured the human translator’s stylistic features “much better” than the pre-trained model. Yirmibeşoğlu et al. (ibid.) operationalized ‘style’ using a set of 29 textual features that are relatively easily quantified. These include word-level features (type-token ratio, number of unique words, mean word length, etc.), sentence-level features (mean sentence length, standard deviation of sentence lengths, etc.), morphological features (mean morphemes per sentence, etc.) and the frequency of certain forms or ‘keywords’, that is, forms found to be particularly common or uncommon in the translator’s work. These features were represented as normalized frequencies in a 29-dimensional vector for each text, allowing texts translated under different conditions to be easily compared. Yirmibeşoğlu et al. found that fine tuning on Yeginobalı’s works captured the translator’s style “18-40% better in terms of cosine similarity” than the pre-trained model. We share Yirmibeşoğlu et al.’s (2023) interest in quantifiable style and keywords, but are committed to working with living translators who might be able to interact with any models we build or outputs we produce, just as Nathalie Serval does in Hansen and Esperança-Rodier’s (2022) experiments, and consistent with Kenny’s

work with Hans-Christian Oeser (Kenny and Winters, 2020; Winters and Kenny, 2024).

2.2 LLMs, literary translation and style

With the launch of ChatGPT in late 2022 and the subsequent popularization of generative AI, research into AI-assisted literary translation shifted focus to LLMs, with studies usually: comparing LLM outputs to NMT and/or human translation (HT); testing LLMs’ ability to translate whole paragraphs or documents as opposed to isolated sentences; and testing various LLM finetuning and prompting strategies. Thai et al. (2022) and Karpinska and Iyyer (2023), both of which had Polish as one of their target languages, are good examples of this early work. The former used LLMs to post-edit NMT output; the latter to translate literary paragraphs directly. The applicability of LLMs to multiple tasks in the translation workflow has been embraced by Wu et al. (2025), who argue that multi-agent collaboration, as exemplified by their TRANSAGENTS framework, is particularly effective for translating longer literary texts. Zhang et al. (2025) problematize findings in this and related research on the basis that the evaluation methods they use are not always adequate for literary texts. Even in a source as thoughtful as Zhang et al. (2025), however, the notion of ‘style’ remains vague. The term designates both a category of translation error and a sometimes nebulous property of literary texts, referred to by the authors in their experimental prompts. Zhang et al. (2025: 10970), nevertheless, isolate ‘style’ as one of the features of literary texts to which future automatic evaluation metrics will have to attend.

2.3 Stylometry in translation analysis

Of the hundreds of research articles that now examine the interface of literary translation and NMT and/or LLMs, many mention style in passing, but few are primarily concerned with style. Those that do focus on style tend to make use of the quantitative methods developed in stylometry, the application of which to literary translation was pioneered by Jan Rybicki (Rybicki, 2006, 2012; Rybicki and Heydel, 2013), the translator whose work we study in this paper.

Examples include Lee (2022), who conducted a stylometric study of unigrams, bigrams and trigrams in NMT translations (by Google Translate, Bing Translation and Papago) and human

translations of modern Korean novels. But while Lee (ibid.) differentiated between NMT systems, ‘human translation’ remained an undifferentiated mass. Rybicki (2025), meanwhile, uses stylistic techniques to study English translations of classic French literature done by Google Translate and DeepL, on the one hand, and humans, on the other. While distributions of the most frequent words (MFW) in the texts under comparison were not found to distinguish between machine-generated and human translations, Rybicki finds that sentence length distributions can be used to discriminate between the two modes, for “more traditional” texts at least (ibid. 271). Kong and Macken (2025), for their part, initially considered 447 features in their study of three sets of translations into Chinese of *Peter Pan* (HT, NMT and LLM-based respectively). These features include lexical (e.g. TTR, lexical density, PoS tag distribution), syntactic (average sentence length, dependency structures), readability (measures designed to assess a text’s difficulty and comprehensibility for the target Chinese-speaking audience; measures of concreteness), and *n*-gram features (word-level or PoS uni-, bi- and tri-grams), to which they add features considered important in the translation of creative texts (e.g. related to repetition, rhythm, and translatability, as well as a ‘miscellaneous’ category). They find that their approach can distinguish relatively easily between human translations and automatic translations. Yao et al. (2025), finally, use stylistic techniques to compare human and LLM-based translations of Chinese online literature, finding that “GPT-4 translations closely mirror human translations in lexical, syntactic and content features” (ibid. 1)

While the above stylistic studies are rich in ideas on how to operationalize style, and how to conduct experimental research, providing guidance, for example, on prompting, none of them is interested in investigating whether the styles of an individual, named, human translator can be replicated by either NMT or LLMs. While we look to them for procedural guidance, we break new ground by applying stylistic techniques in the study of how LLMs can be used to mimic human translator styles. It should be noted at this point, however, that individual human translator style has been the subject of much research in translation studies since at least the early 2000s. Space restrictions prevent us from elaborating here. Instead, the interested reader is referred to (Kenny and Winters, 2024).

3 Research Design

This pilot study aims at establishing appropriate prompting strategies to produce AI-assisted, style-aware AI translations. Much recent attention has been dedicated to prompt engineering in LLM-based MT, with Tang et al. (2025) arguing that models are sensitive to prompts and that providing appropriate examples and instructions is essential for obtaining quality outputs. Our analysis thus evaluates the effectiveness of zero-, one- and few-shot prompting strategies in the creation of style-aware literary translation outputs. We structure prompts following the template provided (for few-shot prompts) by Tang et al. (ibid.). Our study further attempts to *quantify* – using stylistic techniques via the *stylo* package (Eder et al., 2016) – the extent to which we can personalize two LLMs using selected prompts. As outlined by Herrmann et al. (2015) and illustrated above, hundreds of features are available for the quantitative characterization of texts. In this study, and drawing on the studies reviewed earlier, we rely on average sentence length, sentence length distribution, lexical diversity based on standardized type-token ratios, and most frequent words. We also provisionally determine some features of Rybicki’s style based on the passages selected for use in our experiment and compare them to those of the two LLMs in use.

3.1 Selected LLMs and prompting strategies

This study compares two LLM systems – ChatGPT 5.4 and Claude Sonnet 4.6 – to assess whether either or both can be prompted to output translations in Jan Rybicki’s style. ChatGPT was selected due to its popularity and large market share, with web traffic analyses suggesting it is the most frequently visited LLM chatbot (Similarweb, 2026). Claude was chosen as it is described by Lokalise as the best for creative translation in testing rounds involving English to Polish translation (Wolff, 2026).

ChatGPT was accessed via Arena (arena.ai), a platform created to evaluate AI models (Chiang et al., 2024). Arena offers free access to the best ChatGPT model. To minimize potential bias resulting from previous interactions, a new Chrome Incognito session was opened to input a new prompt in a new Arena account. After selecting ‘Direct’ chat, ‘gpt-5.4-mini-high’ was chosen. Interactions with Claude were performed directly

on the claude.ai website using a standard free account with cross-chat memory disabled to reduce risk of bias from previous interactions. The best available model for generating translations with Claude was Sonnet 4.6 Adaptive. As this model was also the most advanced one available via Arena, we opted for direct interaction on the claude.ai website.

The prompts used in our interactions with the LLMs are listed in section 3.4, in our explanations of the various texts used in our stylometric analysis. As indicated above, the one-shot and few-shot prompts were designed based on the few-shot prompt in Tang et al. (2025).

3.2 Test data

As our source text (ST), we chose a 1,000-word passage (44 sentences) from Ceridwen Dovey’s debut novel *Blood Kin*, which had been previously translated by Rybicki but never published. This created a unique research opportunity where there is a high certainty that the human translator’s output had never been used to train any LLM systems. LLM training data is often obscured, with suspicions arising that non-open-access data has been used to train these models (Nature, 2026); therefore, any published translation may have already been used as training data, limiting the validity of evaluation of the LLM’s actual capabilities (can a model translate the text well because it’s good at translating or because it has access to an already existing translation of said text?). The ST was accessed through a publicly available ebook.

3.3 Bilingual examples used in LLM prompts

The bilingual examples used in our prompts were selected based on the above-mentioned corpus of Jan Rybicki’s translations. Within the corpus, we identified three translations of contemporary works that used diverse narrative styles (espionage realism, metafiction and speculative fiction). Selecting contemporary translations reduced the risk of diachronic stylistic variation in Rybicki’s work. The three translations were based on source texts (STs) by three different authors to minimise ST-author stylistic influence. All three ST examples were accessed using publicly available ebooks. The excerpts from each ST were selected from the first few pages of the work, in such a way as to ensure the structural integrity of the excerpts, and to allow continuous sections of around 400 words (see Table 1). The

translation of the selected excerpt was then located in Rybicki’s corpus.

source text	author	excerpt length
Agent Running in the Field	John le Carré	359 words
The Gum Thief	Douglas Coupland	407 words
Tanglewreck	Jeanette Winterson	454 words

Table 1. ST-data used with Rybicki’s translation in prompts

3.4 Translated texts

We saved the translated texts on which our stylometric analysis is based under the following filenames (with the prompt that elicited each translation (if applicable) given in the parentheses):

- Rybicki (Jan Rybicki’s human translation)
- GPT_control (ChatGPT’s output prompted as follows: “Translate the below text from English to Polish”)
- GPT_style.prompt (ChatGPT’s output prompted as follows: “Translate the below text to Polish in Jan Rybicki’s translation style”)
- GPT_1shot (ChatGPT’s output prompted as follows: “I will give you examples of text fragments, where the first one is in English and the second one is Jan Rybicki’s translation of the first fragment to Polish. These sentences will be displayed below. ... Translate the below text to Polish in Jan Rybicki’s translation style.” with one example of a passage from Le Carré’s ‘Agent Running in the Field’ and Rybicki’s translation of that passage into Polish).
- GPT_3shot (ChatGPT’s output prompted as follows: “I will give you examples of text fragments, where the first one is in English and the second one is Jan Rybicki’s translation of the first fragment to Polish. These sentences will be displayed below. ... Translate the below text to Polish in Jan Rybicki’s translation style.” with three examples of passages from: Le

Carré’s ‘Agent Running in the Field’, Coupland’s ‘The Gum Thief’ and Winterson’s ‘Tanglewreck’, and Rybicki’s translations of these passages into Polish).

- `Claude_control` (Claude’s output prompted as follows: “Translate the below text from English to Polish”)
- `Claude_style.prompt` (Claude’s output prompted as follows: “Translate the below text to Polish in Jan Rybicki’s translation style”)
- `Claude_1shot` (Claude’s output prompted as follows: “I will give you examples of text fragments, where the first one is in English and the second one is Jan Rybicki’s translation of the first fragment to Polish. These sentences will be displayed below. ... Translate the below text to Polish in Jan Rybicki’s translation style.” with one example of a passage from Le Carré’s ‘Agent Running in the Field’ and Rybicki’s translation of that passage into Polish).
- `Claude_3shot` (Claude’s output prompted as follows: “I will give you examples of text fragments, where the first one is in English and the second one is Jan Rybicki’s translation of the first fragment to Polish. These sentences will be displayed below. ... Translate the below text to Polish in Jan Rybicki’s translation style.” with three examples of passages from: Le Carré’s ‘Agent Running in the Field’, Coupland’s ‘The Gum Thief’ and Winterson’s ‘Tanglewreck’, and Rybicki’s translations of these passages into Polish).

3.5 Stylometric analysis

Eder et al.’s (2016) R Studio package ‘stylo’ was used to quantitatively analyse and compare AI outputs with Rybicki’s translation. A 100 MFW cluster tree was created to assess whether model outputs are stylistically similar to Rybicki’s translation based on their distributions of most frequently-used words. Further inquiry into word frequency resulted in a bootstrap consensus tree (BCT) to validate the findings. To investigate lexical diversity, the type-token ratio metric was applied. Lastly, sentence length was approached from two angles: sentence length distribution and average sentence length analyses were conducted

to discover Rybicki’s stylistic tendencies. All processing was based on default settings in *stylo*.

4 Results

Both 100 MFW (Fig. 1 in the Appendices) and BCT (Fig. 2) analyses clearly confirm that Rybicki’s translation is set completely apart from all LLM-based machine translations; in the ‘corpus cluster analysis’ his text joins the tree at a distance of over 2.0, which is the maximum value possible and shows how far the text is from the MT outputs. The Delta distance matrix is further evidence of Rybicki’s translation’s outlier position with distances between 1.46–1.65. Within the group of MT outputs, both GPT and Claude show a symmetrical pattern: three translations from each system cluster within their own model family. The BCT across multiple word frequency ranges (100–500 MFW) further verifies the findings (Fig. 2). While the MFW100 figure presents some grouping patterns among LLM outputs, the BCT shows weak cluster stability under resampling. This suggests that examined prompts did not produce consistent stylometric profiles, although this may be amplified by output sizes below 1,000 words. The BCT further supports the argument that prompt conditioning does not reliably bring LLM outputs closer to the human translator’s style. In short, based on Burrow’s Delta method, none of the prompts, regardless of the strategy, brought any LLM output significantly closer to Rybicki’s style.

To further inspect the word frequency, an analysis of most frequent words was run. Among the results, two words were of particular interest due to their relatively similar use among models and striking distance from Rybicki’s text, namely, the words ‘jego’ and ‘się’. Both words are pronouns, ‘jego’ is a third-person masculine singular pronoun, whereas ‘się’ is a common reflexive particle, indicative of the Polish language. These stylistic features distinguish Rybicki’s texts from the AI outputs. Possessive pronouns are often naturally omitted in Polish, but AI models tend to translate them literally, suggesting English interference. Moreover, Rybicki translated chapter titles ‘His portraitist’, ‘His chef’ and ‘His barber’ omitting the pronouns (‘Portrecista’, ‘Kucharz’ and ‘Fryzjer’), exercising his translator’s agency by making an active decision – which none of the models replicated in its output. In Rybicki’s translation ‘jego’ has a relative frequency of 0.83%, which is significantly lower than in the

LLM translations, which range from 1.53% to 2.67%, with two of the ChatGPT outputs using the word three times more frequently than Rybicki (Fig. 3).

A reverse phenomenon can be observed in Fig. 4 presenting the frequency of the word ‘się’. As ‘się’ is indicative of the Polish language, less frequent usage may suggest less natural text. Rybicki’s significantly higher usage of this word (2.97%) points to a more idiomatic use of structures such as reflexive verbs or impersonal constructions, characteristic of Polish prose. Typography and punctuation can be indicative of one’s personal style. Some LLM models use certain punctuation marks so much that it becomes an indicative feature of their outputs (Przystalski et al., 2026). Certain stylistic features of LLM outputs are particularly apparent. For example, recent research notes that LLMs are characterised by an increased use of em dashes compared to traditional human writing, potentially resulting from specific training data, tokenization and reinforcement learning processes (Hawkinson, 2025). Indeed, all genAI translations in this corpus contained dashes (see Fig. 5), whereas Rybicki’s text does not use them at all, favouring conjunctions or splitting sentences into shorter ones. Interestingly, ChatGPT outputs use 4 to 5 dashes within the whole text, but Claude’s versions range from 6 to 16 dashes, with the few-shot prompt resulting in the lowest value of 6 (whereas ChatGPT’s few-shot prompt resulted in one more dash than the rest of the outputs).

To analyse the lexical diversity of the outputs and compare them with Rybicki’s translation, the TTR (Fig. 6) measure was applied. The results indicate similarity in terms of lexical variety, with all models’ outputs displaying similar diversity to Rybicki’s version (values between 0.627 and 0.672, with Rybicki’s text having a value of 0.63). Although all texts have similar values, it’s interesting to see Claude’s scores are consistently higher than GPT’s ones, suggesting Claude’s higher vocabulary richness and potential wordiness.

On a sentence level, Rybicki’s translation is distant from all AI outputs across various metrics. On average, as presented in Fig. 7, his sentence length is significantly shorter than LLM ones, with a mean sentence length of 15.38 words for his translation and between 17.87 (one-shot prompted Claude’s output) and 19.68 (few-shot prompted ChatGPT’s output) for MT versions.

This pattern can be further observed in sentence length distribution (see Fig. 8), where the short sentence peak is the highest for Rybicki’s text and a steeper decline toward longer sentences indicates a differentiating density of concise sentences across his translation and a smaller number of longer sentences overall. As for absolute values, Rybicki’s translation has 26 short sentences of 10 words or less and only 8 long sentences of 25 words or more. To contrast it with the AI outputs, all models produced between 19 and 22 short sentences and between 14 and 16 long sentences. This clearly indicates Rybicki’s stylistic preference for shorter sentences in comparison with the LLMs’ more unified sentence length distribution. Lastly, the total number of sentences per translation yet again indicates Rybicki’s stylistic distance from all AI-generated outputs. His text consists of 58 sentences, whereas the AI versions have between 47 and 50 sentences, which is consistent with Rybicki’s tendency to use shorter sentences (for example, through splitting longer sentences into shorter ones) leading to the higher total number. (Recall that the ST contains 44 sentences.)

5 Discussion

Having inspected the common quantitative metrics with stylo, the results suggest that no prompting strategy resulted in stylistic convergence. Apart from the token-type ratio, where all outputs have similar lexical diversity (with Claude showing slightly higher and ChatGPT equal or slightly lower richness compared with Rybicki’s translation), all analysed metrics isolate Rybicki’s style from LLM outputs. As suggested through Burrow’s Delta and BCT metrics, the human translator remains an outlier, with his text being structurally different from eight AI-generated machine translations. AI models, on the other hand, are more stylistically stable internally with both ChatGPT and Claude outputs creating their own clusters in terms of lexical diversity and most frequently used words. The observed stylistic differences are driven more by the system’s internal behaviour than prompt variation, as the generated outputs exhibit a relatively stable style across four prompting strategies.

We investigated various prompts: two zero-shot prompts where one asked LLMs to generate an English to Polish translation and the other asked them to do so in Jan Rybicki’s translation style; one one-shot prompt and one few-shot

prompt with 3 examples of Rybicki's past translations. None of the prompting strategies led to an output that would be stylometrically related to the human translation, maintaining the stylistic space between the HT and MT. These findings suggest that style-related conditioning through prompting does not result in measurable improvement in output distribution, with the prompting instructions failing to override the underlying text-generation processes of the studied models. These findings are consistent with existing research on LLM style in creative texts, such as O'Sullivan (2025), which concludes that human creative writing is distinctive from LLM outputs, which tend to cluster together tightly by model and exhibit less stylistic diversity.

To complement this quantitative analysis, the next step would be to inspect the selected translations qualitatively using non-statistical metrics like storytelling or the use of rhetorical devices, such as irony, to determine whether they are in any way replicated in the MT versions.

6 Conclusions

This pilot study aimed at investigating whether two LLMs, ChatGPT and Claude, are capable of producing stylistically-aware machine translations imitating a specific literary translator's voice – in this case that of Jan Rybicki. The idea was not to reproduce exactly the translator's stylometric profile using LLMs, but rather to use stylometry to test whether prompting LLMs in given ways gets them to produce output that gets closer to the human translator's stylometric profile. If such prompts were effective in getting the LLM in question to approximate our translator's style, then this would suggest an avenue that could be pursued in our ongoing efforts to personalise CALT systems for use by literary translators. The results clearly show that none of the tested methods resulted in an improvement of stylometric distance from Rybicki's text, consistently indicating that human translation remains distinctive from AI translation. This suggests that prompting may not be an effective strategy to improve stylistic awareness of LLM-based translation. The study contributes to ongoing research on AI-assisted literary translation and demonstrates, through the lens of stylometric measures, the existing limitations of zero-, one- and few-shot prompting techniques in AI-generated translation.

References

- Chiang, Wei-Li, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. <https://doi.org/10.48550/arXiv.2403.04132>
- Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. Stylometry with R: A Package for Computational Text Analysis, *The R Journal*, 8: 107–21.
- Equity. 2026. Don't Sign On Set: 8 Steps to protect yourself from AI exploitation. <https://www.equity.org.uk/advice-and-support/know-your-rights/ai-toolkit/dont-sign-on-set-8-steps-to-protect-yourself-from-ai-exploitation> (Accessed: 21 April 2026).
- Hansen, Damien, and Emmanuelle Esperança-Rodier. 2022. Human-Adapted MT for Literary Texts: Reality or Fantasy? In Sheila Castilho, Rocio Caro Quintana, Maria Stasimioti, and Vilelmini Sosoni (eds) *NeTTT 2022*, 178–190, Rhodes, Greece. [hal-04038025](https://doi.org/10.1007/978-94-007-0403-8_025)
- Hawkinson, Eric. 2025. The Em Dash as a Site of Contest Between AI Determinism and Human Agency. *Together Research*, vol. 2, no. 1. <https://doi.org/10.62883/FILJ2955>
- Herrmann, J. Berenike, Karina van Dalen-Oskam, and Christof Schöch. 2015. Revisiting Style, a Key Concept in Literary Studies. *Journal of Literary Theory* 9 (1): 25–52.
- Karpinska, Marzena and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. <https://doi.org/10.48550/arXiv.2304.03245>
- Kenny, Dorothy. In preparation. Do it in Style! Human and automatic translation revisited.
- Kenny, Dorothy. 2025. Literary machine translation: from taboo to controversy. In Stefan Baumgarten and Michael Tieber (eds) *The Routledge Handbook of Translation Technology and Society*. 381–395, Routledge, New York. 10.4324/9781003271314-33
- Kenny, Dorothy, and Marion Winters. 2024. Customisation, Personalisation and Style in Literary Machine Translation. In *Translation, Interpreting and Technological Change: Innovations in Research, Practice and Training*. Ursula Böser, Sharon Deane-Cox, and Marion Winters (eds). 59–79. Bloomsbury, London. <https://doi.org/10.5040/9781350212978>
- Lee, Changsoo. 2021. How do machine translators measure up to human literary translators in stylometric tests? *Digital Scholarship in the Humanities*

- 37(3): 813–829.
<https://doi.org/10.1093/lc/fqab091>
- O’Sullivan, James. 2025. Stylometric comparisons of human versus AI-generated creative writing. *Humanities and Social Sciences Communications*, 12(1), 1708. <https://doi.org/10.1057/s41599-025-05986-3>
- Przystalski, K., J. K. Argasiński, I. Grabska-Gradzińska, and J. K. Ochab. 2026. Stylometry recognizes human and LLM-generated texts in short samples. *Expert Systems with Applications*, 296, 129001.
<https://doi.org/10.1016/j.eswa.2025.129001>
- Rybicki, Jan. 2006. Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz’s Trilogy and Its Two English Translations. *Literary and Linguistic Computing* 21 (1):91–103.
- Rybicki, Jan. 2012. The Great Mystery of the (Almost) Invisible Translator: Stylometry in Translation. In *Quantitative Methods in Corpus-Based Translation Studies*, edited by Michael Oakes and Meng Ji, 231–248. John Benjamins.
- Rybicki, Jan. 2025. Can Machine Translation of Literary Texts Fool Stylometry? *Digital Scholarship in the Humanities* 40 (1): 268–276.
- Rybicki, Jan, and Magda Heydel. 2013. The stylistics and stylometry of collaborative translation: Woolf’s *Night and Day* in Polish. *Literary and Linguistic Computing*, 28(4):708–717.
<https://doi.org/10.1093/lc/fqt027>
- Similarweb. 2026. Top AI Chatbots and Tools Websites Ranking. Available at: <https://www.similarweb.com/top-websites/ai-chatbots-and-tools/> (Accessed: 21 April 2026).
- Tang, Minghua, Chen Bian, Liming Yang, and Xueling Zhong. 2025. Key-concept thinking prompting for improved reasoning in large language models. *Neurocomputing*, 656:130986,
<https://doi.org/10.1016/j.neucom.2025.130986>.
- Thai, Katherine, Marzena Karpinska, Kalpesh Krishna, William Ray, Moira. Inghilleri, John Wieting and Mohit Iyyer. 2022. Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9882-9902, Association for Computational Linguistics, <https://aclanthology.org/2022.emnlp-main.672.pdf>
- Winters, Marion and Dorothy Kenny. 2024. Mark my keywords: a translator-specific exploration of style in literary machine translation. In Andrew Rothwell, Andy Way and Roy Youdale (eds) *Computer-Assisted Literary Translation*. 69—87, Routledge, New York. 10.4324/9781003357391-5
- Wolff, Rachel. 2026. What is the best LLM for translation? A comparison of top AI translation models. Lokalise Blog. Available at: <https://localise.com/blog/what-is-the-best-llm-for-translation/> (Accessed: 21 April 2026).
- Wu, Minghao, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wan, Weihua Luo, Kaifu Zhang. 2025. (Perhaps) Beyond Human Translation: Harnessing Multi-Agent Collaboration for Translating Ultra-Long Literary Texts. *Transactions of the Association for Computational Linguistics*, vol. 13, pp. 901–922. MIT Press, Cambridge, MA.
<https://doi.org/10.1162/tacl.a.25>
- Yao, Xiaofang, Yong-Bin Kang and Anthony McCosker. 2025. Missing the human touch? A computational stylometric analysis of GPT-4 translations of online Chinese literature. *Translation Spaces*.
<https://doi-org.dcu.idm.oclc.org/10.1075/ts.24043.yao>
- Yirmibeşoğlu, Zeynep, Olgun Dursun, Harun Dallı, Mehmet Şahin, Ena Hodzik, Sabri Gürses, and Tunga Güngör. 2023. Incorporating human translator style into English-Turkish literary machine translation, [arXiv preprint arXiv:2307.11457](https://arxiv.org/abs/2307.11457)

Appendices

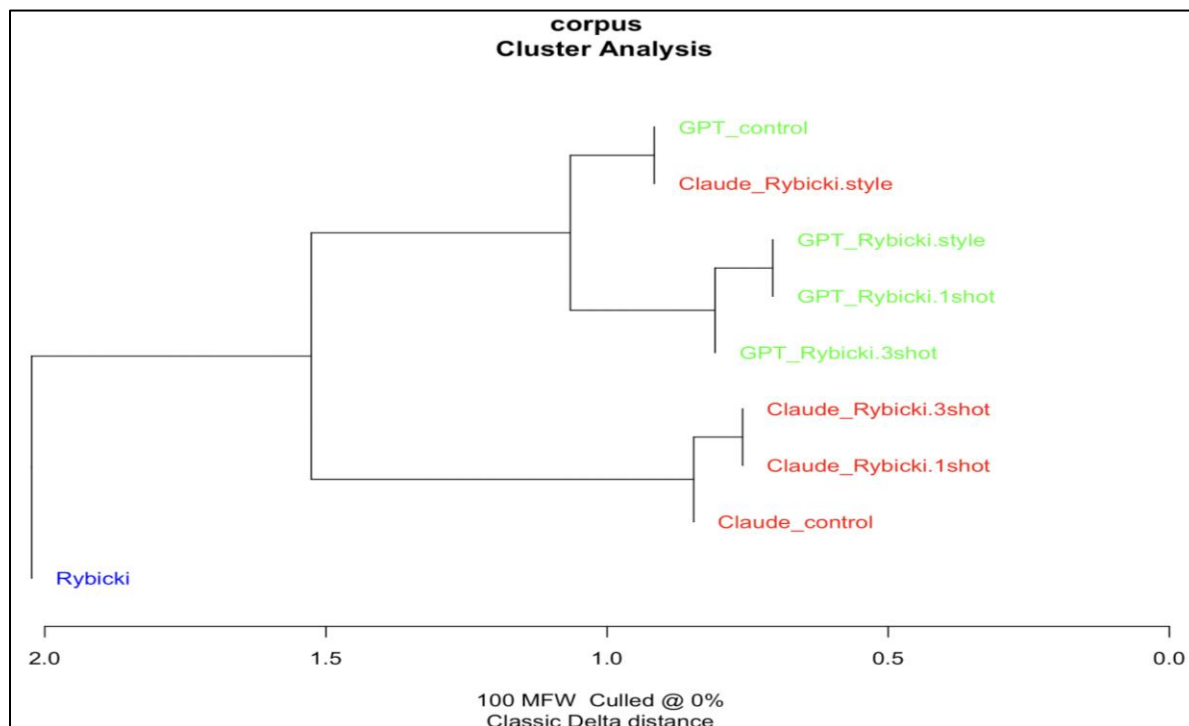


Fig. 1: 100 MFW Cluster Analysis

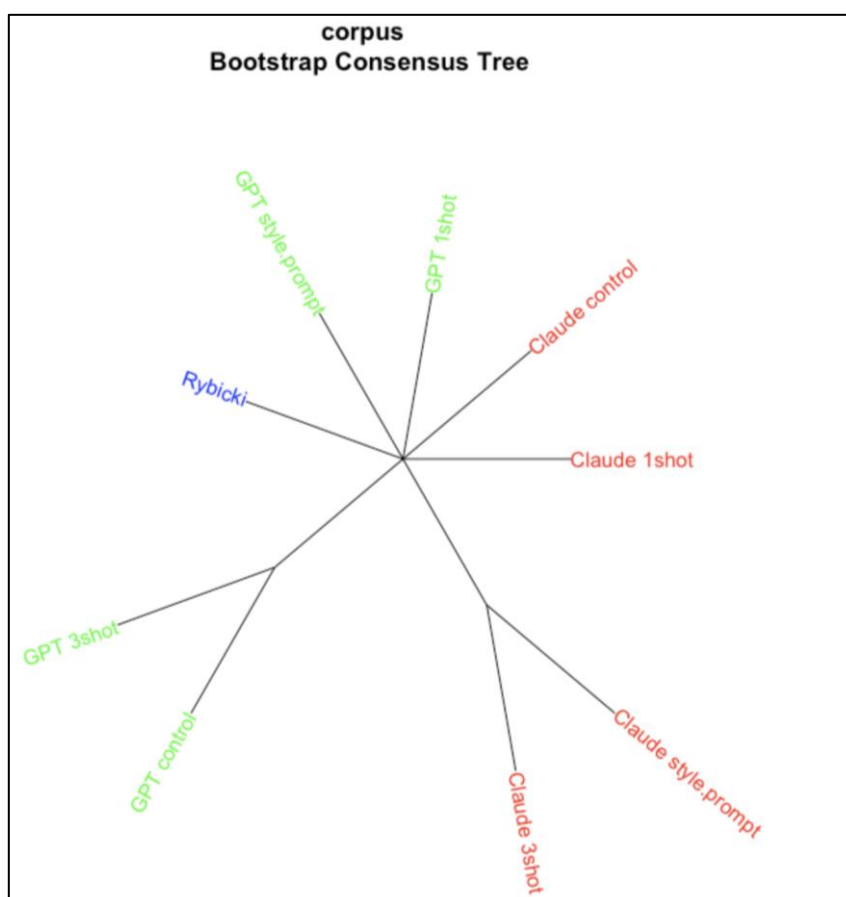


Fig. 2: Bootstrap Consensus Tree (BCT)

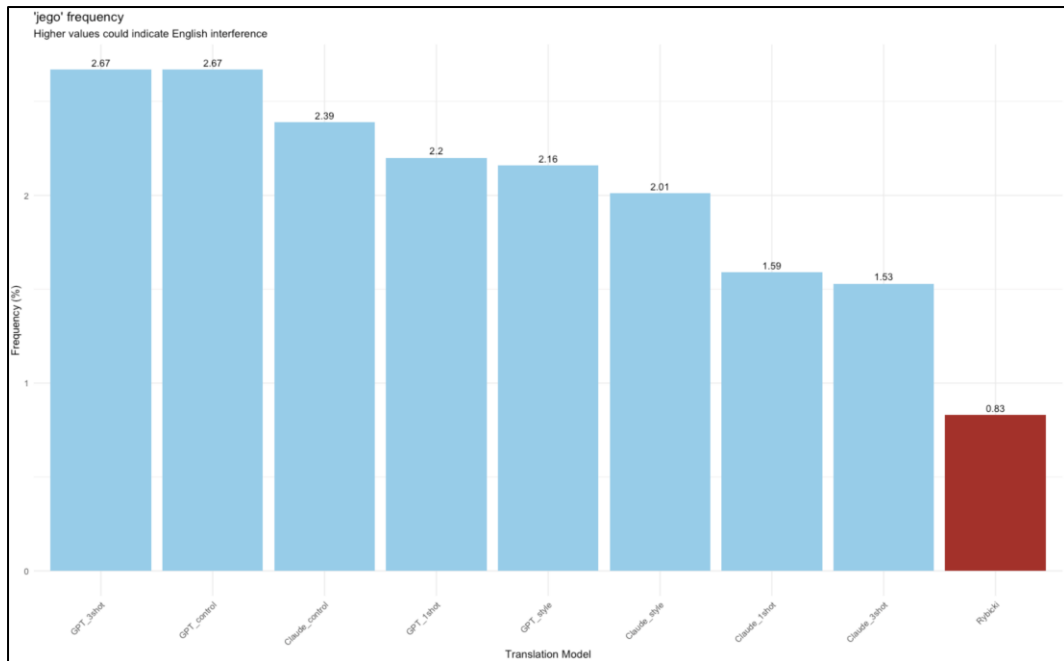


Fig. 3: Frequency of the pronoun 'jego'

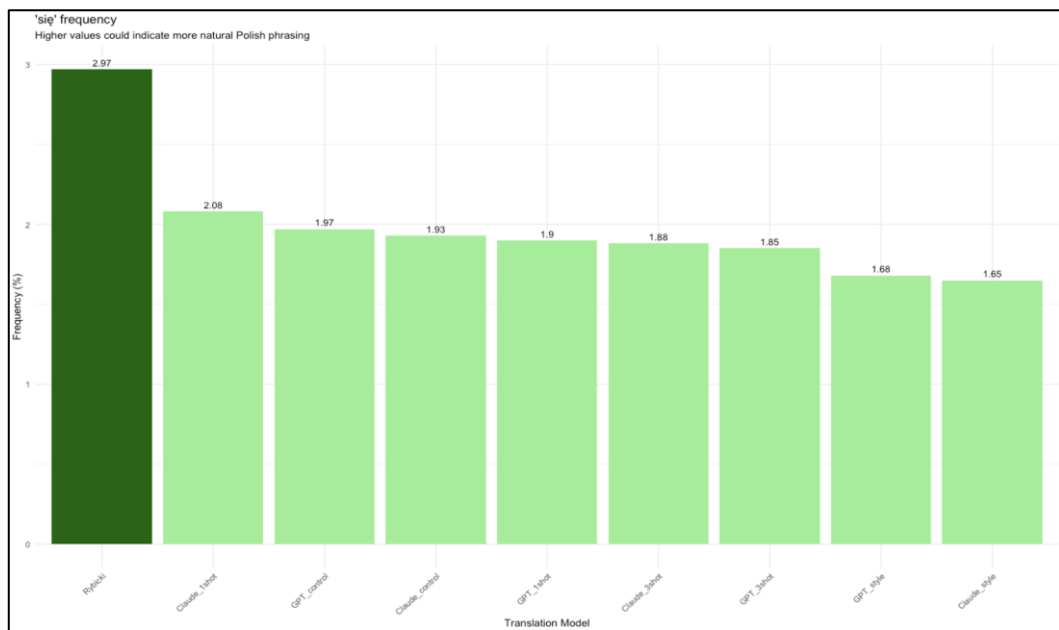


Fig. 4: Frequency of the particle 'się'

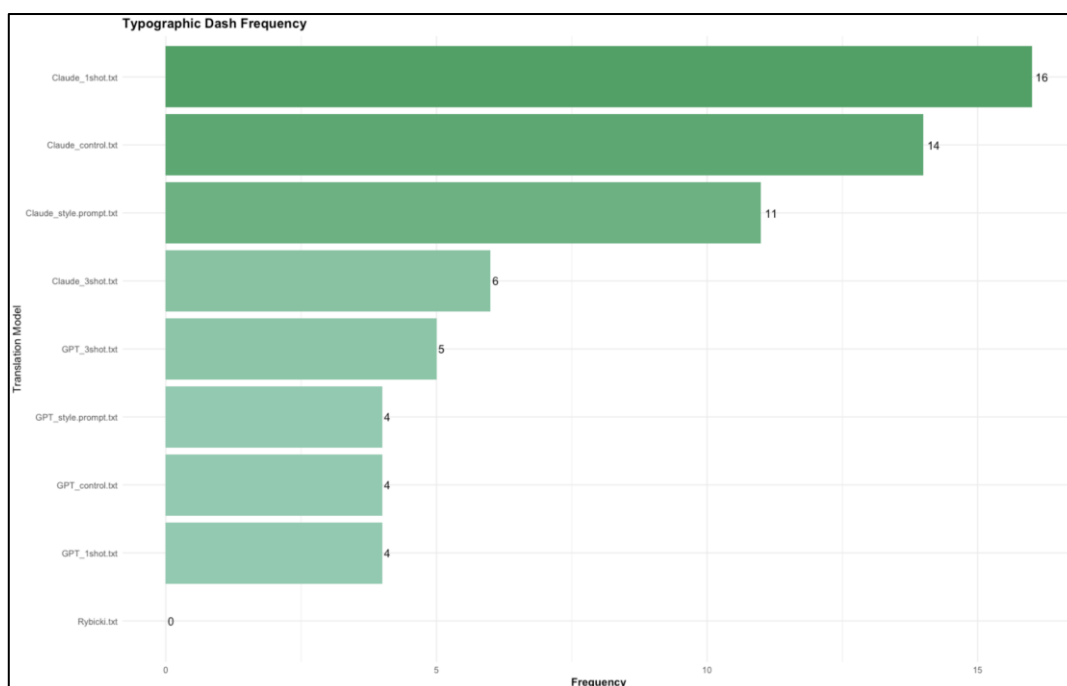


Fig. 5: Frequency of dashes

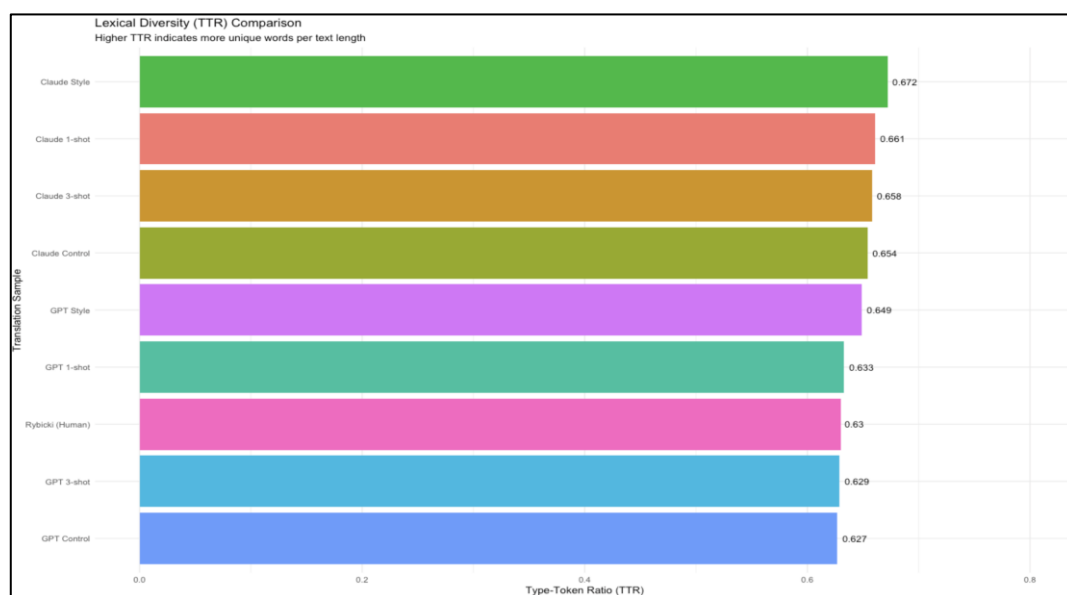


Fig. 6: Model Token-Type Ratio Comparison

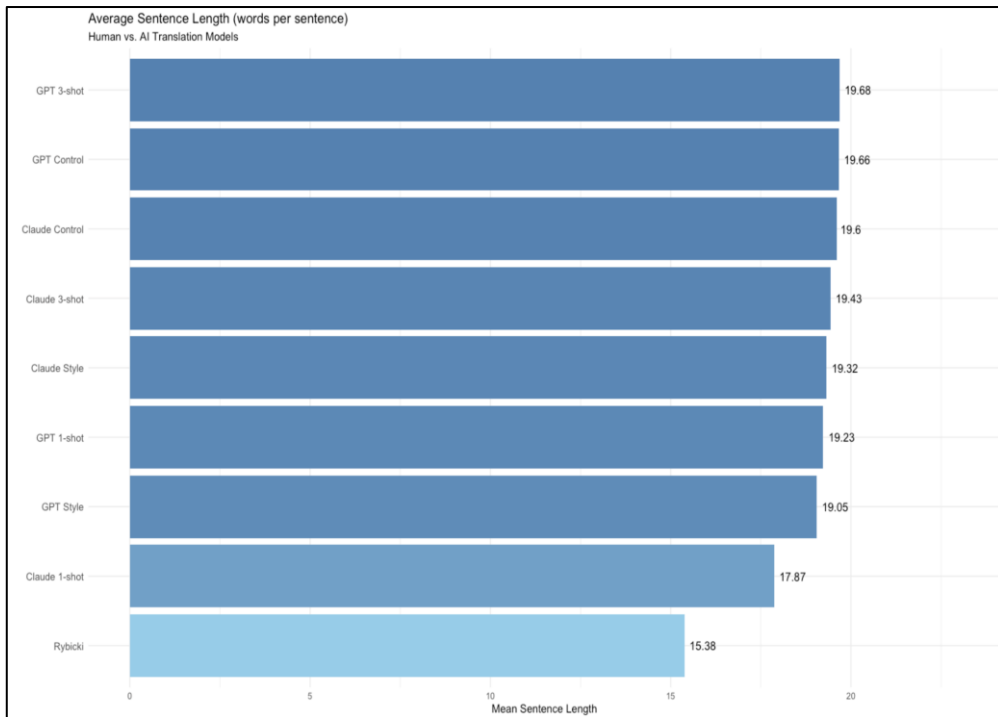


Fig. 7: Average sentence length

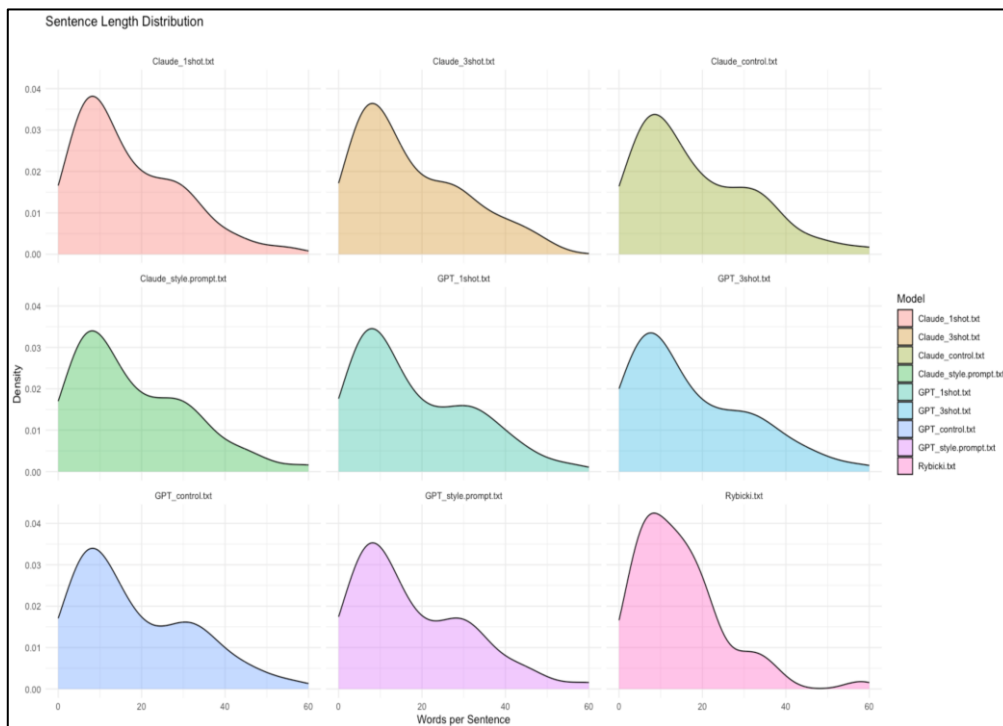


Fig. 8: Sentence length distribution

Lexical Variation in English–Italian News Translation: A Comparative Study of Google Translate and ChatGPT-5.2

Aurora Trapella^{1,2} Lieve Macken² and Alessandra Molino¹

¹Department of Foreign Languages, Literatures, and Modern Cultures, University of Turin, Italy

²Language and Translation Technology Team (LT³), Department of Translation, Interpreting and Communication, Ghent University, Belgium

¹aurora.trapella@unito.it

²lieve.macken@ugent.be

³alessandra.molino@unito.it

Abstract

This study investigates stylistic features in English–Italian machine translation (MT) by comparing the outputs of Google Translate (GT) and ChatGPT-5.2 (GPT). The analysis adopts a quantitative corpus-based approach to studying news reports and opinion articles for which parallel translations produced by both systems were collected via API. In particular, the focus is on the lexical variety of the translation equivalents of three high-frequency mental verbs: *know*, *think*, and *believe*. Concordance lines were analyzed and annotated according to the equivalents used, enabling a systematic comparison of lexical range across systems and the two news genres. The results show that the two systems seem to adopt similar translation solutions, but they differ in the quantitative distribution of translation equivalents. While GT exhibits greater sensitivity to morphological variation, GPT tends to rely on a more restricted set of equivalents and a wider range of phrasal and clausal reformulations. These findings contribute to studies on machine translationese (MTese) for the English–Italian language pair, with implications for the study of variation in MT and Artificial Intelligence (AI) literacy.

1 Introduction

This study investigates the stylistic features of automatic translations of a corpus of news texts comprising the two subgenres of news reports and opinion articles. The analysis focuses on news translated from English into Italian using Google Translate – a Neural MT (NMT) tool – and ChatGPT-5.2 – a general-purpose Large Language Model (LLM). Studies on machine-translated language often refer to it as an “artificially impoverished” language (Vanmassenhove et al., 2021), also named machine translationese (MTese) (Daems et al., 2021; De Clercq et al., 2021), which has become a widely investigated topic in disciplines such as Computational Linguistics, Natural Language Processing (NLP), and Translation Studies. Despite extensive research on the features of statistical and neural MT, the language output produced by LLMs has only recently begun to attract scholarly attention (Kong and Macken 2025b). However, to the best of our knowledge, little to no research has been carried out on the English–Italian language combination, although recently the WMT Shared Task (Kocmi et al., 2025) reported results indicating that news translated from English into Italian by the top-performing systems still ranked in the lowest positions. This underscores the necessity for additional research to be conducted on the characteristics of machine-translated Italian, with a view to enhancing the performance of training models and

supporting the training of translators and post-editors to recognize machine interference and preserve idiomatic language.

Such investigations are also required in light of the increasing use of AI tools in journalistic translation workflows and other non-professional settings. With the increasing integration of AI tools into different phases of the journalistic value chain (Cools and Diakopoulos, 2024), MT tools have become fundamental to journalistic workflows, and more recently, Generative AI (GenAI) tools, especially chatbots like OpenAI's ChatGPT, have been incorporated into journalistic practices for translation purposes (Roush 2023), with reference as well to the Italian context (Cagnan, 2025). While some news outlets rely on professional post-editing (Bullard 2023) or on in-house editing and journalist review ('journalist-translator') (Cagnan, 2025), others rely on fully automated translation (Roush, 2023) and on automatic content generation through transcription and automatic speech recognition (ASR) (The New York Times Trust Team 2024). However, there seems to be a lack of a theoretical framework due to the opportunistic use of AI technologies on a case-by-case basis (Caswell, 2019), together with a difficulty in acquiring AI literacy (Murru and Carlo, 2024).

Without the development of AI literacy (Bowker 2024), over-reliance on these tools by journalists and editors might result in information drift and the unintentional introduction of bias and stance-altered expressions, while lay users who use NMT tools or plugins (e.g., Google Translate) and LLMs (e.g., ChatGPT) to translate online news texts for personal use might encounter a less nuanced and potentially standardized language (Hermand, 2014).

Moreover, news discourse is characterized by frequent references to cognitive and epistemic states, as reported information is often mediated through evaluation and attribution. In this context, cognitive verbs as defined by Biber et al. (1999) play a key role in signaling degrees of certainty and perspective. These verbs may pose challenges for MT systems, which can struggle to capture subtle differences in epistemic force and contextual usage. For this reason, mental cognitive verbs are selected as the focus of the present study.

This study is an exploratory investigation within a larger project examining the linguistic and stylistic features of Italian MTese. The project has two main objectives: (1) to compare raw, non-

post-edited translations generated by NMT systems and LLMs in order to identify whether and how they differ in their rendering of English source texts (ST), considering that NMT is a translation-specific technology whereas LLMs were not originally developed for translation tasks; and (2) to compare the output produced by NMT and LLMs with a comparable corpus of Italian news texts written before the advent of AI. The overall aim of the project is to provide an empirical description of the features of Italian MTese and to identify possible signs of standardisation in machine-translated output (Hermand, 2014). The findings will have implications for translator and post-editor training, as well as for AI literacy development among both professionals (e.g. journalists and editors) and general users.

The present study aims to answer the following research questions:

- (1) How do Google Translate (GT) and ChatGPT (GPT) differ in lexical variation when translating English mental cognitive verbs (*know*, *believe*, *think*) into Italian?
- (2) Does genre play a role in the extent to which these stylistic tendencies differ?

2 Literature review

Style has long been recognized as a central dimension of linguistic expression, commonly defined as a writer's individual mode of expression. As Murtaza and Qasmi (2013) argue, style encompasses a wide range of choices operating along both paradigmatic and syntagmatic axes, including the choice of lexical items, the use of figurative language, syntactic and phrasal structuring, and the organization of discourse in paragraphs. From this perspective, stylistic variation emerges from systematic patterns of linguistic choice, which can be observed and measured across texts.

In the context of MT, several works point to the frequent reduction of lexical and stylistic variation by neural MT tools. Indeed, machine-translated language is often characterized by processes of standardization also referred to as 'lexical convergence' (Hermand, 2014), whereby target language (TL) output tends to use a simplified range of vocabulary and to conform to more neutral and conventional patterns. This tendency is closely related to findings that MT systems frequently reduce lexical and stylistic diversity. For instance, Vanmassenhove et al. (2019) show that NMT systems tend to generalize and avoid colloquialisms,

leading to a loss of nuance, while also exhibiting biases that favor more frequent lexical items at the expense of less common ones. This “algorithmic bias” contributes to a narrowing of the lexical repertoire available in translated output.

Further evidence of reduced diversity in MT comes from studies comparing translated texts with original language data. Translations produced by different MT architectures have been shown to be consistently less lexically diverse than their training data (Vanmassenhove et al., 2021) and to contain fewer low-frequency items such as hapax legomena, alongside occasional non-standard or non-sense forms (De Clercq et al., 2021). This reduction in lexical variation has been described as a form of lexical impoverishment, whereby the range of available equivalents is simplified (Espinosa-Giménez and Forcada, 2024). Similarly, Looock (2020) and Jiang and Niu (2022) observe that MT output tends to over-represent certain linguistic features and lexical items, often due to source language interference or a lack of sensitivity to TL norms. Moreover, as Webster et al. (2020) note, MT systems may produce translations that are semantically comprehensible but stylistically inadequate, failing to capture appropriate register and more idiomatic formulations.

At the same time, genre has been shown to play a significant role in lexical variation (Niu and Jiang, 2024). Moreover, recent research has begun to explore differences between NMT and LLMs. While NMT has been associated with tendencies toward standardization and reduced variation, LLM-based systems appear to exhibit greater stylistic flexibility. Kong and Macken (2025a) observe that LLM outputs contain a higher frequency of descriptive lexical items, which are generally linked to increased textual vividness and specificity. In a subsequent study, Kong and Macken (2025b) report higher levels of lexical diversity in LLM outputs compared to NMT, suggesting a broader and more varied lexical repertoire. Despite these advances, capturing stylistic nuance remains a challenge for MT systems, which continue to struggle in modelling context-sensitive stylistic choices (Daems, 2022). The majority of these studies focus on language combinations such as English–French, English–Dutch, or Chinese–English, while the English–Italian pair remains underexplored in this respect. Against this background, the present study contributes to the growing body of research on stylistic variation in MT by comparing how different systems realize lexical variation across news genres.

3 Methodology

3.1 Corpus description

The English STs corpus contains approximately 340,000 words of published news retrieved from two pre-existing resources: the Dutch Parallel Corpus (DPC) (Macken et al., 2010) and the EMA (“editorial material”) corpus (Molino, 2010), which were compiled for previous research projects. The texts contained in these two resources were published in similar time periods: DPC collects texts published between 2003 and 2008, while EMA contains texts published between 2002 and 2004. As the texts were published before AI-assisted writing and translation became widespread, their selection avoids potential contamination from machine-generated language. To ensure stylistic and register consistency, only news published by British broadsheet newspapers present in these two corpora was selected for the study. Table 1 provides a description of the ST corpus.

As the texts selected for the English STs corpus were encoded differently, they were converted to UTF-8 using a Python script, and a formatting function was added to insert missing spaces after full stops, improving readability and facilitating the subsequent analysis. Further preliminary processing produced clean plain-text versions, removing metadata from the DPC and manually cleaning the EMA corpus of headlines, dates, and author information.

The corpus was pre-processed through Python using the spaCy (Honnibal et al., 2023) NLP framework. The script implemented language-specific models (en_core_web_lg for the English STs and it_core_news_lg for the Italian translations) to perform tokenization, lemmatization, Part-of-Speech (POS) tagging and Named Entity Recognition (NER), while the dependency parser and sentence segmenter were disabled, as the input data was already pre-segmented into a one-sentence-per-line format. The data was then serialized into a structured JSON format to facilitate subsequent quantitative and comparative analysis.

	DPC (n. words)	EMA (n. words)	Overall ST corpus (n. words)
News reports	182,622	0	182,622
The Guardian	71,337	0	71,337
The Independent	111,285	0	111,285
Opinion articles	77,057	79,370	156,427
Financial Times	0	2,544	2,544
The Daily Telegraph	0	18,072	18,072
The Guardian	17,516	31,608	49,124
The Independent	59,541	14,922	74,463
The Spectator	0	12,224	12,224
Total n. of words	259,679	79,370	339,049

Table 1. ST corpus description: total number of words across news genres (news reports and opinion articles) and each of the original corpora (DPC and EMA).

The TT corpora to be examined within the quantitative analysis of concordance lines comprise translations performed automatically via Python scripts for the free version of Google Translate API and the licensed ChatGPT-5.2 API in March 2026. Python scripts included a two-second pause between requests and a retry mechanism of up to three times. These functions prevent overloading the systems and reduce the risk of incomplete translations. For large files, the scripts for both systems implement a paragraph-based chunking strategy to preserve contextual coherence during processing. The following prompt was given to perform the translation using the licensed ChatGPT-5.2 API: “You are a professional translator. Translate the following English text to Italian. Preserve the formatting, line breaks, and structure exactly. Only provide the translation, no explanations or additional text.”

3.2 Research design

This study adopts a quantitative corpus-based approach to analyze lexical and stylistic variation in MT output. The analysis is based on concordance lines extracted for a set of selected ST verbs and focuses on the distribution of translation equivalents. Specifically, it assesses the extent to which the two systems differ in their lexical variability, operationalized as the number of distinct translation equivalents associated with each verb, and what differences can be observed between the MT systems and the news genres. In addition, differences in the structural realization of translation choices across the two systems and genres are discussed.

As this research serves as an exploratory study to assess whether a quantitative distribution of translation equivalents can effectively distinguish between the stylistic features of tools with different architectures, we focus on a small set of verbs

to ensure accurate manual analysis. Such methodology will then be applied to a broader semantic range of verbs and translations produced by other tools in subsequent research phases.

3.3 Selection of linguistic variables

Mental verbs were selected as the focus of analysis based on their distribution across written registers in the *Longman Grammar of Spoken and Written English* (Biber et al., 1999). In their seminal work, Biber et al. (1999) argue that verbs can be classified according to their semantic domain, identifying seven categories: activity verbs, communication verbs, mental verbs, causative verbs, verbs of simple occurrence, verbs of existence or relationship and aspectual verbs. Overall, activity verbs constitute the largest proportion of verbs among the domains of conversation, fiction, news and academic prose, but their primarily descriptive function makes them less informative for a stylistic comparison. Similarly, among the two written domains, verbs of existence and relationship are more characteristic of academic writing and fulfil largely structural functions. By contrast, mental verbs describe activities and states experienced by humans, including cognitive, perceptual, and affective meanings. This category of verbs occurs more frequently in news discourse than in academic prose, making them particularly suitable for analysis, as they offer a more productive domain for observing stylistic variation in how information is presented and interpreted across different systems. The dominant sub-category in terms of frequency is that of mental verbs describing cognitive states, which are here labelled “stative cognitive verbs”.

Following pre-processing of the ST, all verb lemmas contained in the corpus were extracted using a custom Python script to create a frequency-based wordlist. From the corpus used for this

study, concordance lines were extracted for the three most frequent stative cognitive verbs shared across the two news subgenres, namely *know*, *think* and *believe*. These three verbs appear particularly interesting as their semantic generality provides a ‘blank canvas’ for lexical choice; because a single English verb can be mapped to a wide array of Italian equivalents and reformulations: i.e., *sapere*, *conoscere*, *intendersi di* for *know*; *pensare*, *credere*, *ritenere*, and *trovare* + adjective for *think*, and *credere*, *pensare*, *ritenere*, *secondo* + pronoun for *believe*. Moreover, some Italian equivalents are shared across the three English verbs, resulting in potential interchangeability in translation. Therefore, these verbs might provide interesting insights into the stylistic flexibility and lexical variety of the tools used for translations, assessing their potential for over-generalization or reformulation.

For each occurrence of these verbs, the corresponding translation equivalents were identified from concordance lines extracted from the TT corpora produced by both systems. The translation equivalents were subsequently annotated in a spreadsheet with part-of-speech information, and quantified to enable a systematic comparison of lexical variation across MT systems and news subgenres.

Results

Across the corpus, the three selected verbs occur 1,110 times in total (*know* = 478; *think* = 430; *believe* = 202). Semi-automatic sentence alignment issues excluded some concordances from the analysis; therefore, only 1,073 concordance lines were successfully analyzed (*know* = 454; *think* = 431; *believe* = 199). Translation equivalents are reported only for concordance lines in which the target verbs were successfully identified as such, due to occasional lemmatization and part-of-speech tagging issues in the automatic extraction process. In addition, instances in which the verbs formed part of proper names, such as song titles or movement names, were excluded from the analysis, as they were all retained in English. Based on these considerations, the frequencies of the three verbs are distributed as displayed in Table 2:

Verb	News reports	Opinion articles	Total frequency
know	268	184	452
think	196	227	423
believe	104	94	198
			1,073

Table 2. Distribution of the three selected verbs across the ST corpus.

Table 3, Table 4 and Table 5 present a frequency-based comparison of translation equivalents produced by the two MT tools (Google Translate and ChatGPT-5.2) across opinion articles and news reports, focusing on three high-frequency verbs (*know*, *think*, *believe*). ST frequencies of the verbs, as well as relative frequencies (RF) of the translation equivalents, are reported in the three tables.

The analysis reveals differences in how GT and GPT handle the translation of the three cognitive verbs across genres. A first general observation concerns the distribution of core equivalents. Across all datasets, both systems opt for the same translation equivalents, particularly *sapere* for *know*, *pensare* for *think* and *credere* and *ritenere* for *believe*.

For the verb *know* (Table 3), in opinion articles, both systems rely primarily on *sapere*, *noto*, and *conoscere*, but GT shows a wider variety of low-frequency items. GPT, on the other hand, maintains a tighter distribution while also introducing functionally equivalent paraphrases such as the verbal phrases *avere notizia* (literally ‘to have news on’, in this context ‘to hear from [someone]’) and *non avere idea* (‘to have no idea’). In news reports, GT again displays greater variety, including more reformulations.

know		News reports (n = 268)				Opinion articles (n = 184)					
GT		RF	GPT		RF	GT		RF	GPT		RF
sapere	V	[0.664]	sapere	V	[0.653]	sapere	V	[0.663]	sapere	V	[0.459]
conoscere	V	[0.134]	noto	ADJ	[0.138]	conoscere	V	[0.158]	conoscere	V	[0.168]
noto	ADJ	[0.123]	conoscere	V	[0.134]	noto	ADJ	[0.082]	noto	ADJ	[0.092]
conosciuto	ADJ	[0.049]	conosciuto	ADJ	[0.045]	conosciuto	ADJ	[0.022]	conosciuto	ADJ	[0.022]
capire	V	[0.007]	capire	V	[0.015]	capire	ADJ	[0.016]	capire	V	[0.016]
chissà	ADV	[0.007]	chiamare	V	[0.007]	riconoscere	V	[0.011]	chissà	ADV	[0.011]
scoprire	V	[0.004]	chissà	ADV	[0.004]	chissà	ADV	[0.011]	avere notizia	PHR	[0.005]
ignorare	V	[0.004]	individuare	V	[0.004]	conoscenza	N	[0.005]	definire	V	[0.005]
consapevolezza	N	[0.004]				risaputo	ADJ	[0.005]	non avere idea	PHR	[0.005]
essere a conoscenza	PHR	[0.004]				consapevolezza	ADJ	[0.005]	omission		[0.005]
						chiamare	V	[0.005]			
						essere a conoscenza	PHR	[0.005]			
						consapevole	N	[0.663]			
						omission		[0.663]			

Table 3. Translation equivalents for the verb *know* with GT and GPT across news reports and opinion articles.

Overall, *think* (Table 4) displays the greatest dispersion in both genres, and it is therefore the verb that lends itself more to a more varied type of translation. In news reports, *think* is translated predominantly as *pensare* in both systems, with very similar frequencies. Differences emerge in the distribution of secondary options. GT shows a preference for verbal alternatives such as *ritenere*, *credere*, and *considerare*, along with other lower-frequency verbs. On the other hand, in GPT, the main equivalents *credere* and *ritenere* are followed by more phraseological or interpretive renderings, such as *secondo* + subject (literally ‘according to...’), *avere dubbi* (‘to have doubts’), and *si direbbe* (‘one would say’). As for opinion articles, the main equivalent for *think* is *pensare*, with GPT showing a higher frequency (0,806 vs. 0,744 RF in GT). In contrast with news reports, here GT displays a wider range of lexical and phraseological alternatives, including *secondo* + subject, *venire in mente* (‘to come to mind’), *trovare* + adjective (literally ‘to find [sth] ...’, meaning ‘to think it is ...’), and *per come la vedo io* (‘the way I see it’), suggesting greater lexical dispersion also at the level of reformulations if compared to GPT.

The verb *believe* (Table 5) is the least variable item with the least pronounced differences between systems. Both GT and GPT rely predominantly on *credere* and *ritenere* in both genres. In the GPT equivalents for news reports, while *credere* reaches a RF of 0,615 the second most frequent option, *ritenere*, drops to 0,375. On the other hand, in GT, the difference between the frequencies of the same two equivalents is not so marked (*credere* 0,528 vs. *ritenere* 0,423). The data therefore suggest that GPT relies more heavily on a single dominant verb, whereas GT maintains a more balanced use of available alternatives, distributing frequencies across a wider range of alternatives. Independent of the two genres, both systems include reporting-style phrases such as *a suo avviso* (‘in his/her view’), and *secondo* + subject in GT, and *che piaccia o no* (‘like it or not’) in GPT for opinion articles.

think		News reports (n = 196)				Opinion articles (n = 227)					
GT		RF	GPT		RF	GT		RF	GPT		RF
pensare	V	[0.775]	pensare	V	[0.796]	pensare	V	[0.744]	pensare	V	[0.806]
credere	V	[0.071]	credere	V	[0.086]	credere	V	[0.114]	credere	V	[0.114]
ritenere	V	[0.082]	ritenere	V	[0.056]	consid- erare	V	[0.04]	ritenere	V	[0.441]
consid- erare	V	[0.03]	secondo + subject	PHR	[0.015]	ritenere	V	[0.035]	trovare + adjective	V	[0.013]
riflettere	V	[0.01]	considerare	V	[0.01]	riflettere	V	[0.013]	consid- erare	V	[0.009]
trovare + adjective	V	[0.01]	trovare + ad- jective	V	[0.01]	secondo + subject	PHR	[0.009]	riflettere	V	[0.004]
stimare	V	[0.005]	sbagliarsi	V	[0.005]	venire in mente	PHR	[0.009]	venire in mente	PHR	[0.004]
sbagliarsi	V	[0.005]	illudersi	V	[0.005]	trovare + adjective	V	[0.009]	chiedersi	V	[0.004]
illudersi	V	[0.005]	avere dubbi	PHR	[0.005]	affrontare	V	[0.004]	con- vinzione	N	[0.004]
<i>omission</i>		[0.005]	si direbbe	PHR	[0.005]	immagi- nare	V	[0.004]			
						concepire	V	[0.004]			
						chiedersi	V	[0.004]			
						per come la vedo io	PHR	[0.004]			
						chiedersi	V	[0.004]			
						<i>omission</i>		[0.004]			

Table 4. Translation equivalents for the verb *think* with GT and GPT across news reports and opinion articles.

believe		News reports (n = 104)				Opinion articles (n = 94)					
GT		RF	GPT		RF	GT		RF	GPT		RF
credere	V	[0.528]	credere	V	[0.615]	credere	V	[0.680]	credere	V	[0.723]
ritenere	V	[0.423]	ritenere	V	[0.375]	ritenere	V	[0.255]	ritenere	V	[0.244]
a suo av- viso	PHR	[0.019]	secondo + subject	PHR	[0.009]	secondo + subject	PHR	[0.021]	secondo + subject	PHR	[0.01]
pensare	V	[0.009]				a mio/suo avviso	PHR	[0.021]	pensare	V	[0.01]
secondo + subject	PHR	[0.009]				convinto	ADJ	[0.021]	che piaccia o no	PHR	[0.01]
convinto	ADJ	[0.009]									

Table 5. Translation equivalents for the verb *believe* with GT and GPT across news reports and opinion articles.

As regards structural variation in the translation strategies adopted by the two systems (Table 6), the data show that GT more frequently preserves a direct verb-based (V-V) correspondence (GT 16 vs. GPT 14) across both genres. At the same time, GT demonstrates a higher tolerance for morphological variation, including nominalization (verb to noun; V-N), adjectival (verb to adjective, V-ADJ) and adverbial forms (verb to adverb; V-ADV). While these forms are not absent in GPT, they tend to occur less frequently and with lower diversification. For instance, it is interesting to point out that GPT uses nominalization only once across both genres. By contrast, it more readily introduces phraseological reformulations. This is

visible in the use of prepositional phrases (PP) (*secondo + subject*, *a suo avviso*) and clausal constructions (CLAUSE) (*si direbbe*, *che piaccia o no*), particularly in the translation of *think*; meanwhile, GT produces a less varied output in terms of phraseological reformulations, with the most frequent of them being the construction *secondo + subject*.

	News reports		Opinion articles	
	GT	GPT	GT	GPT
V-V	16	14	16	14
V-ADJ	4	3	5	2
V-N	1	0	2	1
V-ADV	1	1	1	1
V-	3	4	5	5
PP/CLAUSE				

Table 6. Structural variation of translation strategies, measured as the number of distinct translation equivalents provided by each MT system across the two news genres.

Moreover, some differences with respect to genre conventions might also be observed. In news reports, both systems show a strong preference for reporting verbs such as *sapere*, *credere*, *ritenere*, and *pensare*, with limited variation and a high degree of overlap between GT and GPT outputs. Therefore, the data might be interpreted as reflecting the informational and objective style of the news report genre. In opinion articles, by contrast, the output is more varied with regard to reporting verbs, also including more word-class transformation and the presence of phrasal or clausal constructions. The use of a wider range of lexical items and more frequent fixed phrases and expressions can be observed in the output of both systems, such as for *per come la vedo io*, *a suo avviso*, or *venire in mente*. Variation increases particularly in the GT output for *think*, in which many more equivalents are provided compared to GPT.

Overall, these results point to minor differences in MT output. GT shows lexically diverse equivalents, often relying slightly more on word-class transformation. GPT, by contrast, exhibits a more interpretive style, favouring reformulations while reducing lexical dispersion. Genre conventions seem to be reflected in the output of both systems, with opinion articles having a wider range of equivalents and especially phrasal and clausal reformulations, which refer to the more argumentative structure of the ST.

4 Conclusion

This study has examined lexical variation in English–Italian MT by analyzing the distribution of translation equivalents for three high-frequency stative cognitive mental verbs in outputs produced by GT and GPT by answering the following research questions:

(1) How do GT and GPT differ in lexical variation when translating English mental cognitive verbs - *know*, *believe*, *think* –into Italian?;

(2) Does genre play a role in the extent to which these stylistic tendencies differ?

The results show that the two systems differ slightly in their stylistic realization, with GT exhibiting greater sensitivity to morphological variation and producing a wider range of translation equivalents. GPT, by contrast, tends to rely on a more restricted set of equivalents, resulting in a more lexically concentrated output and a wider range of phrasal and clausal reformulations. While GT tends to distribute meaning across a broader set of lexical choices, GPT appears to achieve variation more frequently through restructuring at the phrase and clause level rather than through expansion of lexical alternatives. Furthermore, the results might suggest some sensitivity to news genre: opinion articles display greater lexical and structural variation than news reports in both systems, indicating that textual conventions present in the ST might be relevant enough to shape the range of translation choices adopted by the systems. This is particularly evident with the equivalents of the verb *think*, which show the highest degree of variation, while *believe* remains comparatively stable across both systems.

This study has several limitations. First, the analysis is restricted to a small set of verbs, which, albeit frequent and representative of stative cognitive meaning, cannot capture the full range of lexical behavior in MT. Second, technical constraints might have influenced the results, such as the parameters set in the Python script when performing the API and in the prompt. Third, the focus on quantitative measures of lexical variety does not account for contextual or discourse-level factors that may influence translation choices.

Future work within the broader research project will focus on expanding the TT corpora with the translations performed by other NMT systems and LLMs to assess whether their outputs provide similar results and whether the results obtained can be generalized. Subsequently, the TT corpora will be compared with a comparable Italian news corpus. Moreover, the analysis could be integrated by examining a broader range of lexical items and performing a qualitative analysis of concordance lines, which would allow for a more detailed investigation of how lexical choices interact with context and discourse. Furthermore, including other indicators of variation such as syntactic complexity, could provide a more comprehensive account of stylistic variation in MT.

The findings of this study contribute to existing research on machine translation for this specific language combination, which has not been extensively researched until now. They provide a quantitative account of stylistic variation in English-Italian translations and offer a comparative perspective on NMT and LLM outputs, showing how different systems and genres interact to produce lexical variation.

References

- Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. Longman Grammar of Spoken and Written English.
- Bowker, L. (2024). Risks for lay users in machine translation and machine translation literacy. In *The Social Impact of Automating Translation* (pp. 60-76). Routledge.
- Bullard, G. (2023). Smart ways journalists can exploit artificial intelligence. *Nieman Reports*. <https://niemanreports.org/artificial-intelligence-newsrooms/> (niemanreports.org). Last accessed: 2026-05-26.
- Cagnan, P. (2025). Uso dell'intelligenza artificiale nelle redazioni italiane. *Ordine dei Giornalisti*. <https://www.odg.it/uso-dellintelligenza-artificiale-nelle-redazioni-italiane/59863>. Last accessed: 2026-05-26.
- Caswell, David (2019). Structured Journalism and the Semantic Units of News. *Digital Journalism*, 7(8), 1134-1156.
- Cools, H., and Diakopoulos, N. (2024). Uses of Generative AI in the Newsroom: Mapping Journalists' Perceptions of Perils and Possibilities. *Journalism Practice*, 1-19. <https://doi.org/10.1080/17512786.2024.2394558>
- Daems, J. (2022). Dutch literary translators' use and perceived usefulness of technology: The role of awareness and attitude. In *Using technologies for creative-text translation* (pp. 40-65). Routledge. doi: 10.4324/9781003094159-3
- Daems, J., De Clercq, O., and Macken, L. (2017). Translationese and post-edits: How comparable is comparable quality? *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 16, 89-103. doi: 10.52034/lanstts.v16i0.434.
- De Clercq, O., De Sutter, G., Loock, R., Cappelle, B., and Plevoets, K. (2021). Uncovering machine translationese using corpus analysis techniques to distinguish between original and machine-translated French. *Translation Quarterly*, 101, 21-45. (hal-03406287).
- Espinosa-Giménez, M., and Forcada, M. L. (2024). Lexical impoverishment in neural machine translation: automatic and human study of Spanish-English noun translation. *Tradumàtica technologies de la traducció*, (22), 162-180. doi: 10.5565/rev/tradumatica.428.
- Hermans, M. H. (2014). The Euroregional discourse. Increasing evidence towards legitimizing an institutional space. *Mots. Les langages du politique*, (106). doi: 10.4000/mots.21839.
- Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. <https://doi.org/10.5281/zenodo.1212303>
- Jiang, Y., and Niu, J. (2022). A corpus-based search for machine translationese in terms of discourse coherence. *Across Languages and Cultures*, 23(2), 148-166. doi: 10.1556/084.2022.00182.
- Kocmi, T., Artemova, E., Avramidis, E., Bawden, R., Bojar, O., Dranch, K., Dvorkovich, A., Dukanov, S., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Lakougn, H., Lundin, J., Monz, C., Murray, K., Nagata, M., Perrella, S., Proietti, L., Popel, M., Popović, M., Riley, P., Shmatova, M., Steingrímsson, S., Yankovskaya, L., Zouhar, V. (2025). Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets. In B. Haddow, T. Kocmi, P. Koehn, and C. Monz (Eds), *Proceedings of the Tenth Conference on Machine Translation* (pp. 355-413). Association for Computational Linguistics. doi: 10.18653/v1/2025.wmt-1.22.
- Kong, D., and Macken, L. (2025a). Can Peter Pan Survive MT? A Stylometric Study of LLMs, NMTs, and HTs in Children's Literature Translation.
- Kong, D., and Macken, L. (2025b). Decoding Machine Translationese in English-Chinese News: LLMs vs. NMTs. In *20th Machine Translation Summit*.
- Loock, R. (2020). No more rage against the machine: How the corpus-based identification of machine-translationese can lead to student empowerment. *Journal of Specialised Translation*, (34), 150-170. (halshs-02913980).
- Macken, L., De Clercq, O., Paulussen, H. (2011). "Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus." *META* 56 (2): 374-90.
- Molino, A. (2011). *A Corpus-Based Study of Stance Adverbials in Learner Writing*. Roma: Aracne Editrice.
- Murru, M. F., and Carlo, S. (2024). AI e newsmaking: Un'indagine esplorativa nelle redazioni nazionali e locali italiane. *Mediascapes Journal*, 23(1), 184-198. <https://rosa.uniroma1.it/rosa03/mediascapes/article/view/18714>
- Murtaza, G., and Qasmi, N. Q. (2013). Style and stylistics: An overview of traditional and linguistic approaches. *Galaxy: international multidisciplinary research journal*, 2(3), 2278-9529.
- Ngai, J. (2022). Evaluation across newspaper genres: Hard news stories, editorials and feature articles. Routledge. doi: 10.4324/9781003150640.
- Niu, J., and Jiang, Y. (2024). Does simplification hold true for machine translations? A corpus-based anal-

- ysis of lexical diversity in text varieties across genres. *Humanities and Social Sciences Communications*, 11(1), 1-10. doi: 10.1057/s41599-024-02986-7.
- Roush, Chris. (2023). What Reuters is telling its journalists about using artificial intelligence. *Talkingbiz*. <https://talkingbiznews.com/?s=What+Reuters+is+telling+its+journalists+about+using+artificial+intelligence>. Last accessed:2026-05-26.
- The New York Times Trust Team. (2024). How The New York Times uses AI in journalism. *The New York Times*. <https://www.nytimes.com/2024/10/07/reader-center/how-new-york-times-uses-ai-journalism.html>. Last accessed:2026-05-26.
- Vanmassenhove, E., Shterionov, D., and Gwilliam, M. (2021). Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2203–2213). Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.188.
- Vanmassenhove, E., Shterionov, D., and Way, A. (2019). Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track* (pp. 222-232). European Association for Machine Translation.
- Webster, R., Fonteyne, M., Tezcan, A., Macken, L., and Daems, J. (2020). Gutenberg Goes Neural: Comparing Features of Dutch Human Translations with Raw Neural Machine Translation Outputs in a Corpus of English Literary Classics. *Informatics*, 7(3), 32. doi:10.3390/informatics7030032.

Lexical and Syntactic Diversity: Still Lost in Machine Translation?

Lise Volkart

FTI/TIM, University of Geneva
Switzerland
lise.volkart@unige.ch

Pierrette Bouillon

FTI/TIM, University of Geneva
Switzerland
pierrette.bouillon@unige.ch

Abstract

This paper investigates whether recent developments in machine translation (MT) have affected two well-documented stylistic characteristics of MT: reduced lexical diversity and increased source syntax mirroring. Using a corpus of English into French translations, we compared MT outputs produced by a widely used system over time (2024 vs. 2026) and under different models (NMT vs. LLM-based). Lexical diversity is measured using word translation entropy (HTra), and source syntax mirroring using four syntactic similarity metrics. Results show that lexical and syntactic characteristics of NMT output remain stable over time. The LLM-based MT exhibits a slight increase in lexical diversity, but still presents strong signs of lexical overgeneralisation. Syntactic similarity to the source remains largely unchanged. Overall, the findings suggest that recent MT developments have a very limited impact on these two stylistic characteristics.

1 Introduction

Machine translation (MT) is generally considered a rapidly evolving field. Yet, since 2016, and the so-called neural revolution, progress has been more gradual, and despite the emergence of large language models (LLMs), conventional neural machine translation (NMT) remains a relevant technology (Lazarov, 2026). In terms of output quality,

both technologies achieve state-of-the-art performance, particularly for high-resource languages, typically written in Latin scripts (Ataman et al., 2025; Lazarov, 2026). In the past few months, several traditional MT providers announced the implementation of so-called LLM-based MT architectures or the integration of LLM-related features, with claims of improved output quality.¹ Now, beyond general quality, it remains unclear whether such a technology shift affects the stylistic properties of the MT output. Previous work has shown that NMT output differs from human translation (HT), particularly through reduced lexical diversity and increased literalness, with higher syntactic similarity to the source (Vanmassenhove et al., 2019; Toral, 2019; Tezcan et al., 2019; Webster et al., 2020; Vanmassenhove et al., 2021; Volkart and Bouillon, 2024; Volkart, forthcoming). It is timely to determine whether such findings remain valid over time and in light of recent developments of LLM-based MT. In this paper, we address this question through a corpus-based study of English into French MT outputs produced by a widely used online MT system at different points in time, and under different models. Specifically, we compare outputs generated by NMT in 2024 and 2026 with those produced in 2026 under an LLM-based architecture. Using established metrics of lexical diversity and syntactic similarity, we aim to assess both the diachronic evolution of NMT and the potential stylistic impact of LLM-based MT.

Our main objective is to determine whether pre-

© 2026 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

¹See for instance: <https://www.supertext.com/en/technology> [Accessed: April 2026]
<https://www.deepl.com/en/blog/next-gen-language-model> [Accessed: April 2026]
<https://blog.google/products-and-platforms/products/search/gemini-capabilities-translation-upgrades/> [Accessed: April 2026]

vious findings on the lexical and syntactic characteristics of MT output are still relevant in the context of recent technological developments. More precisely, we pursue two complementary goals: (i) to evaluate the evolution of NMT output over time with respect to lexical diversity and syntactic similarity, and (ii) to assess whether the introduction of LLM-based MT leads to substantial stylistic changes.

We formalise these goals under the following research question:

Did lexical and syntactic features of MT output evolve over time and with the integration of LLM-based MT?

The lexical and syntactic analysis conducted shows that, despite recent technological developments, the stylistic characteristics previously observed in NMT output remain largely unchanged, suggesting that LLM-based MT does not appear to constitute a major shift on the stylistic level.

The remainder of the paper is structured as follows: Section 2 presents related work, Section 3 describes the background of our study as well as the data and metrics. Our results are presented in Section 4 and discussed in Section 5. Finally, Section 6 concludes the paper and Section 7 outlines limitations and directions for future research.

2 Related work

Since the advent of NMT systems around 2016, numerous studies, such as Martikainen and Kübler (2016), Toral (2019), Webster et al. (2020), Vanmassenhove et al. (2019), Vanmassenhove et al. (2021) or Volkart and Bouillon (2024), have shown that MT tends to exhibit a lower lexical variety than HT and to mirror source syntax, therefore producing more literal translations. More recent work has begun exploring translation produced by LLMs, mainly in terms of output quality. A number of studies have demonstrated that LLMs can achieve comparable or higher performance than conventional NMT systems, especially on high-resource languages (Hendy et al., 2023; Ataman et al., 2025; Lazarov, 2026). Regarding style, LLMs are often argued to achieve a higher level of fluency and produce less literal translations than NMT (Li et al., 2025). Raunak et al. (2023), for instance, have shown that LLMs produce less literal translation than NMT when translating out

of English, particularly for the translation of idiomatic expressions. Other works, such as Li et al. (2025), have shown that LLMs still exhibit significant amounts of *translationese*, i.e. translations that are too literal and therefore sound unnatural (Li et al., 2025). To the best of our knowledge, there are no studies yet focusing specifically on these questions for LLM-based MT output, that is translation produced by MT tools relying on an LLM architecture.

3 Methodology

3.1 Background

The analysis conducted in Volkart and Bouillon (2024) and Volkart (forthcoming) on texts collected in professional contexts have shown that NMT tends to exhibit a lower level of translation solutions variation and a higher level of syntactic similarity than HT. Specifically, these previous studies show that, on the lexical level, NMT tends to overuse the most frequent translation solutions and to underuse the less frequent ones, leading to a lexical overgeneralisation phenomenon, as already demonstrated by other authors. On the syntactic level, these experiments demonstrated that NMT tends to reproduce the source syntactic structure, leading to more literal translations. For the present study, we use the raw MT corpora produced in 2024 for these experiments and compare it to more recent MT outputs (NMT and LLM-based), to estimate whether MT has made any progress on those characteristics over the past two years.

3.2 Corpus

This work is based on two corpora (C1 and C2) compiled and described in Volkart and Bouillon (2024) and Volkart (forthcoming), containing authentic HT and PEMT from English into French collected in professional contexts, alongside a baseline raw MT generated using the online and publicly available NMT system DeepL² in February 2024. Original HT and PEMT corpora were collected between 2021 and 2022, C1 contains translations shared with us by the European Investment Bank (EIB), and C2 contains translations shared by a sports organization based in Switzerland. For the present experiment, we include only the baseline NMT and its source corpus. Using DeepL again, we generated two new raw MT corpora in February 2026, using the

²<https://www.deepl.com>

classic DeepL model (NMT) and the so-called next-gen language model. In DeepL’s documentation³, the classic model is presented as the standard NMT model, while the next-gen model is described as based on an LLM-based MT model. In doing so, we obtain two multiple translation corpora containing a collection of source sentences with multiple MT outputs (NMT from 2024, NMT from 2026 and LLM-based MT from 2026). Tables 1 and 2 summarize the main corpus statistics.

3.3 Metrics

The first metric we rely on is the word translation entropy HTra (Carl et al., 2016), which measures the diversity of translation solutions in a corpus. HTra is computed as the sum over all observed word translation probabilities $p(s \rightarrow t_i)$ of a given source text word s into target text word $t_i \dots n$ multiplied with their information content $I(p) = -\log_2(p)$ (Carl et al., 2016; Volkart and Bouillon, 2024) as in the following equation:

$$H(s) = -\sum_{i=1}^n p(s \rightarrow t_i) \times \log_2 p(s \rightarrow t_i)$$

The score reflects the amount of different translation solutions present in the target for a given source lemma, as well as the distribution of these solutions (if they are equiprobable, or if some are more represented than others). In that sense, HTra is a good indicator of the lexical diversity level in a translation corpus (for a more extensive discussion on this aspect, see Volkart (forthcoming)). The higher the variety and the more balanced the distribution, the higher HTra (Carl et al., 2016; Bangalore et al., 2016; Volkart, forthcoming).

Our second axis of research concerns the source sentence structure mirroring, which we estimate using the following metrics from the *ASTrED*⁴ library (Vanroy et al., 2021):

- Syntactically Aware Cross (SACr): quantifies the degree of reordering of word sequences that occurred between source and target (Vanroy et al., 2021)
- Aligned syntactic tree edit distance (ASTrED): computes the tree edit distance between source and target dependency trees

while taking word alignments into account (Vanroy et al., 2021).

- Label change for part-of-speech (POS) (Label POS): corresponds, for a given source-target sentence pair, to the number of source-target word-aligned pairs that have different POS tags, normalised by the total number of alignments for that sentence (Vanroy et al., 2021)
- Label change for dependency labels (Label DEP): corresponds, for a given source-target sentence pair, to the number of source-target word-aligned pairs that have different dependency labels, normalised by the total number of alignments for that sentence (Vanroy et al., 2021)

Each of these four metrics captures a specific aspect of the syntactic similarity between a source sentence and its translation (Volkart, forthcoming). Altogether, these metrics provide a comprehensive estimate of the degree of literality of a translation on the syntactic level.

HTra is computed for all content lemma appearing at least three times in the source corpus, based on the word-alignment generated using *awesome-align* (Dou and Neubig, 2021). Syntactic similarity metrics are computed using the *ASTrED* python library (Vanroy et al., 2021). Tagging and lemmatization are performed using SpaCy’s transformer models for English, French and German.⁵

4 Results

4.1 HTra

HTra scores for the different MT outputs for all POS and for each POS category are presented in Table 3. We obtain very similar results for C1 and C2. First of all, HTra scores do not differ between NMT24 and NMT26, indicating that the level of lexical diversity has not improved in the translation produced by DeepL’s NMT model over the past two years. Regarding the translation produced by the next-gen model, it demonstrates a slightly higher word translation entropy, meaning that the transition to an LLM architecture seems to have improved the lexical diversity of the MT output, and this across almost all the POS categories. The differences between NMT24 and LLM-MT26 are statistically significant according to the Wilcoxon

³<https://support.deepl.com/hc/en-us/articles/14241705319580-About-DeepL-language-models>

⁴<https://github.com/BramVanroy/astred>

⁵<https://spacy.io/models>

Subcorpus	# segments	# tokens	mean segment length
NMT24	1 852	47 294	25,54
NMT26	1 852	47 338	25,60
LLM-MT26	1 852	47 351	25,61

Table 1: Descriptives statistics for corpus C1.

Subcorpus	# segments	# tokens	mean segment length
NMT24	2 280	48 760	21,39
NMT26	2 280	52 390	22,98
LLM-MT26	2 280	52 732	23,13

Table 2: Descriptives statistics for corpus C2.

signed-rank test for all POS together, for nouns and verbs, though effect sizes⁶ are generally low (ranging from -0.09 to -0.17), indicating limited practical significance. While NMT has remained stable over the past two years in terms of translation solutions diversity, the introduction of an LLM-based architecture is associated with a modest improvement that, although limited, deserves further investigation.

On the technical level, a higher HTra score can be explained by two different factors: either a higher number of unique translation solutions or a better distribution of them, and, eventually, a combination of both. To better understand the improved HTra observed for the nouns and the verbs in the LLM-based MT, we attempt to isolate the main factor contributing to this score increase. To do so, we extract all the nouns and verbs presenting a higher score for the LLM-based MT with all their translation solutions, and manually annotate the main explaining factor for the score increase (either quantity or distribution). The cases for which LLM-MT26 has an equal or lower number of translation solutions compared to NMT24, but a more balanced distribution are annotated as "due to distribution" ("distribution" hereafter). The cases for which LLM-MT26 has more unique translation solutions compared to NMT24 are annotated as "due to variety" ("variety" hereafter) and split into three subgroups according to the number of additional solutions (1, 2 and 3 or

more). Results of this classification are presented in Tables 4 and 5.

Again, C1 and C2 exhibit very similar tendencies. For the nouns, in about 80% of the cases, the higher HTra observed in the LLM-based translation is due to a greater translation solutions variety rather than to a more balanced distribution. Amongst those, around 50% of cases exhibit only one additional translation solution. This is very similar for the verbs, with almost 90% of the increase cases due to a larger variety and a bit more than 50% of them having only one additional solution. This clearly indicates that the LLM-based MT, while having a slightly higher HTra score, still follows the same tendency as the NMT: it exhibits the same centralisation and generalisation tendencies on the lexical level, but with occasional use of less frequent translation solutions.

4.2 Syntactic similarity

Table 6 presents the sentence similarity scores for the different MT outputs. For all metrics, a lower score indicates a higher degree of source sentence similarity. NMT24 and NMT26 present identical scores for the four metrics. The Wilcoxon signed-rank test, though, indicates some statistically significant differences between them for some of the scores (Label DEP and AStrED for C1 and SACr Label DEP and AStrED for C2). As this may seem counter-intuitive, it can be easily explained: the Wilcoxon signed-rank test reveals a systematic discrepancy between the paired differences, which

⁶computed as $r = \frac{z}{\sqrt{n}}$

	C1			C2		
	NMT24	NMT26	LLM-MT26	NMT24	NMT26	LLM-MT26
Adjectives	0,97	0,97	1,00	1,01	1,01	1,04
Adverbs	1,10	1,09	1,04	1,25	1,26	1,27
Nouns	0,71	0,72	0,75*	0,86	0,86	0,94[◊]
Verbs	1,27	1,27	1,39[◊]	1,45	1,45	1,56[◊]
All POS	0,92	0,92	0,97[◊]	1,07	1,07	1,14[◊]

Table 3: HTra scores for the different MT outputs (NMT24, NMT26 and LLM-MT26) for corpus C1 and C2. [◊] indicates a significant difference with NMT24 at $p < 0,001$, and * at $p < 0,05$. Significance was tested using Wilcoxon signed-rank test.

Main HTra increase factor	Percentage of nouns	Percentage of verbs
↗ distribution	20,67%	12,30%
↗ variety	79,33%	87,70%
of which +1	56,98%	55,74%
of which +2	18,44%	22,95%
of which > +3	3,91%	9,02%

Table 4: Percentage of nouns and verbs according to the main HTra increase factor for LLM-MT26 for corpus C1

Main HTra increase factor	Percentage of nouns	Percentage of verbs
↗ distribution	17,62%	11,11%
↗ variety	82,38%	88,19%
of which +1	50,08%	52,08%
of which +2	15,33%	20,83%
of which > +3	11,11%	15,28%

Table 5: Percentage of nouns and verbs according to the main HTra increase factor for LLM-MT26 for corpus C2

means that a large number of sentences have scores varying in the same direction. Considering the large sample size, it is sufficient to produce a statistically significant result. Nevertheless, the effect sizes remain very limited (ranging from 0.05 to 0.06). We interpret this as the NMT26 not producing the exact same translation as the NMT24, but still with a very similar average level of syntactic similarity with the source.

The LLM-based MT presents marginal differences with NMT24, with a slightly lower SACr score and a moderately higher ASTRaED for C1, indicating that LLM-based MT presents a higher similarity in terms of word order, but a lower similarity at the level of deep syntactic structure.

For C2, some differences (Label POS, Label DEP and ASTRaED) emerge as statistically signif-

icant according to the Wilcoxon signed-rank test, despite identical average scores. However, the associated effect sizes remain very small (ranging from 0.08 to 0.09), again suggesting limited practical significance. Taken together, these results indicate that, although the LLM-based MT actually differs from the NMT24 at the sentence level, the overall degree of syntactic similarity remains largely unchanged. Given the marginal differences between the subcorpora and the lack of consistency across C1 and C2, these findings are difficult to interpret. More importantly, the observed differences are substantially smaller than those reported between human translation (HT) and raw MT in other studies (Tezcan et al., 2019; Shaitarova et al., 2023; Volkart and Bouillon, 2024). This suggests that the different MT outputs are consider-

ably more similar to one another than to HT. In this respect, the well-documented tendency of MT systems to adhere more closely to source syntax than HT appears to persist.

5 Discussion

This experiment shows that the stylistic characteristics of MT output tend to be stable over time, even when switching from NMT to LLM-based MT. This may indicate that modern MT systems have reached a plateau in terms of the level of naturalness they can achieve, and the integration of LLMs does not fundamentally change the picture. This has several implications: first of all, it means that post-editing (PE) by professionals still holds significant value to ensure translations remain stylistically rich and fluent in the target language. On the other hand, it also means that the MT interference that could be measured in post-edited MT (Volkart, forthcoming) is also likely to remain unchanged. Finally, those results show that other strategies may be needed to recover lexical and syntactic diversity in translation produced by automated systems. These strategies might include directly prompting LLMs for translation or leveraging them to perform editing targeted on stylistic features. To this end, LLMs may bring significant value.

6 Conclusion

In this experiment, we compared the levels of lexical diversity and syntactic similarity of different MT outputs produced using DeepL. Our main goal was to estimate whether NMT had evolved substantially on these characteristics between 2024 and 2026, and whether the introduction of LLM-based MT constitutes a significant game-changer.

Our results show that the NMT generated by DeepL in 2026 exhibits the same level of lexical diversity as in 2024. The NMT tendency toward lexical overgeneralisation and the reduction of the translation solutions variation compared to HT remains constant and this for the two analysed corpora. The LLM-based MT exhibits a slightly higher level of word translation entropy, mainly due to occasional addition of translation variants, but without a significant reduction of the lexical overgeneralisation phenomenon.

On the syntactic level, our experiment demonstrates that the level of source sentence similarity has not changed substantially over the past two

years, neither with the transition to an LLM-based architecture. While the more recent systems do not generate a translation identical to older outputs, the overall level of source sentence similarity remains very similar.

Those results show that the lexical and syntactic features of raw MT output generated by one of the most popular online MT engines have not evolved substantially over time, neither with the integration of LLM-based MT. This confirms that the MT tendency to overgeneralise on the lexical level and to follow the source syntax structure is stable over time and that results from previous studies remain highly relevant. Our work therefore shows that, on the stylistic levels, MT systems still struggle to produce human-like, diverse and natural sounding translations.

7 Limitations and Future Work

While this study highlights relevant concerns about some stylistic features of MT output, it remains within the scope of the case study. The analysed corpus, while presenting the advantage of being collected from authentic translation workflows, remains very specific and the analysed MT outputs originate from a single general-purpose online MT provider, without extensive and detailed information on model characteristics. In that sense, our results are not strictly generalisable and may reflect system-specific behaviour. Nonetheless, they provide a snapshot of the capabilities of a system widely used by professional translators and the general public. Furthermore, they invite us to consider MT progress with caution and to keep in mind that some limitations and drawbacks of MT systems might silently persist even if the technology seems to improve. Of course, it is now necessary to explore these aspects further by conducting diachronic studies on MT style on a larger scale and with a larger variety of systems. It is also essential to explore more widely translations done by LLMs and to measure the influence of different prompting strategies on the lexical and syntactic levels.

8 Acknowledgement

We would like to thank the language services who kindly accepted to share their translation memories for this research project.

	C1			C2		
	NMT24	NMT26	LLM-MT26	NMT24	NMT26	LLM-MT26
SACr	0,19	0,19	0,17 [◊]	0,17	0,17*	0,16
Label POS	0,13	0,13	0,13	0,13	0,13	0,13 [◊]
Label DEP	0,19	0,19*	0,19	0,20	0,20*	0,20 [◊]
ASTrED	0,55	0,55*	0,57[◊]	0,57	0,57 [◊]	0,57 [◊]

Table 6: Syntactic similarity scores for the different MT outputs (NMT24, NMT26 and LLM-MT26) for corpus C1 and C2. [◊] indicates a significant difference with NMT24 at $p < 0,001$, and * at $p < 0,05$. Significance was tested using Wilcoxon signed-rank test.

References

- Ataman, Duygu, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. 2025. Machine Translation in the Era of Large Language Models: A Survey of Historical and Emerging Problems. *Information*. <https://doi.org/10.3390/info16090723>.
- Bangalore, Srinivas, Bergljot Behrens, Michael Carl, Maheshwar Ghankot, Arndt Heilmann, Jean Nitzke, Moritz Schaeffer, and Annegret Sturm. 2016. Syntactic variance and priming effects in translation. In *New directions in empirical translation process research*, pages 211–238. Springer.
- Carl, Michael, Moritz Schaeffer, and Srinivas Bangalore. 2016. The CRITT translation process research database. In *New directions in empirical translation process research*, pages 13–54. Springer. https://doi.org/10.1007/978-3-319-20358-4_2.
- Dou, Zi-Yi and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128. ACL. <https://doi.org/10.18653/v1/2021.eacl-main.181>.
- Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *Preprint : arXiv:2302.09210*. <https://doi.org/10.48550/arXiv.2302.09210>.
- Lazarov, Todor. 2026. LLMs and AI in the Language Industry: An Overview. In *2026 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, pages 1–6. <https://doi.org/10.1109/ACDSA67686.2026.11468231>.
- Li, Yafu, 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. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12875–12894, Vienna, Austria, July. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.acl-long.630>.
- Martikainen, Hanna and Natalie Kübler. 2016. Ergonomie cognitive de la post-édition de traduction automatique : enjeux pour la qualité des traductions. *ILCEA*, (27), November. <https://doi.org/10.4000/ilcea.3863>.
- Raunak, Vikas, Arul Menezes, Matt Post, and Hany Hassan Awadalla. 2023. Do GPTs Produce Less Literal Translations? pages 1041–1050, Toronto, Canada, July. <https://doi.org/10.18653/v1/2023.acl-short.90>.
- Shaitarova, Anastassia, Anne Göhring, and Martin Volk. 2023. Machine vs. Human: Exploring Syntax and Lexicon in German Translations, with a Spotlight on Anglicisms. In *The 24th Nordic Conference on Computational Linguistics*. <https://aclanthology.org/2023.nodalida-1.22/>.
- Tezcan, Arda, Joke Daems, and Lieve Macken. 2019. When a ‘sport’ is a person and other issues for NMT of novels. In *Machine Translation Summit XVII*, pages 40–49. European Association for Machine Translation. <https://aclanthology.org/W19-7306/>.
- Toral, Antonio. 2019. Post-editeese: an Exacerbated Translationese. In *Proceedings of MT Summit XVII*, volume 1, pages 273 – 281, Dublin, Ireland. <https://aclanthology.org/W19-6627/>.
- Vanmassenhove, Eva, Dimitar Shterionov, and Andy Way. 2019. Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In *Proceedings of MT Summit XVII*, volume 1, pages 222 – 232, Dublin, Ireland. <https://aclanthology.org/W19-6622/>.

- Vanmassenhove, Eva, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. ACL. <https://aclanthology.org/2021.eacl-main.188>.
- Vanroy, Bram, Orphée De Clercq, Arda Tezcan, Joke Daems, and Lieve Macken. 2021. Metrics of syntactic equivalence to assess translation difficulty. In *Explorations in empirical translation process research*, pages 259–294. Springer. https://doi.org/10.1007/978-3-030-69777-8_10.
- Volkart, Lise and Pierrette Bouillon. 2024. Post-editors as Gatekeepers of Lexical and Syntactic Diversity: Comparative Analysis of Human Translation and Post-editing in Professional Settings. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, Sheffield. <https://aclanthology.org/2024.eamt-1.33/>.
- Volkart, Lise. forthcoming. *Traduction automatique neuronale et post-édition : vers une uniformisation des textes traduits ? Étude sur la langue post-éditée en contexte professionnel*. Ph.D. thesis, University of Geneva.
- Webster, Rebecca, Margot Fonteyne, Arda Tezcan, Lieve Macken, and Joke Daems. 2020. Gutenberg Goes Neural: Comparing Features of Dutch Human Translations with Raw Neural Machine Translation Outputs in a Corpus of English Literary Classics. *Informatics*, 7(3). <https://doi.org/10.3390/informatics7030032>.