Received: 09 Aug. 2026 | Revised: 10 Aug. 2026 | Accepted: 13 Aug. 2026 | Published online: 19 Aug. 2026

RESEARCH ARTICLE

Journal of Global Humanities and Social Sciences
Generative AI in the Humanities: Methods, Applications, and Critical Perspectives
2026,Vol. 1 1-9
DOI: 10.61360/BoniGHSSAH262020650101

Lexicalizing Affective Meaning in Published and LLM-Generated Literary Translation: A Corpus-Based Study of Chinese Translations of Margaret Mitchell's *Gone with the Wind*

Dandan Wu[1,*] & Yuxi Chang[1]

[1]*Beijing International Studies University, China.*

Abstract: This study examines how affective and evaluative meaning is lexicalized in published Chinese translations and large language model (LLM) outputs of literary prose. The corpus comprises 98 purposively selected English sentences from Margaret Mitchell's *Gone with the Wind*, three published Chinese translations, and twelve LLM outputs generated by six systems under a general translation prompt and an emotion-and-style-oriented prompt. Affective explicitness is operationalized through the density and type range of contextually validated entries in the DUTIR Chinese Emotion Ontology; lexical convergence through word- and category-level Jaccard overlap among versions aligned to the same source sentence. The comparisons use 36 complete-case sentence groups in which all 15 versions contain at least one retained DUTIR item. In this sensitivity sample, LLM outputs show higher dictionary-recognized affective-word density and broader type ranges than the published translations, alongside greater within-condition lexical convergence. The second prompt leaves aggregate density largely unchanged, increases distinct-type counts, and reduces cross-model word overlap. The findings treat lexical visibility and convergence as properties of output form rather than proxies for affective intensity, literary quality, or translation superiority.

Keywords: literary translation, affective meaning, LLM translation, lexical explicitness, prompting

1. Introduction

Literary translation does not merely transfer propositional content. It also reorganizes the linguistic resources through which a text represents feeling, evaluation, social attitude, and degrees of interpretive openness. In a literary sentence, affective meaning may be named directly, implied through action, distributed across syntax, or refracted through irony and narrative perspective. A target text may therefore preserve a comparable affective orientation while changing how visibly that orientation is lexicalized and how many alternative readings remain available to the reader.

The emergence of large language models makes this problem newly consequential. Recent comparative studies have examined machine- and LLM-mediated translation in subtitles, canonical fiction, poetry, and *literary prose* (Calvo-Ferrer, 2024; Chen et al., 2026; Gao et al., 2024; Hu & Li, 2023). Related work on pain, suffering, and sentiment in literary and cross-cultural translation also shows that affective material is shaped through contextual mediation rather than through direct labels alone (Liu, 2026; Terando, 2023). LLMs can produce fluent *literary prose* and can be instructed to attend to particular textual features, yet fluency does not by itself reveal how a system chooses among explicit emotion words, evaluative expressions, figurative formulations, and context-dependent implications. Published translations are the outcome of

Corresponding Author: Dandan Wu
Beijing International Studies University, China

individual interpretation, revision, editorial mediation, and sustained engagement with a whole work, whereas an LLM output is generated under a specified technical condition. A human-centred account of augmented translation therefore treats these conditions as complementary but non-equivalent rather than as evidence for a simple human–machine hierarchy (O'Brien, 2024, pp. 391–393).

This article focuses on three related phenomena. The first is affective explicitness, namely the extent to which affective or evaluative meaning is realized through dictionary-recognized lexical items rather than through implication, imagery, syntax, or narrative context. The second is lexical convergence, the degree to which different versions choose the same emotion words or emotion categories. The third is prompt-conditioned variation, understood as changes in lexical range and overlap associated with an emotion-and-style-oriented instruction. These distinctions prevent a higher frequency of emotion words from being mistaken for greater emotional intensity, and prevent a higher Jaccard coefficient from being treated as evidence of translation quality.

The corpus brings together 98 emotionally or evaluatively marked sentences from *Gone with the Wind*, three published Chinese versions, and twelve LLM outputs generated by six systems under two prompts. The article asks whether published translations and LLM outputs differ in the explicit lexical realization of affective meaning, whether LLM outputs converge more strongly on common lexical choices, and whether an emotion-and-style-oriented prompt changes lexical diversity and convergence without necessarily changing aggregate density or broad category distribution.

The contribution is methodological as well as descriptive. It places a dictionary-based measure within an Appraisal-informed interpretive framework, distinguishes lexical visibility from affective intensity, and treats prompting as a condition that may redistribute lexical choices rather than as a universal quality-improvement device. The study therefore contributes evidence about the form and regularity of LLM-generated literary translation while keeping lexical counts separate from claims about literary adequacy.

2. Theoretical Orientation and Analytical Distinctions

2.1 Appraisal and affective lexicalization

Appraisal theory, developed within systemic functional linguistics, models evaluation as a discourse-semantic resource rather than as a list of isolated emotion words (Martin & White, 2005, pp. 34–40). Within the subsystem of Attitude, Affect concerns the representation of emotion, while Judgment and Appreciation capture evaluations of people, behaviour, and things that may be emotionally consequential without naming a psychological state (Martin & White, 2005, pp. 45–56). The framework is used here to identify the affective or evaluative function of a source expression and to interpret its target-text realization; it is not used to equate every dictionary entry with an Appraisal category.

Martin and White show that Affect may be realized as a quality, a process, a comment, or a nominalized form (2005, pp. 45–46). They also distinguish attitude that is directly inscribed through attitudinal lexis from attitude that is indirectly invoked through ideational meanings and contextual patterning (2005, pp. 61–67). These distinctions are decisive for the present design: a DUTIR match is best treated as evidence of one visible form of affective lexicalization, while invoked, figurative, syntactic, and prosodic meanings may remain outside the count. In corpus-based emotion research, emotion talk can evaluate entities and behaviour and characterize a participant rather than merely label a momentary emotional state (Bednarek, 2008, p. 2).

Graduation provides a further contextual dimension. Force scales intensity or amount, whereas Focus sharpens or softens category boundaries (Martin & White, 2005, pp. 135–141). A choice such as "可怕" may evaluate a situation, describe an affective response, or characterize a social norm depending on context. Similarly, "兴趣" may encode inclination, while "乐意" foregrounds willingness. The analysis therefore treats DUTIR matches as operational observations of lexicalization. Their semantic role remains dependent on co-text, negation, degree marking, narrative perspective, and the relation between explicit wording and implied meaning. The broader use of Appraisal for tracing evaluative positioning in discourse provides a

methodological precedent for this contextual reading (Hood, 2010, pp. 1–6).

2.2 Affective explicitness, lexical convergence, and translation quality

Corpus-based translation studies make it possible to examine translated language as an empirical object and to compare recurring choices across aligned texts (Baker, 1993, pp. 233–244). The present study adopts that empirical orientation but limits its claims to observable output form. Affectively explicit translation is not necessarily more intense translation. A source emotion may be reconstructed through an explicit emotion noun, an evaluative adjective, a verb of feeling, an image, a syntactic construction, or a pragmatic implication. A dictionary count captures only one of these resources. The present study consequently uses "emotion-word density" as an operational measure of dictionary-recognized lexicalization, not as a direct measure of emotional force or reader response.

The same restriction applies to quality. Translation is a re-contextualizing linguistic-textual operation, and quality assessment requires textual description and culturally informed comparison before a value judgement is made (House, 2015, pp. 1–3). A lexical count cannot therefore establish literary adequacy, stylistic value, or reader response. The present analysis uses convergence as a descriptive property of aligned outputs: a high Jaccard coefficient can indicate shared probable equivalents, a shared prompt-induced tendency, or a genuinely constrained lexical field. It does not by itself indicate poor translation. Following corpus-based emotion profiling, automated or dictionary-assisted analysis is combined with contextual manual scrutiny (Bednarek, 2008, pp. 27–28).

3. Corpus and Methods

3.1 Source text and published translations

The source text is Margaret Mitchell's *Gone with the Wind*, first published by The Macmillan Company in 1936; the working edition used in the project is the 1947 Macmillan edition recorded in the original manuscript. Candidate sentences were selected when an emotion-related or evaluative expression was central to the sentence and could be aligned across the published and generated versions. Contextual screening removed structurally ambiguous and analytically unsuitable cases. The resulting 98 sentences constitute a purposive sample of affectively and evaluatively marked prose, not a statistical sample of the novel's emotional profile. Each source sentence was paired with three published Chinese translations: Fu Donghua's version, Li Meihua's version, and the collaborative edition translated by Dai Kexin, Li Yuzhong, and Zhuang Yuxin. Their publication histories and editorial conditions differ from those of the LLM outputs. The comparison therefore concerns translation patterns under different production conditions; it does not compare isolated translators and models as equivalent agents.

3.2 LLM outputs and prompt conditions

The LLM outputs were generated through the systems' APIs in May 2026. The recorded identifiers were GPT-5.5, Claude Opus 4.7 Thinking, Gemini 3.1 Pro Preview Thinking, DeepSeek V4 Pro, Doubao Seed 2.0 Pro, and Kimi K2.6. Temperature was set to 0 where supported, no additional system prompt was supplied, and the source sentence was accompanied by surrounding context. The output instruction required the final Chinese translation only.

Prompt A was: "Please translate the following literary sentence into Chinese. The context is given for reference. (Output the final version only.)" Prompt B added: "Pay particular attention to preserving the emotional intensity, affective tone, irony, and stylistic tension of the original content." Because Prompt B simultaneously invokes emotion, irony, and style, it is referred to here as an emotion-and-style-oriented prompt, not as a pure emotion manipulation. The current design contains one generation per model, sentence, and prompt. Accordingly, prompt effects are interpreted descriptively and should be replicated with repeated generations before strong causal claims are made.

3.3 DUTIR processing and coverage

The lexical analysis uses the DUTIR Chinese Emotion Ontology (Xu et al., 2008), which contains seven broad categories: joy, positive evaluation, anger, sadness, fear, disgust, and surprise. Chinese texts were segmented with Jieba 0.42.1 after DUTIR entries had been added to the segmentation dictionary. A candidate was retained only after contextual manual review.

Personal names, grammatical fragments, segmentation errors, polysemous items used non-emotionally, neutral event descriptions, and other false matches were removed. Negation was checked within the preceding three tokens; an odd number of relevant negators reverses polarity. Degree adverbs were not assigned numerical weights.

The lexical comparisons reported here use the 36 source-sentence groups in which all 15 versions contained at least one validated DUTIR match after contextual review. The remaining 62 groups are outside the complete-case comparisons because the study does not assign zero counts to unmatched outputs. The complete-case results are therefore interpreted as sensitivity evidence about groups with observable dictionary-recognized affective lexicalization, rather than as estimates of the novel as a whole. A zero-inclusive extension would incorporate the remaining groups and report coverage rates alongside lexical density.

3.4 Measures

Emotion-word density is the number of valid DUTIR emotion-word tokens divided by the number of segmented tokens, multiplied by 100. Distinct-type counts indicate the number of different validated DUTIR word types in a version; they are not a complete measure of lexical diversity and are not called vocabulary richness. They are used here as a simple indicator of the range of dictionary-recognized lexical choices within the aligned sample. For a complete corpus analysis, type-token ratio, MTLD or HD-D, and zero-inclusive coverage rates should be added.

For two translations of the same source sentence, lexical convergence is calculated as $J(A,B)=|A\cap B|/|A\cup B|$, once for emotion-word sets and once for category sets. Values are first calculated within source-sentence groups and then averaged. The analysis reports unweighted Jaccard means. Because Jaccard ignores frequency and graded semantic proximity, future research should add a weighted or frequency-sensitive similarity measure and should report sentence-level distributions with cluster-bootstrap intervals. The corpus flow and the split between the 36 complete-case groups and the 62 groups reserved for a zero-inclusive extension are summarized in Figure 4.

4. Results

4.1 Analytical scope and complete-case coverage

The lexical results are based on a manually reviewed complete-case subset of 36 source-sentence groups in which all 15 versions contained at least one retained DUTIR item. The remaining 62 groups are outside the complete-case comparisons because the study does not assign zero counts to unmatched outputs. The reported density, type-range, and overlap values therefore describe lexical behaviour in groups with observable dictionary matches across all conditions, rather than the affective profile of the novel as a whole. The complete-case restriction also determines the interpretation of the comparisons. It does not test whether a version contains more affective meaning in an absolute sense, because implicit, figurative, and non-dictionary realizations are not counted. It tests how readily selected affective meanings become visible through the operational lexicon under the same sentence-level alignment.

4.2 Affective categories and polarity profiles

As shown in Table 1, negative classifications were the largest polarity category in the three groups: 53.7% for published human, 44.0% for Prompt A, and 50.5% for Prompt B. Positive classifications accounted for 29.6%, 35.2%, and 35.2%. The equal positive proportions in the two LLM conditions indicate no aggregate shift in this coarse polarity measure. Prompt B was associated with fewer balanced cases, declining from 19.0% to 12.5%, and with a corresponding increase in negative cases. These are descriptive proportions, not estimates of the novel's overall emotional composition.

At the seven-category level, positive evaluation and disgust/negative evaluation were the two most frequent categories in all three groups. Together they represented 54.8% of validated matches in the published-human group, 55.4% in Prompt A, and 56.4% in Prompt B. Fear accounted for 15.8%, 13.8%, and 13.8%, respectively. The stability of category distributions alongside changes in individual word choice is important: prompts may alter lexical realization without changing the broader affective field.

Table 1. Descriptive DUTIR results.

Group	Versions	Words/100	Mean distinct types	Positive	Negative	Balanced	Neutral
Published human	3	7.991	64.7	29.6%	53.7%	11.1%	5.6%
LLM–Prompt A	6	9.291	84.5	35.2%	44.0%	19.0%	1.9%
LLM–Prompt B	6	9.294	89.0	35.2%	50.5%	12.5%	1.9%

4.3 Affective-word density and type range

The published-human translations contain 7.991 validated DUTIR matches per 100 segmented tokens, compared with 9.291 under Prompt A and 9.294 under Prompt B. As shown in Figure 1, the complete-case summaries therefore indicate more explicit dictionary-recognized affective lexicalization in the LLM outputs. The mean number of distinct DUTIR types is 64.7 for published human, 84.5 for Prompt A, and 89.0 for Prompt B. Prompt B increased the distinct-type count in five of the six models, although density increased in only three. The prompt's more consistent effect on type range than on token density suggests redistribution of lexical alternatives rather than a uniform increase in affective wording.

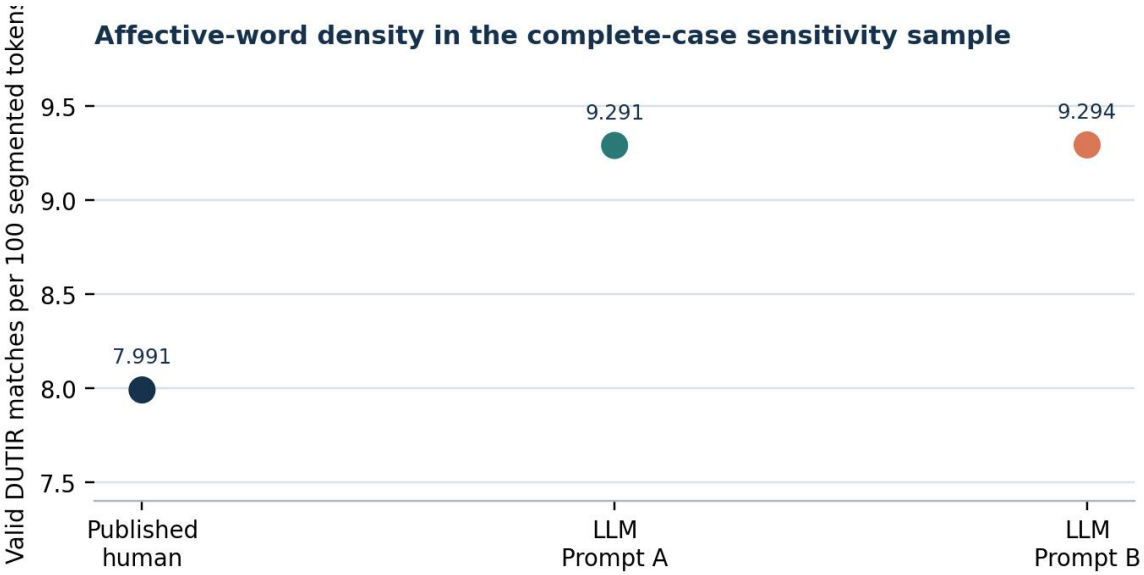

Figure 1. Group-level affective-word density in the 36-group complete-case sensitivity sample. The figure reports descriptive means for dictionary-recognized affective-word tokens per 100 segmented tokens.

4.4 Lexical convergence

The mean emotion-word Jaccard coefficient was 0.226 among the three published-human versions, 0.415 among Prompt-A LLM pairs, and 0.322 among Prompt-B LLM pairs. Category-level Jaccard values were 0.583, 0.753, and 0.723, respectively. Prompt B therefore reduced cross-model word overlap by approximately 22.4%, whereas the reduction in category overlap was approximately 4.0%. Human–LLM word overlap declined from 0.237 under Prompt A to 0.164 under Prompt B. The result indicates that Prompt B diversified model-specific word selection without moving the outputs closer to any one published translation. The full comparison is reported in Table 2 and visualized in Figure 2.

Table 2. Mean Jaccard similarity.

Pairing group	Number of pairings	Emotion-word Jaccard	Emotion-category Jaccard
Human–Human	3	0.226	0.583
LLM A–A	15	0.415	0.753
LLM B–B	15	0.322	0.723
Human–LLM A	18	0.237	0.625
Human–LLM B	18	0.164	0.585

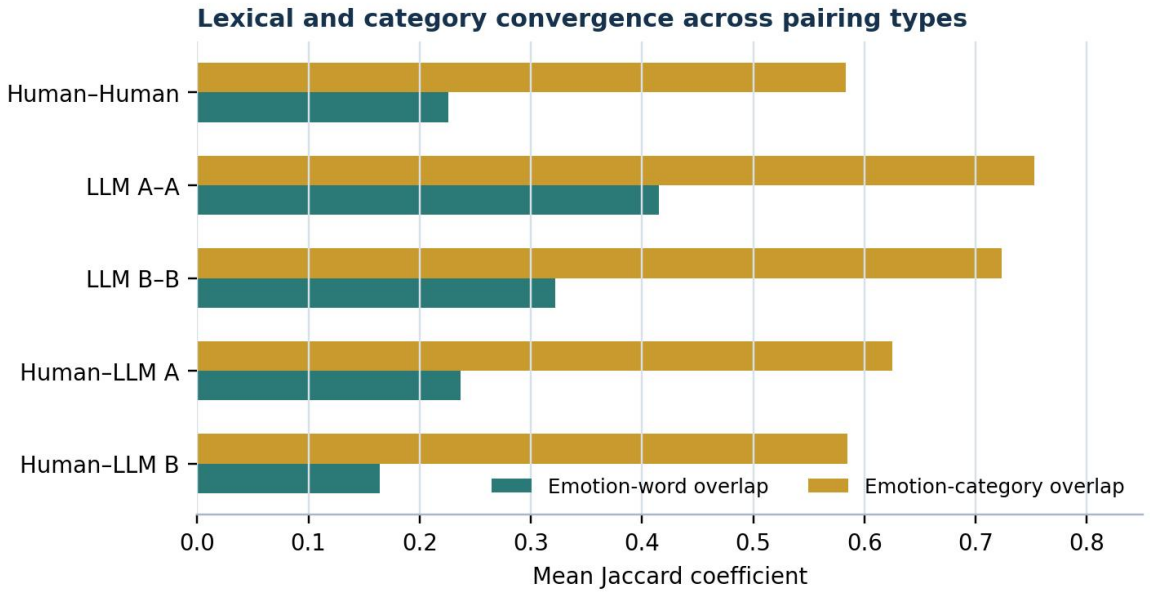

Figure 2. Mean lexical and category convergence by pairing type in the complete-case sensitivity sample. Jaccard values are overlap measures, not quality scores.

4.5 Prompt-conditioned lexical variation

The mean similarity between the Prompt-A and Prompt-B outputs of the same model was 0.428 at the word level and 0.762 at the category level. This relatively high within-model overlap indicates modification rather than wholesale replacement of each model's lexical pattern. At the group level, density changed from 9.291 to 9.294, distinct types from 84.5 to 89.0, and LLM cross-model word Jaccard from 0.415 to 0.322. As summarized in Figure 3, the prompt thus appears more consequential for lexical alternatives and cross-model convergence than for aggregate density or broad polarity.

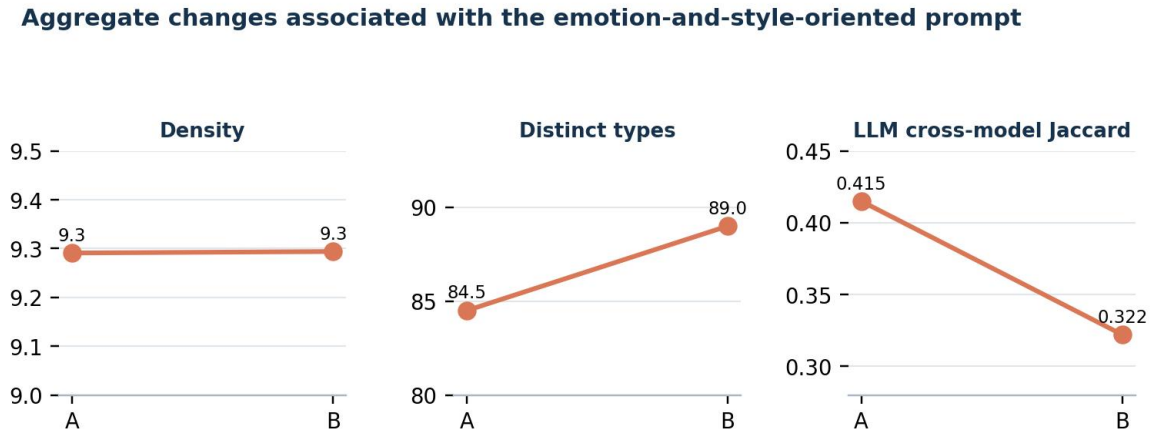

Figure 3. Aggregate changes associated with the emotion-and-style-oriented prompt. Panel scales differ; the figure is descriptive and does not supply inferential uncertainty.

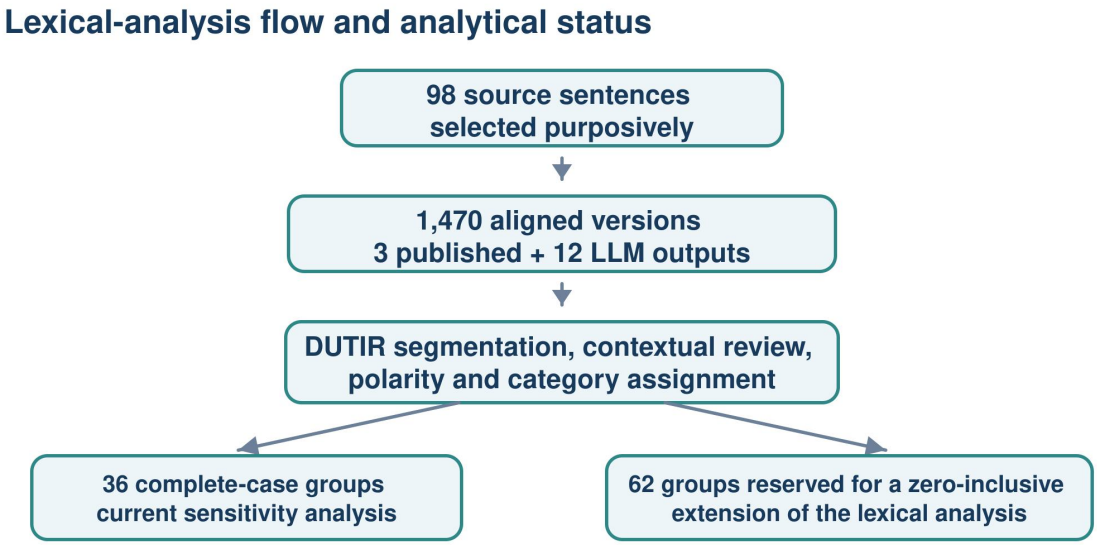

Figure 4. Corpus flow and analytical status in the lexical analysis.

5. Discussion

5.1 Lexical explicitness should not be read as intensity

The higher DUTIR density in the LLM outputs is compatible with a tendency toward more explicit affective lexicalization. It is not evidence that the outputs create stronger emotional effects or reproduce the source emotion more faithfully. Sentence 27, as represented in the aligned sample, presents passion through male madness and describes consequences as painful and embarrassing. The published forms include "疯狂", "奴役", "尴尬", "狂热", and "痛苦", while several LLM outputs cluster terms such as "疯狂", "难堪", "屈从", and "痛苦". A cluster of recognizable emotion words makes the affective structure easier to detect computationally, but it can also reduce ambiguity or stylistic restraint. The defensible conclusion is greater lexical explicitness within the operational dictionary, not greater affective intensity.

5.2 Cross-model convergence as output regularity

The higher LLM–LLM Jaccard values suggest that different systems more often select the same or closely

related lexical realizations than the three published versions do. Sentence 18 offers the clearest reported example. For the statement that it was "most dreadful" for young women to show interest in gentlemen, all six Prompt-A outputs reportedly contained the validated set "可怕" and "兴趣". Two human versions also used these items, while another used "乐意" for part of the meaning. The LLM agreement is compatible with reliance on highly probable, standardized equivalents. It does not prove a specific training-data mechanism, and it should not be treated as a defect without contextual examination.

In literary translation, convergence matters because lexical repetition can reduce the visibility of individual interpretive decisions. Yet repeated choices may also be accurate and stylistically appropriate. A publishable version of the study should therefore combine Jaccard with close reading and, ideally, a human judgment of contextual appropriateness. The current result supports a claim about output regularity, not a claim about inferior creativity.

5.3 What the emotion-and-style prompt changes

Prompt B reduces word-level convergence and increases distinct-type counts in most models, while leaving mean density and positive polarity almost unchanged. Sentence 25 provides a plausible illustration: the reported Prompt-B forms include "气愤", "勃然大怒", "恼怒", "怒不可遏", and "火冒三丈". These items occupy a broadly similar negative field but differ in intensity, register, and stylistic colouring. The prompt appears to encourage alternative lexical realizations rather than to redirect the sentence's affective orientation.

The effect is not uniform across systems. Density rose in only three models, and one model showed a small reduction in distinct-type count. Moreover, the prompt combines emotional intensity, affective tone, irony, and stylistic tension. Its effect cannot be attributed to "emotion" alone. Repeated generation with a fixed seed or documented stochastic settings, together with a prompt-only control design, would be required for a stronger causal account.

5.4 Implications and limitations

The study suggests that evaluation of LLM literary translation should distinguish at least three dimensions that are often collapsed: the visibility of affective lexicalization, the range of lexical alternatives, and the regularity of choices across systems. A human-centred workflow can use LLMs to generate candidate formulations, but human review remains necessary for implication, irony, narrative voice, register, and the distribution of meaning across a whole work. Lexical convergence may identify places where model outputs deserve closer comparison; it cannot replace contextual literary analysis.

The limitations are substantial. The source sample is purposive; the published and LLM outputs arise under non-equivalent production conditions; the lexical analysis is restricted to 36 complete-case groups; Jaccard does not represent frequency or graded semantic similarity; and each model–prompt condition was generated once. Future research should add a zero-inclusive 98-group lexical audit, sentence-level confidence intervals or cluster-bootstrap estimates, frequency-sensitive similarity measures, and repeated generations.

6. Conclusion

This study has examined how affective meaning becomes lexically visible in three published Chinese translations and twelve LLM-generated versions of selected sentences from Gone with the Wind. In the 36-group complete-case sensitivity sample, LLM outputs contain more dictionary-recognized affective-word tokens and broader type ranges than the published translations, while also displaying greater within-condition lexical convergence. The emotion-and-style-oriented prompt increases distinct lexical types and lowers word-level cross-model overlap, while producing little aggregate change in density or broad category distribution.

The findings should be understood as evidence about affective explicitness, lexical regularity, and prompt-conditioned variation, not as a ranking of machine and human literary translation. Published translators and LLM systems operate under different conditions, and dictionary counts cannot capture all affective resources. A fuller account should combine zero-inclusive corpus measures with close reading of implication, irony, narrative voice, and register. Within those limits, the study shows why lexical visibility and lexical convergence should be treated as analytically

separate from affective intensity and literary quality.

Conflict of Interest

The authors declare that they have no conflicts of interest to this work.

References

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). John Benjamins. https://doi.org/10.1075/z.64.15bak

Bednarek, M. (2008). *Emotion talk across corpora*. Palgrave Macmillan. https://doi.org/10.1057/9780230285712

Calvo-Ferrer, J. R. (2024). Can you tell the difference? A study of human vs machine-translated subtitles. *Perspectives*, 32(6), 1115–1132. https://doi.org/10.1080/0907676X.2023.2268149

Chen, L., Varga, L., & Mehdizadkhani, M. (2026). Taming AI for The Little Prince: A comparative analysis of NMTs and LLMs in Hungarian translation. *Humanities and Social Sciences Communications*, 13, 46. https://doi.org/10.1057/s41599-025-06343-0

Gao, R., Lin, Y., Zhao, N., et al. (2024). Machine translation of Chinese classical poetry: A comparison among ChatGPT, Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications*, 11, 835. https://doi.org/10.1057/s41599-024-03363-0

Hood, S. (2010). *Appraising research: Evaluation in academic writing*. Palgrave Macmillan. https://doi.org/10.1057/9780230274662

House, J. (2015). *Translation quality assessment: Past and present*. Routledge. https://doi.org/10.4324/9781315752839

Hu, K., & Li, X. (2023). The creativity and limitations of AI neural machine translation: A corpus-based study of DeepL's English-to-Chinese translation of Shakespeare's plays. *Babel*, 69(4), 546–563. https://doi.org/10.1075/babel.00331.hu

Liu, Y. (2026). Unmasking sentiment disguise in cross-cultural literary translation: Analyzing Oscar Wilde's *The Happy Prince* and its heartbroken audience. *Humanities and Social Sciences Communications*, 13, 193. https://doi.org/10.1057/s41599-026-06491-x

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan.

Mitchell, M. (1947). *Gone with the Wind*. The Macmillan Company.

Mitchell, M. (2001). *飘 [Gone with the Wind]* (L. Meihua, Trans.). Yilin Press.

Mitchell, M. (2004). *飘 [Gone with the Wind]* (D. Kexin, L. Yuzhong, & Y. Zhuang, Trans.). People's Literature Publishing House.

Mitchell, M. (2008). *飘 [Gone with the Wind]* (F. Donghua, Trans.). Zhejiang Literature and Art Publishing House.

O'Brien, S. (2024). Human-centred augmented translation: Against antagonistic dualisms. *Perspectives*, 32(3), 391–406. https://doi.org/10.1080/0907676X.2023.2247423

Terando, L. (2023). Recasting the unspeakable: Pain, suffering and loss. *The Translator*, 29(1), 17–27. https://doi.org/10.1080/13556509.2021.2022853

Xu, L., Lin, H., Pan, Y., Ren, H., & Chen, J. (2008). 情感词汇本体的构造 [Construction of an emotion lexicon ontology]. *Journal of the China Society for Scientific and Technical Information*, 27(2), 180–185.

How to Cite: Wu, D. & Chang, Y. (2026). Lexicalizing Affective Meaning in Published and LLM-Generated Literary Translation: A Corpus-Based Study of Chinese Translations of Margaret Mitchell's Gone with the Wind. *Journal of Global Humanities and Social Sciences*, *1(Special Issue)*, 1-9 https://doi.org/10.61360/BoniGHSSAH262020650101