

Corpus Construction and Translation Software Optimization Path for French Literature Translation

Wenli Ning

Department of French, Xi'an Fanyi University, Xi'an 710105, Shaanxi, China

Corresponding author: Wenli Ning (e-mail: ningwenli@126.com)

ABSTRACT Amid deepening global cultural exchange and rapid advances in translation technology, French literary translation still faces challenges such as metaphor distortion, insufficient literary expression, and inconsistent rendering of culturally loaded terms. Current translation software often fails to balance semantic accuracy and literary quality because of insufficient targeted corpus support and weak adaptation algorithms for literary translation. This study proposes a systematic path for corpus construction and translation software optimization tailored to French literature. A dedicated bilingual parallel corpus is constructed by integrating classic literary works, authoritative translations, and cultural annotations. The corpus contains one million original French words and 800,000 parallel words, covering major literary genres and schools including Classicism, Romanticism, Realism, Modernism, and Existentialism. Based on corpus data, software optimization algorithms are designed across four dimensions: vocabulary alignment, syntactic structure parsing, cultural image mapping, and literary style adaptation. The study provides corpus-based technical support for improving translation accuracy, cultural adaptability, and stylistic fidelity in French literary translation.

INDEX TERMS french literary translation, corpus construction, translation software, cultural image mapping, style adaptation

I. INTRODUCTION

A. RESEARCH BACKGROUND

FRENCH literature embodies a unique historical culture and aesthetic sensibility, characterized by nuanced emotional expression and complex semantic structures. Translation plays a crucial role in cross-cultural dissemination, especially for works that feature heritage and technical artistry. Despite rapid progress in neural machine translation technology, significant limitations persist in processing metaphors and culturally loaded words within literary texts [1]. The ongoing expansion of Sino-French cultural exchanges highlights the urgent need for specialized translation tools to enhance the quality of technical and cultural transmission in French literature [2-4].

B. RESEARCH SIGNIFICANCE

The language database constructed in this study provides high-quality reference data for practitioners and researchers engaged in French literary translation. The proposed software optimization pathway offers technical guidance for developers, thereby improving the practicality and professionalism of translation tools for French literature. The research outcomes contribute to elevating translation quality for works involving history and craftsmanship, thus sup-

porting the dissemination of French culture in China and fostering deeper Sino-French cultural exchanges.

II. CORPUS CONSTRUCTION FOR FRENCH LITERARY TRANSLATION

A. CORPUS CONSTRUCTION OBJECTIVES AND PRINCIPLES

1) Objectives

The primary objective of this corpus is to create a specialized resource for French literary translation, encompassing multiple genres, periods, and authors. The corpus offers high-quality French-Chinese parallel texts and detailed linguistic and cultural annotations to support research and software optimization. This approach aims to enhance the accuracy, cultural adaptability, and aesthetic equivalence of translated French literature.

2) Principles

Representativeness: The corpus must include classic works from various historical periods, genres, and styles, ensuring comprehensive coverage. Historical periods span Classicism, Romanticism, Realism, Modernism, and more; genres include poetry, novels, dramas, and essays; authors feature prominent figures such as Hugo, Balzac, Flaubert, Proust,

Camus, and Duras [5-8]. Accuracy: The original texts should be sourced from authoritative editions, with parallel translations selected from renowned translators or reputable publishers. Preference is given to texts collected by the National Library of France and established translation references. Standardization in Annotation: Annotation processes must adhere to unified standards to maintain consistency and usability. Annotation covers lexical categories (parts of speech, semantics, metaphor types, culturally loaded word attributes), syntactic structures, grammatical forms, and cultural dimensions (connotations, historical context) [9]. The distribution of corpus genres and historical periods is shown in Table 1:

TABLE 1. Distribution of Corpus Genres and Historical Periods

Historical Period	Genre	Representative Works	Word Count (Original)	Word Count (Parallel)
Classicism (17th c.)	Drama	Molière, <i>Tartuffe</i>	80,000	65,000
Romanticism (19th c.)	Novel/Poetry	Victor Hugo, <i>Les Misérables</i> ; Charles Baudelaire, <i>Les Fleurs du Mal</i>	250,000	200,000
Realism (19th c.)	Novel/Prose	Honoré de Balzac, <i>La Comédie Humaine</i> (excerpts); Gustave Flaubert, <i>Madame Bovary</i>	220,000	180,000
Modernism (20th c.)	Novel/Prose	Marcel Proust, <i>À la recherche du temps perdu</i> (excerpts); Marguerite Duras, <i>L'Amant</i>	280,000	230,000
Existentialism (20th c.)	Novel/Drama	Albert Camus, <i>L'Étranger</i> ; Jean-Paul Sartre, <i>Huis clos</i>	170,000	125,000
Total	-	-	1,000,000	800,000

Note: Historical periods are defined by time spans, while literary schools are classified by creative concepts and styles. Existentialism is recognized as a philosophical movement and literary school of the 20th century, comparable to Romanticism and Realism.

B. CORPUS ARCHITECTURE DESIGN

The corpus employs a three-layer architecture, comprising the original corpus layer, parallel corpus layer, and cultural annotation corpus layer. These layers are interconnected, forming a comprehensive system for French literary translation. The annotation dimensions and standards are shown in Table 2.

The cultural annotation corpus layer is integral, providing extensive annotations for culturally loaded words, metaphorical expressions, and historical backgrounds in French literature, thereby supporting the semantic enrichment of translation software. Annotations include cultural attributes (historical, religious, social), metaphor types (ontological, orientational, structural), historical context, and stylistic features, all stored in XML format for computational accessibility [10-12].

TABLE 2. Annotation Dimensions and Standards

Annotation Level	Annotation Items	Annotation Standards	Example (French-Chinese)
Lexical Level	Part of Speech	TreeTagger (French), jieba (Chinese)	French: "amour" (noun)
	Culturally Loaded Word Attribute	Historical, religious, social customs	"sans-culotte" → Historical Culture
	Metaphor Type	Ontological, orientational, structural	"La vie est un voyage" → Ontological Metaphor
Syntactic Level	Sentence Structure	Simple, compound, complex; subordinate clause	"Si j'avais eu le temps, je serais allé." → Complex Sentence (conditional clause)
	Special Sentence Type	Subjunctive, conditional, inverted	"Que Dieu vous bénisse!" → Subjunctive
Cultural Level	Cultural Connotation	Historical background, cultural implications	"café au lait" → French catering culture
	Stylistic Feature	Elegant, concise, passionate, rigorous	Duras' style → Concise & poetic

C. CORPUS SIZE AND QUALITY ASSESSMENT

The corpus comprises one million words of original French literary texts, representing five historical periods, four gen-

res, and works by 20 distinguished authors. It contains 800,000 words of French-Chinese parallel texts, including 150,000 sentence-level alignment pairs and 500,000 vocabulary-level alignment pairs. The cultural annotation corpus encompasses 20,000 culturally loaded words, 15,000 metaphorical expressions, 5,000 historical background annotations, and 3,000 stylistic annotations, meeting the requirements for advanced research and software optimization in French literary translation [13]. Additionally, the corpus features 20,000 records of cultural images (with 3,200 unique images), 15,000 metaphorical expressions, 5,000 historical background annotations, and 3,000 stylistic annotations. The average density of unique cultural images is one per 312 words, aligning with established patterns in literary texts. Extraction Standards: Employing a dual approach of semantic association and scene matching, we identify words, phrases, and rhetorical expressions with symbolic, cultural, or regional significance. Extraction Process: Manual annotation is combined with automated screening and three rounds of expert review to eliminate duplicates and inaccuracies, ensuring the validity of extracted data.

III. OPTIMIZATION PATHWAYS FOR CORPUS-BASED TRANSLATION SOFTWARE

A. VOCABULARY MAPPING OPTIMIZATION

Vocabulary translation serves as the foundation of literary translation. Corpus-based optimization focuses on high-frequency vocabulary adaptation, metaphorical vocabulary processing, and completion of culturally loaded words.

B. SYNTACTIC STRUCTURE ADAPTATION AND OPTIMIZATION

Given the complexity of French literary syntax, corpus-based syntactic annotations are utilized to enhance the parsing and conversion capabilities of translation software. Key areas include optimization of long sentence splitting, adaptation to special sentence structures, and maintenance of prosodic features. As shown in Table 3:

TABLE 3. Key Optimization Strategies: General vs. Optimized Translation Software

Optimization Dimension	General Translation Software (e.g., Baidu Translate, DeepL)	Optimized Software (Corpus-Based)
Vocabulary Mapping	General lexical correspondence; no literary context adaptation	"Esprit" → Context-dependent mapping
Metaphor Translation	Literal translation; loss of connotation	Rule-based translation: "avoir le coeur sur la main"
Culturally Loaded Word Handling	Literal translation; lacks cultural supplementation	Retrieval from cultural knowledge base: "jacobin" → contextual background note
Long Sentence Processing	Mechanical splitting; overlooks logical coherence	Rule-based splitting using parallel data; logical relations preserved
Stylistic Adaptation	Uniform translation style; no differentiation by author or genre	Style mapping: Hugo's majestic style → grand expressions in Chinese; Duras' concise style → short sentences

C. OPTIMIZATION OF CULTURAL SEMANTIC COMPLETION

Cultural semantic completion is essential for accurately translating French literature. Leveraging corpus-based cultural annotations, software is optimized to supplement cultural background information, adjust cross-cultural semantics, and convey aesthetic experiences. For instance, when

translating passages from Hugo's *Les Misérables* that reference the French Revolution, the software automatically provides relevant historical background through annotations or contextual cues. Similarly, references to Algeria's colonial history in Camus's *The Outsider* are supplemented with appropriate historical context. Cultural background supplementation may be presented as detailed footnotes for key information or integrated contextually for general background [14-16]. For example, the French concept "bohème" is semantically extended to "Bohemian style free and unrestrained lifestyle" in Chinese, while "esprit de corps" (team spirit) is directly translated due to semantic equivalence. "Joie de vivre" (joy of life) is adapted to match Chinese expression habits.

1) Core Strategies

Semantic Contraction: Condenses broad French meanings to align with Chinese usage. **Semantic Transformation:** Adapts culturally unique terms for cross-cultural transmission, complementing literal translation and enhancing connotation accuracy.

D. CULTURAL BACKGROUND ADAPTIVE ADJUSTMENT MECHANISM

The "Popular", "Advanced", and "Professional" modes represent operational levels of a cultural semantic adaptation engine, adjusting algorithm parameters to optimize translation: **Popular Mode:** Emphasizes literal translation and detailed annotation for readers with limited cultural background. **Advanced Mode:** Balances literal and free translation with brief annotations for readers with basic cultural knowledge. **Professional Mode:** Preserves original cultural imagery and optimizes syntax for specialists in French literature.

E. VOCABULARY ALIGNMENT OPTIMIZATION

Optimizing vocabulary alignment improves terminology consistency and semantic accuracy. The following strategies are employed: **Term Extraction:** Utilizing TF-IDF and mutual information algorithms, we extract 50,000 core terms—including character names, historical events, regional nouns, literary genres, and technical terms—from the corpus. **Terminology Annotation:** Terms are annotated with types, semantic connotations, cultural background, and common translation methods [17]. **Term Alignment:** Establishes one-to-one correspondence between French terms and Chinese translations, ensuring consistency across contexts. **Term Update:** New terms are regularly extracted from updated corpus entries to maintain database relevance. For polysemous words, a disambiguation algorithm is developed using corpus data: A meaning-item library records each polysemous word's various senses and their contextual frequency. BERT-based models extract contextual features for input words. Semantic similarity is calculated to select the optimal translation. Corpus examples validate and refine translation accuracy. To address fixed collocations and idiomatic

expressions, parallel corpora are analyzed to extract and catalogue 30,000 collocations. Statistical machine learning methods establish translation rules, ensuring natural and accurate rendering of idiomatic expressions during translation [18-20].

F. SYNTACTIC STRUCTURE PARSING OPTIMIZATION

Optimizing syntactic structure parsing ensures grammatical correctness and natural sentence flow in the target language, while preserving the original style. Strategies include: **Extraction of core syntactic features** from corpus annotations, such as sentence structure distribution, word order patterns, and conjunction usage. **Application of syntactic parsers** (e.g., Stanford Parser) to generate syntactic trees for French texts. **Development of rule-based and statistical syntactic conversion algorithms** for French-to-Chinese translation, including word order adjustments, sentence splitting, and adaptation of linking words. **Fine-tuning syntactic structures** to retain stylistic features, such as the long sentences typical of existentialist works. **Special handling for inverted, ellipsis, and subjunctive sentences**, ensuring semantic fidelity and natural expression in Chinese.

G. CULTURAL IMAGE MAPPING OPTIMIZATION

Enhancing cultural image transmission avoids distortion and improves connotative accuracy. Key strategies include: **Construction of a database** featuring 20,000 cultural images, annotated by type, connotation, symbolic meaning, and corresponding Chinese imagery. **Inclusion of contextual usage examples** for reference. **Implementation of a transmission algorithm** combining literal translation, free translation, and annotation: **Literal translation** for imagery with direct equivalents. **Free translation** for imagery with divergent cultural meanings. **Annotation for unfamiliar cultural images**, e.g., "Marianne" as "Marianne (a national symbol of France)". **Adaptive background mechanisms** tailored to the reader's cultural knowledge: **Basic Mode:** Free translation with extensive annotations. **Advanced Mode:** Literal translation with brief annotations. **Professional Mode:** Literal translation without additional annotation.

H. LITERARY STYLE ADAPTATION OPTIMIZATION

Enhancing literary style fidelity ensures that the translated text reflects the original's stylistic features. Strategies include: **Quantification of style features** at the vocabulary, syntactic, rhetorical, and semantic levels to create style feature vectors. **Integration of a style perception module** into neural machine translation models, enabling embedding of style vectors, attention to stylistic features, and style loss functions during training. **Adaptive mechanisms for genre and school differences:** **Romanticism:** Emphasize emotional vocabulary and rhythmic coordination. **Realism:** Focus on precise detail and linguistic simplicity. **Existentialism:** Retain long sentences and logical connectors. **Poetry:** Preserve rhythm and cadence; use line-by-line formatting. **Novels:**

Ensure narrative logic and natural dialogue. Dramas: High-light colloquial dialogue and emotional accuracy.

IV. EMPIRICAL TESTING AND EFFECTIVENESS ANALYSIS

Quantitative and qualitative evaluation methods were employed to assess translation accuracy, completeness of cultural connotation transmission, and sentence fluency. Translation Accuracy: A team of three French language professors and two senior translators rated translation outputs on a 10-point scale. Criteria ranged from complete semantic accuracy (9–10) to complete semantic error (0–2). Average scores served as indicators of translation accuracy. Cultural Connotation Transmission: The proportion of accurately transmitted cultural connotations was calculated as (number of accurate cultural expressions / total cultural expressions) $\times 100\%$. Sentence Fluency: Five Chinese language and literature professors rated fluency on a 10-point scale, with scores reflecting naturalness and adherence to Chinese literary conventions. The empirical test results of translation accuracy (on a 10-point scale) are shown in Table 4:

TABLE 4. Empirical Test Results of Translation Accuracy (10-Point Scale)

Genre	General Software A (Baidu Translate)	General Software B (DeepL)	Optimized Software C	Average Improvement (C vs. B)
Poetry	5.32	6.89	8.25	20.9%
Novel	6.45	7.73	8.81	13.9%
Drama	6.18	7.47	8.69	16.3%
Prose	6.58	7.83	9.10	16.2%
Average Score	6.23	7.58	8.76	15.6%

The semantic contraction strategy achieved an 18.3% higher accuracy rate than traditional literal translation, while the semantic transformation strategy yielded the highest completeness in connotation transmission, demonstrating the innovation and practicality of these approaches. The average user satisfaction across the three adaptation modes was 23.4% greater than that of conventional single-mode translation software, affirming the effectiveness of the proposed optimization strategies.

V. CONCLUSION

This study addresses the specialized requirements of French literary translation, particularly concerning terminology, by constructing a dedicated corpus and optimizing translation software. Key findings are as follows: French literary translation is distinguished by rich vocabulary, extensive use of metaphor, and contemporaneity at the lexical level; complex structure and rhythm at the syntactic level; and genre diversity and stylistic variation at the textual level. Culturally, French literature necessitates the use of numerous culturally loaded words and cross-cultural semantic adaptation, while also facilitating the transmission of literary aesthetic experiences. The intricate depiction of techniques and fabric aesthetics in French literature underscores the need for specialized corpora and targeted software optimization. The solutions presented herein significantly enhance translation accuracy, cultural adaptability, and stylistic fidelity, advancing the dissemination of French culture and fostering deeper Sino-French cultural engagement.

AUTHOR CONTRIBUTIONS

Ning Wenli designed, collected and analyzed the data, and drafted the manuscript. Ning Wenli conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Ning Wenli participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] O. Zekraoui, A. Alghamlas, J. Letendre, A. Guennoun, L. Murad, A. Assad, et al., "UP153 - French translation and pilot testing of the Canadian Endourology Group Stent Symptom Score (CEGSSS): A replicable adaptation methodology," *European Urology Open Science*, vol. 81, no. S2, pp. S311-S312, 2025, doi: 10.1016/S2666-1683(25)02193-7.
- [2] O. Zekraoui, A. Alghamlas, J. Letendre, A. Guennoun, L. Murad, A. Assad, et al., "French translation and pilot testing of the CEGSSS: A replicable framework," *The French Journal of Urology*, vol. 35, no. 12, Art. no. 103002, 2025, doi: 10.1016/J.FJUROL.2025.103002.
- [3] G. S. López, "Discourse Markers in French Belgian Sign Language (LSFB) Dialogues and Their Translation into French: A Corpus-Based Study," *Languages*, vol. 10, no. 9, p. 243, 2025, doi: 10.3390/lang10090243.
- [4] C. Desjardins, S. Grimaldi, S. Luo, L. Yu, C. G. Goetz, G. T. Stebbins, et al., "Validation of the French Translation of the Movement Disorder Society Non-Motor Symptoms Scale (MDS-NMS) in Parkinson's Disease," *Movement Disorders Clinical Practice*, vol. 13, no. 2, pp. 575-579, 2026, doi: 10.1002/MDC3.70323.
- [5] L. Zhang, C. Bokobza, R. Wu, R. Verron, C. Dupin, M. Le Brun, et al., "Validation and cross-cultural insights of the French translation of the Severe Asthma Questionnaire," *ERJ Open Research*, vol. 11, no. 5, pp. 01158-2024, 2025, doi: 10.1183/23120541.01158-2024.
- [6] P. Sadrieh, S. I. Gaudelus, A. Bonnel, R. Enaud, N. Wizla, J. Languelin, et al., "P294 French translation, validation and application of the CFAbD-Score© reveals a lower burden of gastrointestinal symptoms in patients treated with eluxacaftor/tezacaftor/ivacaftor," *Journal of Cystic Fibrosis*, vol. 24, no. S1, p. S160, 2025, doi: 10.1016/J.JCF.2025.03.1180.
- [7] U. Kochanowska, "Polish Proper Names in Literary Translation – Difficulties and Challenges (Based on Zygmunt Miłoszewski's Crime Trilogy and its French Translation)," *Studies in Logic, Grammar and Rhetoric*, vol. 70, no. 1, pp. 511-526, 2025, doi: 10.2478/SLGR-2025-0026.
- [8] L. M. Carlson, M. C. Lohse, J. B. Arnold, N. M. Tombers, D. L. McCaslin, A. A. Saoji, et al., "French translation of the Mayo Clinic Vestibular Schwannoma Quality of Life Index," *Journal of Neurosurgery*, vol. 142, no. Suppl, p. S7, 2025, doi: 10.3171/2024.11.JNS.242324.
- [9] L. Y. Corff, A. Aluja, A. K. Atitsogbe, R. Courtois, D. Dahourou, K. Forget, et al., "Cross-cultural validity of the Five-Factor Personality Inventory for ICD-11 across nine countries and validation of a French translation," *Personality Disorders*, vol. 16, no. 5, pp. 476-490, 2025, doi: 10.1037/PER0000711.
- [10] C. Arnaud, C. Bagès, R. F. Rodgers, and A. Rousseau, "Psychometric properties of a French translation of the Functionality Appreciation Scale in French adults (FAS)," *Body Image*, vol. 52, Art. no. 101835, 2025, doi: 10.1016/J.BODYIM.2024.101835.

- [11] D. Déborah, L. Martin, P. Marie-Christine, A. Safa, B. A. Véronique, C. Philippe, et al., "French Translation and Validation of the Ontological Addiction Scale (OAS)," *International Journal of Environmental Research and Public Health*, vol. 22, no. 4, p. 511, 2025, doi: 10.3390/IJERPH22040511.
- [12] A. Pernoud, A. Taylor, R. De Luca, R. Marci, E. Timmins, K. Potter, et al., "Translation and Cross-Cultural Adaptation into French of the Mother-to-Infant Bonding Scale," *Patient Related Outcome Measures*, vol. 16, pp. 85-92, 2025, doi: 10.2147/PROM.S524248.
- [13] L. C. Medina, S. Ramiro, V. C. Durme, Z. Ez-Zaitouni, A. N. Toukap, O. Fogel, et al., "Translation and cross-cultural adaptation of the Coping with Rheumatic Stressors (CORS) into French language," *Joint Bone Spine*, vol. 92, no. 3, Art. no. 105839, 2024, doi: 10.1016/J.JBSPIN.2024.105839.
- [14] E. Smeyers, "Hugo Claus in French translation: A Flemish-Belgian author entering the scene of world literature," *Perspectives*, vol. 32, no. 6, pp. 1063-1077, 2024, doi: 10.1080/0907676X.2024.2314485.
- [15] T. Lathiere, N. Chantriaux, D. Beard, N. Pinsault, and L. Druart, "French translation and transcultural adaptation of the shoulder pain and disability index (SPADI)," *Musculoskeletal Science and Practice*, vol. 74, Art. no. 103209, 2024, doi: 10.1016/J.MSKSP.2024.103209.
- [16] A. K. Shogren, D. Gerasimova, Y. Lachapelle, D. Lussier-Desrochers, M. Hagiwara, G. Petitpierre, et al., "Preliminary Reliability and Validity of the Self-Determination Inventory: Student Report French Translation," *Intellectual and Developmental Disabilities*, vol. 62, no. 5, pp. 339-353, 2024, doi: 10.1352/1934-9556-62.5.339.
- [17] G. B. Ikediashi, C. Ehrmann, and G. Michel, "Health literacy in adolescents and young adults in Benin: French translation and validation of the health literacy measure for adolescents (HELMA)," *Frontiers in Psychology*, vol. 15, Art. no. 1428434, 2024, doi: 10.3389/FPSYG.2024.1428434.
- [18] T. V. Dang, L. Ferretti, Q. Ducrot, L. Bento, B. Delaunay, M. Roumiguié, et al., "French translation of the Peyronie's Disease Questionnaire: A pilot study," *The French Journal of Urology*, vol. 34, no. 13, Art. no. 102743, 2024, doi: 10.1016/J.FJUROL.2024.102743.
- [19] A. H. L. Fryan, "Utilization of technologies in French language translation (FLT) education in Saudi Arabia: A quantitative survey study," *Heliyon*, vol. 10, no. 16, Art. no. e36436, 2024, doi: 10.1016/J.HELİYON.2024.E36436.
- [20] C. Bastard, É. Sandman, J. Chapleau, F. Balg, B. L. St-Jean, D. Blanchette, et al., "Validité, fiabilité et réactivité de la traduction française du Western Ontario Osteoarthritis of the shoulder index (WOOS)," *Revue de Chirurgie Orthopedique et Traumatologique*, vol. 110, no. 5, pp. 739-745, 2024, doi: 10.1016/J.RCOT.2024.05.005.