# How Decoding Settings Shape Lexical Variation in Iterative LLM-Translation Workflows: A Case Study of Chinese-English Literary Translation

**Lulu Wang**
The Hong Kong Polytechnic University
Hong Kong
lulu99.wang@connect.polyu.hk

**Hengtian Zhang**
The Hong Kong Polytechnic University
Hong Kong
heng-tian.zhang@connect.polyu.hk

**Lieve Macken**
Ghent University
Ghent, Belgium
lieve.macken@ugent.be

**Kanglong Liu***
The Hong Kong Polytechnic University
Hong Kong
kl.liu@polyu.edu.hk

## Abstract

Large language models (LLMs) are increasingly used in automated translation workflows involving iterative refinement, yet the role of decoding settings in shaping lexical variation within such workflows remains underexplored. This study examines how temperature and top-p settings influence lexical diversity, lexical frequency distribution, and inter-stage lexical similarity within a three-stage literary translation pipeline (initial translation, self-revision, and stylistic refinement). Using GPT-5.1, we translated contemporary Chinese fiction texts under controlled parameter settings. Results show that less restrictive decoding (higher temperature and top-p) leads to increased lexical diversity, higher proportions of low-frequency vocabulary, and more extensive inter-stage revision, with these effects amplified across successive refinement stages. Compared with professional human translations, LLM outputs exhibit higher lexical diversity and a preference for lower-frequency vocabulary across all conditions. The findings reveal how decoding parameters shape lexical variation in LLM-generated translations, with implications for LLM-based translation workflow design.

## 1 Introduction

Large language models (LLMs) demonstrate strong performance in translation tasks, producing fluent outputs across a wide range of genres (Gao et al., 2024; Jiang et al., 2024). Recent work has explored automated LLM-based translation workflows, with some studies reporting that they can rival or even surpass professional human translations in perceived quality (Wang et al., 2026; Wu et al., 2024; Wu et al., 2025). A common feature of these workflows is iterative refinement, where models revise their initial outputs based on self-reflection and feedback (Chen et al., 2024; Madaan et al., 2023; Xu et al., 2024).

Despite improvements in overall quality, little is known about how the lexical properties of LLM-generated translations evolve across refinement stages. This issue is particularly consequential for literary translation, where lexical patterns underlie stylistic expression and creativity, influencing tone, voice, and perceived literary quality (Leech & Short, 2007; Munat, 2007). As texts undergo successive revision and refinement, cumulative lexical changes may alter stylistic characteristics in systematic ways. Understanding these dynamics is therefore essential for evaluating LLM-based translation workflows beyond holistic quality assessments.

A key but underexplored factor in this process is the setting of decoding parameters. LLMs generate text by sampling from a probability distribution over candidate tokens at each step, and outputs are sensitive to the parameters governing this process (Wiher et al., 2022; Zhou et al., 2025). Among these parameters, temperature rescales the probability distribution prior to sampling: higher values flatten the distribution and increase the

* Corresponding author



*Proceedings of the 3rd International Conference on New Trends in Translation and Interpreting Technology*, pages 98–108
June 24 – 27, 2026, Dubrovnik, Croatia. ©NeTTIT 2026
https://doi.org/10.26615/issn.2815-4711.2026_014

likelihood of selecting lower-probability tokens, introducing greater randomness, whereas lower values sharpen the distribution and produce more deterministic outputs (Arias et al., 2025; Troshin et al., 2025). Top-p sampling (also known as nucleus sampling) restricts generation to the smallest set of tokens whose cumulative probability exceeds a threshold $p$, effectively truncating the low-probability tail of the distribution. These two mechanisms operate on different aspects of the sampling process: temperature affects distribution shape, while top-p constrains candidate set size (Holtzman et al., 2020; Ippolito et al., 2019).

Although the computational mechanisms of temperature and top-p are well understood, their stylistic consequences are often assumed rather than empirically verified. Existing studies have primarily examined their impact on output quality in single-pass generation settings (Ippolito et al., 2019; Peng et al., 2023), providing limited insight into how decoding parameters shape lexical stylistic features. Their role in multi-stage refinement workflows has received comparatively little attention.

The iterative refinement process in LLM-based translation workflows introduces additional complexity. In refinement pipelines, the same model repeatedly revises its own outputs, implicitly applying internal criteria for improvement at each stage. Prior work suggests that such processes can steer outputs toward model-preferred stylistic patterns (Madaan et al., 2023; Xu et al., 2024), yet it remains unclear whether lexical variation introduced through stochastic sampling persists across refinement stages or is gradually normalized through successive revisions. More broadly, lexical diversity is a text-level property emerging from many sequential decisions, and the relationship between token-level stochasticity and document-level outcomes under iterative revision is not fully understood.

To address these gaps, this study examines how two decoding parameters, temperature and top-p, influence lexical stylistic features within a three-stage refinement pipeline for Chinese-to-English literary translation (initial translation, self-revision, and stylistic refinement). We track changes in lexical diversity, frequency distribution, and inter-stage lexical similarity under controlled sampling conditions. By situating decoding parameters within a realistic iterative workflow, the study provides a process-oriented account of how lexical variation emerges and evolves in LLM-generated literary translation, with further implications for the design of LLM-based translation pipelines.

## 2 Related Work

Research in natural language generation has demonstrated that decoding strategies play a central role in shaping model outputs (Arias et al., 2025). Early work established a notable but task-specific trade-off between output diversity (specifically lexical variation and $n$-gram richness across generated sequences) and overall text quality, whereby higher sampling temperatures increase variability but may compromise output quality, particularly fluency, contextual coherence, and accuracy (Ippolito et al., 2019; Wiher et al., 2022).

A growing body of research has further examined how sampling hyperparameters influence LLM performance across a range of tasks, including code generation (Arora et al., 2024; Donato et al., 2025), reasoning (Troshin et al., 2025), open-ended creative writing (Peeperkorn et al., 2024), genre analysis (Grullon-Polanco, 2026), and machine translation (Peng et al., 2023). Across these domains, results consistently highlight the sensitivity of LLM outputs to decoding configurations. For instance, Donato et al. (2025) show that temperature and top-p substantially affect both output variability and correctness in code generation. Similarly, Peeperkorn et al. (2024) find that temperature has modest but measurable effects on novelty and coherence in narrative generation. These findings suggest that decoding parameters are not merely technical settings but fundamental determinants of LLM output behavior.

In the context of machine translation, prior work has primarily examined decoding effects in terms of accuracy. Peng et al. (2023) report that increasing temperature leads to lower BLEU and COMET scores, while also increasing the risk of hallucination in non-English-centric translation tasks. Likewise, Li et al. (2025) recommend near-deterministic decoding (i.e., temperature close to zero) to maximize translation accuracy. However, these studies focus on reference-based automated evaluations of quality, leaving the linguistic properties, such as stylistic features, largely overlooked.

Despite these advances, most studies focus on single-pass generation, leaving the effects of

decoding settings in multi-stage workflows underexplored. Given that decoding effects are task-dependent (Wiher et al., 2022), it is uncertain whether prior findings extend to stylistically constrained tasks such as literary translation, particularly with respect to how decoding settings influence lexical characteristics within emerging iterative refinement pipelines (Chen et al., 2024; Wang et al., 2026; Wu et al., 2024; Wu et al., 2025).

## 3 Methods

### 3.1 Corpus

This study uses a corpus of 200 Chinese fiction texts extracted from contemporary published works across diverse genres (science fiction, *wuxia*, romance, humor, crime and detective, and rural fiction). Source passages average 2,626 Chinese characters ($SD$ = 160), ranging from 2,229 to 3,138 characters, with a total size of 525,200 Chinese characters. Each passage has a corresponding published professional English translation for baseline comparison.

### 3.2 LLM Translation Workflow

The LLM translation process follows a three-stage iterative pipeline that progressively shifts from interlingual semantic transfer to intralingual stylistic refinement (Jakobson, 1959; Macken et al., 2022; Whyatt, 2017; Wang et al., 2026).

**Stage 0 (Initial Translation):** The model receives the Chinese source text and is prompted to produce a faithful and fluent literary translation. Quality criteria emphasize accuracy, fluency, and stylistic appropriateness for fiction translation.

**Stage 1 (Self-Revision):** The model receives its Stage 0 output together with the source text and revises for accuracy, fluency, and lexical consistency, following criteria adapted from ISO 18587 post-editing guidelines (International Organization for Standardization, 2017).

**Stage 2 (Stylistic Refinement):** The model receives only its Stage 1 output, without access to the source text, and refines fluency, style, and narrative flow while preserving semantic content.

### 3.3 Experimental Design

All translations were generated using GPT-5.1 via the Microsoft Azure[1] OpenAI API in March 2026. To examine the effects of sampling parameters on lexical characteristics across refinement stages, we manipulated temperature and top-p settings in two separate experiments, avoiding confounds between distribution flattening (temperature) and probability mass truncation (top-p).

**Experiment 1 (Temperature):** All 200 source texts were translated across three temperature conditions (0, 0.5, 1) with top-p fixed at 1. Although GPT-5.1 supports temperature values up to 2.0, we restricted our investigation to the 0–1 range, which encompasses the settings most commonly adopted in practical applications and prior studies on LLM-based translation and text generation (Cai et al., 2024; Zhang et al., 2025). Temperature values exceeding 1.0 increasingly flatten the probability distribution toward uniform sampling, typically compromising output coherence in translation tasks. The unrestricted top-p setting ensures that temperature alone determines the degree of sampling variability within this range.

**Experiment 2 (Top-p):** A randomly sampled subset of 100 source texts was translated across three top-p conditions (0.5, 0.8, 1) with temperature fixed at 0.7, a conventional default that balances output diversity and coherence (Cai et al., 2024). The subset enables controlled comparison within practical resource constraints, while the top-p conditions maintain sufficient stochasticity for top-p to meaningfully constrain the candidate token space, avoiding near-deterministic outputs that would occur at lower temperatures. The selected top-p values represent a gradient from restricted (0.5) to unrestricted (1) sampling, enabling systematic examination of how probability mass truncation affects lexical outcomes.

Under each condition, the same parameter settings were applied across all three workflow stages.

### 3.4 Lexical Style Metrics

In computational linguistics and stylometry, lexical style functions as a measurable "linguistic fingerprint" representing idiosyncratic patterns of language use (Alrabaee et al., 2019). In this study, it is examined through three dimensions: lexical diversity, lexical frequency distribution, and inter-stage lexical similarity.

Lexical diversity is measured using Type-Token Ratio (TTR), Moving-Average Type-Token Ratio (MATTR; Covington & McFall,

[1] https://azure.microsoft.com/

2010), Measure of Textual Lexical Diversity (MTLD; Malvern & Richards, 1997; McCarthy, 2005), and lexical entropy. TTR is included as a conventional baseline measure of vocabulary richness, while MATTR and MTLD provide more robust estimates by reducing sensitivity to text length. Lexical entropy serves as a complementary indicator by capturing the probability distribution, average information content, and unpredictability of word usage within a text (Mohseni et al., 2022).

Lexical frequency distribution is examined through lexical frequency profile (LFP) analysis, which measures the extent to which texts rely on high- or low-frequency vocabulary (Laufer & Nation, 1995; Van Parys et al., 2025). This dimension is particularly relevant given recent concerns that LLMs, as statistical next-word predictors, may systematically favor high-frequency forms while pruning away "rich, long statistical tails" of language (Vanmassenhove, 2025, p. 1). LFP analysis can thus reveal whether iterative refinement amplifies or mitigates such frequency-based biases.

Finally, lexical similarity is calculated using Levenshtein edit distance at both character and word levels (Anson et al., 2019), capturing morphological variations and structural reorganizations across refinement stages. Taken together, these measures provide a multidimensional account of lexical style, covering vocabulary diversity, frequency-based lexical choice, and inter-stage similarity across refinement stages.

### 3.5 Lexical Diversity

This dimension captures the range and variability of vocabulary usage within a text, reflecting both the breadth of lexical choice and the distributional evenness of word frequencies (McCarthy & Jarvis, 2010). In this study, it is operationalized using four complementary metrics.

**TTR (Type-Token Ratio):** The ratio of unique word types to total tokens. TTR is the most fundamental measure of lexical diversity but is sensitive to text length, as the proportion of new types inevitably decreases in longer texts due to repetition of high-frequency words (Vermeer, 2000).

**MATTR (Moving-Average Type-Token Ratio):** To mitigate length sensitivity, MATTR calculates TTR over a fixed-size sliding window (e.g., 100 tokens) and averages these scores across the text (Covington & McFall, 2010).[2]

**MTLD (Measure of Textual Lexical Diversity):** MTLD calculates the mean length of sequential word strings (factors) that maintain a specific minimum TTR threshold (typically 0.72) before dropping below it, effectively evaluating how long a text can sustain diverse vocabulary without repeating words excessively (McCarthy, 2005). The algorithm evaluates the running TTR as it reads sequentially left-to-right; when the TTR drops below 0.72, that segment is counted as one factor, and the calculation restarts. The procedure is usually repeated both forward and backward, and the final score is the total number of tokens divided by the total number of factors (averaged across both passes) (Koizumi, 2012).

**Lexical Entropy:** Rooted in information theory, lexical entropy quantifies the average uncertainty or information density in a text based on the relative frequency probability distribution of its words (calculated as $H = -\sum p_i log_2 p_i$; Shannon, 1951). Higher entropy indicates a more uniform distribution and greater textual unpredictability. A highly repetitive, simplistic text will feature a steep word frequency distribution and low entropy, whereas canonical literature utilizing varied, unpredictable word choices possesses a flatter distribution and substantially higher entropy values.

### 3.6 Lexical Frequency Profile Analysis

LFP analysis was used to examine the distribution of vocabulary across frequency bands. It measures the extent to which a text relies on vocabulary drawn from different frequency levels and is widely used to assess lexical sophistication and vocabulary use in corpus-based research (Laufer & Nation, 1995; Van Parys et al., 2025). Based on the BNC/COCA 25,000 word list (Nation, 2017), words in each translation were assigned to frequency levels and grouped into a three-band profile following Vanmassenhove et al. (2021): B1 (the 1,000 most frequent words), B2 (words ranked 1,001–2,000), and B3 (words beyond the 2,000-frequency threshold). For each translation, the proportion of tokens in each band was calculated to examine whether refinement stages

[2] For example, with a window size of 100, the algorithm computes TTR for tokens 1–100, then 2–101, 3–102, and so on until the end of the document, finally averaging these values.

and decoding settings favor more frequent or less frequent lexical choices.

### 3.7 Inter-Stage Lexical Similarity

To quantify how extensively translations are revised across stages, we computed the Levenshtein edit distance between outputs at successive stages (S0–S1, S1–S2) and cumulatively (S0–S2). This metric captures the minimum number of single-unit operations (insertions, deletions, and substitutions) required to transform one version into another (Anson et al., 2019).

Distances were calculated at both character and word levels. Character-level distance captures fine-grained orthographic and morphological variation. Word-level distance operates over tokenized sequences, treating each word as an irreducible unit, and thus more directly reflects meaningful lexical and structural changes; it was therefore used as the primary measure in our analysis.

To ensure comparability across texts of varying lengths, raw distances were normalized by the length of the longer text. The resulting proportion reflects the extent of revision between stages.

## 4 Results and Discussion

### 4.1 Text Length

As shown in Table 1, the lengths of LLM translations remain relatively stable from the initial translation stage (S0) to self-revision stage (S1) across all temperature settings (0, 0.5, 1). During the intralingual stylistic refinement stage (S2), which aims to improve fluency and stylistic quality, text length increases dramatically, with the highest temperature producing the longest outputs ($M$ = 2852.65, $SD$ = 404.37). This suggests that higher temperature not only lengthens the text but also introduces greater variability in length during intralingual refinement.

Similar to the temperature conditions, text lengths consistently increase from S0 to S2 across all top-p settings (Table 2). Expanding the candidate token space (top-p = 1) results in slightly longer texts at every stage compared to more restrictive settings (top-p = 0.5 and 0.8), with the effect most pronounced at S2 ($M$ = 2669.08, $SD$ = 393.79). Limiting the candidate token set based on cumulative probability appears to constrain text expansion during refinement.

Across all conditions, LLM translation outputs exceed the professional human translations in length, with the gap widening substantially at S2.

| Temperature Setting | Refinement Stage | | |
|---|---|---|---|
| | S0 (Initial) | S1 (Self-revision) | S2 (Stylistic Refinement) |
| **0** | 2078 (198) | 2137 (201) | **2490 (291)** |
| **0.5** | 2064 (195) | 2128 (206) | **2484 (376)** |
| **1** | **2110 (198)** | **2212 (214)** | **2853 (404)** |
| **Human Translation** | 1901 (418) | | |

Table 1: Mean word counts (*SD*) of LLM translations across refinement stages and temperature settings compared with human translation ($N$ = 200).

| Top-p Setting | Refinement Stage | | |
|---|---|---|---|
| | S0 (Initial) | S1 (Self-revision) | S2 (Stylistic Refinement) |
| **0.5** | 2110 (216) | 2167 (223) | **2561 (353)** |
| **0.8** | 2116 (220) | 2189 (225) | **2641 (371)** |
| **1** | **2121 (222)** | **2199 (229)** | **2669 (394)** |
| **Human Translation** | 1883 (459) | | |

Table 2: Mean word counts (*SD*) of LLM translations across refinement stages and top-p settings compared with human translation ($N$ = 100).

### 4.2 Lexical Diversity

This section examines how decoding parameters, specifically temperature (0, 0.5, and 1) and top-p (0.5, 0.8, and 1), influence lexical diversity across the iterative refinement workflow.

As detailed in Table 3 and Table 4 and visualized in the radar plots (Figure 1 and Figure 2), both parameters exert a systematic and monotonic influence on lexical diversity. Across all three refinement stages (S0, S1, and S2), increasing either the sampling randomness (temperature) or the candidate token pool (top-p) consistently yields higher scores across all measured metrics (TTR, MATTR, MTLD, and lexical entropy).

The results show that while lexical diversity increases across the iterative refinement process, less restrictive parameter settings amplify this trajectory. Diversity scores remain relatively clustered during S0, but the gaps between different settings widen considerably by S2. This indicates that granting the model more sampling freedom encourages the exploration of a much broader and more varied vocabulary, particularly during the stylistic refinement phase.

Compared with professional human translations, the LLM outputs display higher lexical diversity across all parameter settings and stages. Even at the most deterministic settings (T = 0 or top-p = 0.5) in the initial draft, the LLM generates a more diverse vocabulary (e.g., MATTR ≈ 0.73–0.74, MTLD ≈

113–114) than the human baseline (MATTR ≈ 0.728, MTLD ≈ 106). By S2 at maximum settings (T = 1 or top-p = 1), this gap expands dramatically, suggesting a tendency for the LLM to produce increasingly dispersed lexical choices under less restrictive sampling.

This finding should be interpreted with caution. Notably, the human translation sample in this study is relatively limited in size, whereas LLMs draw on vastly larger training corpora, potentially exposing them to a wider range of lexical alternatives. It also remains uncertain whether the increasing lexical diversity across refinement stages reflects genuine stylistic enrichment or problematic variation, such as unnecessary synonymic substitution, register inconsistencies, or semantic drift. As higher diversity does not necessarily indicate higher quality (Shypula et al., 2025), without quality evaluation, we cannot determine whether these lexical choices are contextually appropriate or functionally equivalent to those characteristic of professional human translation.

| Refinement Stage | Temperature Setting | Lexical Diversity | | | |
|---|---|---|---|---|---|
| | | TTR | MATTR | MTLD | Lexical Entropy |
| S0 | 0 | 0.3699 | 0.7417 | 114.3634 | 8.3378 |
| | 0.5 | 0.3721 | 0.7446 | 117.4163 | 8.3551 |
| | 1 | **0.3823** | **0.7518** | **125.7408** | **8.4204** |
| S1 | 0 | 0.3768 | 0.7459 | 118.4793 | 8.3898 |
| | 0.5 | 0.3806 | 0.75 | 123.1024 | 8.4173 |
| | 1 | **0.3915** | **0.7588** | **133.7715** | **8.5029** |
| S2 | 0 | 0.3842 | 0.7617 | 139.6092 | 8.5526 |
| | 0.5 | 0.3926 | 0.7676 | 147.9877 | 8.5968 |
| | 1 | **0.3991** | **0.7807** | **168.8523** | **8.7729** |
| Human Translation | | 0.3549 | 0.7281 | 105.948 | 8.145 |

Table 3: Lexical diversity of LLM translations across refinement stages and temperature settings compared with human translation.

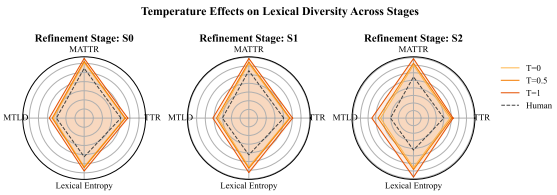


Figure 1: Radar plots of lexical diversity of LLM translations across refinement stages and temperature settings compared with human translation.

| Refinement Stage | Top-p Setting | Lexical Diversity | | | |
|---|---|---|---|---|---|
| | | TTR | MATTR | MTLD | Lexical Entropy |
| S0 | 0.5 | 0.3675 | 0.741 | 113.457 | 8.3446 |
| | 0.8 | 0.37 | 0.7426 | 115.2529 | 8.3609 |
| | 1 | **0.3721** | **0.7452** | **117.8178** | **8.3732** |
| S1 | 0.5 | 0.3754 | 0.746 | 118.061 | 8.4006 |
| | 0.8 | 0.3776 | 0.7483 | 121.4817 | 8.4235 |
| | 1 | **0.38** | **0.7508** | **123.7239** | **8.437** |
| S2 | 0.5 | 0.3848 | 0.7635 | 141.9736 | 8.5818 |
| | 0.8 | 0.3872 | 0.7672 | 146.912 | 8.6259 |
| | 1 | **0.3919** | **0.7697** | **151.5484** | **8.6569** |
| Human Translation | | 0.3595 | 0.7285 | 106.5089 | 8.1305 |

Table 4: Lexical diversity of LLM translations across refinement stages and top-p settings compared with human translation.

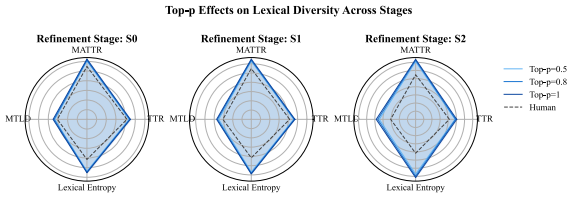


Figure 2: Radar plots of lexical diversity of LLM translations across refinement stages and top-p settings compared with human translation.

### 4.3 Lexical Frequency Profile

Overall, iterative refinement consistently shifted vocabulary toward lower-frequency words across all decoding settings (Figure 3 and Figure 4). Across refinement stages, the proportion of the most frequent 1,000 words (B1) declined from the initial draft (S0) to the final draft (S2), while the proportion of words outside the top 2,000 most frequent words (B3) increased correspondingly. At T = 0.5, for instance, B1 dropped from 69.56% at S0 to 63.00% at S2, while B3 rose from 14.96% to 20.04%.

Temperature had a clear effect on lexical frequency distribution (Figure 3). Higher temperatures produced translations with lower proportions of high-frequency words and greater use of lower-frequency vocabulary. At S2, B1 decreased from 64.09% at T = 0 to 61.59% at T = 1, while B3 increased from 19.13% to 21.49%. This pattern suggests that higher temperatures increase the likelihood of selecting less common lexical items from a broader range of candidates.

Top-p showed a similar directional pattern (Figure 4). Increasing top-p from 0.5 to 1 at S2 reduced B1 from 64.44% to 63.50% and increased B3 from 19.22% to 20.04%. As with temperature, expanding the candidate token space was associated with a shift toward lower-frequency vocabulary.

Compared with human translations, LLM outputs contained fewer high-frequency words and more low-frequency vocabulary at every stage. This tendency intensified with iterative refinement, with final drafts (S2) showing the greatest reliance on lower-frequency items. Rather than converging toward human lexical norms, successive refinement therefore appears to increase divergence from the human lexical frequency

profile, pushing LLM translations toward more atypical vocabulary choices.

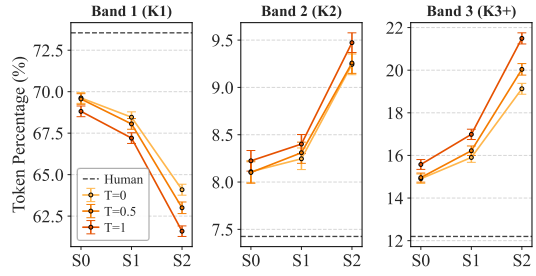


Figure 3: Lexical frequency profiles of LLM-generated translations across refinement stages and temperature settings. Error bars indicate standard error.

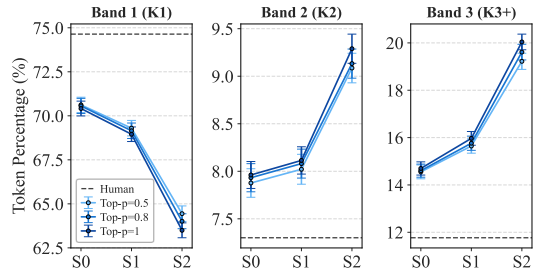


Figure 4: Lexical frequency profiles of LLM-generated translations across refinement stages and top-p settings. Error bars indicate standard error.

The differences in lexical frequency profiles between LLM-generated and human translations suggest distinct underlying orientations toward stylistic quality. Even at the initial stage, and under deterministic decoding, LLM outputs exhibit a higher proportion of lower-frequency vocabulary. This pattern appears to challenge theoretical expectations that LLM-generated texts converge toward high-probability forms due to the probabilistic nature of language modeling (Uhlmann, 2025; Vanmassenhove, 2025). However, the tendency toward sophisticated vocabulary may instead reflect biases in the model's training data or its learned associations between literary translation and elevated register. Specifically, what constitutes a high-probability token is determined by the model's training distribution. If the training data overrepresents certain registers or genres, lower-frequency vocabulary in general usage may in fact be high-probability within the model's learned distribution. Iterative refinement appears to strengthen this tendency, with each stage shifting outputs away from high-frequency vocabulary toward more marked forms. One possible interpretation is that the model implicitly associates lexical sophistication with stylistic improvement, though this remains speculative without further investigation.

The implications of these findings for translation quality, however, remain uncertain. Without a benchmark based on original non-translated English literary fiction, it is difficult to determine whether LLM outputs are genuinely over-sophisticated or whether human translations are comparatively simplified. The model's preference for lower-frequency vocabulary might align more closely with native literary prose than with translated text, which complicates straightforward quality judgments.

### 4.4 Inter-Stage Lexical Similarity

As shown in Table 5 and Table 6, both temperature and top-p are positively associated with edit distance: higher settings lead to more extensive textual modification at each stage (Figure 5 and Figure 6). While word-level distance serves as the primary measure, character-level distances follow a similar pattern, indicating that revisions involve both broad lexical restructuring and fine-grained orthographic adjustments.

During the self-revision stage (S0–S1), higher parameter settings result in substantially more editing. Normalized word-level distance more than doubles from 0.19 at T = 0 to 0.42 at T = 1, and increases from 0.21 at top-p = 0.5 to 0.32 at top-p = 1. The stylistic refinement stage (S1–S2) introduces even greater changes: at maximum settings (T = 1 or top-p = 1), normalized word distance reaches 0.67 and 0.59, respectively.

In the cumulative refinement process (S0–S2), texts undergo substantial lexical shifts. Normalized word distance reaches 0.72 at T = 1 (compared to 0.54 at T = 0) and 0.65 at top-p = 1 (compared to 0.58 at top-p = 0.5). These values indicate that the number of edit operations approaches or exceeds two-thirds of the total token count under less restrictive decoding settings, reflecting extensive rewriting across the pipeline.

| Stage Pair | Temperature Setting | Levenshtein Distance | | | |
|---|---|---|---|---|---|
| | | Char. Dist. | Norm. Char. Dist. | Word Dist. | Norm. Word Dist. |
| S0-S1 | 0 | 1961.23 | 0.1599 | 401.22 | 0.1888 |
| | 0.5 | 2582.14 | 0.2095 | 542.93 | 0.2548 |
| | 1 | **4396.69** | **0.3403** | **937.82** | **0.4231** |
| S1-S2 | 0 | 6084.5 | 0.4107 | 1232.42 | 0.4905 |
| | 0.5 | 6567.87 | 0.4358 | 1331.04 | 0.5248 |
| | 1 | **9480.09** | **0.5481** | **1920.58** | **0.6683** |
| S0-S2 | 0 | 6666.16 | 0.4507 | 1360.75 | 0.5423 |
| | 0.5 | 7247.23 | 0.4821 | 1484.79 | 0.5869 |
| | 1 | **10180.65** | **0.5895** | **2076.06** | **0.7234** |

*Note.* Char. = Character, Dist. = Levenshtein edit distance, Norm. = Normalized by the length of the longer text in each comparison.

Table 5: Inter-stage edit distance of LLM translations across temperature settings.

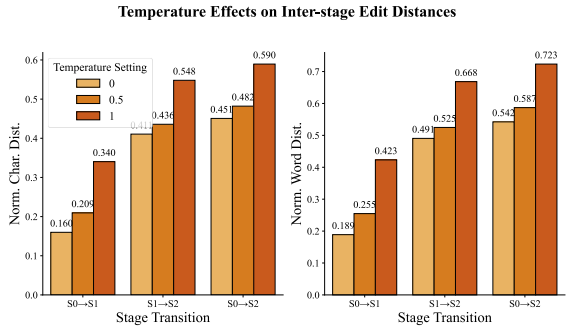


Figure 5: Bar plots of normalized edit distances across refinement stages and temperature settings.

| Stage Pair | Top-p Setting | Levenshtein Distance | | | |
|---|---|---|---|---|---|
| | | Char. Dist. | Norm. Char. Dist. | Word Dist. | Norm. Word Dist. |
| S0-S1 | 0.5 | 2144.04 | 0.1727 | 440.14 | 0.2045 |
| | 0.8 | 2831.98 | 0.2247 | 597.02 | 0.2737 |
| | 1 | **3283.07** | **0.259** | **696.09** | **0.3175** |
| S1-S2 | 0.5 | 6651.39 | 0.4384 | 1345.54 | 0.5241 |
| | 0.8 | 7316.58 | 0.4647 | 1484.33 | 0.5596 |
| | 1 | **7826.9** | **0.4888** | **1586.15** | **0.5898** |
| S0-S2 | 0.5 | 7239.19 | 0.4773 | 1476.78 | 0.5753 |
| | 0.8 | 8006.92 | 0.509 | 1637.37 | 0.6177 |
| | 1 | **8516.1** | **0.5327** | **1743.8** | **0.6492** |

*Note.* Char. = Character, Dist. = Levenshtein edit distance, Norm. = Normalized by the length of the longer text in each comparison.

Table 6: Inter-stage edit distance of LLM translations across top-p settings.

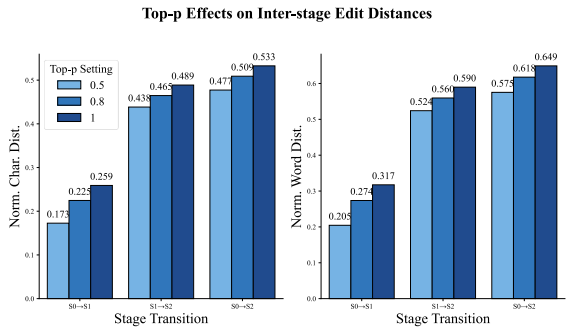


Figure 6: Bar plots of normalized edit distances across refinement stages and top-p settings.

The edit distance patterns have implications for understanding how parameter settings modulate the extent to which an LLM alters its output across refinement stages. Given the same prompt, more deterministic settings lead the model to preserve much of its prior output while making adjustments, whereas less restrictive settings produce more substantial departures from previous iterations, generating extensive rewrites. Whether this represents a more thorough revision or simply increased stochastic variation remains unclear; higher temperature and top-p values mechanically expand the candidate pool at each token position, which may produce surface-level instability without corresponding gains in quality or coherence. Disentangling these possibilities would require complementary quality evaluation.

## 5 Conclusion

This study examined how decoding parameters (temperature and top-p) influence the lexical characteristics of LLM-generated translations across an iterative refinement workflow. The findings reveal a consistent pattern: less restrictive parameter settings produce greater lexical diversity, more low-frequency vocabulary, and more extensive inter-stage revision. These effects are amplified through iterative refinement, with the stylistic refinement stage showing the most pronounced divergence from both earlier drafts and human translations. Notably, LLM outputs displayed higher lexical diversity and larger proportions of lower-frequency vocabulary than human translations across all settings, and these differences widened through successive refinement stages.

These findings carry tentative practical implications. Translators working with LLMs may benefit from understanding how parameter settings shape lexical output rather than relying on defaults or intuition. More restrictive settings may suit tasks requiring accessible, high-frequency vocabulary, while less restrictive settings might benefit texts where lexical richness is valued. Awareness that iterative refinement progressively increases lexical diversity while shifting vocabulary toward lower-frequency bands could inform decisions about how many revision stages to employ before such effects become counterproductive.

However, several limitations of this study should be acknowledged. First, this study focused exclusively on lexical profiles without evaluating translation quality; a richer vocabulary does not necessarily indicate better translation, and the trade-offs between output diversity and quality warrant further investigation. Second, the analysis was limited to a single LLM (GPT-5.1) and a single broad domain (literary fiction), which constrains generalizability. Finally, only Chinese-to-English translation was examined; parameter effects may differ for other language pairs. Future research should address these gaps by integrating quality evaluation, expanding sample sizes, and extending the analysis across models, text types, and language pairs.

## References

Gao, Ruiyao, Yumeng Lin, Nan Zhao, and Zhenguang G. Cai. 2024. Machine translation of Chinese classical poetry: A comparison among ChatGPT,

Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications*, 11(1):1–10. https://doi.org/10.1057/s41599-024-03363-0.

Jiang, Zhaokun, Qianxi Lv, Ziyin Zhang, and Lei Lei. 2024. Convergences and divergences between automatic assessment and human evaluation: Insights from comparing ChatGPT-generated translation and neural machine translation. *arXiv preprint arXiv:2401.05176*.

Wang, Lulu, Sanjun Sun, Xing Wang, Jinghang Gu, and Kanglong Liu. 2026. Workflow matters: Comparing human translators and multi-agent LLMs in literary translation. *Target*. https://doi.org/10.1075/target.25081.wan.

Wu, Minghao, Jiahao Xu, and Longyue Wang. 2024. TransAgents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141. https://doi.org/10.18653/v1/2024.emnlp-demo.14.

Wu, Minghao, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. (Perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922. https://doi.org/10.1162/tacl.a.25.

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. https://aclanthology.org/2024.eamt-1.17.

Madaan, Aman, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Xu, Wenda, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492. https://doi.org/10.18653/v1/2024.acl-long.826.

Leech, Geoffrey N. and Mick Short. 2007. *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. Pearson Education. https://doi.org/10.2307/1772012.

Munat, Judith. 2007. Lexical creativity as a marker of style in science fiction and children's literature. In *Lexical Creativity, Texts and Contexts*, pages 163–185. John Benjamins, Amsterdam & Philadelphia, PA. https://doi.org/10.1075/sfsl.58.20mun.

Wiher, Gian, Clara Meister, and Ryan Cotterell. 2022. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012. https://doi.org/10.1162/tacl_a_00502.

Zhou, Yuxuan, Margret Keuper, and Mario Fritz. 2025. Balancing diversity and risk in LLM sampling: How to select your method and parameter for open-ended text generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26352–26365. https://doi.org/10.18653/v1/2025.acl-long.1278.

Arias, Esteban Garces, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2025. Decoding decoded: Understanding hyperparameter effects in open-ended text generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9992–10020. https://aclanthology.org/2025.coling-main.668.

Troshin, Sergey, Wafaa Mohammed, Yan Meng, Christof Monz, Antske Fokkens, and Vlad Niculae. 2025. Control the Temperature: Selective Sampling for Diverse and High-Quality LLM Outputs. arXiv:2510.01218.

Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. ICLR 2020.

Ippolito, Daphne, Reno Kriz, Joao Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762. https://doi.org/10.18653/v1/P19-1365.

Peng, Keqin, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633. https://doi.org/10.18653/v1/2023.findings-emnlp.373.

Arora, Chetan, Ahnaf Ibn Sayeed, Sherlock Licorish, Fanyu Wang, and Christoph Treude. 2024. Optimizing LLMs for code generation: Which hyperparameter settings yield the best results? In *2024 31st Asia-Pacific Software Engineering Conference (APSEC)*, pages 281–290. IEEE. https://doi.org/10.1109/APSEC65559.2024.00039.

Donato, Benedetta, Leonardo Mariani, Daniela Micucci, and Oliviero Riganelli. 2025. Studying how configurations impact code generation in LLMs: The case of ChatGPT. arXiv:2502.17450.

Peeperkorn, Max, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? arXiv:2405.00492.

Grullon-Polanco, Larry. 2026. LLM-assisted genre analysis: The effect of sampling temperature on reliability. *Research Methods in Applied Linguistics*, 5(2):100314. https://doi.org/10.1016/j.rmal.2026.100314.

Li, Lujun, Lama Sleem, Niccolo Gentile, Geoffrey Nichil, and Radu State. 2025. Exploring the impact of temperature on large language models: Hot or cold? *Procedia Computer Science*, 264:242–251. https://doi.org/10.1016/j.procs.2025.07.135.

Jakobson, Roman. 1959. On linguistic aspects of translation. In Reuben Arthur Brower, editor, *On Translation*, pages 232–239. Harvard University Press, Cambridge, MA. https://doi.org/10.4159/harvard.9780674731615.c18.

Macken, Lieve, Bram Vanroy, Luca Desmet, and Arda Tezcan. 2022. Literary translation as a three-stage process: Machine translation, post-editing and revision. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 101–110. https://aclanthology.org/2022.eamt-1.13.

Whyatt, Boguslawa. 2017. Intralingual translation. In John W. Schwieter and Li Wei, editors, *The Handbook of Translation and Cognition*, pages 176–192. John Wiley & Sons, Hoboken, NJ. https://doi.org/10.1002/9781119241485.ch10.

International Organization for Standardization. 2017. *ISO 18587:2017: Translation services — Post-editing of machine translation output — Requirements*. International Organization for Standardization, Geneva, Switzerland.

Cai, Chengkun, Xu Zhao, Yucheng Du, Haoliang Liu, and Lei Li. 2024. T² of Thoughts: Temperature Tree Elicits Reasoning in Large Language Models. arXiv:2405.14075.

Zhang, Ran, Wei Zhao, and Steffen Eger. 2025. How Good Are LLMs for Literary Translation, Really? Literary Translation Evaluation with Humans and LLMs. In *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. https://doi.org/10.18653/v1/2025.naacl-long.548.

Alrabaee, Saed, Paria Shirani, Lingyu Wang, Mourad Debbabi, and Aiman Hanna. 2019. Decoupling coding habits from functionality for effective binary authorship attribution. *Journal of Computer Security*, 27(6):613–648. https://doi.org/10.3233/JCS-191292.

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. https://doi.org/10.1080/09296171003643098.

Malvern, David D. and Brian J. Richards. 1997. A new measure of lexical diversity. *British Studies in Applied Linguistics*, 12:58–71.

McCarthy, Philip M. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Mohseni, Mahdi, Christoph Redies, and Volker Gast. 2022. Approximate entropy in canonical and non-canonical fiction. *Entropy*, 24(2):278. https://doi.org/10.3390/e24020278.

Laufer, Batia and Paul Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3):307–322. https://doi.org/10.1093/applin/16.3.307.

Van Parys, Amaury, Vanessa De Wilde, Lieve Macken, and Maribel Montero Perez. 2025. LexPro: A plurilingual lexical profiling tool to assist teachers and researchers in analysing vocabulary of L2 input. *Language Teaching Research*, 13621688251352259. https://doi.org/10.1177/13621688251352259.

Vanmassenhove, Eva. 2025. Losing our tail—again: On (un)natural selection and multilingual large language models. arXiv:2507.03933.

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. https://doi.org/10.18653/v1/2021.eacl-main.188.

Anson, Ian G., Cary Moskovitz, and Chris M. Anson. 2019. A text-analytic method for identifying text recycling in STEM research reports. *The Journal of Writing Analytics*, 3. https://doi.org/10.37514/JWA-J.2019.3.1.07.

Shypula, Alexander, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. 2025. Evaluating the diversity and quality of LLM generated content. arXiv:2504.12522.

McCarthy, Philip M. and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392. https://doi.org/10.3758/BRM.42.2.381.

Vermeer, Anne. 2000. Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, 17(1):65–83. https://doi.org/10.1177/026553220001700103.

Koizumi, Rie. 2012. Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, 1(1):60–69. https://doi.org/10.7820/vli.v01.1.koizumi.

Shannon, Claude E. 1951. Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1):50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x.

Nation, I.S.P. 2017. The BNC/COCA Level 6 word family lists (Version 1.0.0) [Data file]. http://www.victoria.ac.nz/lals/staff/paul-nation.aspx.

Uhlmann, Gyburg. 2025. The even sheen of AI: Kitsch, LLMs, and homogeneity. arXiv:2509.16794.