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 stylometric 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 stylometric 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

[1] See `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:

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

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 stylometric tendencies observed quantitatively manifested in concrete translational choices.

The qualitative categories were defined prior to analysis based on the stylometric 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 stylometric 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, stylometric-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 stylometric 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 stylometric measures by target language and translation condition. Values represent item-level means.

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.

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 coexists 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.

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 stylometric 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 alignment 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 euas-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 يفسّر 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.