Journal of Humanities and Cultural Studies
ISSN: 3080-986X (Print), ISSN: 3080-9878 (Online) | Vol. 3, No. 1, 2026

A Computer-Aided Study on Imagery in Ancient Chinese Poetry

Yuanyuan Mu, Qin Xu

Chaohu University, Hefei, China

Abstract: Researchers often face major challenges when using computers to analyze classical Chinese poetry, mainly because the semantic labels tend to be broad. At the same time, evaluating how well these poems are translated still depends heavily on personal opinions. Because of this, we do not have enough reliable data to make fair comparisons. Based on existing word-level labels, our study expands three basic groups (entities, events, and attributes) into a more detailed tagging system. This system includes physical objects, actions, time, space, emotions, culture, and aesthetics. We selected the *Three Hundred Tang Poems* and two different English translations as our research materials. By tagging both the original imagery and the translated units, we identified four types of mapping: equivalent, near-synonymous, omitted, and added tags. We used the Delphi method to give each tag a specific weight, and we applied a discount factor for near-synonyms. Using these steps, we built an imagery fidelity index ranging from 0 to 1. This index is calculated as the harmonic mean of weighted recall and precision. In addition, we created a dual-layer knowledge network, within which co-occurrence terms define the edges, while overlapping labels define their weights. Finally, we combined these fidelity scores for each poem. We then used rank correlation and mixed-effects models to see how these scores relate to how overseas readers receive the translations. Overall, our method provides a clear and reproducible way to measure how well poetic imagery is translated.

Keywords: Computer-aided research, classical Chinese poetic imagery, semantic annotation, cross-lingual tag alignment.

1. Introduction

Classical Chinese poetry plays a very important role in sharing Chinese culture with the world. In these poems, imagery is the most basic unit for artistic expression and cultural meaning. Words are the building blocks of poetic lines, and they act as the outer shell of the imagery. Therefore, analyzing the imagery is the next logical step after analyzing the words. Most Chinese poems are short lyrical works. They use very few words but contain a high density of imagery, especially images from nature. The meanings of these images have become richer over generations of poetic writing[1]. When translating these poems, whether overseas readers can recognize and understand the imagery directly affects the final outcome of cultural export.

In the past, most studies relied on close reading, comparing different translations, and discussing translation strategies. The evaluation standards were usually based on personal experience. Researchers often judge whether a translated image is good based on their own knowledge. This makes it very hard to perform large-scale, repeatable, and fair comparisons. Computer-aided methods can solve this problem. For example, counting words and finding relationships in digital texts can help us resort to a much larger dataset. This makes our conclusions more accurate, stable, and testable[2]. Quantitative analysis can also be used to test long-held beliefs in traditional research[3]. To solve these problems, we linked several different analytical steps together. Our method combines detailed annotation, cross-lingual alignment, and quantitative evaluation. We also included network mining and reader reception tests. Our study aims to answer three main questions: how to expand the tag system, how to compare Chinese and English tags, and how to turn these comparisons into a computable score.

2. Literature Review and Research Gaps

Research on classical poetry using computers has developed in a few different directions. One main area is sorting poems by theme. To do this, researchers usually build datasets with thousands of annotated poems. They use pre-trained language models and ensemble learning to classify them[4]. However, these studies still have some clear limitations. The datasets are often too small. Also, using a single label cannot capture all the multiple themes a poem might have. As a result, their real-world performance is still not good enough for practical use.

Another important area is analyzing imagery and emotions. Scholars use pre-trained models for ancient Chinese and sequence tagging to find imagery entities[5]. When they combine sentiment tagging with spatial-temporal maps, the data shows something interesting. The main emotion linked to a single image can change significantly in different dynasties.

Besides classifying and analyzing, recent studies have started to focus on generating text and automated evaluation. Large language models can now write poetry based on specific formats, poets, themes, emotions, or imagery. However, automated scoring still has inherent biases[6]. These models often praise themselves and reinforce each other's outputs. As a result, their scores are systematically different from human expert judgments. This means that humans and machines must work together for reliable evaluation.

Turning to translation studies, previous work has predominantly centered on specific texts, individual translators, or typical images. Scholars frequently describe strategies such as literal translation, paraphrase, substitution, addition, and deletion. While these studies remain closely grounded in textual analysis, their results are inherently

difficult to quantify.

From our review of the literature, we can identify three unresolved gaps. For one thing, nobody has yet formally systematized the semantic structure of imagery, which makes systematic grouping extremely difficult. For another, cross-textual comparisons often lack a computable alignment basis. Because of this, translation loss remains unmeasured. Finally, studies on translation behavior and overseas reader reception remain isolated from each other. We still lack empirical work to verify if our proposed metrics actually match how readers understand the text.

3. Research Design

3.1. Research Subjects and Corpus

We define classical poetic imagery as a composite cognitive unit. It is activated by words within specific contexts and integrates elements such as entities, actions, attributes, allusions, and emotional inclinations. For lexical definitions, we rely on the Dictionary of Ancient Chinese (2nd Edition)[7] as our baseline. The word list was drawn from roughly 3,000 high-frequency characters in the *Three Hundred Tang Poems*, sorted them by frequency, and narrowed to common entries.Examples include *ren* (person), *feng* (wind), *hua* (flower), and *yue* (moon), among others. At the contextual level, the *Three Hundred Tang Poems*[8] serves as the source text. We selected two English translations for comparison: Yuanchong Xu's bilingual version[9] and the co-translated text by Bynner and Kiang Kang-Hu[10].

It is necessary to clarify the nature of this corpus. The two translations differ across three variables simultaneously: historical era, poetic form, and translator identity. The former is a rhymed translation by a contemporary Chinese translator, whereas the latter is a prose translation by a 20th-century overseas collaborative team. Consequently, our comparison involves two distinct translation paradigms. This is a case study comparison rather than a single-variable controlled experiment. Our findings are not intended to determine which translator group is superior, and future corpus expansions should prioritize fixing one or two of these variables. Furthermore, while bilingual editions of Tang poetry in German, Spanish, Japanese, and Russian are known to exist in the Library of Chinese Classics, we could not verify a corresponding Chinese-English volume. As a result, we reference this series only for formatting purposes.

3.2. Multi-tiered Multi-label System

Three top-level groups are retained: entity, event, and attribute, each carrying secondary and tertiary sub-tags. An imagery unit may bind to several tags at once, so single-label classification is not used. Table 1 lists the main branches and examples.

Table 1. Examples of the Multi-Level Label System for Poetic Imagery

Top-level category	Second-level label	Third-level label examples	Verse examples
Entity	Natural objects	celestial bodies, mountains and rivers, vegetation, birds and beasts, wind and mist	moon, wild goose, willow, river
Entity	Human-associated entities	buildings, carriages and horses, implements, clothing, figures	carriages and horses, tall tower, lone boat
Entity	Allusion-based entities	historical figures, allusive designations	wild goose, piebald horse
Event	Physical action	travel, gallop, gaze, return, part	rush forward, look back
Event	Social behavior	farewell, banquet, ascent, expression of feelings	seeing a guest off, climbing high
Event	Metaphorical behavior	homesickness, meditation on the past, withdrawal	gazing homeward, mourning at a site
Attribute	Spatiotemporal attribute	season, hour, spatial scale, direction	spring, dusk, large, south
Attribute	Emotional attribute	desolation, ease, solitude, high spirits	sorrowfully, leisurely
Attribute	Cultural attribute	symbolic meaning, ritual connotations, folk-custom connotations	breaking a willow at parting, climbing high and thinking of kin
Attribute	Aesthetic attribute	vast distance, secluded serenity, majestic vigor	open and expansive, secluded stillness

Tag semantics are handled by language. For Chinese, HowNet's sememe system builds the hierarchy. This structure makes tag hypernyms, hyponyms, and neighbors computable. HowNet covers modern Chinese well but lacks ancient senses. Therefore, we strictly use the Dictionary of Ancient Chinese (2nd ed.) to define senses. If ancient and modern meanings differ, we create a separate sense and register it. Modern senses are never copied over unchecked. For English, WordNet synsets serve as the baseline. The matching synset IDs are recorded, and the Chinese-English bridge starts from bilingual-dictionary equivalents. Bilingual coders compare senses from both sides to verify them. They log verified cases into the manual. This practice rules out intuition-based equivalence judgments.

We split tags into literal and context-derived types. Consider Meng Haoran's poem "Written on a Trip to the Grand Embankment for Wan Qi" (《大堤行寄万七》): "Where pleasure is sought on the grand embankment, / Carriages and horses gallop in haste." The literal tag for "che ma" (carriages and horses) is Entity-Human Artifact-Transportation. Its derived tags include Event-Outing/Excursion and Attribute-Bustling/Noisy. For the word "da" (grand/large), its literal tag falls under "Attribute," specifically "Spatiotemporal Attribute" and "Spatial Scale." However, in this line, it

derives into an "Entity" tag, functioning as a "Human Artifact Accessory" that refers to the embankment itself.

3.3. Bilingual Annotation, Mapping, and Consistency

For the target text, we tag semantic chunks. First, we find the target unit that matches the source imagery. This unit can be a single word, a phrase, or a many-to-one chunk. Then, we label it using a parallel English tag system. We allow soft alignment rather than strict word-for-word matches. This approach handles paraphrase, expansion, and merging.

The four mapping types are scored differently. For equivalent mapping (e.g., yue 月 to moon), senses and sememe hierarchies align on both sides. Therefore, we count the full weight. For near-synonym mapping (e.g., yan 雁 to wild goose), it keeps the avian entity. However, it loses the cultural meanings of the messenger and homesickness. Therefore, we apply a discounted weight here. We derive this score via the Delphi method, defaulting to 0.5. If a source tag is omitted, the target text has no matching tag. It is excluded from the numerator and therefore lowers recall alone. If the target text adds tags absent from the source, we exclude them from the numerator. This process only lowers precision.

Cohen's κ only suits single-label data, so we cannot use it for multi-label sets. Instead, we test rater overlap using three metrics. Krippendorff's α captures overall stability at the set level. The mean Jaccard score measures tag overlap at the unit level. Macro-averaged F1 checks stability across classes. The manual explains how to score partial overlaps. After merging tags from both coders via equivalent mapping, we count intersections as consistent. Remaining differences count as inconsistent. We allow no "partially correct" middle ground. If the score falls below the threshold, we check the manual and revise the rules.

3.4. Construction of the Imagery Fidelity Index

Let *S* be the tag set for a specific imagery unit in the source poem. Let *T* be the tag set for the matching target unit. *w(t)* denotes the weight of tag *t*. The Delphi method assigns and stabilizes these weights. Both tag sets share this unified system to ensure fair comparison. We give cultural and emotional attributes higher weights than spatiotemporal or aesthetic ones. If the former are lost, the imagery loses its cultural meaning.

Equation (1) combines equivalent tags and discounted near-synonym tags into a total effective match. From this, we derive weighted recall and precision in Equations (2) and (3). Finally, we calculate the fidelity index as their harmonic mean in Equation (4).

$$M=\sum_{t\in S\cap T_{eq}} w(t)+\alpha\sum_{t\in S\cap T_{sim}} w(t) \tag{1}$$

$$R_w=\frac{M}{\sum_{t\in S} w(t)} \tag{2}$$

$$P_w=\frac{M}{\sum_{t\in T} w(t)} \tag{3}$$

$$F=\begin{cases}\frac{2P_wR_w}{P_w+R_w}, & P_w+R_w>0\\ 0, & P_w=R_w=0\end{cases} \tag{4}$$

Where Teq is the subset of equivalent tags in the target text, Tsim is the subset of near-synonymous tags, and α is the discount coefficient for near-synonymy.

All three measures fall between zero and one. We do not use the subtractive form, which takes a penalty from the match score, for two reasons. First, its lower bound moves down as the discount factor drops. Heavy addition in a translation can then push the value below zero. Such a score is hard to explain, and it neither averages well nor enters a regression model. Second, the harmonic mean penalizes omission and addition at once, so no further penalty factor is needed. Where the tendency to add needs watching, a drift index may be reported alongside, namely the weighted share of the labels a translation adds. It serves as an aid only and does not enter the main measure.

Fidelity is scored per image, while reader data are mostly kept per poem. We must therefore settle a rule for aggregation first. Let *ni* be the number of labels carried by the *i*-th image, and *N* the total number of labels in the poem. Equation (5) weights the image scores by their label load to give one value for the poem. A simple mean is avoided, since it would let a minor word with a single object label count as much as a core image.

$$F_{poem}=\sum\left(\frac{n_i}{N}\right)\times F_i \tag{5}$$

3.5. Two-Layer Imagery Knowledge Network

The network has two layers, each with a distinct task. The first layer is a co-occurrence backbone for comparing translations. Two images are linked only if they occur in the same line or poem. Edge weight is the weighted overlap of their shared tags. Real co-occurrence forms the backbone; tags only set edge strength. Only then can structural change reflect how translation reshapes image structure, not merely tag detail. To help others reproduce the results, we set the pruning threshold at the 80th percentile of all edge weights. All edges below it are pruned, and both the threshold and script are released publicly.

The second layer is a semantic similarity network. Its edges depend only on tag overlap. Images in the same natural category will be densely linked. This layer shows proximity between cultural concepts and checks if a cultural tag has lost its linked partners. However, it does not test whether the source and translated texts differ in their links.

The network can answer three questions. Which images act as hubs and repeatedly organize the whole poem? Which images rely on cultural tags rather than literal links? Do the source and translated networks break or reorganize on this backbone? Both types of edges reflect statistical links, not markers of literary origin or allusion. Such claims still need documentary and textual evidence.

4. Results

Across 421 translations, mean poem-level fidelity was 0.556 for Xu Yuanchong (median 0.571, SD 0.190). Bynner's mean was 0.583 (median 0.587, SD 0.161). The tag-count-weighted means were 0.544 and 0.577. Both are close to the simple means. A few long poems therefore do not dominate the results.

Table 2. Distribution of Poem-Level Fidelity (Fpoem) in the Two Translations

Metric	Yuanchong Xu (n=125)	Bynner (n=296)
Mean Fpoem	0.556	0.583
Median	0.571	0.587
Standard Deviation	0.19	0.161
Interquartile Range (Q1–Q3)	0.451–0.667	0.493–0.679
Tag-Count-Weighted Mean	0.544	0.577

The scores of both translations concentrate between 0.4-0.8.Yuanchong Xu reaches 72.8% in this range, 41.6% in 0.4–0.6 and 31.2% in 0.6–0.8. In comparison, Bynner's translation reached 82.8%, with 43.9% and 38.9% in the two sub-ranges. Bynner has fewer low scores, 1.4% falling below 0.2 against 3.2% for Xu. Above 0.8, the two are close, at 8.0% and 7.8%. Their overall distributions are alike, and the main gap lies in the 0.2–0.4 interval, where Xu records 16.0% but Bynner only 8.1%.

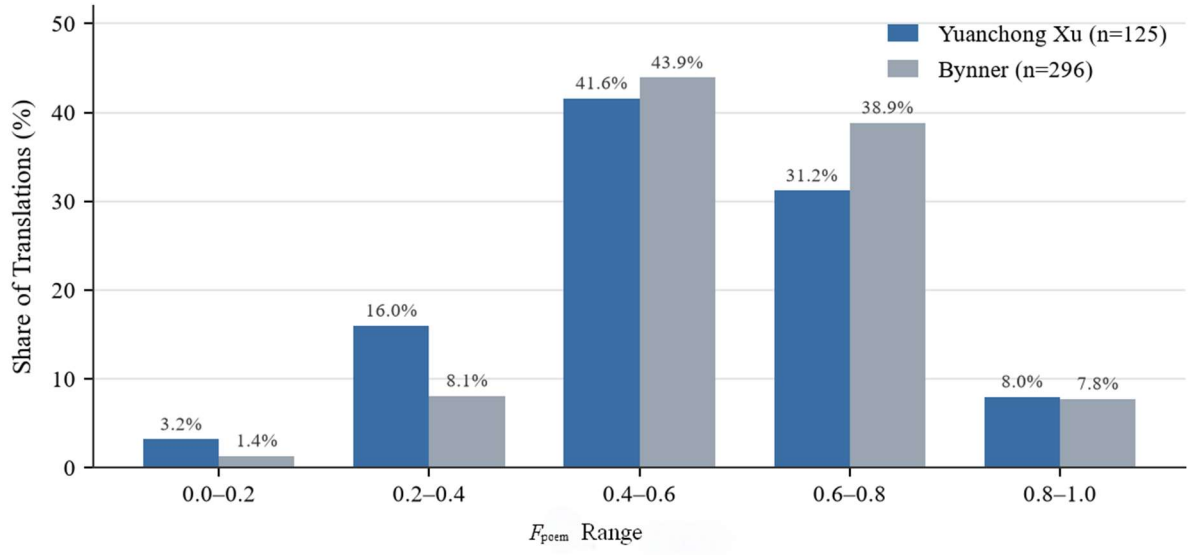

Fig. 1 Distribution of poem-level fidelity F_{poem} across intervals for the two translations

5. Conclusion

This study aims to make the evaluation of poetic imagery translation computable, reproducible, and empirically grounded. By expanding word-level labels into a multi-tiered, multi-label system spanning entities, events, and attributes, and by aligning Chinese and English tags through HowNet, WordNet, and bilingual sense verification, we converted an interpretive judgment into a structured and inspectable representation. The empirical results show that across 421 translations, mean poem-level fidelity clustered between 0.4 and 0.8 for both translations, with Bynner's version showing a slightly higher mean (0.583 versus 0.556) and a notably smaller share of low-scoring poems. The gap is concentrated in the 0.2–0.4 interval rather than at the top of the distribution, suggesting that the two translation paradigms differ less in their best renderings than in how often they fall considerably short. These figures should be read as a descriptive comparison of two translation paradigms rather than as evidence of the superiority of either.

The contribution of this work lies in the end-to-end workflow it provides: annotation and cross-lingual alignment feed into a weighted fidelity index, which in turn supports graph-based structural analysis. The pruning threshold and scripts are released so that others can replicate and extend the procedure. Because the design is modular, new annotation schemes, additional languages, and refined weighting principles can be accommodated as the digital-humanities study of classical poetry continues to develop.

Acknowledgments

This paper is funded by Anhui Province Outstanding Young Scholar Project (2022AH020092), “Knowledge Network Construction and Mining for the Cross-Cultural Translation and Dissemination of Classical Chinese Poetry Imagery in the Digital Humanities Context”.

References

[1] Yuan, X. (2009). A study of Chinese poetic art (3rd ed.). Peking University Press. (in Chinese)

[2] Liu, S., & Yin, X. (2022, March 23). Mapping classical literature through digital means. Guangming Daily. (in Chinese)

[3] Liu, Z., & Wang, Z. (2012). A quantitative analysis of Tang and Song ci poetry. Peking University Press. (in Chinese)

[4] Hou, J., & Zhang, S. (2024). Exploring thematic diversity in classical Chinese poetry: A novel dataset and a BERT-enhanced ensemble learning approach. Journal on Computing and Cultural Heritage, 17(4), Article 60, 1–19. https://doi.org/10.1145/3685679

[5] Jiang, Z., & Li, L. (2025). A data-driven visualization study of imagery and sentiment of classical Chinese frontier poetry. In Proceedings of the 2025 2nd International Conference on

Intelligent Computing and Data Analysis (ICDA 2025). ACM. https://doi.org/10.1145/3772726.3772741

[6] Ma, B., Yao, Y., & Haensch, A. C. (2026). Capabilities and evaluation biases of large language models in classical Chinese poetry generation: A case study on Tang poetry. In Findings of the Association for Computational Linguistics: ACL 2026 (pp. 16933–16957). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.findings-acl.836

[7] Dictionary Research Centre, The Commercial Press. (2014). Dictionary of ancient Chinese (2nd ed.). The Commercial Press. (in Chinese)

[8] Sun, Z. (2005). Three hundred Tang poems. People's Literature Publishing House. (in Chinese)

[9] Xu, Y. (2000). Three hundred Tang poems: Chinese-English bilingual edition. Higher Education Press. (in Chinese)

[10] Bynner, W., & Kiang, K. H. (1929). The jade mountain: A Chinese anthology, being three hundred poems of the T'ang dynasty, 618–906. Alfred A. Knopf.