An Empirical Corpus Study on Word-Formation and Translation Strategies in Chinese Medicinal Nomenclature

Xiaoqing Hu[a], Dechun Lu*

Chongqing University of Chinese Medicine, Chongqing, 402760, China

[a]20228299@qq.com, *cqldczxf@163.com
*Corresponding author

Abstract. This multidisciplinary study integrates linguistics, TCM, and computer science to explore the relationship between word-formation patterns and translation strategies of Chinese medicinal names. A large-scale, high-quality Chinese-English parallel corpus was constructed. Chinese medicinal names were classified by morphological types, and their English translations by strategy. Statistical analyses including chi-square tests and correspondence analysis were conducted. Results show that word-formation significantly influences translation strategy selection ($p < 0.01$). Part-based names tend to use Latin names (with plant parts) or literal translation; efficacy-based names prefer Pinyin + generic term or Latin names; compound names mainly adopt Pinyin + annotation. An evidence-based translation decision-making model is proposed. Word-formation is critical in predicting and determining translation strategy selection.

Keywords: Chinese medicinal names; corpus; translation strategy

1 Introduction

Traditional Chinese Medicine culture is one of the most valuable cultural heritage, and life-saving value embodied by Chinese civilization . Since the Chinese Materia Medica (CMM) has been one of its major carriers, translation accuracy and consistency are now a bottleneck in the international communication, the exchange of information and practice [1, 2]. Quality translation of CMM name has direct impact on clinical safety (risks of misidentification of drugs, dosage errors, etc.) and scientific collaboration (imposing international replication of pharmacological and clinical studies). Poor translation can also lead to cultural misinterpretation (cultural misinterpretation), restricting global transmission of TCM's unique cultural image and cognitive models [3].

Names of CMM contain linguistic, cultural and pharmaceutical information, posing translation challenges [4]. Scholars have conducted multi-dimensional research and systematized the compositional logic of CMM names from etymological, linguis-

T. Chamaratana et al. (eds.), *Proceedings of the 2026 12th International Conference on Digital Humanities and Frontiers in Social Sciences (DHFSS 2026)*, Advances in Social Science, Education and Humanities Research 1031,
https://doi.org/10.2991/978-2-38476-611-6_119

tic and cultural semantic perspectives [5], revealing rich information such as medicinal parts, morphology, efficacy, origin and mythology, and proving that word formation constitutes a complex system of natural observation, experience and cultural metaphors [6]. International standards and principles have been proposed, including priority to Latin scientific names, pinyin with annotations and functional equivalence [7 – 9]. However, most studies remain speculative or qualitative. A key gap exists: no large-scale high-quality corpus has been used to quantify the correlation between word-formation features and translation strategies. As the “linguistic gene” of CMM names, word formation remains to be explored for its guidance in translation strategy selection.

Traditional TCM translation studies mostly depend on manual analysis, while recent AI advances offer new tools to automatically identify word-formation rules and translation strategies of Chinese medicinal names. Although large-scale corpus-based analysis enhances research systematicity and efficiency, relevant applications remain scarce in the TCM field because general models cannot capture its unique word-formation and translational features. This study adopts corpus-driven methods and deep learning to build an automatically annotated Chinese – English parallel corpus; it applies a domain-specific BERT-CRF model for key information extraction and a multi-model fusion framework to construct a word-formation-based translation strategy prediction model, which transforms TCM translation from empirical experience to scientific rule-based selection, facilitates TCM terminology standardization, and boosts the global dissemination of TCM knowledge.

2 Methods

Six major sources of data were incorporated, including Chinese Pharmacopoeia, PubMed, CNKI collection of TCM academic papers and Chinese Medical Database of the Chinese Medical Association. For these sources we developed hierarchical targeted acquisition system. Structured web data were extracted using Beautiful Soup and Scrapy, PDF documents were parsed in depth using PyPDF2 library and Tesseract OCR engine. Dynamic proxy pool coupled with multibrowser identification rotations avoided anticrawling constraints and allowed for the stable harvesting of average 12000 data entries per day.

Data quality was ensured through automated verification and manual review by two independent TCM and translation annotators. Automated scripts were used to check data formats and translation consistency. Missing translations were supplemented via semantic similarity imputation using a word-embedding model. A fine-tuned generative model was then applied to generate domain-compliant translation pairs.

2.1 Data Analysis Methods

SPSS 23.0 and R (version 4.0.3) were used. Statistical statistics for word form types and translation strategies were expressed as frequency (n) and percentage (%).

Chisquare tests and multinomial logistic regression were used to identify the associations between the two variables. Association rule mining was performed using R package Arles to find strong association rules. Pvalue 0.05 was considered statistically significant for all tests.

3 Results

3.1 Data Processing

A three-level quality control system was adopted. First, duplicate entries were removed using Pandas deduplication and Levenshtein similarity matching (threshold = 0.85). Second, spelling errors were corrected with a customized TCM glossary and tools, reaching 91.7% accuracy. Finally, data formats were standardized, and missing translations (15.3%) were supplemented by cross-database tracing. This produced a high-quality parallel corpus containing 1,850 core herbal entries and more than 3,200 translation instances.

We combine rule-based annotation and deep learning techniques for word-formation annotation. A domain-specific glossary is created by TCM names first. We also introduce bidirectional learning based on complementary strengths of two models: batch annotation is handled by MOE and fine-grained corrections of unclear cases can be performed by Transformer model. By optimization, word-formation recognition accuracy is 99.3%.

3.2 Model Architecture

We constructed a multi-feature fusion machine learning model for translation strategy matching by extracting a 12-dimensional feature vector covering word-formation types, entity attributes and corpus frequency. An ensemble learning framework integrating SVM, Random Forests and Logistic Regression achieved an overall accuracy of 89.6%, with the highest accuracy of 93.2% for compound names adopting the Pinyin + generic term strategy. A bidirectional feedback mechanism was further designed to optimize entity recognition, forming a closed-loop system with a matching accuracy of 92.4%.

3.3 Descriptive Statistics

Based on semantic components extracted via lexical analysis, functional attribute tagging was performed by matching the components against dedicated lexicons. Six categories of functional attributes were automatically annotated, including body-part-based types and efficacy-based types. Inter-annotator agreement was evaluated using Cohen's kappa ($\kappa = 0.87$), which indicated an "almost perfect" level of consistency.

The most common type of word forms (38.2% of entries) is the type T5 (Compound). This indicates that many descriptive words are incorporated in Chinese medicinal herbs name. T1 (Part) and T2 (Morphology/Color) are the most common types with 24.8% and 19.5% respectively.

In terms of the number of translation strategies, S3 (Pinyin + Generic Noun) was the most common strategy (44.1% of translations) and the most popular approach (25.9% Latin). S5 (Literal/Descriptive Translation) was applied to some herbs (S4 (Pure Pinyin), S6 (Efficiency Based Translation) and S7 (Hybrid Translation with Annotation).

3.4 Correlation Analysis Results

We performed a chi-square test of independence with word-formation type as the independent variable and translation strategy choice as the dependent variable. The test was 867.41 ($p < .001$) and strongly related variables, cramer V = 0.41, which according to Cohen interpretation [10], is moderate to strong effect size between word-formation type and translation strategy.

The major spatial observation is that part-based is close to S2 (Latin Scientific Name with Part) and S3 (Pinyin + Generic Noun) and efficiency-based is closer to S6 (Efficiency Based Translation) than S3 is close to S1 (Pure Latin Scientific Name). S6 is marginally close to other word-formation type clusters. First two dimensions provide a representative solution, 48.3% variance explained by Dimension 1 and 32.1% explained by Dimension 2. Two interpretable quadrants emerge from point distribution: upper-left “Scientific-Nativization” quadrant (including Latin-based strategies S2 and S1) and lower-right “Cultural-Foreignization” quadrant, where T5 is strongly associated to S3 (Pinyin + Generic Noun) and S7 (Hybrid Translation with Annotation).

We modelled translation strategy selection (S1–S7) by word-formation types (T1–T6) with T1 as the reference category. Word-formation type proved to be the strongest predictor of strategy choice and its predictive effect is quantified by odds ratios. For example, a T5 (Compound) name is 11.29 times more likely to be translated with S3 strategy than T1 name, but only 0.08 times more likely to get translated with S2 strategy than T1 name.

4 Discussion

Based on authoritative sources including the 2020 Chinese Pharmacopoeia, we established systematic annotation schemes for word-formation types and translation strategies, and identified stable correlations between them. Compound names (T5, 38.2%) were the most common, followed by part-based (T1, 24.8%) and morphology/color-based names (T2, 19.5%). For translation strategies, Pinyin + generic noun (S3, 44.1%) was dominant, while Latin-based strategies (S1 + S2) accounted for 38.1%, and efficacy-based translation (S6) was rarely used (2.0%). Part-based names (T1) preferred Latin with part (S2); compound names (T5) favored S3; morphology/color-based names (T2) inclined to literal translation (S5); efficacy-based names (T3) avoided S6 to reduce clinical risks. Word-formation significantly determines strategy selection along the scientific–cultural and domestication–foreignization dimensions. We therefore propose a word-formation-based translation decision model that recommends prioritized strategies for different name types.

5 Conclusion

This present study established a specialized English-Chinese parallel corpus for Chinese Materia Medica (CMM) and constructed a multi-feature fusion prediction model to quantitatively analyze the correlation between word-formation attributes and translation strategies of CMM names. By integrating corpus-driven methods and domain-specific deep learning techniques, this research effectively remedies the over-reliance on qualitative analysis and subjective experience in traditional TCM translation studies, realizing intelligent and data-based translational decision-making for CMM terminology. The empirical results confirmed that word formation acts as a decisive factor affecting strategy selection. Specifically, compound names prefer the Pinyin plus generic noun strategy, while part-based and morphology-related terms are inclined to Latin scientific translation and literal translation, respectively. Low utilization of efficacy-focused translation also indicates such approaches may incur potential clinical risks.

The proposed BERT-CRF ensemble framework yields satisfying prediction accuracy, offering a feasible technical reference for standardized TCM terminology translation. Nevertheless, this study still has certain limitations: the corpus fails to cover multi-category medicinal materials, and contextual factors are not incorporated into the model. Future work will expand the corpus coverage, optimize the algorithm framework, and add semantic disambiguation modules for culture-loaded terms. Overall, this research provides systematic translational guidelines for CMM terms, standardizes international term specifications, and further facilitates the cross-cultural dissemination of TCM knowledge worldwide.

Acknowledgment

Chongqing Municipal Education Commission Humanities and Social Sciences Research Project; Project Category: Planning Project; Project Title: A Study on Word-formation and Translation Strategies for the Construction of an English Corpus of Chinese Medicinal Materials; Project Number: 25SKGH274

References

1. Pritzker, S. E .Standardization and Its Discontents: Translation, Tension, and the Life of Language in Contemporary Chinese Medicine[J].East Asian Science Technology & Society, 2014, 8(1):25-42.DOI:10.1215/18752160-2406196. https://www.tandfonline.com/doi/full/10.1215/18752160-2406196.

2. LAI H, MOORE A L, YAN X D, et al. A qualitative analysis of WHO's international standard terminologies on TCM[J]. Chinese Medicine and Culture, 2025, 8(1): 86-95. DOI:10.1097/MC9.0000000000000115. https://journals.lww.com/chinesemedicineandculture/fulltext/2025/03000/a-qualitative-analysis-of-who-s-international.9.aspx.
3. XU Q H. WHO international standard terminologies on traditional Chinese medicine: use in context, creatively[J]. Integrative Medicine in Nephrology and Andrology, 2023, 10(2): e00029.DOI: 10.1097/IMNA-D-22-00029. https://journals.lww.com/integrative-nephrology-and-andrology/fulltext/2023/06000/who-international-standard-terminologies-on.2.aspx.
4. XIAO Y E, ZHANG H X. A history of standardization in the English translation of traditional Chinese medicine terminology[J]. Journal of Integrative Medicine, 2017, 15(5): 344-350.DOI:10.1016/S2095-4964(17)60357-1. https://www.sciencedirect.com/science/article/pii/S2095496417603571.
5. LIM L, CHEN P, WANG V X. Translating TCM nomenclature into English: a bilingual reference tool needed for clinical practice[J]. European Journal of Integrative Medicine, 2022, 54: 102155. DOI:10.1016/j.eujim.2022.102155. https://www.sciencedirect.com/science/article/pii/S1876382022000848.
6. Lin X, Huang C J, Lai J Y, et al. An Analysis of English Translation of Reduplicated Words in Treatise on Cold Damage Diseases from the Perspective of Hermeneutic Translation Theory[J]. Guiding Journal of Traditional Chinese Medicine and Pharmacy, 2026, 32(1): 260-264,270.DOI:10.13862/j.cn43-1446/r.2026.01.044. https://dx.doi.org/10.13862/j.cn43-1446/r.2026.01.044.
7. CHEN K, HU J, KARABULATOVA I. Translation adaptation of TCM (Traditional Chinese Medicine) terminology for speakers of other cultures: features of cultural connotation[J]. Philosophy, Ethics & Humanities in Medicine, 2025, 20(1). DOI:10.1186/s13010-025-00199-7. https://pehmed.biomedcentral.com/articles/10.1186/s13010-025-00199-7.
8. LAI H, MOORE A L, YAN X, et al. TCM-centric innovations in WHO's international standard terminologies: a qualitative exploration[J]. Chinese Medicine and Culture, 2024. DOI:10.1097/mc9.0000000000000115. https://journals.lww.com/chinesemedicineandculture/fulltext/2024/09000/tcm_centric_innovations_in_who_s_international.6.aspx.
9. YE X, DONG M H. A review on different English versions of an ancient classic of Chinese medicine: Huang Di Nei Jing[J]. Journal of Integrative Medicine, 2017, 15(1): 11-18. DOI:10.1016/S2095-4964(17)60310-8. https://www.sciencedirect.com/science/article/pii/S2095496417603108.
10. YE X X, LI Y T, SANG Z. Key factors influencing the TCM nomenclature in international standards[J]. Pharmacological Research, 2021, 170: 105516. DOI:10.1016/j.phrs.2021.105516. https://www.sciencedirect.com/science/article/pii/S1043661821002417.