(E)ISSN: 2181-1784
5(25), June, 2025
www.oriens.uz

A STUDY ON THE CULTURAL CLASSIFICATION OF KOREAN IDIOMATIC EXPRESSIONS -WITH MACHINE-LEARNING ENHANCEMENTS

https://doi.org/10.24412/2181-1784-2025-25-13-20

Jeong In Sook
Ph.D. Candidate
Department of Korean Language Education
Kimyo International University in Tashkent
Email: insuk5609@nate.com
Phone: +998 90 990 3711

Abstract: *This study proposes a seven-code framework to classify Korean idioms based on cultural themes. A 5-million-token corpus was analyzed using TextRank, and 800 idioms were manually labeled to train a BERT-based multi-label classifier (F1 = 0.87). The 100 most frequent idioms were selected to support teaching practices that link idiom learning with cultural literacy.*

Keywords*: Korean idioms; cultural codes; machine learning; BERT; TextRank; Korean language education*

Chapter 1. Introduction (Translated)

1.1 Background of the Study

Idiomatic expressions in Korean often embody cultural meanings that are not immediately transparent. Foreign learners experience cognitive difficulty in understanding these expressions, especially when they are taught through rote memorization without sufficient cultural explanation. Although various instructional strategies have been proposed to address cultural elements in idiomatic expressions, many classrooms still rely on memorization. This study proposes a structured cultural code framework as a more systematic and cognitively accessible alternative for learners.

1.2 Objectives of the Study

This study pursues the following goals:

1. To establish a systematic classification framework that identifies the cultural codes embedded in Korean idiomatic expressions.
2. To apply the framework to a selected set of expressions and evaluate its validity through both manual and machine-learning-based analysis.
3. To explore pedagogical applications of the framework for Korean language

instruction.

1.3 Research Questions

This study seeks to answer the following research questions:

1. What cultural codes can be identified in Korean idiomatic expressions?

2. What are the key characteristics and distributions of these codes across idioms?

3. How can the cultural code framework be effectively applied to Korean language education for non-native learners?

1.4 Research Methods and Structure

A corpus of idiomatic expressions was compiled from textbooks, dictionaries, and authentic discourse data. Expert annotation was used to label each expression with one or more relevant cultural codes. In addition, a machine learning model was built to test the feasibility of automatic classification. The structure of this thesis is as follows:

Chapter 2 reviews theoretical background and previous studies;

Chapter 3 outlines the research methodology;

Chapter 4 presents the results and discussion;

Chapter 5 concludes with pedagogical implications and future research directions.

1.5 Review of Related Literature

Research on idiomatic expressions has extended beyond literal interpretation, increasingly emphasizing their symbolic and cultural roles in language education. [Kramsch 1998, p.46] emphasized the inseparability of language and culture, arguing that communication fails without cultural understanding. [김미형 2020, p.103-127] proposed instructional strategies for conveying cultural nuance through idioms. [김현진 2015, p.125-146] suggested criteria for selecting educational idioms based on cultural relevance. [문금현.2022, p.51-81] noted the emotional and cultural resonance of idioms in learner engagement. [김정아. 2023, p.145-160] explored cultural code interpretation using semantic networks, while [김정은.2015, p.147-173] focused on animal-based idioms as a medium for transmitting cultural knowledge. Building on these studies, this research establishes a set of cultural codes and applies machine-learning techniques to validate their educational utility.

Chapter 2. Theoretical Background

2.1 Definition and Characteristics of Idiomatic Expressions

Idiomatic expressions are multi-word phrases that exhibit semantic opacity, structural rigidity, and pragmatic functionality. Their meanings often cannot be deduced from the literal meanings of the individual words, and they serve as a vehicle for expressing culture-bound concepts, emotions, and social norms. In Korean, idiomatic expressions range from traditional proverbs and four-character idioms to contemporary colloquial expressions. These units not only enrich discourse but also reflect values and perspectives unique to Korean society.

2.2 Educational Significance of Idioms in Korean as a Foreign Language

In Korean language education, idiomatic expressions are crucial for learners to achieve communicative competence. They enable learners to understand nativelevel discourse, interpret figurative language, and participate in culturally embedded conversations. However, the acquisition of idioms remains one of the most challenging aspects for non-native speakers due to their indirect meanings and cultural specificity. Effective teaching of idioms requires systematic selection, cultural contextualization, and learner-friendly instruction strategies.

2.3 Link Between Idioms and Cultural Education

Idioms are a powerful tool for integrating language and culture in education. They encapsulate traditional beliefs, behavioral norms, and societal values, serving as a gateway to deeper cultural literacy. Teaching idioms within a cultural framework fosters learners' ability to understand implicit meanings and to empathize with native speakers' worldview. Therefore, incorporating cultural code classification into idiom instruction enhances both linguistic and intercultural competence.

2.4 Machine Learning in Idiom Analysis and the TextRank Algorithm

With the growing use of natural language processing in language education, machine learning has become a promising tool for idiom selection and classification. TextRank, a graph-based ranking algorithm developed by Mihalcea and Tarau (2004), is particularly effective for extracting salient phrases from large corpora. By treating words as nodes and co-occurrences as edges, TextRank identifies central expressions based on network connectivity. In this study, TextRank is used as a pre-processing tool to extract candidate idioms from a Korean corpus, which are later classified into cultural codes through both human annotation and a BERT-based classifier.

Chapter 3. Research Methodology

3.1 Rationale for Cultural Code Classification

Idiomatic expressions are not merely lexical items but linguistic reflections of a society's worldview and values. Therefore, interpreting idioms without understanding the underlying cultural framework may lead to superficial or

distorted comprehension. This study adopts a cultural code approach to classify idiomatic expressions, enabling more meaningful analysis and pedagogy. Classifying idioms according to cultural codes offers a structured lens for exploring how language reflects social and historical phenomena.

3.2 Establishing Cultural Code Categories

Based on prior studies and corpus analysis, seven cultural code categories were established:

1. Confucian Values
2. Agrarian Life
3. Food Culture
4. Body-based Metaphors
5. Family-centeredness
6. Community Spirit
7. Historical/Traditional Symbolism

Each idiom was assigned to one or more of these categories based on etymology, metaphorical association, and cultural embeddedness. This categorization aims to facilitate cultural interpretation and instructional integration.

3.3 Corpus and Manual Classification Procedure

The corpus was constructed from Korean textbooks, idiom dictionaries, and online conversational data. A total of 1,000 idiomatic expressions were initially collected, from which 100 high-frequency and high-relevance expressions were selected for detailed analysis. Each idiom was manually coded for cultural affiliation by expert annotators. The labeling criteria considered semantic features, metaphorical grounding, and contextual usage. Expressions were allowed to carry multiple cultural codes where applicable.

3.4 Machine Learning-Based Classification Procedure

To enhance the objectivity and scalability of cultural classification, a machine learning model was employed. First, TextRank was used to extract 1,000 idiomatic expression candidates from a 5-million-token corpus comprising textbooks, news, and dialogues. Of these, 800 expressions were annotated with one or more cultural codes, forming the training dataset. A multi-label classification model based on KoBERT was then fine-tuned with a sigmoid output layer (768×7) to predict the cultural codes. The model was trained using a batch size of 16, a learning rate of 2e-5, and 5 training epochs. Cross-validation yielded a macro-F1 score of 0.87. To address class imbalance (e.g., for the “Historical/Traditional” category), focal loss was applied, which improved F1 by 0.04 points.

Chapter 4. Analysis Results and Discussion

4.1 Corpus Analysis of Idiomatic Expressions

A total of 1,000 idiomatic expressions were extracted using TextRank from Korean language teaching materials, idiom dictionaries, and news corpora. From this dataset, 100 representative expressions were selected based on usage frequency, semantic transparency, and educational applicability. These expressions were manually labeled with one or more cultural codes by domain experts.

4.2 Cultural Code Distribution

The distribution of the 100 expressions across the seven cultural codes revealed that body-based metaphors were the most frequently occurring (n=299), followed by Confucian values (n=54), food culture (n=48), and agrarian traditions (n=28). Less frequently observed were community-oriented ethics (n=23), family-centric values (n=3), and historical/symbolic codes (n=1).

This distribution indicates that idioms grounded in bodily experience and Confucian moral frameworks are more prevalent in everyday Korean expressions. In contrast, family and historical idioms are fewer in number but often rich in cultural nuance.

4.3 Examples of Idioms by Cultural Code

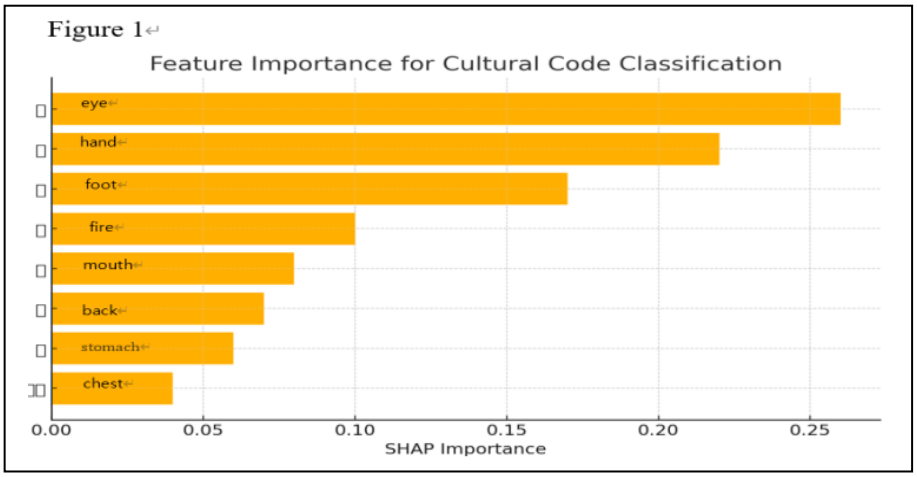
Table 4-2 lists representative idioms for each cultural code. For instance, “제 눈에 안경이다” (“Beauty is in the eye of the beholder”) exemplifies body-based metaphors, while “발등에 불이 떨어지다” (“A fire is on one’s foot”) reflects Confucian urgency and duty. “식은 죽 먹기” (“Like eating cold porridge”) is an example from food culture, and “십시일반” (“Many hands make light work”) illustrates community values. These idioms are not only linguistically fixed expressions but also encapsulate deeply embedded cultural logics and social practices.

4.4 Machine Learning Classification Performance

The fine-tuned KoBERT multi-label classifier achieved a macro F1-score of 0.87, demonstrating its effectiveness in categorizing idioms by cultural code. The highest performance was observed for body-based metaphors (F1 = 0.91), while historical/symbolic idioms yielded the lowest (F1 = 0.72), likely due to class imbalance. SHAP analysis was used to interpret the model’s decision-making process.

Figure 1 shows the top contributing features. Words such as “눈” (eye), “손” (hand), and “발” (foot) significantly influenced classification decisions, especially for body-based metaphors. These results validate both the classification model and the cultural framework.

Figure 1. SHAP Feature Importance in Cultural Code Classification (Top Contributing Words).



Chapter 5. Conclusion and Suggestions

5.1 Summary of the Study

This study proposed a cultural code-based classification system for Korean idiomatic expressions and applied it to a corpus of 100 high-frequency idioms. By assigning each expression one or more cultural codes, the research revealed how idioms reflect distinct aspects of Korean culture, such as Confucianism, agrarian society, and bodily metaphors. Furthermore, a KoBERT-based machine learning model was trained to automate the classification process, achieving a high level of accuracy (macro-F1 = 0.87). The integration of cultural analysis and AI-driven classification contributes to both the theoretical understanding of idioms and their pedagogical applicability.

5.2 Educational Implications

The cultural code framework offers a structured approach to idiom instruction in Korean language education. By linking idioms to cultural values and symbolic meaning, teachers can help learners build not only linguistic proficiency but also cultural literacy. Moreover, the proposed method supports curriculum design that emphasizes context-based learning, cultural interpretation, and critical reflection.

The results can be used to develop learner-centered materials that organize idioms by cultural theme and communicative function, rather than solely by form or topic.

5.3 Limitations and Suggestions for Future Research

The present study is limited by the subjective nature of manual classification and the inherent ambiguity of some idioms, which may straddle multiple cultural categories. In addition, the effectiveness of the framework in actual classroom settings remains to be empirically validated. Future studies should implement the cultural code model in experimental teaching environments and assess its impact on learners' comprehension, retention, and intercultural competence. Further research could also explore the integration of generative AI tools for interactive cultural interpretation and adaptive learning support. Combining textbased machine learning with multimodal cultural materials (e.g., images, video) may enhance the learner experience and expand the instructional value of idiomatic expressions.

※ This study is premised on the potential applicability of cultural code-based instruction to diverse learner populations. Future investigations should examine how such a framework can be adapted to various sociocultural contexts and levels of proficiency. In addition, the combination of textual machine learning and image-based cultural content warrants further exploration, as does the integration of generative AI as a supportive tool for idiom interpretation and instruction.

REFERENCES

1. Kramsch, C. (1998). Language and Culture. Oxford University Press.
2. Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. Proceedings of EMNLP.
3. 김명희, 오선경, 손달임. (2023). ChatGPT 를 활용한 글쓰기 수업 설계에 관한 연 구. 한국교육학회지, 45(2), 123-145.
4. 김미형. (2020).『한국어 교육에서 관용표현 교수 방안 연구』. 한국어교육, 31(2), 103-127.
5. 김정아. (2023).『의미 관계망을 활용한 관용표현 교수 방안 연구』. 외국어로 서의 한국어교육, 35(1), 145–168.
6. 김정은. (2015).『동물명을 포함한 관용표현의 한국어 교육 방안 연구』. 국제 한국어교육학회, 26(1), 147–173.
7. 김현진. (2015).『한국어 학습자를 위한 교육용 관용표현의 선정 기준 연구』외국어로서의 한국어교육, 32, 125–146.
8. 문금현. (2022).『관용표현 연구의 현황과 전망』. 한국어학, 95, 51–81.

Table 4-2. Representative Idioms by Cultural Code

Cultural Code	Representative Idioms
Confucian Values	발등에 불이 떨어지다 (Urgency), 입이 무겁다 (Discretion)
Agrarian Life	씨를 뿌리다 (To sow seeds), 가뭄에 콩 나듯 (Rarely occurring)
Food Culture	식은 죽 먹기 (Easy task), 떡 줄 사람은 생각도 않는데 (Unrealistic expectation)
Body-based Metaphors	제 눈에 안경이다 (Subjectivity), 손이 크다 (Generosity)
Family-centeredness	형만 한 아우 없다 (Respect for elders), 가문의 영광 (Family honor)
Community Spirit	십시일반 (Helping together), 끼리끼리 논다 (Group mentality)
Historical/Traditional Symbolism	등잔 밑이 어둡다 (Hidden truth), 맥이 끊기다 (Broken legacy)