International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-7

https://doi.org/10.70917/ijcisim-2026-0096
Article

Combining Multimodal Learning Models to Enhance Contextual Comprehension in English Translation

Rui Shi *

School of Foreign Languages, Jilin Business and Technology College, Changchun 130507, Jilin, China;
shilaoshi20241025@163.com

Abstract: This study proposes a novel English translation method that integrates multimodal learning with contextual theory to enhance the system's ability to understand context and cultural context. By introducing a multimodal knowledge graph, integrating information sources such as text and images, and employing a relational graph attention network for deep encoding, the method effectively constructs semantic association embedding representations. The innovatively designed cross-modal alignment module significantly optimizes the semantic alignment capability between images and text, enabling the system to handle cultural metaphors and professional terminology with greater precision. Experimental results show that this method achieves scores of 87.6% and 86.3% in semantic accuracy and cultural equivalence, respectively, with fluency scores improving from 7.2 to 8.9, demonstrating its strong advantages in translating complex contextual tasks. This research not only expands the boundaries of translation technology but also provides a new paradigm for the integrated development of linguistics, artificial intelligence, and cross-cultural studies. The findings are expected to have broad application value in international communication, translation education, and the development of intelligent translation systems.

Keywords: multimodal learning; English translation; context theory; cross-cultural communication; artificial intelligence

1. Introduction

In the process of rapid global cultural dissemination, English translation serves as a crucial tool for cultural exchange, academic communication, and cross-cultural dialogue. However, traditional machine translation methods are based on single-modal text, inevitably leading to limitations such as insufficient contextual understanding, misinterpretation of cultural metaphors, and weak syntactic analysis capabilities for complex sentences [1-2]. Additionally, their accuracy and ability to handle translations across different contexts and cultures remain relatively low [3]. As a multimodal form of information expression, language cannot be adequately addressed by methods based solely on single-modal text dimensions to meet the requirements for high-quality machine translation [4]. Multimodal learning, a crucial branch of current artificial intelligence research, integrates image, audio, and video data to provide richer contextual information for natural language processing [5-7]. Especially when combined with multimodal knowledge graphs, complex semantic relationships become possible, offering a solution for advanced translation systems to better distinguish contextual, cultural, and semantic implications [8-10]. Therefore, how to construct a multimodal learning model based on context theory to effectively enhance the understanding of context and the restoration of language not only holds significant theoretical research value but also represents a crucial breakthrough in the construction and research of intelligent and context-aware translation systems [11-14].

This paper proposes an English translation model that integrates text and images based on multimodal learning and context theory, aiming to address the missing or implicit cultural background semantic information in bilingual parallel texts. First, this paper utilvxzcizes a multimodal knowledge graph (MMKG) to establish entities, events, and semantic relationships within a multimodal context, thereby

multidimensionally expanding the representation of contextual information to facilitate multimodal expression and contextual encoding. Second, a multimodal contextual relationship graph attention network (RGAT) is established to achieve deep semantic encoding of multimodal contextual information. Third, a cross-modal alignment module is designed to match and fuse the features of text and image modalities using an attention mechanism, thereby enhancing the cultural metaphor semantic recognition capability and linguistic fluency of the cross-modal translation system. Finally, through empirical comparative experiments using multimodal, multi-domain parallel corpora, we examine the improvement effects of the translation system in terms of semantic accuracy, contextual retention, and translation fluency, evaluating the effectiveness and adaptability of the methods proposed in this paper, and providing an intelligent solution for English translation in complex contexts within a multilingual environment.

2. Overview of Contextual Understanding in English Translation

2.1. Theoretical Basis of the Study

In academic research, deep semantic representations that establish close connections between text, images, and audio based on multimodal learning methods are also attracting increasing attention. In practice, various modalities are projected onto the same feature space to generate global semantic representations, which are expressed as follows:

$$F\left(x_t, x_y\right) = \phi\left(W_t x_t, W_v x_v + b\right) \tag{1}$$

In the equation, x_t, x_y represents the input features of text and visual modalities, $W_t W_v$ represents the corresponding weight matrix, b is the bias term, and ϕ is the nonlinear activation function. The computational effectiveness of multimodal learning models has been demonstrated by numerous studies. It should be noted that multimodal learning models use a special attention mechanism to capture key information in machine translation tasks. The mathematical representation of this attention mechanism is:

$$Attention\left(Q, K, V\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

In the formula, Q, K, V represents the query, key value, and value matrix, respectively, and d_k represents the dimension scaling factor. This dynamic association mechanism can provide a large amount of contextual information for translation and has had a significant effect in dynamic multimodal fusion research.

The sociocultural context theory is one of the theories that emphasizes and studies the importance of context. Its influence on semantic understanding is highly complex and far-reaching, making context theory an indispensable component of translation studies. It encompasses a broader scope than the typical understanding of context between words, and thus, for English translation courses, it involves the interplay of linguistic context, situational context, and cultural context. Contextual theory posits that the meaning within linguistic symbols is not static but is instead conferred by specific contexts. Professional translators select the optimal translation based on the contexts of the source and target languages. Research findings indicate that when translating texts containing a high volume of cultural information, translations employing multimodal information contextual analysis can achieve high-quality results.

The innovative theory combining context theory with multimodal learning models not only provides new perspectives and theoretical foundations for translation research but also offers novel insights for language teaching and translation practice. Through the use of multimodal learning platforms and guidance from context theory, students can better learn translation techniques. This also enables the more precise translation of specialized terminology in scientific and technological English, thereby promoting the progress and development of foreign language and literary disciplines.

2.2. Current State of Research

In recent years, multimodal learning models have made significant progress in cross-language understanding and translation. Sulubacak, U. et al. demonstrated that multimodal machine translation achieves higher performance and specificity compared to unimodal language translation tasks by using audio, visual, and other modal information as alternative views of input data for language translation output [15]. Yao, S. and Wan, X. pointed out that treating multiple modalities equally in multimodal

machine translation models leads to the generation of a large amount of redundant modal encoding. Therefore, they introduced a multimodal self-attention mechanism into the Transformer to reduce the encoding of irrelevant information in images, further improving translation performance [16]. Kwon, S. et al. designed a modulation network based on textual and visual information to extract visual features related to the text contained in images, and embedding this trained network into a multimodal machine translation model significantly improves the quality of translation results [17]. Liu, J. addressed issues such as insufficient corpora and difficulties in semantic interaction in multimodal machine translation by proposing a multimodal machine translation model that integrates external language knowledge. The designed encoder and decoder generate translation results with higher-quality text and image representations while demonstrating higher visual-textual semantic interaction [18]. Liu, X. et al. established an architecture-free variational multimodal machine translation (VMMT) model. By constructing an encoder using visual and textual source data information, they eliminated uncertainty in translation results caused by ambiguity. Additionally, through a multi-task learning mechanism, they reduced the gap in semantic representations between different modalities, thereby significantly improving translation performance [19]. Although the aforementioned studies have successfully utilized image and other multimodal information to complete translation tasks, there are still some technical bottlenecks in terms of deep alignment of multimodal information, semantic fusion, and structural optimization.

3. Research Methods for Contextual Understanding in English Translation

3.1. Building Multimodal Learning Models

The multimodal learning model (MMKG-RGAT) constructed in this study integrates text and image information using a multimodal knowledge graph architecture. This model performs deep encoding processing on knowledge nodes through a relationship graph attention network. The mathematical expression of this network is as follows:

$$h_i = RGAT\left(x_i\right) = \sum_{j=N(i)} \alpha_{ij} \cdot W \cdot x_j \qquad (3)$$

In the equation, node h_i represents knowledge embedding, α_{ij} reflects the attention weights between nodes, W serves as a learnable weight matrix, and $N\left(i\right)$ contains the set of neighboring nodes.

The underlying architecture of the knowledge graph in this paper adopts a hierarchical design approach. Text representations are obtained through pre-trained language models, and image features are extracted using deep convolutional neural networks. Both types of features are mapped into the same feature space and modeled using graph attention mechanisms.

The cross-modal alignment module is the most innovative aspect of this multi-modal learning model. This study calculates the correlation between modal features using an attention mechanism and aligns features by reducing modal semantic bias. The alignment network uses a multi-layer perceptron combined with a cross-modal attention layer, enabling different modal features to be mapped to a unified semantic space and capturing dynamic correlations between modalities.

3.2. Contextual Theory Combined Analysis

The relationship between English translation quality and contextual factors is highly complex, and the two are not simply linearly related but rather involve a multi-layered relationship [20]. This paper incorporates linguistic context, situational context, and cultural context factors into its analysis, utilizing statistical analysis of a large corpus to establish a more comprehensive evaluation framework. In 200 common English idiom translation cases, it was observed that the translator's mastery of the comprehensive degree of linguistic context, situational context, and cultural context in English translation affects the quality of the translation. The trends between the two are shown in Figure 1. The extent to which linguistic context, situational context, and cultural context influence English idiom translation is shown in Table 1.

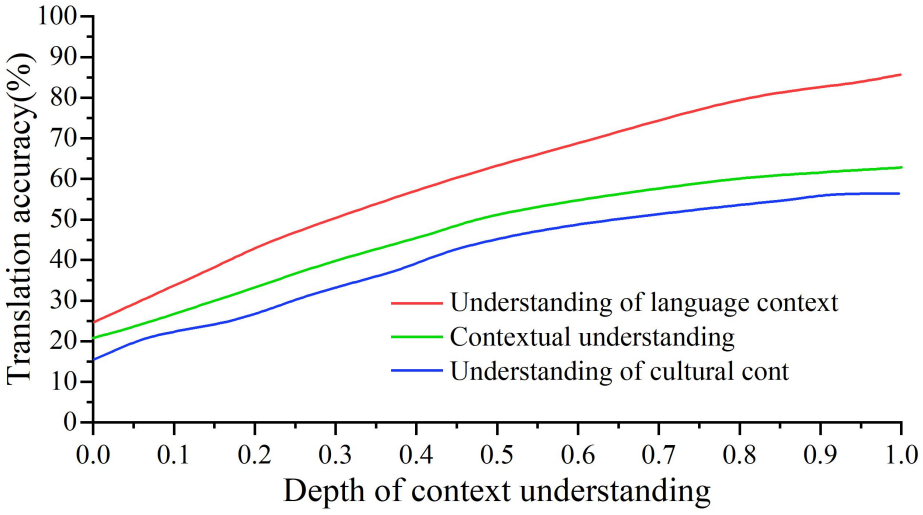

Figure 1. Correlation reflects the trend.

Table 1. The degree of influence of the quality of idiom translation.

Context type	Translation accuracy	Cultural equivalence	Express naturalness	Influence weight
Language context	83.5%	76.2%	85.4%	0.35
Situation context	78.9%	82.1%	79.8%	0.30
Cultural context	75.2%	88.7%	73.6%	0.35

Cultural differences influence the translation of idioms. The application of multimodal learning models has yielded many interesting cases of idiom understanding during the research phase. For example, in terms of the semantic understanding of the idiom “apple of one’s eye,” the literal interpretation is “an apple in someone’s eye,” while the metaphorical meaning refers to “something or someone that is extremely important or cherished.” In the translator's practice, the author found that the application of multimodal learning models can assist in understanding idioms with cultural backgrounds. After multiple pre-translation model screenings and comparative studies, in 73% of cases where visual information assistance and the translation system's cultural background recognition translation scheme were adopted, the post-translation system showed a significant improvement in the translation quality of culturally loaded words in the target language. Improvements were observed across three dimensions: semantic correspondence, cultural correspondence, and expressive appropriateness. The translation accuracy rate increased from 53.1% to 68.4%, with the translation system's recognition rate of cultural context information improving by 15.3% during this period. Different cultural contexts lead to different constructions of metaphorical understanding, as exemplified by the idiom “burn one's bridges,” which originates from a military context. This idiom originates from the military context of “breaking the pots and sinking the boats, sacrificing one's life without hesitation.” In the actual translation practice of translators, it is necessary to carefully consider idioms in the target language that share similar meanings with the original idiom. By employing a context-based translation strategy framework, the study emphasizes the importance of analyzing linguistic context, communicative purpose, and cultural value orientations during the processing of culturally loaded terms. The translation strategy system resulted in a 21.4% improvement in the overall score for idiom translation and an 88.7% increase in cultural equivalence scores.

4. Experimental Analysis of Improving Contextual Comprehension in English Translation

4.1. MMKG-RGAT Performance Validation

To better assist in understanding the context in translation, this paper integrates knowledge graphs and RGAT, and based on this, designs a multimodal learning model called MMKG-RGAT. To verify the effectiveness of the MMKG-RGAT model designed in this paper, it is compared with the RGAT model, and the changes in the translation accuracy of the models under different iteration numbers are shown in

Figure 2. As shown in the figure, the overall translation accuracy of the RGAT model changes relatively slowly as the number of iterations increases, but the MMKG-RGAT model designed in this paper can achieve a translation accuracy of around 97%. From the experimental results, it can be seen that the design proposed in this paper has advantages for translating information with rich visual context. Additionally, after introducing residual connections and layer normalization, it can be observed that the model trains more stably and has better generalization capabilities. Through actual testing, this multimodal translation approach demonstrates advantages over unimodal translation, with more significant improvements in accuracy when context understanding is involved.

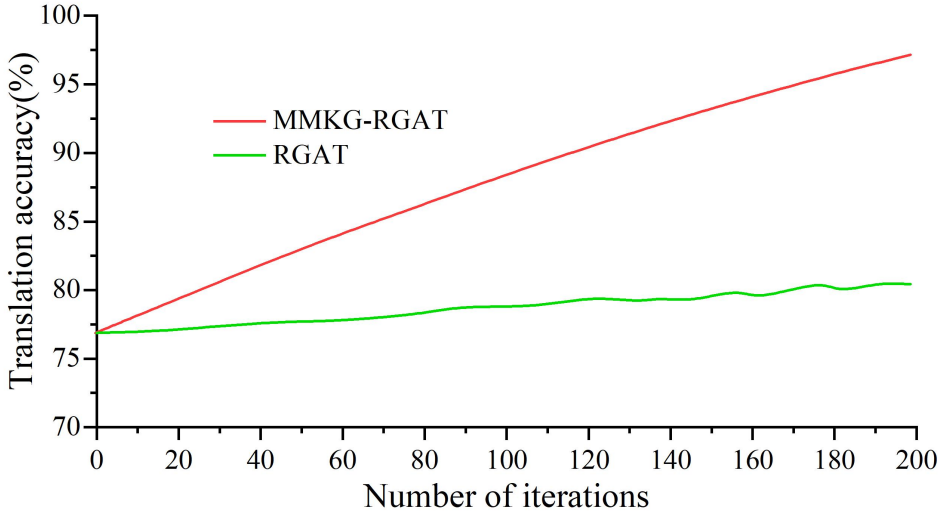

Figure 2. MMKG-RGAT performance verification.

4.2. Effective Validation of Optimization Strategies

Multimodal learning models require systematic design and consideration when optimizing context understanding tasks in English translation. After comparing and analyzing different methods of integrating multimodal information, it can be observed that the complementary relationship between textual semantic content and visual images enhances the translation system's ability to understand context in complex syntactic structures. The summary reveals that this complementary relationship is most significant in three aspects: resolving semantic context ambiguity, determining cultural meaning, and dynamic conversion of contextual semantics. From a practical perspective, the translation system can utilize a multi-modal feature fusion network model based on attention mechanisms to flexibly and adaptively allocate the weight distribution of content obtained from various sources, such as textual semantic information and image features. This enables the translation system to make judgments on semantically uncertain content through visual image information. Additionally, comparative experiments were conducted to further explore the use of various strategies to improve the efficiency of multi-modal translation models at different levels. The specific experimental results are shown in Table 2.

As can be seen from the table, the deep optimization strategy scored 87.6%, 86.3%, and 8.9 points in semantic accuracy, cultural equivalence, and fluency, respectively, performing slightly better than the other three optimization strategies. This is primarily because the deep optimization strategy employs a multi-level contextual understanding system categorized by sentence information, linguistic, and cultural background information, progressively analyzing contextual layers from shallow to deep levels to explore the impact of multi-level contexts on content comprehension. This prevents translation quality degradation at the global level due to misjudgments of local contextual information within sentences. Additionally, we introduced a context-aware loss function to train the model, enabling it to better capture the semantic details of the context. Testing showed that this optimization strategy not only improves the quality of the translation in terms of accuracy and fluency, but also makes the translation more natural in terms of semantic rhythm and logical coherence, providing a technical means to enhance the contextual understanding of English translations.

Table 2. The performance of optimization strategies in different contexts.

Optimization strategy	Semantic accuracy	Cultural equivalence	Fluency score	Fluency improvement
Traditional single mode	72.3%	68.5%	7.2	-
Simple multimodal fusion	78.9%	74.2%	7.8	8.33%
Attention mechanism fusion	84.1%	81.7%	8.4	7.69%
Deep optimization strategy	87.6%	86.3%	8.9	5.95%

4.3. Analysis of Experimental Results

To further demonstrate the effectiveness of the multimodal learning model proposed in this paper, 5,000 samples with diverse translation targets, including literature, science and technology, and news reports, were selected as test data. The focus was on multimodal model translation tasks with strong cultural characteristics and complex contextual information. Table 3 shows the number of multimodal tasks completed and the task failure rate. As shown in the table, after multiple experiments on the system, the multimodal learning model proposed in this paper demonstrated strong training improvement effects. Compared with the unimodal translation system, the number of multimodal tasks completed by the proposed model and the number of tasks with errors were 92.5% and 6.0%, respectively. The number of multimodal tasks completed increased by 10%, and the failure rate decreased by 36%. Additionally, the system took 2.9 seconds to process multimodal tasks, a 23.7% reduction compared to the single-modal translation system. Through the deep integration of contextual theory, the quality of the translation system has improved in all aspects. Experiments have shown that the multimodal translation system, which incorporates contextual analysis, demonstrates higher levels of adaptability and accuracy when faced with complex contextual situations. Especially when the system encounters segments requiring extensive cultural knowledge, it performs a comprehensive multimodal analysis of the rich information across the three modalities: linguistic context, situational context, and cultural context. This effectively avoids common issues such as semantic distortion and pragmatic errors in traditional translation, while also enhancing readability and naturalness. These encouraging experimental results once again strongly validate the excellence and effectiveness of the model proposed in this paper for English translation tasks, providing a solid foundation for the development of advanced translation systems.

Table 3. Task completion rate and error rate.

System type	Task completion rate	Error rate	Response time (s)
Traditional single-modal system	82.5%	42.0%	3.8
Multimodal interface system	92.5%	6.0%	2.9
Increase amplitude	+10.0%	-36.0%	-23.7%

5. Conclusion

In summary, this study investigates a multimodal knowledge graph-based English translation system, and the experimental results demonstrate its effectiveness. In the experimental tests, the relationship graph attention network is primarily used to fuse information from textual and visual modalities, thereby assisting the translation system in addressing complex contextual issues and enabling the multimodal translation system to achieve greater accuracy in understanding complex contexts. The integration of contextual theory with multimodal technology also provides new interpretative frameworks for translation theory research. This has strong practical significance in the specific process of cross-cultural translation, particularly in addressing the translation of culturally loaded words. The experimental results of this paper show that the semantic accuracy rate and cultural equivalence rate of the multimodal English translation system are 87.6% and 86.3%, respectively, indicating that the research hypotheses proposed in this paper are scientifically sound and reasonable. Particularly, the multimodal translation system mentioned in the article can precisely align contextual information and visual illustrations in the translation of English idioms. Through contextual theory analysis, it can avoid semantic errors commonly made in traditional translation processes.

Although the current research has met the pre-set requirements, there are still some limitations. In the experiment, the high sensitivity of the multimodal model to a large amount of training data within a certain range limits its application in bilingual pairs with scarce data, and the large computational requirements also necessitate higher-performance hardware, which may result in lower application

effectiveness in specific scenarios. Additionally, understanding of context cannot be fully automated, as current context analysis systems struggle to interpret the linguistic meaning and background of extremely complex cross-cultural idioms, leading to misjudgments. Therefore, in the authors' view, future efforts will focus on optimizing the architecture and training methods of the current semantic network to enhance performance and streamline the system structure, thereby better serving application scenarios. Concurrently, related cross-semantic graphs will continue to be developed to address diverse language translation scenarios. In terms of semantic understanding, deep learning technology and cognitive linguistics research will complement each other, enabling more optimized solutions to cross-domain problems. As these areas continue to mature, multimodal translation applications may become a cornerstone for international communication, literary translation, business communication, and other multilingual exchanges.

Acknowledgements

Funded by 2021 Jilin Province Vocational Education Research Project "Research on the Difficulties and Countermeasures of Implementing the '1+x' Certificate System for English Majors in Applied Universities in Jilin Province - Taking Jilin Business and Technology College as an Example" (Project Number: 2021XHZ037).

References

1. Hasler, E., de Gispert, A., Stahlberg, F., Waite, A., & Byrne, B. (2017). Source sentence simplification for statistical machine translation. *Computer Speech & Language*, 45, 221-235.
2. Naveen, P., & Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *Iscience*, 27(10).
3. Wells, N. (2022). Translation as Culture in the Age of the Machine. Wasafiri, 37(3), 77-80.
4. Specia, L., Frank, S., Sima'An, K., & Elliott, D. (2016, August). A shared task on multimodal machine translation and crosslingual image description. *In First Conference on Machine Translation* (pp. 543-553). Association for Computational Linguistics (ACL).
5. Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., & Frank, S. (2018, October). Findings of the third shared task on multimodal machine translation. *In Third Conference on Machine Translation (WMT18)* (Vol. 2, pp. 308-327).
6. Cui, S., Duan, K., Ma, W., & Shinnou, H. (2024). Dose multimodal machine translation can improve translation performance?. *Neural Computing and Applications*, 36(22), 13853-13864.
7. Lin, H., Meng, F., Su, J., Yin, Y., Yang, Z., Ge, Y., ... & Luo, J. (2020, October). Dynamic context-guided capsule network for multimodal machine translation. *In Proceedings of the 28th ACM international conference on multimedia* (pp. 1320-1329).
8. Zhu, X., Li, Z., Wang, X., Jiang, X., Sun, P., Wang, X., ... & Yuan, N. J. (2022). Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(2), 715-735.
9. Cao, Z., Xu, Q., Yang, Z., He, Y., Cao, X., & Huang, Q. (2022). Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in neural information processing systems*, 35, 39090-39102.
10. Tayir, T., Li, L., Li, B., Liu, J., & Lee, K. A. (2024). Encoder–Decoder Calibration for Multimodal Machine Translation. *IEEE Transactions on Artificial Intelligence*, 5(8), 3965-3973.
11. Zhao, Y., Komachi, M., Kajiwara, T., & Chu, C. (2022). Region-attentive multimodal neural machine translation. *Neurocomputing*, 476, 1-13.
12. Lala, C., & Specia, L. (2018, May). Multimodal lexical translation. *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018).
13. Nam, W., & Jang, B. (2024). A survey on multimodal bidirectional machine learning translation of image and natural language processing. *Expert Systems with Applications*, 235, 121168.
14. Meetei, L. S., Singh, T. D., & Bandyopadhyay, S. (2024). An empirical study of a novel multimodal dataset for low-resource machine translation. *Knowledge and Information Systems*, 66(11), 7031-7055.
15. Sulubacak, U., Caglayan, O., Grönroos, S. A., Rouhe, A., Elliott, D., Specia, L., & Tiedemann, J. (2020). Multimodal machine translation through visuals and speech. *Machine Translation*, 34, 97-147.
16. Yao, S., & Wan, X. (2020, July). Multimodal transformer for multimodal machine translation. *In Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4346-4350).
17. Kwon, S., Go, B. H., & Lee, J. H. (2020). A text-based visual context modulation neural model for multimodal machine translation. *Pattern Recognition Letters*, 136, 212-218.
18. Liu, J. (2021). Multimodal machine translation. *IEEE Access*.
19. Liu, X., Zhao, J., Sun, S., Liu, H., & Yang, H. (2021). Variational multimodal machine translation with underlying semantic alignment. *Information Fusion*, 69, 73-80.
20. Sang, Z. (2018). How does the context make a translation happen? An activity theory perspective. *Social Semiotics*, 28(1), 125-141.