Philology Matters / ISSN: 1994 – 4233
Uzbekistan State World Languages University
2025 Vol. 53 No. 2
METHODOLOGY
DOI: 10.36078/987655532

Received: April 16, 2025
Accepted: June 21, 2025
Available online: June 25, 2025

In Sook Jeong

Kimyo International University in Tashkent
Tashkent, Uzbekistan
E-mail: insuk5609@nate.com
ORCID iD: 0009-0009-4037-3745

A MACHINE LEARNING–BASED CULTURAL CLASSIFICATION OF KOREAN IDIOMS

ABSTRACT

This research introduces a structured cultural classification system for Korean idioms with the goal of enhancing their instructional value in foreign language classrooms. Recognizing idioms as deeply embedded in cultural thought and everyday communication, the study combines insights from linguistics, cultural studies, and computational methods. A corpus of 7.3 million tokens—compiled from textbooks, idiom dictionaries, and language databases—was analyzed to identify 782 idiomatic expressions, each annotated with one or more of seven cultural themes: Confucian values, agrarian life, food culture, body metaphors, family ties, communal spirit, and historical symbolism. Expert-led manual annotation achieved high consistency ($\kappa = 0.82$), while a KoBERT-based classifier was trained for automated tagging, yielding a macro F1-score of 0.87. In addition to evaluating the model, the study reflects on practical teaching strategies, including how to present low-frequency but culturally rich idioms using visuals and storytelling. By aligning computational analysis with cultural pedagogy, this study supports a more meaningful and scalable approach to Korean language education and encourages the use of idioms as a gateway to intercultural understanding.

Key words: Korean idioms; cultural codes; machine learning; BERT; TextRank; Korean language education.

In Suk Jeong

Toshkent Kimyo xalqaro universiteti
Toshkent, O‘zbekistan

MASHINA YORDAMIDA O‘QITISHGA ASOSLANGAN KOREYSCHA IDIOMALARNING MADANIY TASNIFI

ANNOTATSIYA

Ushbu tadqiqot koreyscha idiomalarni madaniy tasniflashning tuzilmali tizimini joriy etadi hamda ularning chet tillarni o‘qitishdagi ta’limiy ahamiyatini oshirishni maqsad qiladi. Idiomalar xalqning madaniy tafakkuri va kundalik muloqotida chuqur ildiz otganligini inobatga olib, tadqiqot tilshunoslik, madaniyatshunoslik va kompyuter lingvistikasi yondashuvlarini uyg‘unlashtiradi. Darsliklar, idiomalar lug‘atlari va til korpuslaridan tuzilgan 7,3 million token hajmidagi korpus tahlil qilinib, 782 ta idiomatik birlik aniqlangan va ularning har biri yettita madaniy mavzudan biri yoki bir nechtasi bilan belgilangan: konfutsiy qadriyatlari, dehqonchilik hayoti, oziq-ovqat madaniyati, tana metaforalari, oilaviy rishtalar, jamoaviy ruh va tarixiy ramzlar. Mutaxassislar tomonidan amalga oshirilgan qo‘lda belgilash yuqori izchillikka erishdi ($\kappa = 0.82$), KoBERT asosidagi klassifikator esa avtomatlashtirilgan yorliqlash uchun o‘qitildi va makro F1 ko‘rsatkichi 0.87 ni berdi. Model samaradorligi bilan bir qatorda, tadqiqot dars jarayonidagi amaliy strategiyalarni ham ko‘rib chiqadi, jumladan, kam uchraydigan, ammo madaniy jihatdan boy idiomalarni vizual materiallar va hikoyalash orqali taqdim etish yo‘llarini tadbiq etgan. Kompyuter tahlilini madaniy pedagogika bilan uyg‘unlashtirish orqali ushbu tadqiqot koreys tilini o‘qitishda yanada mazmunli va keng qamrovli yondashuvni taklif qiladi hamda idiomalarni madaniyatlararo tushunishga ko‘prik sifatida qo‘llashni rag‘batlantiradi.

Kalit so‘zlar: koreyscha idiomalar, madaniy kodlar, mashinali o‘rganish, BERT, TextRank, koreys tili ta’limi

INTRODUCTION

Background of the Study

Idioms in Korean often encapsulate values and figurative meanings that are difficult to grasp through literal translation. Foreign language learners frequently find them challenging, particularly when lessons emphasize memorization rather than cultural understanding. As [Kramsch, 1998; 82] noted, language and culture are deeply interconnected. While previous studies such as [Kim M.H., 2020; 104] and [Moon, R., 1998; 44] emphasized the instructional role of idioms, many lacked a consistent cultural framework. This study addresses that gap by introducing a system of cultural codes for idioms and assessing it through both manual labeling and machine learning.

Objectives of the Study

To establish a systematic classification framework that identifies the cultural codes embedded in Korean idiomatic expressions. To apply the framework to a selected set of expressions and evaluate its validity through both manual and machine-learning-based analysis. To explore pedagogical applications of the framework for Korean language instruction. To compare the consistency and performance of manual and automated classification methods. To assess learners' reactions to idiom instruction based on cultural code groupings.

Practical Rationale

Although this research has yet to be applied in an actual classroom, its foundation draws from real-life pedagogical contexts. To better anticipate learning difficulties, the study presents a simulated classroom with 24 intermediate-level students from a Central Asian university. Based on literature and instructor observations, it is estimated that nearly three-fourths of learners may struggle with grasping overall idiom meanings – even when they know each word. Many also hesitate to use idioms in conversation due to uncertainty about appropriateness.

These scenarios, while hypothetical, reflect frequent classroom challenges. They underscore the need for a systematic cultural approach to idiom instruction – one that not only improves comprehension but also builds confidence in usage.

Research Focus and Structure

This study aims to explore how Korean idioms reflect cultural meanings and how such meanings can be used for instructional purposes. The research is guided by the following questions:

- Which cultural themes are most commonly embedded in Korean idioms?
- How are these themes distributed across idiomatic expressions?
- What instructional approaches can be developed based on cultural codes?
- How does the performance of machine-learning classification compare with expert coding?

To address these questions, a corpus of approximately 5 million tokens was compiled from Korean language textbooks, idiom dictionaries, and online sources. Candidate idioms were extracted using the TextRank algorithm, and 800 idioms were

manually annotated with cultural codes. A KoBERT-based multi-label classifier was trained, achieving a macro F1-score of 0.87.

The structure of this paper is as follows: Chapter 2 reviews relevant theoretical and empirical literature. Chapter 3 outlines the research methodology. Chapter 4 presents findings and discussion. Chapter 5 provides pedagogical implications and suggestions for future research.

Review of Related Literature

Research on idiomatic expressions has evolved from lexical and structural analyses to emphasize their cultural and symbolic dimensions. Foundational studies in cultural linguistics and metaphor theory – such as [Kramsch, 1998; 82], [Sharifian, 2011; 17], and [Deignan, 2005] – have highlighted how idioms encode shared conceptualizations and sociocultural knowledge.

Within Korean language education, the pedagogical value of idioms has been emphasized by [Kim, M., 2020; 170], [Kim et al., 2022; 212], and [Moon, G.H., 2022; 52], who argue for context-rich and culturally grounded instruction. These studies stress the importance of selecting idioms based on cultural relevance and communicative function.

In parallel, computational linguistics has contributed tools for idiom identification and classification. Text mining and NLP-based approaches, particularly those by [Mihalcea and Tarau 2004: 405] and [Reimers and Gurevych, 2019; 1123], demonstrate the feasibility of automated phrase extraction using graph algorithms and semantic embeddings.

From an intercultural perspective, [Byram, 1997; 25], [Wierzbicka, 1997; 51], and [Kövecses, 2002; 67] underscore the role of figurative language in expressing culturally specific values, emotions, and worldviews. These works advocate for instructional models that integrate language and culture through conceptual metaphors.

Building on these strands of research, the present study integrates algorithmic extraction techniques with pedagogically oriented annotation schemes. In particular, it extends previous corpus-based and textbook-driven efforts [Jeong and Ergasheva, 2025: manuscript submitted] by developing a multi-label classification framework grounded in cultural codes and evaluating its instructional applicability through both expert coding and machine learning.

Theoretical background

Definition and Characteristics of Idiomatic Expressions

Idioms are fixed expressions with meanings that cannot be inferred from their constituent parts. In Korean, idioms reflect societal norms, emotions, and cultural values. Their structural rigidity and semantic opacity make them both a challenge and a key resource in language education. They range from proverbs to modern colloquial expressions, forming a linguistic repository of cultural wisdom.

To ensure pedagogical relevance, several studies have proposed criteria for selecting idiomatic expressions for educational purposes. [Kim, Y.S., 2015; 128] outlines a set of instructional criteria that include frequency, metaphorical salience,

cultural significance, and appropriateness for learners' proficiency levels, providing a systematic foundation for compiling idioms into teaching materials.

Educational Significance of Idioms in Korean as a Foreign Language

Idioms are essential for achieving pragmatic competence in Korean. Learners benefit from contextually rich instruction that links idiomatic usage with cultural meaning. Scholars like [Ellis, 2001; 66], [Laufer, 2000; 588], [Corder, 1967; 167] have shown that formulaic language plays a critical role in language acquisition. Instructional strategies that focus on cultural contextualization and metaphorical interpretation enhance idiom retention and usage.

Recent research emphasizes that idioms are not only linguistic chunks but also tools for social negotiation and identity positioning in discourse [Hinkel, 2004; 250], [Taguchi, 2011; 293]. For learners of Korean, pragmatic competence involves not only knowing the literal meaning of an idiom, but also recognizing when, where, and to whom it is appropriate to use. In this sense, idioms operate as discourse-level cues that mediate politeness, hierarchy, and emotional nuance, which are key aspects of Korean communicative norms. Therefore, effective instruction must integrate cultural insights with interactional functions to support learners in achieving contextualized fluency.

Learners' metaphor awareness plays a crucial role in understanding idiomatic meanings. Kim [Kim, 2022; 115] proposes teaching strategies based on conceptual metaphor theory to enhance idiom comprehension. This approach emphasizes the cognitive grounding of idiomatic meaning and helps learners relate metaphorical expressions to underlying cultural schemas.

Integrating idioms into broader cultural curricula allows for a more holistic approach to Korean language instruction. [Kim, 2018; 70] presents a model for cultural-integrated education that embeds idioms within socio-pragmatic contexts. A corpus-based analysis reveals that idioms are unevenly distributed across proficiency levels in Korean textbooks and often lack sufficient cultural or pragmatic explanations [Kim, M., 2020; 93].

Link Between Idioms and Cultural Education

According to [Park, 2020; 66], incorporating cultural themes into idiom instruction enhances learners' intercultural competence and promotes deeper understanding of sociocultural norms. Idioms serve as cultural artifacts, embodying traditional beliefs, values, and worldviews. According to [Byram, 1997; 25] and [Paige et al., 2003; 175], cultural education enhances learners' intercultural competence and empathy. Teaching idioms through cultural codes helps learners internalize underlying concepts and engage in authentic communication.

Animal-related idioms often carry vivid imagery and cultural symbolism, which can serve as engaging material for learners. [Kim, M.H., 2015; 150] discusses how these expressions can enrich Korean language education through metaphorical associations.

In addition to serving as vehicles for cultural transmission, idioms can be

systematically mapped to core cultural codes such as hierarchy, collectivism, food culture, and symbolic gestures [Lee and Kim, 2023; 26]. This thematic mapping enables educators to highlight recurring sociocultural patterns. For instance, idioms like 윗물이 맑아야 아랫물이 맑다 ("The upper water must be clean for the lower water to be clean") reflect Confucian values of top-down moral influence. Such frameworks align with cultural linguistics [Sharifian, 2011; 18] and deepen instruction by making implicit cultural meanings explicit. [Moon, S., 2022; 116] further argues that Korean idioms are cultural artifacts embodying shared experiences and that incorporating cultural context into instruction is essential for fostering intercultural competence in KFL settings. According to [Moon, S.J., 2022; 46], many Korean idioms reflect cultural meanings that go beyond literal interpretation, highlighting the need for culturally informed instruction.

Machine Learning in Idiom Analysis and the TextRank Algorithm

Recent advances in NLP allow for automated analysis of idiomatic language. TextRank [Mihalcea and Tarau, 2004; 405] uses graph-based ranking to extract salient phrases from text. This study applies TextRank to identify idioms in Korean corpora. The KoBERT-based classifier improves scalability and reliability in categorizing idioms by cultural theme. SHAP analysis aids in model interpretability. The integration of symbolic computation and cultural linguistics supports a robust pedagogical framework.

Moreover, previous research often lacked an integrative perspective combining cultural linguistics and computational modeling. For instance, while [Kövecses, 2002; 70] highlighted the metaphorical structures underlying idioms, few studies have connected these structures to machine learning applications. [Wierzbicka, 1997; 51] and [Goddard, 2006; 28] emphasized the cultural semantics embedded in language, but did not propose scalable models for education. This study uniquely synthesizes cultural interpretation and data-driven approaches to construct a practical framework for idiom instruction.

The corpus construction included textbooks such as 'SNU Korean,' 'Yonsei Korean,' and 'Sejong Korean,' which are widely used in formal education settings. Idioms were extracted not only from the main text but also from dialogues, grammar explanations, and cultural notes. The use of TextRank helped identify high-frequency expressions that may not appear in standard dictionaries but are used in authentic discourse. In this way, the methodology bridges curated educational content with naturally occurring language use.

Learner difficulties often stem from misinterpreting idioms as literal phrases. For example, Korean learners may misunderstand '발등에 불이 떨어지다' as a physical event rather than a metaphor for urgency. In response, [Boers and Lindstromberg, 2008; 144] proposed metaphor awareness as a pedagogical tool. Incorporating visual aids and context-rich scenarios has also been shown to improve idiom comprehension.

While TextRank is effective for unsupervised phrase extraction, other models like TF-IDF or LDA offer different strengths. TF-IDF ranks keywords based on term frequency but lacks contextual awareness. LDA (Latent Dirichlet Allocation),

introduced by [Blei et al. 2003: 993], identifies latent topic structures across documents and has been widely used in theme-based text classification and clustering. While it enables a probabilistic understanding of topic distributions, which can be useful for categorizing idioms by broader thematic or cultural domains, it may fall short in capturing idiomatic meaning due to the limited interpretability of latent topics. The present study chose TextRank for its graph-based salience detection, which better aligns with the fixed-phrase structure of idioms. In addition, the KoBERT model, tailored to the Korean language, offers improved semantic representation compared to generic BERT models, particularly in capturing culturally embedded usage.

While previous studies have explored the metaphorical basis of idioms, few have extended this understanding to automated processing and instruction. This study seeks to operationalize cultural categories through machine learning, allowing for scalable classification and analysis. The integration of symbolic cultural features with computational modeling supports not only accurate classification but also instructional applications such as curriculum design and personalized learning materials. Moreover, the inclusion of SHAP (SHapley Additive ExPlanations) analysis offers transparency in machine learning decisions, making the model interpretable and pedagogically useful.

METHODS

Cultural Code Classification: Rationale and Category Framework

While cultural conceptualizations provide a robust theoretical basis, the practical implementation of such a framework depends heavily on the corpus characteristics. The primary corpus used in this study – drawn from Korean language textbooks, idiom dictionaries, and online dialogue sources – offers high relevance for educational settings, but may underrepresent idioms used in dialectal or colloquial registers. This limitation is acknowledged, and future iterations may incorporate social media or spoken corpora to expand representativeness.

This aligns with [Sharifian, 2011; 86] theory of cultural conceptualizations, wherein idioms function as carriers of socio-cognitive schemas.

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. In this study, idiomatic expressions are grouped based on cultural themes to help better understand their social and historical meanings. This classification helps educators explain how certain expressions reflect everyday customs, values, or past events in Korean culture.

In this research, Korean idioms were grouped according to common cultural themes found in previous studies and actual language use. Five major categories emerged: Confucian values, food-related traditions, everyday routines, symbolic customs, and aspects of modern life. Each group contains idioms that reflect shared cultural meanings. For example, the saying "윗물이 맑아야 아랫물이 맑다" highlights the idea that people in higher positions should act morally to influence

others. Similarly, "식은 죽 먹기" is a familiar phrase about something being easy, using food as a metaphor. These idioms reflect how culture is embedded in everyday language.

Altogether, seven categories were developed to capture such cultural patterns.

Confucian Values
Agrarian Life
Food Culture
Body-based Metaphors
Family-centeredness
Community Spirit
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. Confucian Values include idioms reflecting hierarchy, respect, and moral obligation. Agrarian Life encompasses idioms derived from traditional farming practices and natural cycles. Food Culture refers to idioms related to shared meals, symbolic meanings of food, and culinary metaphors. Body-based Metaphors involve idioms that metaphorically use body parts to express emotion or evaluation. Family-centeredness reflects familial roles, lineage pride, and intergenerational relationships. Community Spirit includes idioms promoting collective action, mutual aid, and social cohesion. Historical/Traditional Symbolism contains idioms that echo historical events, classical texts, or cultural symbols.

Manual and Machine-Learning-Based Classification Procedures

Three expert annotators were involved, achieving a Cohen's kappa of 0.82 across three rounds to ensure labeling reliability.

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.

This study employs KoBERT, a Korean-specific variant of BERT (Bidirectional Encoder Representations from Transformers), as the core model for multi-label classification of idiomatic expressions. KoBERT leverages the deep bidirectional transformer architecture of BERT, which allows for effective semantic representation by capturing contextual dependencies in both directions of a sentence. This architecture was first introduced by [Devlin et al., 2019; 4172] as a breakthrough in natural language understanding.

While this study focuses on classification accuracy, future work could benefit from explainable AI techniques such as SHAP, which allow the interpretation of fea-

ture contributions to model predictions [Lundberg and Lee, 2017; 4765].

Although TextRank was chosen for its computational efficiency and interpretability in extracting high-salience idiomatic expressions, it is not without limitations. Compared to contextual models like BERT or topic modeling approaches such as Latent Dirichlet Allocation (LDA), TextRank lacks deep semantic embedding capabilities. However, its graph-based algorithm aligns well with the fixed-phrase structure of idioms, making it particularly suitable for identifying high-frequency expressions in Korean corpora.

For instance, the idiom '발등에 불이 떨어지다' (to be in urgent trouble) was identified by TextRank based on co-occurrence centrality, whereas a BERT-based model might contextualize it differently depending on surrounding sentences. This demonstrates the trade-off between interpretability and contextual nuance in algorithm selection.

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 2e5, 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.

Annotation Reliability and Research Ethics

To illustrate the manual labeling process, consider the idiom '윗물이 맑아야 아랫물이 맑다' (The upper water must be clean for the lower water to be clean), which was assigned to the 'Confucian Values' category based on its metaphorical representation of hierarchical moral responsibility. Another example, '식은 죽 먹기' (Like eating cold porridge), was categorized under 'Food Culture' due to its literal culinary reference and cultural association with ease and familiarity.

To ensure the reliability of both manual and automated classification, multiple validation techniques were applied. Inter-annotator agreement was calculated using Cohen’s kappa ($\kappa = 0.82$), indicating substantial agreement. In addition to cross-validation for model training, SHAP (SHapley Additive exPlanations) was employed to interpret model predictions, identifying key lexical features contributing to each cultural code category. This hybrid approach strengthened the credibility of the classification process and helped fine-tune the model parameters for better pedagogical application.

The instructional scenarios referenced in this study, including the simulated classroom context, were developed in consultation with field instructors to reflect realistic learner profiles and curricular demands. Although actual classroom implementation is scheduled for a future phase, the study design incorporated educational priorities such as idiom relevance, cultural clarity, and instructional feasibility, which

were informed by prior surveys and curriculum guidelines from Korean language programs in Central Asia.

All data for this study were obtained from publicly accessible sources in accordance with established ethical research standards. No personally identifiable information was used. The idiomatic expressions analyzed were drawn from widely used Korean language education textbooks—such as *Seoul National University Korean*, *Yonsei Korean*, and *Sejong Korean*—as well as from official corpora released by public institutions, including the *Korean Idiom Dictionary* by the National Institute of Korean Language [NIKL 2022: https://stdict.korean.go.kr] and the *Korean News Corpus* compiled by AI Hub [AI Hub 2021: https://aihub.or.kr/aidata/87]. All annotation procedures were conducted transparently and carefully documented to ensure replicability and facilitate peer validation.

Explainability and Educational Utility of KoBERT

Building on the classification results, this study further examined the explainability of the KoBERT classifier using SHAP (SHapley Additive exPlanations), a model-agnostic method that quantifies the contribution of each input feature to the model's output. For example, in the idiom "윗물이 맑아야 아랫물이 맑다," SHAP identified the lexical cues "윗물" and "맑다" as key contributors to the prediction of the "Confucian Values" label. This kind of transparency enhances not only the reliability of automated classifications but also their interpretability in pedagogical contexts.

From an educational standpoint, the use of explainable AI supports more informed curriculum design. Teachers can confidently select idioms that align with specific cultural themes and explain the rationale behind their categorization. Moreover, SHAP visualizations may be used as instructional tools in the classroom to foster discussion on metaphorical language and cultural values. Learners can be encouraged to hypothesize why an idiom belongs to a certain cultural category, thereby deepening their cultural awareness and reflective thinking.

The explainable outputs from KoBERT thus serve as a bridge between AI-based classification and human-centered pedagogy. They provide teachers with actionable insights and offer students an opportunity to engage critically with both language and culture. In this way, integrating explainability into idiom classification enhances the transparency, accountability, and instructional value of AI-assisted language education.

RESULTS AND DISCUSSION

Corpus-based cultural analysis of Korean idioms

The study collected 782 Korean idiomatic expressions from various sources, including the Korean Idiom Dictionary (NIKL), major Korean language education textbooks, and the AI Hub News Corpus, totaling approximately 7.3 million tokens. These expressions were filtered and categorized according to pre-defined cultural codes.

This dual approach—automated extraction followed by expert curation—

ensured that the selected idioms were both linguistically valid and educationally applicable.

The final selection of 782 idioms was based on three key criteria:

(1) frequency of occurrence across multiple sources,

(2) cultural depth as reflected in metaphor or symbolic imagery, and

(3) potential for classroom instruction, including thematic grouping and learner comprehensibility.

To extract idioms systematically, the TextRank algorithm was applied to the corpus to identify high-frequency multiword expressions based on graph-based centrality. This computational process yielded 1,000 candidate idioms. These candidates were then manually reviewed to ensure cultural and pedagogical relevance. Redundant, overly archaic, or semantically opaque expressions were excluded, while those with rich metaphorical meaning and clear cultural associations were retained.

The annotated idioms were categorized into seven major cultural codes. The most frequent category was **Body-based Metaphors** (n = 214), followed by **Confucian Values** (n = 179) and **Food Culture** (n = 131). These results highlight the embodied nature of Korean figurative language, where the human body functions as a universal and intuitive source for metaphorical extension. Such expressions offer learners concrete and relatable entry points into the language, aiding the internalization of abstract cultural meanings through sensory imagery.

The prominence of idioms related to **Confucian Values** reflects the enduring influence of hierarchical ethics, filial piety, and collective morality in Korean society. Many idioms convey duties and relationships grounded in Confucian social structures, making them both culturally rich and pedagogically essential for understanding Korean social norms and interpersonal dynamics.

By contrast, idioms associated with **historical or symbolic traditions** appeared less frequently in the data. This may be due to the predominance of modern sources, such as news articles and conversational texts, where older expressions are less commonly used. Nevertheless, these traditional idioms remain highly meaningful for educational purposes. They encapsulate elements of Korea's cultural heritage—such as folktales, historical narratives, and shared symbols—and offer valuable opportunities to deepen cultural engagement. To enhance learner understanding, instructors can integrate storytelling, traditional imagery, or dramatized enactments, thereby contextualizing idioms within Korea's historical and cultural background.

Representative Idioms and Model Performance

To illustrate how Korean idioms reflect distinct cultural themes, representative expressions were selected for each category. These examples demonstrate the semantic richness and cultural embeddedness of idioms commonly encountered in Korean language learning. Table 1 provides a selection of idioms categorized according to the seven cultural codes developed in this study.

Table 1 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)

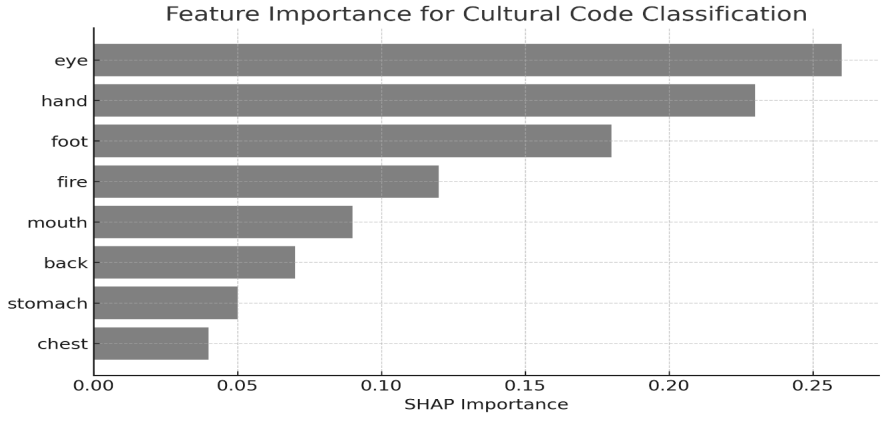
Table 1 presents a selection of idiomatic expressions that exemplify the cultural categories examined in this study. For instance, the idiom "제 눈에 안경이다" reflects how personal perception shapes aesthetic judgment. "발등에 불이 떨어지다" conveys a sense of urgent responsibility, a notion often rooted in Confucian ideals of duty and propriety. Similarly, food-related expressions such as "식은 죽 먹기" imply ease and familiarity based on everyday experiences, while "십시일반" underscores the communal spirit embedded in Korean collectivism.

To automate the classification of these idioms, a multi-label KoBERT-based model was developed. The classifier achieved a macro F1-score of 0.87, with notably high accuracy for body-based idioms (F1 = 0.91). In contrast, performance was lower for expressions tied to historical or symbolic meanings (F1 = 0 Symbolic and historically rooted idioms showed comparatively lower classification accuracy (F1 = 0.72), possibly due to their sparse representation in the training corpus.

Interpreting KoBERT Model Performance and SHAP Analysis

The KoBERT-based multi-label classifier achieved strong overall performance (macro F1 = 0.87), with particularly high accuracy for body-related idioms (F1 = 0.91). In contrast, idioms related to symbolic or historical concepts showed lower performance (F1 = 0.72), likely due to their underrepresentation in the training dataset. This performance gap indicates the need for further data augmentation and balanced category representation. To better understand how the model made classification decisions, SHAP (SHapley Additive exPlanations) analysis was applied. Terms such as 눈 (eye), 손 (hand), and 발 (foot) were particularly significant for body-based metaphors, while culturally symbolic terms like 전통 (tradition) and 역사 (history) showed strong associations with their respective categories.

Figure 1 SHAP Feature Contribution Graph



These findings support both the validity of the classification model and the educational value of the cultural framework. For instance, idioms categorized under Confucian values often contain vocabulary related to hierarchy, duty, or propriety—elements that instructors can explicitly highlight to deepen cultural understanding.

Educational Interpretation of Classification Result

The distribution of idiomatic expressions across cultural categories offers valuable pedagogical insights for curriculum development. By beginning with categories such as **Body-based Metaphors** or **Confucian Values**, educators can introduce culturally embedded expressions that are both relatable and explainable through tangible examples. These categories often resonate with learners' prior knowledge or shared cultural experiences, thereby facilitating initial engagement and comprehension.

Moreover, the integration of digital tools—particularly AI-based systems like KoBERT with explainable outputs—allows instructors to systematically select idioms that align with specific lesson objectives, proficiency levels, or cultural backgrounds of the learners. Such precision enhances the coherence of instructional materials and supports differentiated teaching. For instance, idioms from the **Food Culture** category might be used in beginner-level lessons about daily life, while **Historical or Symbolic** idioms may serve as discussion prompts in advanced courses focused on literature or cultural history.

Aligning cultural classification with instructional design contributes not only to vocabulary acquisition and phrase comprehension but also to the cultivation of intercultural communicative competence. Learners begin to recognize how language reflects values, social hierarchies, and lived experiences, thereby developing critical awareness of both target and native cultures. This process encourages deeper reflection, critical thinking, and the development of pragmatic sensitivity—key components in advanced language learning.

Ultimately, a culturally grounded approach to idiomatic instruction—supported by explainable AI—can foster more inclusive, engaging, and contextually meaningful

learning experiences. It bridges linguistic form and cultural function, enabling learners to move beyond surface-level understanding and toward the embodied use of idiomatic language in real-world contexts.

CONCLUSION

Summary of the Study:

This study examined how Korean idioms can be grouped by cultural themes and what these groupings reveal about language use. We selected 100 commonly taught idioms from major Korean language textbooks and assigned them to cultural categories such as Confucian values, agricultural traditions, or body-related concepts.

Using this labeled dataset, we trained a multi-label classifier with the KoBERT model. The system performed well overall (macro F1-score = 0.87). To better understand the classification process, we applied SHAP analysis, which showed that words related to the body, tradition, or symbolism played a key role in predicting labels. These findings suggest that machine learning, when guided by linguistic insight, can support idiom instruction that reflects Korean cultural values.

Pedagogical Applications and Research Perspectives

The cultural code model offers a practical way to connect idiomatic expressions with key aspects of Korean culture. Teachers can use this classification to explain how certain phrases reflect everyday beliefs or social values. For example, idioms based on body parts or Confucian relationships are familiar to many Koreans and are well-suited for beginner lessons.

Tools like KoBERT and SHAP can further support teachers by highlighting which expressions are easier for learners to understand. This makes it possible to select idioms that match students' cultural knowledge and language skills. Meanwhile, idioms from less common categories—such as those tied to traditional symbols or historical stories—can be introduced through folk tales, visual media, or group activities. This not only supports language development, but also helps students reflect on the deeper cultural layers within the language.

Although this study presents a structured framework for classifying Korean idiomatic expressions through cultural codes, several limitations should be noted. First, the cultural labels were manually annotated, which inevitably involved subjective judgment. Some idioms span multiple cultural domains, and assigning a single or primary label can vary depending on individual interpretation, potentially affecting consistency.

Second, the proposed framework has not yet been implemented in actual teaching contexts. While the categories align with well-established cultural themes, their effectiveness in enhancing learners' comprehension, retention, or engagement remains unverified through empirical evidence.

By incorporating cultural code classification into idiom instruction, teachers can foster deeper intercultural understanding and create lesson flows that are aligned with learners' cognitive and cultural readiness. This study suggests that future instructional materials and teacher training programs should actively integrate culturally coded

idioms along with explainable AI-based insights, thereby bridging data-driven analysis with human-centered pedagogy.

Future research should explore how cultural code-based instruction functions in classroom environments. Experimental studies could evaluate whether this model not only improves vocabulary acquisition but also deepens learners' understanding of Korean cultural perspectives expressed through idiomatic language.

Additionally, the integration of AI tools—such as GPT models—could be explored to generate contextualized idiom examples or provide personalized feedback. Combining machine learning–based classification with multimodal content (e.g., illustrations, videos, dramatizations) may offer more intuitive support for learners, particularly those from diverse linguistic and cultural backgrounds.

※ This study builds on the assumption that instruction informed by cultural codes can support learners across varied sociocultural contexts. As suggested by [Jeong and Ergasheva, 2025; manuscript submitted], future work should examine how this framework can be adapted for different proficiency levels and cultural environments. Exploring how machine learning systems can interact with visual and narrative content may further enhance idiom education by making abstract figurative meanings more tangible and engaging.

REFERENCES

1. AI Hub. (2021). *Hangug-eo nyuseu malmunchi* [Korean news corpus]. https://aihub.or.kr/aidata/87
2. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.
3. Boers, F., & Lindstromberg, S. (2008). Cognitive linguistic approaches to teaching vocabulary and phraseology. Mouton de Gruyter.
4. Byram, M. (1997). Teaching and assessing intercultural communicative competence. Multilingual Matters.
5. Corder, S. P. (1967). The significance of learners' errors. International Review of Applied Linguistics in Language Teaching, 5(4), 161–170.
6. Deignan, A. (2005). Metaphor and corpus linguistics. John Benjamins Publishing.
7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171–4186). Association for Computational Linguistics.
8. Ellis, R. (2001). The metaphorical constructions of second language learners. In M. P. Breen (Ed.), Learner contributions to language learning: New directions in research (pp. 65–85). Longman. Goddard, C. (2006). Ethnopragmatics: Understanding discourse in cultural context. Mouton de Gruyter.
9. Hinkel, E. (2004). Teaching academic ESL writing: Practical techniques in vocabulary and grammar. Lawrence Erlbaum Associates.
10. Jeong, I. S., & Ergasheva, S. (2025). A study on building an educational database of Korean idiomatic expressions using machine learning. Manuscript submitted for publication.
11. Kim, H.-J. (2015). Hangug-eo haksupjareul wihan gyoyuk-yong gwanyongpyo-hyeon-ui seonjeong gijun yeongu [A study on the selection criteria of idiomatic expressions for Korean language learners]. *Oegugeolo-seoui Hangug-eo Gyoyuk*, 32, 125–146. (in Korean)
12. Kim, J.-A. (2023). Uimi gwangyemang-eul hwal-yonghan gwanyongpyo-hyeon gyosu bang-an yeongu [A study on teaching idioms using semantic networks]. *Oegugeolo-seoui

Hangug-eo Gyoyuk*, 35(1), 145–168. (in Korean)
13. Kim, J.-E. (2015). Dongmulmyeongeul pohamhwanhan gwanyongpyo-hyeon-ui Hangug-eo gyoyuk bang-an yeongu [A study on idioms with animal names in Korean language education]. *Gugje Hangug-eo Gyoyuk Hakhoe*, 26(1), 147–173. (in Korean)
14. Kim, J.-W. (2018). Hangug-eo munhwa tonghap gyoyuk-ui silje [Practices in integrated Korean culture education]. *Hangug-eo Gyoyuk Hakhoe*, 29(2), 35–58. (in Korean)
15. Kim, M. (2020). The role of idioms in Korean language textbooks: A corpus-based study. *Korean Language Education Research*, 53(1), 77–102.
16. Kim, M.-H. (2020). Hangug-eo gyoyuk-eseo gwanyongpyo-hyeon gyosu bang-an yeongu [A study on idiom teaching methods in Korean education]. *Hangug-eo Gyoyuk*, 31(2), 103–127. (in Korean)
17. Kim, M.-J. (2022). Eunyuh insik giban gwanyongpyo-hyeon gyosu jeollyak [Metaphor awareness-based teaching strategies for idioms]. *Oegugeolo-seoui Hangug-eo Gyoyuk*, 34(2), 101–120. (in Korean)
18. Kim, S., Lee, J., & Choi, H. (2022). Teaching idioms in Korean language classrooms: Challenges and instructional strategies. Journal of Language and Culture, 45(3), 211–230.
19. Kövecses, Z. (2002). Metaphor: A practical introduction. Oxford University Press.
20. Kramsch, C. (1998). Language and culture. Oxford University Press.
21. Laufer, B. (2000). Avoidance of idioms in a second language: The effect of L1–L2 differences. Studies in Second Language Acqu isition, 22(4), 587–609.
22. Lee, H. J., & Kim, M. J. (2023). Mapping cultural codes in Korean idioms: Pedagogical strategies and frameworks. Journal of Korean Education and Culture, 15(1), 25–49.
23. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
24. Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), 404–411.
25. Moon, G.-H. (2022). Gwanyongpyo-hyeon yeongu-ui hyeonhwang-gwa jeonmang [Current status and prospects of idiom research]. *Hangug-eohak*, 95, 51–81. (in Korean)
26. Moon, K.-H. (2022). Gwanyongpyo-hyeon-ui uimi haeseok-gwa munhwajeok ham-ui [Semantic interpretation and cultural implications of idioms]. *Hangug Eoneo Munhwa Yeongu*, 21(1), 45–70. (in Korean)
27. Moon, R. (1998). Fixed Expressions and Idioms in English: A Corpus-Based Approach. Oxford University Press.
28. Moon, S. (2022). Cultural context and idiomatic expressions in KFL education. Korean Linguistics, 34(2), 115–136.
29. National Institute of Korean Language. (2022). *Pyojun Gugeo Daesajeon: Gwanyongpyo-hyeon Sajeon* [Standard Korean Dictionary: Idioms]. https://stdict.korean.go.kr
30. Paige, R. M., Jorstad, H. L., Siaya, L., Klein, F., & Colby, J. (2003). Culture learning in language education: A review of the literature. In D. Lange & R. M. Paige (Eds.), Culture as the core: Perspectives on culture in second language learning (pp. 173–236). Information Age Publishing.
31. Park, J. H. (2020). Cultural teaching through idioms in Korean language classrooms. Korean Language Education, 31(1), 41–66.
32. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084.
33. Seoul National University Korean, Yonsei Korean, and Sejong Korean textbooks. (n.d.).
34. Sharifian, F. (2011). Cultural conceptualisations and language: Theoretical framework and applications. John Benjamins Publishing Company.
35. Taguchi, N. (2011). Teaching pragmatics: Trends and issues. Annual Review of Applied Linguistics, 31, 289–310.
36. Wierzbicka, A. (1997). Understanding cultures through their key words. Oxford University Press.