Standardization of Translation of Green Leather Processing Terminology: Taking Chrome-Free Tanning and Bio-Based Tanning Agents as Examples

Hengzhen Zhang

Luohe Medical College, Luohe 462000, Henan, China

Corresponding author: Hengzhen Zhang (e-mail: 18639546004@163.com)

ABSTRACT Accurate cross-linguistic representation of technical terminology is essential for knowledge sharing and international collaboration in emerging sustainable manufacturing technologies. This study addresses semantic ambiguity and conceptual inconsistency in the translation of green leather processing terminology by focusing on chrome-free tanning and bio-based tanning agents. A Domain-Adaptive Semantic Mapping (DASM) framework is proposed to integrate terminology theory with deep semantic learning. The method constructs a structured bilingual corpus and employs an improved XLM-RoBERTa-based semantic component parser to decompose terms into core concept, process attribute, and ecological attribute subspaces for fine-grained cross-language alignment. A semantic density–driven decision strategy further optimizes standardized translation selection by balancing conceptual equivalence and system consistency. Experimental evaluation demonstrates translation accuracy of 99.1% for Level 1 bio-based tanning terminology, terminology network modularity of 0.72, and a relation misalignment rate of only 3.2%, while maintaining robust performance for structurally complex concepts. Beyond supporting standardized terminology management in leather science, the proposed framework provides a transferable methodology for multilingual technical communication, semantic knowledge organization, and international standardization in engineering disciplines where precise information representation and interoperable terminology are critical, including digital communication and intelligent engineering systems.

INDEX TERMS terminology standardization, domain-adaptive semantic mapping, semantic parsing, cross-lingual alignment, technical communication

I. INTRODUCTION

THE concept of sustainable development in the leather sector has achieved international agreement, and enhancing innovation and usage of green leather processing technologies is destined to be the direction for this sector's transformation, a trajectory mirrored in the textile industry's urgent focus on developing innovative, resource-efficient, and non-toxic finishing and dyeing technologies. In this context, avant-garde technological methods such as chromium-free tanning and bio-based tanning agents are being exchanged more between the international academic and industrial community [1-3]. A common and precise system of professional terminology is a necessary foundation for knowledge dissemination and for the transfer of technology, and of paramount importance for the promotion of deep cooperation and coordinated development in this specific field.

However, the cross-linguistic translation of terminology associated with green leather processing encompasses a number of technical challenges, which are primarily evidenced by two major aspects: ambiguous translations and a lack of systematic knowledge. First, substantial ambiguities are present in the Chinese and English translations of new concepts. For example, the translations are interchangeable, leaving recoverable concepts ambiguously defined, with "aldehyde tanning" and "organic tanning" are used interchangeably, resulting in blurred concepts [4,5]. Second, the lack of a systematic understanding of the terminology is striking. Distinct researchers promote different definitions of the referents of "bio-based tanning agent" and "renewable tanning agent." Existing translation activity is primarily based on the personal experience of the domain expert and not based on objective and unified standardized criteria, leading to a highly subjective and arbitrary exercise which does not accomplish the goal of standardized knowledge services [6,7].

The existing studies on methodology of how to standardize terminology has some shortcomings. Dictionary based,

literal translation cannot manage polysemy or conceptual equivalence, for it can correctly translate "white tanning," which diverges from the actual meaning, too much English it means a specific style of soft leather [8,9]. Models of statistical machine translation lack efficacy in narrow domains with limited training datasets, they can create semantic errors, for example, "metal-free tanning agent tanning" mistranslated as "metal-free tanning" [10-12]. Manual compilation methods based in terminology theory can provide accuracy, but they take excessive time and effort, are not efficient, and are hard to flexibly adjust to the fast pace of concept evolution [13,14]. None of these methods fundamentally solve the problems of systematicity, accuracy, and consistency in the translation of green leather processing terminology.

To address the lack of standardized terminology translation in green leather processing due to a lack of systematic methodology, this paper proposes DASM, a systematic framework integrating terminology principles and deep learning techniques. First, this method automatically constructs a bilingual sentencelevel aligned corpus from authoritative international standards and Chinese academic literature, and annotates semantic relationships according to a concept hierarchy. Then, a multi-granularity semantic parsing module is designed, using a semantic component parser to decompose terms into core and limiting morphemes, and employing an attention-based cross-lingual word embedding model to achieve precise alignment and semantic similarity calculation of source and target language morpheme units in vector space. Finally, an optimal standardized translation is selected from the candidate translation set through a decision function that integrates concept subordination and contextual compatibility. This research establishes the first systematically validated standard set of terminology translation in this field, providing theoretical support and practical tools for knowledge organization and cross-lingual information retrieval in sustainable textile and leather science, thereby enhancing global collaboration and conceptual consistency across both industries.

II. TERMINOLOGY STANDARDIZATION METHOD BASED ON DASM MODEL

CONSTRUCTION OF A STRUCTURED BILINGUAL TERMINOLOGY DATABASE FOR GREEN LEATHER

A high-quality bilingual terminology database is the foundation for subsequent semantic alignment and normative decisions. The data collection in this study strictly followed the principles of authority and domain relevance. The scope of data collection focused on terms in key areas such as chromium-free tanning, biobased tanning agents, and green auxiliaries. The source language terminology was limited to standard documents published by the International Federation of Leather Chemists and Engineers and authoritative journals in the field, such as the Journal of the American Leather Chemists Association [15,16]. To ensure a comparable level of academic rigor and terminological relevance for the target language, Chinese terms were exclusively sourced from core Chinese journals that are recognized as the leading and most authoritative publications in the field of leather science and engineering in China, namely Leather Science and Engineering and Collagen and Leather [17-19]. Initial biases arise because Chinese corpora may contain non-standardized translations. However, the DASM model does not treat Chinese candidate terms as pre-standardized; instead, it uses authoritative source terms as semantic anchors. The subsequent cross-linguistic alignment process, based on multi-subspace semantic analysis, rigorously evaluates and filters candidate translations to match these anchors. This ensures that the final standardized output depends on its conceptual consistency with the authoritative source terms, effectively overcoming potential noise and biases in the initial target corpus. The original corpus was also collected through direct web crawling and manual retrieval, and represents the initial unit heterogeneous set of a corpus.

In order to automatically identify candidate terms and make initial alignment relationships from the raw corpus, this article utilizes an extraction approach that includes a combination of syntactic rules and statistical features [20,21]. For the English corpus, a custom pipeline based on the SpaCy library is used to identify candidate terms by identifying noun phrases that match pre-defined part-of-speech tagging patterns. At the same time, the term boundaries are determined through a calculation of the mutual information of the left and right adjacent words and the term frequency-inverse document frequency score, thus removing common lexically free collocations. The quantification of this process involves a function of term degree:

$$\begin{aligned}\mathrm{Score}(T) =& \lambda_1 \,\mathrm{MI}_{\mathrm{left}}(T) + \lambda_2 \,\mathrm{MI}_{\mathrm{right}}(T) \\ &+ \lambda_3 \log(\mathrm{TFIDF}(T) + 1)\end{aligned} \tag{1}$$

In Formula 1, T refers to the candidate term, $\mathrm{MI}_{\mathrm{left}}$ and $\mathrm{MI}_{\mathrm{right}}$ refer to the mutual information with the left and right adjacent solids, $\log(\mathrm{TFIDF}(T) + 1)$ is used to increase the weights of frequent and proprietary concepts in the domain. Finally, all solid strings with terminology scores that exceeded the empirical threshold were selected as candidate terms.

Once the bilingual candidate term set is obtained, the next important step is to establish relationships between those candidates. For sentence-level alignment, either metadata or contextual structure is used. For standard documents that are known to have an official bilingual version, these paragraphs can be precisely matched based on hashing values. For published research papers, their natural structural properties are utilized, and sentences can be aligned using a dynamic time warping algorithm. Their BM25 (Best Matching 25) vector similarity is calculated to establish preliminary term translation pairs. After the above processing, a three-level structured terminology database was constructed, containing source language terms, target language candidate terms, and their contexts. This database not only stores term strings but also records their source literature, frequency of

occurrence, and aligned example sentences, providing rich data support for subsequent in-depth semantic analysis. The core statistical information of the terminology database is shown in Table 1.

TABLE 1. Statistical characteristics of the structured bilingual terminology database

Term category	Number of source language terms	Number of candidate translations in the target language	Number of effective alignment context pairs
Chrome-free tanning technology	347	1052	1865
Bio-based tanning agents	291	873	1521
Green auxiliaries process	214	598	1103
Total	852	2523	4489

CROSS-LINGUISTIC ALIGNMENT BASED ON SEMANTIC COMPONENT PARSING

After obtaining a structured bilingual terminology database, the core task is to achieve accurate alignment between source language terms and target language candidate translations at the deep semantic level. This section describes an alignment model based on semantic component analysis, which aims to go beyond traditional surface semantic matching and achieve concept-level mapping by deconstructing the internal semantic structure of terms.

The parser's design stems from an in-depth analysis of the semantic composition of green leather processing terminology. Terms in this field typically encompass core concepts of materials or methods, specific process attributes, and key ecological attributes simultaneously. For example, in "bio-based tanning agent," "tanning agent" is the core concept, while "bio-based" implies both material origin (process attribute) and environmental friendliness (ecological attribute). Simply performing overall vector matching easily overlooks the independent contributions of these key semantic components, leading to alignment bias. Therefore, this study explicitly decomposes terminology vectors into three orthogonal semantic subspaces: "core concepts," "process attributes," and "ecological attributes," aiming to achieve independent representation and refined alignment of the multidimensional semantics of terms. The overall architecture and information flow of the model are shown in Figure 1.

The input to the model consists of source language terms and their contexts from a terminology corpus, as well as target language candidate terms and their contexts. First, an XLM-RoBERTa (Cross-Lingual Language Model - Robustly Optimized BERT Pretraining Approach) model pretrained on multilingual corpora is used as the basic encoder, which can map the input terms and their context sequences into a dense vector representation that incorporates contextual information [22,23].

To achieve fine-grained semantic deconstruction, this paper designs a semantic component parser. This parser projects the complete term vector onto three orthogonal semantic subspaces: the core concept subspace, the process attribute subspace, and the ecological attribute subspace [24,25]. The projection process is implemented through three independent linear transformation matrices, generating corresponding subspace vectors:

$$v_C = W_C h_s, \quad v_P = W_P h_s, \quad v_E = W_E h_s \tag{2}$$

In Formula 2, WC WP, WE are trainable parameter matrices. The target language candidate terms undergo the same encoding and parsing process to obtain their corresponding subspace vectors.

The matrices WC, WP, WE are initialized randomly and optimized during training via a multi-task objective that includes term alignment and semantic component classification. Specifically, we employ an attentionbased gating mechanism to softly assign each token in the term to one of the three semantic subspaces, based on its contextualized representation from XLM-RoBERTa. This allows the model to automatically learn the mapping of morphemes to semantic categories without relying on pre-defined morpheme lists or manual annotations. The training objective encourages the subspaces to be orthogonal, thereby promoting disentangled representations of core concepts, process attributes, and ecological attributes.

Cross-linguistic alignment is performed by taking the cosine similarity in each subspace between the source and target languages and averaging the weighted results [26,27]. The similarity in each subspace is computed in the following manner:

$$\mathrm{sim}_X(S, T_j) = \frac{v_X \cdot u_{X,j}}{\|v_X\| \|u_{X,j}\|} \tag{3}$$

In Formula 3, all subspaces of {C, P, E} are equally represented by X. Lastly, the overall semantic alignment score, for a given source language term from the source language S and a candidate target language term Tj, is expressed in Formula 4:

$$\mathrm{AlignScore}(S, T_j) = \alpha \cdot \mathrm{sim}_C + \beta \cdot \mathrm{sim}_P + \gamma \cdot \mathrm{sim}_E \tag{4}$$

In this formula, the weight coefficients α, β, and γ serve to balance the relative importance of different semantic elements in making the alignment decision. Their sum equals 1 and they are trained via a validation set. This way, their translation equivalence can rank highly in the core concept matter while providing exacting consistency in process particulars and environmental characteristics.

DECISION-MAKING AND GENERATION OF STANDARD TRANSLATIONS BASED ON SEMANTIC DENSITY

After deriving the semantic alignment scores of candidate translations, the best standard translation must be selected

FIGURE 1. Semantic component parsing and cross-language alignment model architecture

from multiple high-scoring candidates. Traditional methods may simply rank and select on the basis of alignment scores, neglecting a systematic and consistent terminology system [28,29]. This paper designs a decision mechanism based on semantic density that gives a comprehensive consideration of candidate translation distribution characteristics in the concept space and their compatibility with the embedded terminology system.

With a source term S and k candidate translations T1, T2, ..., Tk, an alignment score is computed for each candidate translation Tk. First, a semantic vector space of the target language terminology can be created that includes all term normalizations and their semantic vectors. For each candidate translation, the candidate translation's local semantic density in the semantic vector space can be calculated:

$$\text{Density}(T_k) = \frac{1}{n}\sum_{i=1}^{n}\exp\left(-\frac{d(T_k, N_i)^2}{2\sigma^2}\right) \tag{5}$$

In formula 5, d(Tk,Ni) indicates the semantic distance between a candidate translation Tk and its i-th i near neighbor Ni, n is the number of nearest neighbors considered, and σ is the bandwidth parameter that determines the density decay rate. This density value indicates the extent to which the candidate translation is integrating into the existing terminology system, and a higher density means that the translation is more semantically aligned with existing terms.

To prevent the introduction of new translations that are semantically overlapping with existing standard terms to the point of causing ambiguity or confusion (e.g., creating near-synonyms without functional distinction), it is necessary to consider the semantic distinctiveness of candidate translations. The goal is not to avoid all semantic proximity, but to penalize candidates that are excessively similar to any existing standard term, which would undermine conceptual clarity. The semantic distinctiveness of candidate translations is defined in the following way:

$$\text{Distinctiveness}(T_k) = 1 - \max_{Q\in G}\sin(T_k, Q) \tag{6}$$

In Formula 6, G denotes the standard-set of terms and sin(Tk,Q) denotes semantic similarity (i.e., a similarity measure) between terms in the standard-set. This measure ensures new translations do not overlap with sufficient difference from the existing standard terms.

By combining alignment score, semantic density, and discriminative power, a final translation quality evaluation function is constructed:

$$\begin{aligned}\text{Quality}(T_k) =& \omega_1 \cdot \tilde{a}_k + \omega_2 \cdot \widehat{\text{Density}}(T_k) \\ &+ \omega_3 \cdot \widehat{\text{Distinctiveness}}(T_k)\end{aligned} \tag{7}$$

In Formula 7, $\tilde{a}_k$is the alignment score, $\widehat{\text{Density}}(T_k)$, and $\widehat{\text{Distinctiveness}}(T_k)$ are the values of the semantic density, and distinguishability, respectively, after max-min normalization, and ω_1, ω_2, and ω_3 are the associated weight coefficients, where $\omega_1 + \omega_2 + \omega_3 = 1$. These coefficients were determined programmatically via a grid search on a validation set, with the objective of maximizing the overall expert-rated translation quality score. This quality assessment function checks the accuracy of the translation, but also the relationship between the new translation, an existing terminological system, and the consequently required distinguishability. The weight coefficients ω_1, ω_2, and
ω_3 are calibrated to balance these factors. Specifically, the relatively lower value of ω_3 (distinctiveness weight) ensures that semantic alignment and systemic integration remain the primary decision drivers. The distinctiveness term acts primarily as a safeguard against introducing redundant or conflicting terms, rather than as a force for maximizing novelty. This design ensures that the selected translation is both accurately aligned and harmoniously integrated into the existing terminology system without unnecessary deviation. Based on the quality score, the candidate translation that maximizes Quality(Tk) is ultimately selected as the standard translation. In cases where multiple candidate translations are judged to have a similar quality score, the candidate translation that best matches the part-of-speech structure of the source language term is preferred in order to maintain consistency in the grammatical structure of the term. This

systematic process ensures that an optimal standard translation has been obtained from several rational candidate translations, creates an accurate and systematic terminology translation system, and ultimately produces a list of the most suitable standard translations of chrome-free tanning agents and bio-based tanning agents.

III. EXPERIMENTAL DESIGN AND DATASET CONSTRUCTION

EXPERIMENTAL DATASET AND EVALUATION METRICS

To validate the efficacy of the method proposed in this study, this paper built a bilingual terminology translation evaluation dataset in the domain of green leather processing. The source language part of the dataset came from official technical standards published by the International Federation of Leather Craftsmen and Chemists, as well as from authority papers published in the Journal of the American Leather Chemists Association. The target language part of the dataset was collected concurrently from academic literature published in core Chinese journals, such as Leather Science and Engineering. The dataset consists of three core subfields, including chromium-free tanning, bio-based tanning agents, and green auxiliaries, to ensure comprehensiveness and representativeness of the research subject on green leather processing. The process of creating the dataset was strictly held to academic standards. First, this paper utilized a combination of specialized extraction tools and domain experts to extract 852 source terms and their corresponding 2523 candidate translations from the original corpus. Each term pair includes full contextual information (the original sentence with the term and surrounding paragraphs). To further increase the reliability of the evaluation, it convened a review committee of independent experts, five individuals who have greater than 10 years of experience in leather chemistry and engineering, to review and annotate the translation quality of all terms in the dataset. Experts scored the translations based on different dimensions, and only translations rated as high-quality by four or more experts were adopted as the benchmark and included in the test set.

Quantitative evaluation system covers the accuracy of terminology translation, which measures the degree of matching between the standardized translations output by the model and the benchmark answers marked by experts. The calculation method is the ratio of the number of correct translations to the total number of terms in the test set. The consistency evaluation model of the terminology system is a systematic performance in the construction of the entire terminology system. It is achieved by calculating the coordination of translations of terms with the same root. Specifically, it is quantified by the standardized ratio of the translation structure of terms with the same root [30,31].

BASELINE MODEL AND PARAMETER SETTINGS

The experiment selected three representative baseline methods for comparative analysis. The statistical machine translation baseline adopted the phrase-based Moses system, which was adaptively trained by injecting a small number of leather-related professional terms on the basis of a general domain parallel corpus[32]. The neural machine translation baseline adopted the Transformer architecture and tested the mBART (Multilingual Bidirectional and Auto-Regressive Transformers) pre-trained on a general large-scale corpus. Multilingual models, as well as the SciBERT (Scientific Bidirectional Encoder Representation from Transformers) model specifically pre-trained on scientific literature [33,34]. Terminological methods implement a manual compilation process based on conceptual patterns, with domain experts performing systematic translations according to terminological principles [35].

The proposed model has undergone thorough parameter optimization. The semantic encoder uses the XLM-RoBERTa-Large version with a hidden layer dimension of 1024 and 16 attention heads. The projection dimensions of the three subspaces of the semantic component parser are all set to 768. The weight coefficients in Equation (7) are determined through grid search, with optimal values of $\omega_1 = 0.60$, $\omega_2 = 0.25$, and $\omega_3 = 0.15$. The initial learning rate was set to 2×10^{-5}, the batch size to 32, and the number of training epochs to 50. Training was terminated early when the performance on the validation set no longer improved. The model parameter configurations are shown in Table 2.

TABLE 2. Configuration of Key Model Parameters

Parameter category	Parameter name	Setting values
Model architecture	Semantic encoder	XLM-RoBERTa-Large
	Hidden layer dimensions	1024
	Number of attention heads	16
Subspace configuration	Core concept dimension	768
	Process attribute dimension	768
	Ecological attribute dimension	768
Optimization parameters	Optimizer	AdamW
	Initial learning rate	2×10^{-5}
	Batch size	32
	Alignment score weight ω_1	0.60
	Semantic density weight ω_2	0.25
	Discrimination weight ω_3	0.15

Table 2 presents the essential parameter settings of each model. The parameter distribution indicates the heavily-parameterized large-scale pre-trained model in the semantic encoder generates a solid understanding of terms. The equal dimension setup for the three subspaces is to ensure more effective representation of the different semantic components. Again, the weight coefficients demonstrate the importance of alignment scores to provide context in the decision making, while semantic density and discriminablilty added necessary constraints. This parameter structure retains the systematicity and consistency of terminology systems and assures translation performance.

And the key hyperparameter configurations for other baseline models are shown in Table 3, including the number of layers and hidden dimensions.

TABLE 3. Baseline Model Parameter Configuration

Model / Parameter	Architecture / Approach	Key Configuration Details
Transformer (Our Impl.)	Encoder-Decoder	6 layers, 512 hidden dimensions, 8 attention heads
Statistical Machine Translation (SMT)	Phrase-based (Moses)	Phrase table limit: 100, Reordering: msd-bidirectional- fe
mBART	Pre-trained Multilingual Seq2Seq	12 layers, 1024 hidden dimensions (mBART-large)
SciBERT	Pre-trained Encoder (BERT-based)	12 layers, 768 hidden dimensions

IV. EXPERIMENTAL RESULTS AND ANALYSIS

COMPARATIVE ANALYSIS OF TERMINOLOGY TRANSLATION ACCURACY

To systematically evaluate the performance of the proposed method in translation tasks of varying complexity, this study designed a hierarchical accuracy test experiment. The experiment divided the test samples into four difficulty levels based on the number of sub-word tokens in the terms, as generated by the XLM-RoBERTa tokenizer. This metric directly reflects the computational complexity faced by the encoder when processing terms of varying lengths, covering three core professional fields: chromium-free tanning, bio-based materials (bio-based tanning agents), and green auxiliaries. By comparing the translation accuracy of the proposed method with two mainstream baseline models, the model's ability to translate simple and complex compound terms was examined. The performance distribution characteristics of each method under different difficulties and fields are shown in Figure 2.

Figure 2 illustrates the performance of three translation methods across three professional fields. Proposed method of this paper achieves the highest accuracy of 99.1% in translating Level 1 terms in the bio-based tanning agent field, and maintains an excellent performance of 94.5% in translating Level 4 complex terms, significantly outperforming the Transformer model's 87.6% and the Statistical Machine Translation (SMT)model's 81.2%. This performance difference stems from the semantic component parsing mechanism introduced in this paper, which effectively deconstructs the internal semantic structure of complex terms, avoiding the semantic information loss caused by traditional end-to-end translation models for long terms. In the chrome-free tanning process field, as the terminology difficulty increases from Level 1 to Level 4, the accuracy of proposed method decreases from 98.2% to 91.7%, a drop of 6.5 percentage points, while the Transformer model and the Statistical Machine Translation model show decreases of 12.8 and 16.7 percentage points, respectively. This suggests that the proposed method has greater robustness with greater semantic complexity from an increased number of morphemes. The overall accuracy level in green additives is relatively low. This field usually contains specific chemical modifiers and functional descriptors, which require more stringent accuracy in terms of the semantic understanding of the modifiers and descriptors of those functionalists. Still, the method suggested here remains at a benchmark level of over 90%. From the experimental results, it shows that the semantic parsing framework described here can successfully keep stable translation performance across difficulty levels and professional domains, providing a reliable technical solution to the translation challenges facing complex terminology in green leather processing.

TERMINOLOGY SYSTEM CONSISTENCY ASSESSMENT

In order to assess the internal logical coherence and consistency of the conceptual structure of the terminology systems generated by different translation methods, this study employs complex network analysis to systematically investigate the output results. The experiment entails a semantic relationship network of terms, computation of topological characteristic indicators like modularity, average clustering coefficient and average path length, and the quantitative evaluation of the relative capacity of each method to sustain systematicity and conceptual semantic relevance of terms. Modularity measures the extent to which the system is partitioned into clearly defined, semantically cohesive concept clusters, reflecting the logical organization of domain knowledge. The average path length captures the overall semantic connectivity and efficiency of conceptual navigation within the term network. A lower value indicates that semantically related terms are closely linked, facilitating efficient knowledge association and retrieval. The number of connected components inversely indicates the conceptual integrity of the system, with fewer components suggesting a more unified and less fragmented knowledge structure. Together, these metrics provide a holistic assessment of whether the translated terminology forms a well-structured, coherent, and navigable conceptual system, which is the ultimate goal of standardization. The network structure comparison and the distribution of indicators are provided in Figure 3.

Figure 3 presents the four translation methods on the horizontal axis, and the standardized network metrics on the vertical axis. Five important metrics were analyzed in this study: the modularity, average clustering coefficient, average path length, number of connected components, and relation misalignment rate. The method proposed in this study has a modularity of 0.72 and a relation misalignment rate of 3.2%, while the Transformer method has a modularity of 0.58 and a relation misalignment rate of 16.3%. The advantage of the proposed method is due to the semantic density decision mechanism proposed in this paper. This mechanism is based on quantifying the distribution characteristics of candidate translations in the space of meanings to ensure the level of meaning consistency between new terms and existing

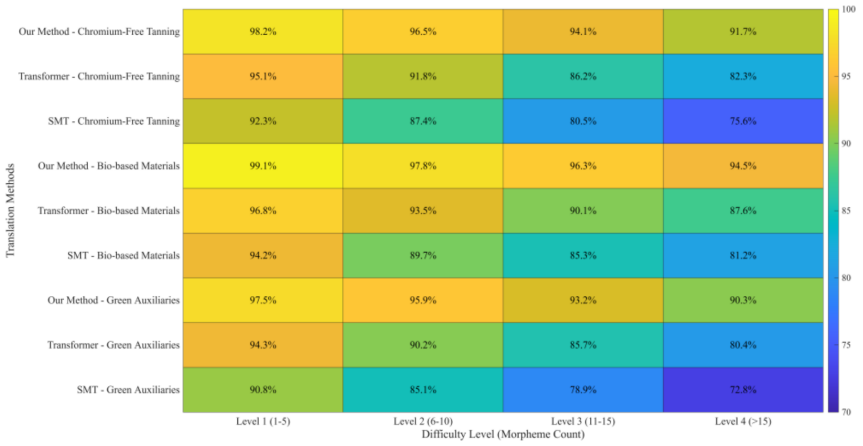

FIGURE 2. Accuracy of each method under different terminology categories and difficulty levels

systems, to reduce relation misalignment, and to enhance network modularity. The average path length metric displays a similar pattern: the proposed method has an average path length of 2.15 and the Transformer method has an average path length of 3.42. A longer path length represents weaker semantic relations between terms and reduced effectiveness in concept retrieval due to the traditional neural translation models that do not model the structure of the terminology system explicitly. In terms of connected components, proposed method produces 8 distinct sub-networks. This result does not indicate undesirable fragmentation but reflects the inherent modularity of the domain knowledge. These eight connected components correspond to major, relatively self-contained conceptual clusters within green leather processing (e.g., chrome-free tanning technologies, bio-based materials, green auxiliaries). The absence of direct semantic links between these clusters is semantically justified and aligns with the domain's logical structure, demonstrating that the standardized system preserves the natural conceptual boundaries. Whereas, the Transformer and SMT (Statistical Machine Translation) methods produce 15 and 23 connected components, respectively. The large number of connected components indicates a fragmentation problem of the terminology system, caused by the lack of a stable semantic relationship between concepts during translation. The experiments demonstrate that the terminology decision-making method based on the constraints of semantic density can effectively establish a compact and logically consistent terminology system to better organize domain knowledge.

VISUAL ANALYSIS OF CROSS-LINGUISTIC SEMANTIC SPACE ALIGNMENT

In order to evaluate the efficacy of the suggested methodology for concept alignment in the cross-linguistic semantic spaces, this research quantitatively assesses the quality of translation according to the distribution relationships of terms in the source language and translated terms in the target language in a highdimensional semantic space. The experiment utilizes a method of semantic encoding via a pre-trained multilingual model to position the bilingual terms in a collective vector space, and measures their relative distance and distribution characteristics. The semantic space projection and distance measurement is visually presented in Figure 4.

Figure 4 includes four complementary visualizations which illustrate the effect of semantic alignment from the following aspects: spatial distribution, statistical characteristics, typical cases, and density function. The scatter plot shows the semantic space projection after t-SNE (t-distributed stochastic neighbor embedding) dimensionality reduction, where the horizontal and vertical axes provide the two semantic dimensions with the distribution of correct, incorrect, and partially correct translation pairs represented with different colors. As shown in Figure 4, correct translation pairs cluster in a tightly pressing area producing an average pairwise Euclidean distance of 0.15, while incorrect translation pairs are scattered with an average distance of 0.43, which had a large variance. The very difference is the mechanism of analyzing semantic components, as previously detailed in this paper. The mechanism of separating three subspaces: core concepts, technological attributes, and ecological attributes maintains closer correspondence of equivalent terms across all semantic dimensions throughout the analysis of the chosen terms. The semantic distance distribution across all correct translation pairs is concentrated with narrow bins with inter-quartile ranges smaller than the average distance of incorrect pairs. The semantic distance distribution of the incorrect translation pairs produced very irregular organization of scatter plots that implicated great instability in semantic understanding. The underlying problem stems from the absence of direct modeling of the internal semantic structure of terms in a end-to-end model. Various case studies indicate the source term "biomimetic tanning" has a distance of 0.08 from its equivalent target term, while "metal-free tanning" has a distance of 0.41. The latter has a conceptual bias because the functional semantic meaning of "metal-free tanning" was lost in direct translation. The density distribution curves substantiate that correct translations cluster in a low-distance region with a unimodal distribution, while incorrect translations have

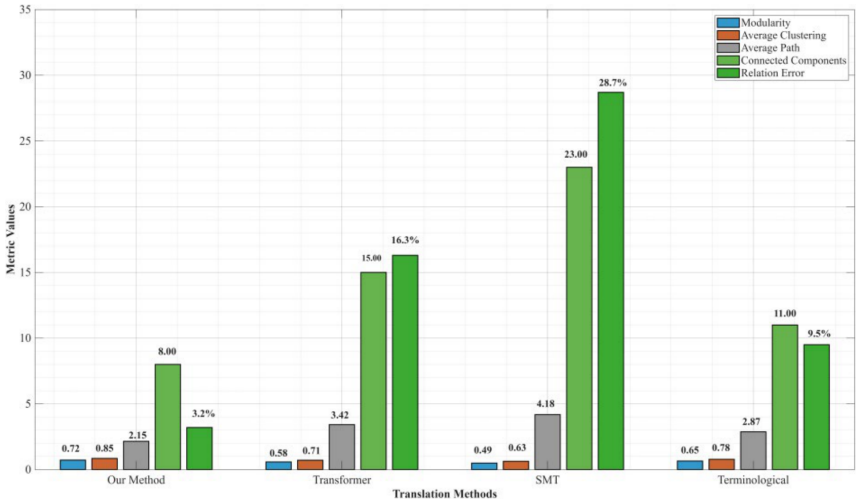

FIGURE 3. Comparison of network topology features of terms

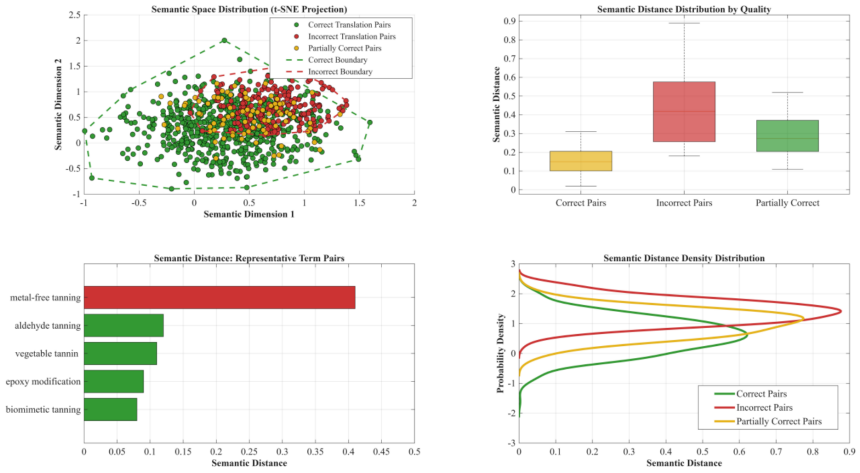

FIGURE 4. Cross-linguistic semantic alignment results

a right-sided skew. Experiments demonstrate the cross-linguistic alignment methodology leveraged from semantic component analysis can effectively reduce the semantic distance of equivalent terms for terminology translation, providing a credible quantitative basis for quality assessment of terminology translation.

CASE STUDIES ON THE TRANSLATION OF COMPLEX CONCEPTS

To further validate the translation quality of the proposed method in relation to meaning-rich terms with structural complexity, this study offers case studies using complex concepts knowingly characterized by structural complexity. This study drew from the two multi-morpheme terms that embodied chemical modifications and descriptions of the process: "metal-free organic tanning system" and "epoxy modified plant tannin." Domain experts scored systematically for four translation approaches over six evaluation domains. The quality ratings over multiple domains and overall performance are provided in Figure 5 and Figure 6.

Figure 5 assesses the performance of the method in terms of conceptual completeness, professional fidelity, contextual applicability, systemic consistency, originality in language, and usability, with extreme paths representing percentage scores. Data indicates that the method achieved superior performance in translating "metal-free organic tanning system," with 96 for conceptual completeness and 95 for systemic consistency, compared to 85 and 80 respectively for the Transformer method. This is primarily due to the mechanism of semantic component analysis described in this paper, which can successfully recognize the functional semantics of "metal-free" rather than its literal meaning, properly mapping "metal-free" to "metalfree tanning agent" rather than a literal translation. In the case of "epoxy modified plant tannin," the method scored 96 for systemic consistency compared to 87 for traditional terminology, which indicates that the mechanism of density of semantic components was able to retain the logical connection between "epoxy modification" and its chemical modification concept group from the traditional terminology system, while the manual compilations, constrained by expert subjective judgment,

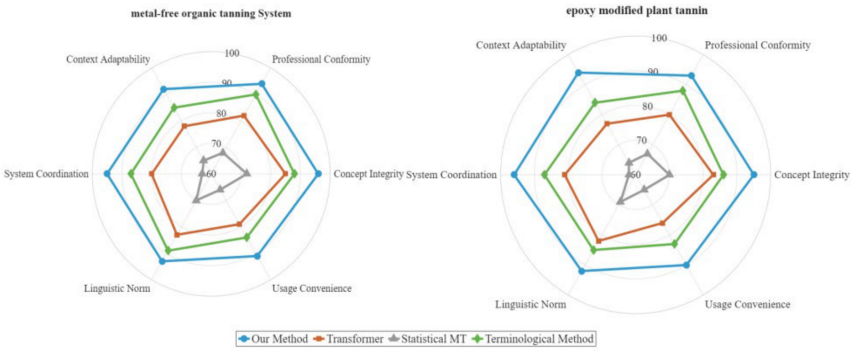

FIGURE 5. Six-dimensional assessment scores for complex terminology concepts

were unable to retain systematic consistency. In Figure 6, the average leaderboard scores of the proposed method in both cases were 93.5 and 93.3, compared to leading the terminology-based methods by 6.3 and 7.8 points, respectively. This stable advantage demonstrates that the multi-subspace semantic alignment strategy universally improves the translation of complex concepts. The experimental results show that semantic parsing-based normalization can balance professional accuracy, systems coordination, and use, providing a systematic solution for the translation of complex terms in the field of green leather processing.

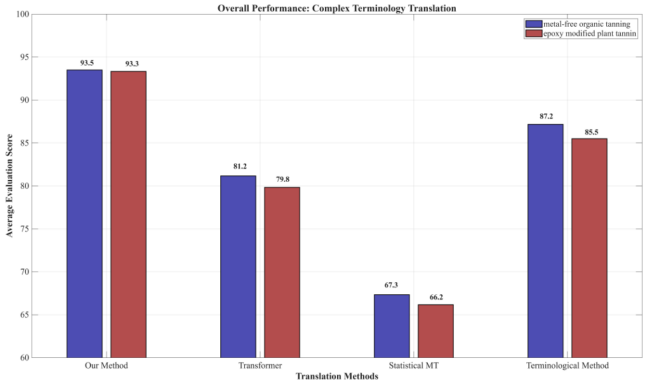

FIGURE 6. Average performance of each method in the case study

V. CONCLUSION

The paper suggests a standardized translation approach for the terminology of green leather processing based on terminological principles, combined with the area of deep learning technology and terminology databases, a methodology that is directly applicable to establishing conceptual precision and terminological clarity within the evolving field of sustainable textile chemistry and processing. A structured bilingual terminology database was created, employing cross-language alignment based on semantic component analysis and a standardized translation decision based on semantic density, effectively addressing issues of translation ambiguity and issues of conceptual inconsistencies. Experimental results show standardized translation accuracy in terms at level 1 difficulty in the bio-based tanning agent domain, 99.1%, with a modularity of 0.72 and a relation misalignment rate of 3.2% in the terminology system consistency assessment, while retaining an average expert score of $\geq$ 93 in complexities of conceptual case studies. Data that demonstrates the cross-language alignment mechanism based on semantic component analysis can effectively capture and align the corresponding multidimensional semantic features of terms within the given cross-language; while the semantic density decision criterion leads to systematicity and consistency of translated names in the conceptual system. The standardized translation framework put forth in this study serves as reliable decision making support for international academic exchange and technology dissemination in the area of green leather processing and had practical application value for the standardization and cross-language dissemination of knowledge systems and understanding in the area of green leather processing.

AUTHOR CONTRIBUTIONS

Hengzhen Zhang designed, collected and analyzed the data, and drafted the manuscript. Hengzhen Zhang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Hengzhen Zhang 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 was supported by Construction and Empirical Research on the Quality Model of Innovative Science and Technology Talents in Henan Province,No:142400411341

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] M. Jiang, Y. Xiao, J. Sang, C. Wang, J. Zhou, and W. Lin, “Biomass-based Tanning Agent for Sustainable Leather Manufacture via Cya-

nuric Chloride Modified Chitooligosaccharide," Journal of the American Leather Chemists Association, vol. 118, no. 10, pp. 417-427, 2023, doi: 10.34314/jalca.v118i10.8230.
[2] R. Wise W, J. Davis S, E. Hendriksen W, D. J. von Behr, S. Prabakar, and Y. Zhang, "Zeolites as sustainable alternatives to traditional tanning chemistries," Green Chemistry, vol. 25, no. 11, pp. 4260-4270, 2023, doi: 10.1039/D3GC00381G.
[3] I. Quaratesi, M. C. Micu, E. Rebba, C. Carsote, N. Proietti, V. Di Tullio, et al., "Cleaner leather tanning and posttanning processes using oxidized alginate as biodegradable tanning agent and nano-hydroxyapatite as potential flame retardant," Polymers, vol. 15, no. 24, pp. 4676, 2023, doi: 10.3390/polym15244676.
[4] S. Radetska, "CHALLENGES AND INNOVATIONS IN SCIENTIFIC AND TECHNICAL TRANSLATION: TERMINOLOGICAL COMPLEXITIES AND 'FALSE FRIENDS'," The Modern Higher Education Review, vol. (9), pp. 119-131, 2024, doi: 10.28925/2617-5266/2024.97.
[5] C. Bertulessi, "Terminology, popularisation and ideology in contemporary China: The 'Scientific & Technical Term of the Day'column," Terminology, vol. 30, no. 1, pp. 58-80, 2024, doi: 10.1075/term.00077.ber.
[6] B. Li, "Conceptual deviation in terminology translation: A case study on translating COVID-19 terminology in multilingual news media," Terminology, vol. 29, no. 2, pp. 351-382, 2023, doi: 10.1075/term.00073.li.
[7] A. A. Matveyuk, A. A. Merker, and Y. A. Nizhelskaya, "ISSUES OF INTERNATIONAL STANDARDIZATION OF TECHNICAL TERMINOLOGY IN THE TRANSLATION OF SCIENTIFIC AND TECHNICAL TEXTS," Bectnik nayki, pp. 4(11 (68)):534-539, 2023, [Online]. Available: https://cyberleninka.ru/article/n/issues-of-internationalstandardization-of-technical-terminology-in-the-translation-of-scientific-and-technical-texts/viewer.
[8] Y. Hou, "Pruning translation of logical and accidental polysemy in traditional Chinese medicine terminology," Terminology, pp. 31(2):, 2025, doi: 311-334. 10.1075/term.24007.hou.
[9] O. A. Alkhonini, "Addressing polysemy in translation pedagogy," Asian-Pacific Journal of Second and Foreign Language Education, vol. 10, no. 1, pp. 28, 2025, doi: 10.1186/s40862-025-00328-x.
[10] T. O. Tafa, S. Z. M. Hashim, M. S. Othman, H. Alhussian, M. Nasser, S. J. Abdulkadir, et al., "Machine Translation Performance for LowResource Languages: A Systematic Literature Review," IEEE Access, vol. 13, pp. 72486-72505, 2025, doi: 10.1109/ACCESS.2025.3562918.
[11] A. L. Tonja, O. Kolesnikova, A. Gelbukh, and G. Sidorov, "Low-resource neural machine translation improvement using source-side monolingual data," Applied Sciences, vol. 13, no. 2, pp. 1201, 2023, doi: 10.3390/app13021201.
[12] S. Araghi and A. Palangkaraya, "The link between translation difficulty and the quality of machine translation: a literature review and empirical investigation," Language Resources and Evaluation, vol. 58, no. 4, pp. 1093-1114, 2024, doi: 10.1007/s10579-024-09735-x.
[13] T. Zhao and M. B. Alias, "Automated programming approaches to enhance computer-aided translation accuracy," PeerJ Computer Science, vol. 10, pp. e2396, 2024, doi: 10.7717/peerj-cs.2396.
[14] K. Xu, Y. Feng, Q. Li, Z. Dong, and J. Wei, "Survey on terminology extraction from texts," Journal of Big Data, vol. 12, no. 1, pp. 29, 2025, doi: 10.1186/s40537-025-01077-x.
[15] Y. Yu, H. Wang, Y. Wang, J. Zhou, and B. Shi, "Construction of a chrome-free tanning system based on highlyoxidized starch–zirconium complexes," Journal of the American Leather Chemists Association, vol. 117, no. 3, pp. 87-95, 2022, doi: 10.34314/jalca.v117i3.4887.
[16] C. K. Ozkan and H. Ozgunay, "Alternative tanning agent for leather industry from a sustainable source: dialdehyde starch by periodate oxidation," Journal of the American Leather Chemists Association, vol. 116, no. 3, pp. 89-99, 2021, doi: 10.34314/jalca.v116i3.4249.
[17] W. Ding, X. Wang, J. Remón, Z. Jiang, X. Pang, Z. Ding, et al., "Engineering an epoxy tanning agent via facile functionalization of sucrose with silane coupling agent for sustainable leather production," Collagen and Leather, vol. 7, no. 1, p. 18, 2025, doi: 10.1186/s42825-025-00200-1.
[18] Y. Wang, Y. Zhang, and Z. Wang, "Biodegradability of leather: a crucial indicator to evaluate sustainability of leather," Collagen and Leather, vol. 6, no. 1, pp. 12, 2024, doi: 10.1186/s42825-024-00151-z.
[19] R. K. Das, A. Mizan, F. T. Zohra, S. Ahmed, K. S. Ahmed, and H. Hossain, "Extraction of a novel tanning agent from indigenous plant bark and its application in leather processing," Journal of Leather Science and Engineering, vol. 4, no. 1, pp. 18, 2022, doi: 10.1186/s42825-022-00092-5.
[20] A. Soliman, "An unsupervised linguistic-based model for automatic glossary term extraction from a single PDF textbook," Education and Information Technologies, vol. 28, no. 12, pp. 16089-16125, 2023, doi: 10.1007/s10639-023-11818-1.
[21] A. Nugumanova, D. Akhmed-Zaki, M. Mansurova, Y. Baiburin, and A. Maulit, "NMF-based approach to automatic term extraction," Expert Systems with Applications, vol. 199, p. 117179, 2022, doi: 10.1016/j.eswa.2022.117179.
[22] B. Li, Y. He, and W. Xu, "Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment," preprint arXiv:2101.11112, 2021, doi: 10.48550/arXiv.2101.11112.
[23] J. A. Ortiz-Zambrano, C. H. Espín-Riofrío, and A. Montejo-Ráez, "Deep encodings vs. linguistic features in lexical complexity prediction," Neural Computing and Applications, vol. 37, no. 3, pp. 1171-1187, 2025, doi: 10.1007/s00521-024-10662-9.
[24] D. Nikolaev and S. Padó, "Investigating semantic subspaces of transformer sentence embeddings through linear structural probing," preprint arXiv:2310.11923, 2023, doi: 10.48550/arXiv.2310.11923.
[25] W. Ponwitayarat, P. Limkonchotiwat, E. Chuangsuwanich, et al., "Space decomposition for sentence embedding," preprint arXiv:2406.03125, 2024, doi: 10.48550/arXiv.2406.03125.
[26] S. Banerjee, B. R. Chakravarthi, and J. P. McCrae, "Cross-Lingual Ontology Matching using Structural and Semantic Similarity," LREC-COLING 2024; 2024:11, [Online]. Available: https://aclanthology.org/2024.ldl-1.2/.
[27] D. Witschard, I. Jusufi, R. M. Martins, K. Kucher, and A. Kerren, "Interactive optimization of embedding-based text similarity calculations," Information Visualization, vol. 21, no. 4, pp. 335-353, 2022, doi: 10.1177/14738716221114372.
[28] A. Gašpar, S. Seljan, and V. Kučiš, "Measuring Terminology Consistency in Translated Corpora: Implementation of the Herfindahl-Hirshman Index," Information, vol. 13, no. 2, p. 43, 2022, doi: 10.3390/info13020043.
[29] K. Semenov and O. Bojar, "Automated Evaluation Metric for Terminology Consistency in MT," WMT 2022; 2022:450, doi: 10.18653/v1/2022.wmt-1.41.
[30] K. Vera, K. Svetlana, and R. Oksana, "Methodology and technique for assessing the terminology consistency in translations: Based on Latvian and Russian economic texts," Bectnik Rocciyckogo yniverciteta dryzhby narodov. Ceriya: Teoriya yazyka. Cemiotika. Cemantika, vol. 14, no. 1, pp. 9-35, 2023, doi: 10.22363/2313-2299-2023-14-1-9-35.
[31] A. Anastasopoulos, L. Besacier, J. Cross, M. Gallé, P. Koehn, and V. Nikoulina, "On the evaluation of machine translation for terminology consistency," preprint arXiv:2106.11891, 2021, doi: 10.48550/arXiv.2106.11891.
[32] C. S. Devi, B. S. Purkayastha, and L. S. Meetei, "An empirical study on english-mizo statistical machine translation with bible corpus," International journal of electrical and computer engineering systems, vol. 13, no. 9, pp. 759-765, 2022, doi: 10.32985/ijeces.13.9.4.
[33] Z. Kozhirbayev, "Enhancing neural machine translation with fine-tuned mBART50 Pre-Trained model: An examination with low-resource translation pairs," Ingenierie des Systemes d'Information, vol. 29, no. 3, pp. 831, 2024, doi: 10.18280/isi.290304.
[34] T. Gupta, M. Zaki, N. M. A. Krishnan, and Mausam, "MatSciBERT: A materials domain language model for text mining and information extraction," npj Computational Materials, vol. 8, no. 1, pp. 102, 2022, doi: 10.1038/s41524-022-00784-w.
[35] S. M. Akpaca, "A Practical Approach to Terminological Research in Specialized Translation," Journal of Ecohumanism, vol. 3, no. 6, pp. 86-97, 2024, doi: 10.62754/joe.v3i6.3980.