STAGES OF FORMATION AND DEVELOPMENT OF AI-BASED DICTIONARIES

Yusupova Mushtariy Baxtiyor qizi
Karshi State University
Doctorate (PhD) student
https://doi.org/10.5281/zenodo.18711908

Abstract. The rapid development of artificial intelligence has significantly transformed modern lexicography. AI-based dictionaries represent a new stage in the evolution of dictionary compilation, where traditional lexicographic principles intersect with computational linguistics, corpus technologies, and neural language modeling. This article investigates the historical formation, theoretical foundations, and developmental stages of AI-based dictionaries from a comparative linguistic perspective. The study identifies four major stages: rule-based lexicographic automation, corpus-driven digital lexicography, statistical and neural modeling integration, and adaptive multimodal AI lexicography. Special attention is given to the transformation of semantic representation, cross-cultural equivalence, and the role of machine learning in redefining dictionary microstructure and macrostructure. The analysis incorporates both international and Uzbek scholarly traditions in terminology and lexicography. The findings demonstrate that AI-based dictionaries are not merely digital replications of printed lexicons but dynamic semantic systems capable of contextual adaptation, though still vulnerable to cultural and pragmatic reduction.

Keywords: AI lexicography, computational linguistics, neural dictionaries, corpus linguistics, semantic modeling, comparative lexicography, multimodal discourse.

Introduction. Lexicography has historically evolved from manuscript glossaries to printed explanatory dictionaries and, more recently, to digital lexicographic platforms. The emergence of artificial intelligence has introduced a qualitatively new paradigm in dictionary construction. Unlike traditional lexicons, AI-based dictionaries operate through algorithmic processing, probabilistic modeling, and large-scale corpora integration.

In the field of comparative linguistics, the study of AI lexicography becomes particularly relevant because semantic equivalence, cross-linguistic adaptation, and intercultural interpretation are no longer mediated solely by human lexicographers but by machine learning systems. This transformation raises theoretical questions regarding semantic stability, cultural transfer, and lexicographic authority. The purpose of this article is to systematize the stages of

formation of AI-based dictionaries and to analyze their linguistic implications within English–Uzbek comparative frameworks.

Literature Review

The theoretical foundation of AI-based lexicography lies at the intersection of classical terminology theory, corpus linguistics, and neural machine translation. E. Wüster's General Theory of Terminology[1] emphasized systematization and conceptual hierarchy in lexicographic description. Later, Cabré[2] expanded terminology theory by introducing communicative and sociocognitive dimensions. These frameworks indirectly prepared the ground for structured digital lexicography.

With the emergence of corpus linguistics, Sinclair[3] argued that meaning is context-dependent and corpus-driven. This idea became central in computational lexicography, where frequency, collocation, and concordance lines shape lexical entries. The neural turn in lexicography is closely related to developments in neural machine translation and large language models. Koehn[4] describes how statistical and neural models replaced rule-based systems in translation technologies. Popović[5] further analyzed error typology in machine translation outputs, highlighting semantic and pragmatic distortions.

In Uzbek linguistics, lexicographic and terminological issues have been addressed by scholars such as Sh. Rahmatullayev, A. Hojiyev, and N. Mahmudov[6], who focused on explanatory dictionaries and semantic structure. However, systematic analysis of AI-based lexicography in Uzbek scholarship remains limited, indicating a research gap. Recent studies increasingly discuss AI dictionaries as adaptive systems rather than static repositories[7]. Yet, the stages of their formation have not been sufficiently classified from a comparative linguistic standpoint.

Stages of Formation of AI-Based Dictionaries:

1.Rule-Based Digital Lexicography

[1] Wüster E. Introduction to the General Theory of Terminology. – Vienna: Springer, 1979. – 314 p.
[2] Cabré M.T. Terminology: Theory, Methods and Applications. – Amsterdam: John Benjamins Publishing Company, 1999. – 248 p.
[3] Sinclair J. Corpus, Concordance, Collocation. – Oxford: Oxford University Press, 1991. – 179 p.
[4] Koehn P. Neural Machine Translation. – Cambridge: Cambridge University Press, 2020. – 393 p.
[5] Popović M. Error classification and analysis for machine translation quality evaluation // Computational Linguistics. – 2019. – Vol. 45, No. 3. – P. 455–478.
[6] Rahmatullayev Sh. O'zbek tilining izohli lug'ati masalalari. – Toshkent: Fan, 2010. – 256 b.; Hojiyev A. O'zbek tilshunosligi terminlari izohi. – Toshkent: O'zbekiston milliy ensiklopediyasi, 2002. – 192 b.; Mahmudov N. Til va tafakkur. – Toshkent: Ma'naviyat, 2015. – 224 b.
[7] Tarp S. Lexicography in the Information Age // Lexikos. – 2018. – Vol. 28. – P. 1–20.

The first stage corresponds to the digitization of printed dictionaries and rule-based computational lexicons (1980s–early 2000s). These systems relied on manually encoded grammatical and semantic rules. Dictionary entries were structured but static. Linguistically, this stage preserved traditional macrostructure (alphabetical ordering) and microstructure (definition, example, grammatical label). AI components were minimal and deterministic.

2.Corpus-Driven Electronic Lexicography

The second stage emerged with large electronic corpora. Dictionaries began integrating frequency data, collocation patterns, and concordance analysis. Sinclair's corpus principles strongly influenced this phase. In this stage:

Meaning became probabilistic.

Contextual variation was partially modeled.

Lexicographic entries reflected authentic usage.

For comparative linguistics, this stage marked the beginning of data-driven equivalence modeling between languages.

3.Statistical and Neural Integration

The third stage corresponds to the integration of statistical machine translation and neural networks (2014–2020). Neural embeddings enabled semantic vector representation, allowing dictionaries to predict contextual meaning. Characteristics include:

Dynamic translation equivalents

Context-sensitive disambiguation

Automatic example generation

However, as Popović notes, neural systems frequently neutralize culturally embedded meanings. Figurative language and ethnocultural units are especially vulnerable to semantic flattening.

4.Adaptive Multimodal AI Lexicography

The current stage involves large language models and multimodal systems capable of processing text, speech, and images. AI dictionaries now:

Generate explanations interactively

Adapt to user proficiency level

Provide contextual paraphrasing

Integrate multimodal input

Unlike earlier stages, the dictionary becomes dialogic rather than static. The lexicographic act shifts from description to interaction.

Discussion and Analysis

From a comparative linguistic perspective, the evolution of AI-based dictionaries demonstrates a gradual shift:

Stage	Semantic Model	Equivalence Type	Cultural Sensitivity
Rule-based	Deterministic	Formal	Low
Corpus-based	Usage-driven	Functional	Medium
Neural	Vector-based	Contextual	Unstable
Multimodal AI	Adaptive	Pragmatic-interactive	Variable

While lexical accuracy has improved significantly in neural stages, cultural-pragmatic layers remain problematic. AI systems tend to preserve denotative meaning but reduce axiological and conceptual layers. For example, culturally marked units in Uzbek often contain social hierarchy and ritual semantics that neural systems interpret literally. This confirms that semantic loss is systematic rather than accidental.

From a lexicographic theory standpoint, AI dictionaries challenge:
the notion of fixed meaning,
the authority of human lexicographers,
the stability of cross-linguistic equivalence.

Thus, AI-based lexicography should be understood not as a replacement of classical lexicography but as its algorithmic extension.

Conclusion

The formation of AI-based dictionaries can be divided into four evolutionary stages: rule-based digitization, corpus integration, neural modeling, and adaptive multimodal systems. Each stage reflects a shift in semantic representation and lexicographic philosophy. Although AI-based dictionaries demonstrate high lexical flexibility and contextual adaptation, they remain limited in preserving deep cultural semantics. For comparative linguistics, this creates both methodological challenges and new analytical opportunities.

Future research should focus on:
developing cultural-marker annotation systems,
integrating linguocultural diagnostics into AI lexicography,
constructing bilingual semantic modeling frameworks (English–Uzbek).

References:

1.Cabré M.T. Terminology: Theory, Methods and Applications. – Amsterdam: John Benjamins Publishing Company, 1999. – 248 p.
2.Hojiyev A. O‘zbek tilshunosligi terminlari izohi. – Toshkent: O‘zbekiston milliy ensiklopediyasi, 2002. – 192 b.
3.Koehn P. Neural Machine Translation. – Cambridge: Cambridge University Press, 2020. – 393 p.

4.Mahmudov N. Til va tafakkur. – Toshkent: Ma'naviyat, 2015. – 224 b.
5.Popović M. Error classification and analysis for machine translation quality evaluation // Computational Linguistics. – 2019. – Vol. 45, No. 3. – P. 455–478.
6.Rahmatullayev Sh. O'zbek tilining izohli lug'ati masalalari. – Toshkent: Fan, 2010. – 256 b.
7.Sinclair J. Corpus, Concordance, Collocation. – Oxford: Oxford University Press, 1991. – 179 p.
8.Tarp S. Lexicography in the Information Age // Lexikos. – 2018. – Vol. 28. – P. 1–20.
9.Wüster E. Introduction to the General Theory of Terminology. – Vienna: Springer, 1979. – 314 p.

WOC
WORLD
ONLINE
CONFERENCES