International Journal of Language Learning and Applied Linguistics
ISSN: 2835-1924
Volume 04 Number 03 (2025)
Impact Factor: 9.89
https://inter-publishing.com/index.php/IJLLAL

www.inter-publishing.com

Article

Natural Language Processing Techniques for Representing Cultural Concepts in AI-Based Lexicography: A Case Study on English and Uzbek

Yusupova Mushtariy Baxtiyor qizi

1. Karshi State University, Doctorate (PhD) Student
* Correspondence: mushtariyyusupova1999@gmail.com

Abstract: Artificial intelligence (AI) and Natural Language Processing (NLP) have transformed lexicographic practices by enabling scalable and automated language analysis. While these advancements allow for the efficient processing of linguistic data, they often fall short in accurately capturing and representing culturally embedded concepts, particularly in underrepresented languages like Uzbek. Despite progress in semantic models and named entity recognition, there remains a significant lack of cultural sensitivity in AI-based lexicographic systems, primarily due to limited annotated corpora and challenges in modeling context-specific meanings. This study examines how NLP techniques—specifically tokenization, word embedding, and named entity recognition—function in representing cultural concepts in English and Uzbek, and evaluates their strengths and limitations. Findings show that while English-language models handle idiomatic and cultural terms with moderate success, models for Uzbek exhibit considerable deficiencies due to morphological complexity and corpus scarcity. Both languages face issues with accurately capturing idiomatic expressions and culturally loaded entities, leading to semantic distortion in automated outputs. The paper introduces a comparative framework grounded in semantic theory and cultural linguistics, providing practical examples of misrepresentation and highlighting the need for culturally annotated corpora and cross-cultural NLP modeling. To achieve culturally competent AI lexicography, interdisciplinary collaboration is essential. Future systems must integrate domain-specific resources, cultural annotations, and linguistic diversity to ensure that AI technologies do not reduce language to mechanistic processing but preserve its cultural and emotional richness.

Citation: Baxtiyor qizi, Y. M. Axiological and Linguistic Concepts in AI-Based Lexicography: A Case Study on English and Uzbek. International Journal of Language Learning and Applied Linguistics 2025, 4(3), 36-40.

Received: 10th March 2025
Revised: 25th March 2025
Accepted: 30thApr 2025
Published: 04th May 2025

Keywords: Artificial intelligence, natural language processing, cultural concepts, AI lexicography, English language, Uzbek language, semantic analysis, named entity recognition

1. Introduction

The integration of artificial intelligence (AI) into lexicography has profoundly reshaped the ways in which linguistic and cultural phenomena are represented and documented. Traditional lexicographic methods often relied on manual data collection and human intuition, but AI-driven systems now enable dynamic and scalable analysis, allowing for the rapid processing of vast corpora and the uncovering of subtle cultural nuances.

Natural Language Processing (NLP), a subfield of AI, has emerged as the primary technological tool for this transformation. NLP encompasses a range of computational techniques, including tokenization, word embedding, semantic parsing, and named entity recognition (NER), each of which contributes uniquely to the automated interpretation of human language.

However, despite these advancements, accurately representing cultural concepts remains an open challenge. Cultural concepts are deeply embedded in historical, social, and emotional contexts that computational models often fail to capture adequately. Particularly, languages like Uzbek, which have relatively limited annotated corpora compared to English, pose additional barriers.

2. Materials and Methods

This study employed a qualitative comparative analysis methodology to evaluate how current Natural Language Processing (NLP) techniques represent cultural concepts in AI-based lexicographic tools, using English and Uzbek as case studies [1]. The research was grounded in a systematic document analysis and system review approach [2]. Selected AI-driven lexicographic platforms such as Google Dictionary, Oxford Lexico [3], and emerging UzbekBERT-based tools were examined with regard to three core NLP components: tokenization accuracy, word embedding effectiveness, and named entity recognition (NER) performance[4]. A corpus of culturally significant terms—idioms, festivals, traditional foods, and symbolic expressions—was compiled in both English and Uzbek [5]. These data points were evaluated for their semantic representation, token boundaries [6], and entity classification success rates [7]. The analysis incorporated real-world translation and recognition outputs to assess fidelity in capturing nuanced meanings [8], with attention given to semantic loss and cultural misinterpretations [9]. For Uzbek, particular emphasis was placed on the challenges posed by agglutinative morphology and low-resource NLP environments [10]. The methodology also integrated a theoretical lens informed by semantic prime theory and cultural scripts [11], enabling deeper interpretation of how lexical meaning intersects with cultural context [12]. This cross-linguistic and interdisciplinary method allowed the study to not only reveal technical limitations in current NLP systems but also propose culturally grounded recommendations for future AI lexicographic development [13]. The insights gained provide a foundation for designing models that are more sensitive to linguistic diversity and cultural specificity [14].

3. Results

This article explores the role of NLP techniques in AI-based lexicography through a comparative analysis of English and Uzbek [15]. It aims to assess current methodologies, identify challenges, and propose strategies for enhancing the cultural sensitivity of AI lexicographic tools.

The application of AI in lexicography has gained momentum since the late 1990s. Initial works, such as WordNet, demonstrated the feasibility of computational semantic networks for representing language knowledge. Subsequent efforts integrated machine learning to automatically categorize and define words based on context.

Recent developments, particularly in deep learning, have led to sophisticated models like BERT and GPT, which offer new possibilities for lexicography. In the Uzbek context, resources like UzbekBERT are emerging but remain in their infancy, limiting their effectiveness for nuanced cultural analysis.

Furthermore, researchers like Wierzbicka and Goddard highlight the intrinsic difficulties of translating cultural meanings across languages. They argue that universal semantic primes exist, but culturally specific meanings require tailored attention, an insight particularly relevant for AI lexicography.

This article employs a qualitative comparative analysis based on document analysis and system review. Selected AI-based lexicographic tools for English (Google Dictionary, Oxford Lexico) and for Uzbek (UzbekBERT-based resources) were examined to assess:

Tokenization accuracy

Word embedding quality for cultural terms

Named Entity Recognition (NER) performance for culturally significant entities

Data points included common idioms, festivals, traditional foods, and symbolic terms across both languages

Analysis and Discussion:

Tokenization Challenges: English benefits from clear word delimiters, but issues arise with idiomatic expressions such as "kick the bucket" or "spill the beans", where literal interpretation fails.

In Uzbek, challenges are compounded by agglutinative morphology and multi-word expressions like "koʻnglini olmoq" (to please someone), where incorrect tokenization disrupts the intended meaning.

Word Embedding Limitations: Standard embedding models such as Word2Vec and GloVe fail to adequately capture cultural richness. In English, concepts like "Thanksgiving" are relatively well embedded due to abundant data. In contrast, Uzbek embeddings often struggle with concepts like "navroʻz" or "soʻfi", reflecting corpus limitations and underrepresentation.

Named Entity Recognition (NER) Difficulties: While English-language NER systems successfully recognize major cultural entities, they often misinterpret lesser-known references. Uzbek NER models face even greater challenges; for instance, distinguishing between common nouns and cultural references like "Asrlar Sadosi" (a cultural festival) requires domain-specific annotation that is currently sparse.

4. Discussion

Practical Examples: Misinterpretations abound when AI models translate culturally laden terms:

English-Uzbek Example: "Black-eyed Susan" is translated literally but loses its botanical meaning.

Uzbek-English Example: "Qorakoʻz" is translated as "black-eyed" without capturing its poetic meaning of beauty and belovedness.

Furthermore, Google Translate and other systems frequently misinterpret expressions involving honorifics or kinship terms in Uzbek.

Semantic Representation of Cultural Concepts

Understanding how cultural concepts are encoded in language requires a theoretical foundation rooted in semantics. Wierzbicka's theory of semantic primes suggests that universal concepts exist across all languages, but their specific cultural realizations vary significantly. According to Goddard , cultural scripts guide social behavior and are deeply embedded in linguistic expressions. Natural Language Processing systems aiming to represent cultural units must account for these underlying semantic structures to avoid misinterpretations. Without a robust semantic framework, NLP models risk simplifying or distorting culturally rich terms during tokenization, embedding, or classification processes.

Challenges in Representing Cultural Concepts

Despite advancements in NLP, several critical challenges hinder the accurate representation of cultural concepts:

Data Imbalance: High-resource languages like English have extensive, annotated corpora, whereas Uzbek remains underrepresented, limiting the quality of AI models.

Idiomatic Complexity: Expressions like "koʻnglini olmoq" (to appease someone's heart) or "qovoq solmoq" (to sulk) are difficult to segment and interpret correctly.

Cultural Uniqueness: Certain cultural practices or artifacts, such as "Asrlar Sadosi" festival or "qozon kabob," lack direct equivalents in English, complicating automatic translation and recognition.

Low-Resource Challenges: The scarcity of domain-specific corpora and annotated datasets for Uzbek hampers the development of accurate cultural embeddings and NER systems.

These challenges demonstrate the necessity for culturally sensitive computational approaches.

Several practical examples reveal the shortcomings of current AI systems:

Navro‘z Holiday: In some NLP systems, "Navro‘z" is interpreted merely as "New Year," omitting its rich historical and cultural significance tied to Persian traditions, renewal, and communal festivities.

Choyxona: The term "choyxona" (teahouse) often gets translated simply as "tea shop," failing to convey its cultural role as a central social institution in Uzbek society, much like a traditional community hub.

Thanksgiving: When translating "Thanksgiving" into Uzbek, systems often render it as "minnatdorchilik kuni," which, while literal, misses the American historical context involving pilgrims and Native Americans.

These examples illustrate that direct translation without cultural grounding can lead to significant semantic loss.

Future of AI-Based Lexicography

Future efforts to enhance cultural representation in AI lexicography should focus on:

Developing culturally annotated corpora with detailed metadata about the socio-cultural contexts of terms.

Creating cross-cultural NLP models trained on parallel datasets that respect linguistic diversity and cultural specificity.

Involving interdisciplinary collaboration among linguists, cultural anthropologists, and AI researchers to build more nuanced models.

Training AI on cultural sensitivity tasks to help models better detect, differentiate, and preserve cultural meanings.

Such initiatives can contribute to AI systems that not only process language mechanically but also appreciate the depth of human cultural experience.

5. Conclusion

While NLP techniques have significantly advanced AI-based lexicography, the accurate representation of cultural concepts remains a formidable challenge. Tokenization errors, embedding limitations, and NER inaccuracies contribute to the loss of nuanced meanings.

To mitigate these issues, future research should prioritize:

Developing larger, culturally annotated corpora

Designing domain-specific NLP models for underrepresented languages like Uzbek

Integrating interdisciplinary expertise from cultural studies, linguistics, and AI engineering

Addressing these limitations in future research will be essential to provide a more comprehensive evaluation. Building AI lexicographic tools that truly reflect the cultural and emotional depth of languages will require collaborative, multi-faceted efforts.

REFERENCES

[1] R. Navigli и S. P. Ponzetto, «BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network», *Artif. Intell.*, 217–250, 2012.

[2] J. Devlin, M.-W. Chang, K. Lee, и K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding», *ArXiv Prepr. ArXiv181004805*, 2019.

[3] T. Mikolov, K. Chen, G. Corrado, и J. Dean, «Efficient Estimation of Word Representations in Vector Space», *ArXiv Prepr. ArXiv13013781*, 2013.

[4] R. Jackendoff, *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, 2002.

[5] V. Evans, *How Words Mean: Lexical Concepts, Cognitive Models, and Meaning Construction*. Oxford University Press, 2006.

[6] A. Radford, K. Narasimhan, T. Salimans, и I. Sutskever, «Improving Language Understanding by Generative Pre-Training», *OpenAI Tech. Rep.*, 2018.

[7] E. F. Tjong Kim Sang и F. De Meulder, «Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition», в *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, 142–147.

[8] N. Ide и J. Véronis, «Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art», *Comput. Linguist.*, 1–40, 1998.

[9] E. Cambria и B. White, «Jumping NLP Curves: A Review of Natural Language Processing Research», *IEEE Comput. Intell. Mag.*, 48–57, 2020.

[10] E. M. Bender, T. Gebru, A. McMillan-Major, и S. Shmitchell, «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?», *Proc. 2021 ACM Conf. Fairness Account. Transpar.*, 610–623, 2021.

[11] P. Piwek, «Presenting Natural Language Generation as a Form of Communication», *Philos. Trans. R. Soc. B*, 2625–2635, 2008.

[12] C. Goddard, *Semantic Analysis: A Practical Introduction*, Oxford University Press, 2011.

[13] A. Wierzbicka, *Understanding Cultures Through Their Key Words: English, Russian, Polish, German, and Japanese*. Oxford University Press, 1997.

[14] B. Tursunov, M. Akhmedov, и N. Xudoyberganov, «UzbekBERT: Pre-trained Language Model for Uzbek Language Understanding». 2021 г.

[15] C. Fellbaum, *WordNet: An Electronic Lexical Database*. MIT Press, 1998.