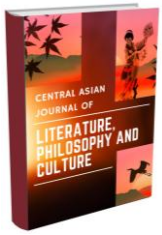


CENTRAL ASIAN JOURNAL OF LITERATURE, PHILOSOPHY, AND CULTURE

https://cajlpc.casjournal.org/index.php/CAJLPC

Volume: 07 Issue: 03 | July 2026 ISSN: 2660-6828

Article

Automatic Detection and Correction of Spelling Errors in Agglutinative Languages (A Case Study of the Uzbek Language)

Nazira Sobirova G'anijon qizi*[1]

1. Basic Doctoral Student, Tashkent State University of Uzbek Language and Literature
*Correspondence: sobirovanazira134@gmail.com

Abstract: One of the key directions in the field of Natural Language Processing (NLP) is the automatic detection and correction of spelling errors in texts. This task becomes particularly challenging in agglutinative languages such as Uzbek, where words are formed through the addition of numerous affixes. This paper analyzes algorithms for detecting and correcting spelling errors in the Uzbek language, their operational principles, and modern machine learning approaches. The study examines dictionary-based methods, the Levenshtein distance algorithm, N-gram models, and context-aware approaches based on neural networks. The findings demonstrate that a hybrid algorithm, combining multiple techniques, provides the most effective solution for Uzbek spell-checking.

Keywords: Spell checking, NLP, Uzbek language, Levenshtein distance, N-gram, language model.

Citation: G'anijon qizi N. S. Automatic Detection and Correction of Spelling Errors in Agglutinative Languages (A Case Study of the Uzbek Language). Central Asian Journal of Literature, Philosophy, and Culture 2026, 7(3), 91-97.

Received: 10th Feb 2026
Revised: 11th Mar 2026
Accepted: 19th Apr 2026
Published: 16th May 2026

Introduction

In recent years, the rapid development of information and communication technologies has led to a significant increase in the volume of digital texts generated across various platforms. The growing number of texts produced through internet networks, social media platforms, electronic document systems, and artificial intelligence-based services has made the tasks of efficient processing, analysis, and quality control of textual data increasingly important [1].

In particular, the rapid detection of spelling errors in written language and their automated correction has become not only a theoretical research problem but also an essential component of practical systems [2].

Although the task of detecting spelling errors may appear simple at first glance, its internal mechanism consists of several complex stages. Initially, the text is segmented into smaller units, after which these units are compared with existing linguistic resources. Based on the identified inconsistencies, possible correct variants are generated, and the most appropriate option is selected among them [3].

For this process to function effectively, a combination of linguistic knowledge, probabilistic modeling, and algorithmic approaches is required. In particular, recent advancements in deep learning methods have created new opportunities in this domain [4].

Spell-checking systems developed for high-resource languages have reached a high level of sophistication, relying on large-scale corpora and extensive lexical databases. Such

systems are capable not only of detecting simple typographical errors but also, in certain cases, identifying contextually inappropriate words [5].

At the global level, numerous advanced systems have been developed based on artificial intelligence and machine learning technologies [6].

Methodology

Popular systems include:

- **Grammarly** — performs contextual grammatical and stylistic analysis
- **Microsoft Word** — includes a built-in spell-checking system
- **Google Docs** — provides real-time AI-based editing
- **LanguageTool** — a multilingual open-source system

These systems utilize the following technologies:

- rule-based approaches
- statistical models
- neural networks (deep learning)

Their main advantage lies in their ability to understand context and detect complex errors.

However, such technologies cannot be uniformly applied to all languages. This limitation is primarily due to the internal structure and grammatical characteristics of each language.

Results and Discussion

Although automatic spell-checking systems are highly developed for English and Russian, this field remains insufficiently developed for low-resource languages, including Uzbek. Due to the agglutinative nature of the Uzbek language, a single root can generate hundreds of word forms, which significantly complicates algorithmic analysis [7].

The Problem of Spelling Errors in the Uzbek Language

In recent years, the development of Natural Language Processing (NLP) has made automatic text analysis, error detection, and error correction an important scientific issue. This problem becomes even more complex in agglutinative languages such as Uzbek [8].

In Uzbek, words are formed through numerous affixes, which allows hundreds of different forms to be generated from a single root [9].

Uzbek Spelling Rules as Linguistic Support

Figure 1. *Linguistic Support Components of the Spell-Checking System.*

For example: **yozmoq → yozdi → yozayotgan → yozilmagan**

Such a structure creates additional challenges for spell-checking systems. As a result, a simple dictionary-based check is not sufficient; it becomes necessary to consider context, while morphological analysis plays an important role [10].

The main features of the Uzbek language are:
a) word formation through suffixes;
b) sequential attachment of morphemes;
c) high morphological variability.

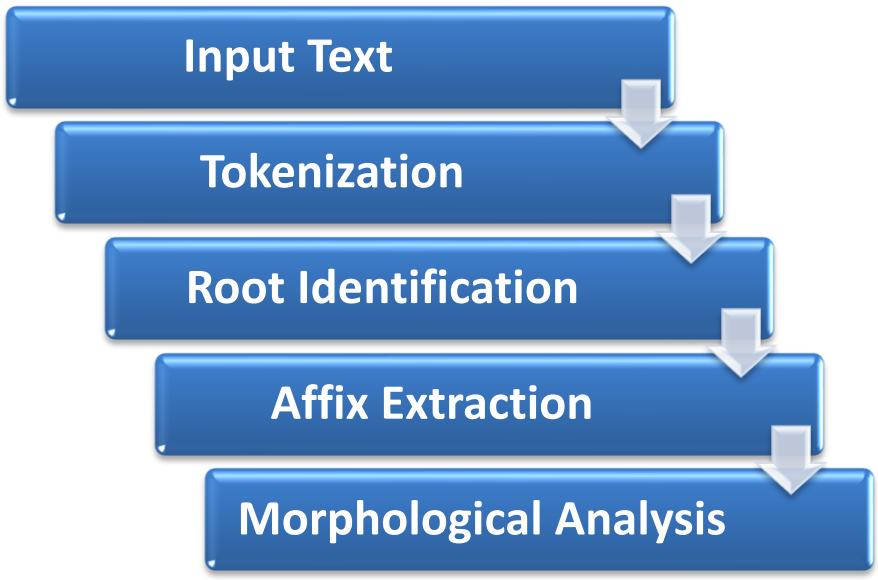
For example:

Table 1. The main features of the Uzbek language.

Root	Suffix	Result
Kitob	-lar	Kitoblar
Kitoblar	-imiz	Kitoblarimiz
kitoblarimiz	-dan	Kitoblarimizdan

This structural complexity makes spell-checking systems more difficult to develop [11].

Morphological analyzers, such as MorphUz, operate through the following stages [12]:



Through this process:

1. the root of the word is identified;
2. its grammatical form is determined;
3. the probability of an error is evaluated.

The MorphUz system classifies words by separating affixes [13].

Table 2. Spelling errors are divided into two main types:

Error Type	Description	Example
Non-word error	The word is not found in the dictionary	Kitobn
Real-word error	The word exists but is used in an incorrect context	men kitob yedi

In Uzbek, spelling errors mainly arise due to the following factors:

1. keyboard errors, such as missing or extra letters;
2. phonetic similarity, such as **x–h** and **o‘–u**;
3. incorrect use of suffixes;
4. morphological complexity.

Studies show that traditional dictionary-based checking systems do not provide sufficient accuracy for Uzbek, because words have numerous grammatical variants [14].

The dictionary-based checking method works quite simply: the input string is checked against a list of accepted words, that is, whether it exists in the dictionary. If the string is not found in the dictionary, it is marked as a misspelled word. However, there are also subtle challenges in constructing a dictionary that is useful for a spell-correction application.

Historically, text recognition systems have relied more on n-gram methods, whereas spell-checkers have mainly used dictionary-based checking. In both cases, problems arise when errors disrupt word boundaries, that is, when words are incorrectly joined or split. The advantage of this approach is that it works quickly and is easy to implement. Its limitation is that it cannot detect real-word errors.

Levenshtein Distance Algorithm

The Levenshtein distance calculates the minimum number of edit operations required to transform one string into another. This algorithm is based on insertion, deletion, and substitution operations [15]. The Norvig model selects the most appropriate word probabilistically based on this distance. Its advantage is that it detects typographical errors effectively. Its limitation is that it does not take context into account.

Table 3. N-gram Statistical Model.

Word	Correct Form	Distance
Kitb	Kitob	1
Kelmadi	Kelmadi	0
Oquvchi	o‘quvchi	1

N-grams are sequences of *n* consecutive characters in a word or string, where *n* is usually equal to 1, 2, or 3.

One-character n-grams are called unigrams or monograms; two-character n-grams are called digrams or bigrams; and three-character n-grams are called trigrams.

In general, n-gram-based error detection methods examine each n-gram in the input string and compare it with a precompiled n-gram statistics table to determine its presence or frequency. Strings containing nonexistent or very rare n-grams are marked as potentially misspelled units.

N-gram methods usually require a dictionary or a large text corpus to build the n-gram table in advance. The N-gram model calculates the probability of a sequence of words:

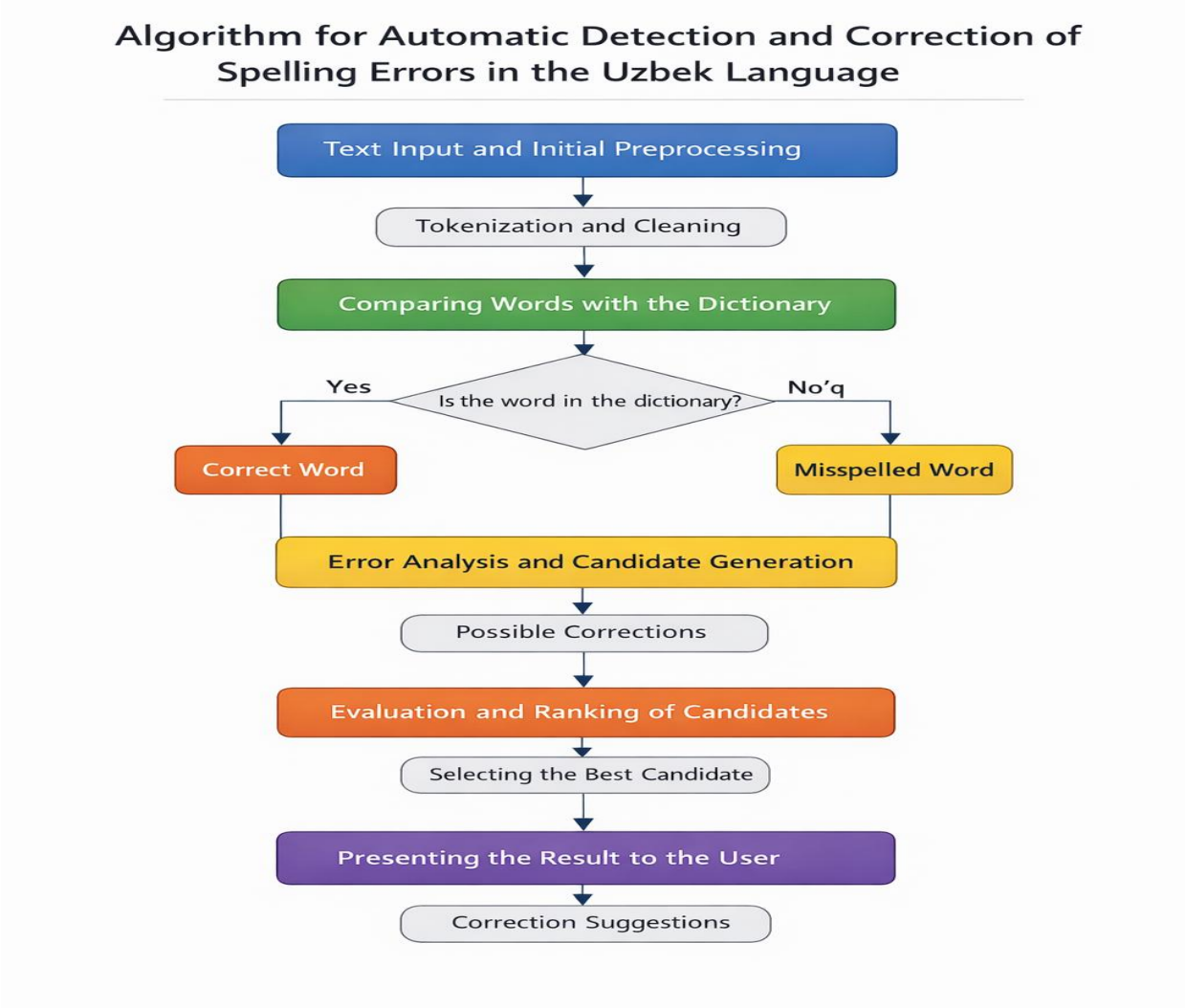
$$P(w_n \mid w_{n-1}, \ldots, w_1)$$

This model makes it possible to select the variant that best fits the sentence context instead of the incorrectly written word [2:56–60].

Neural Networks and Language Models

In recent years, the use of models such as LSTM, BiLSTM, and BERT has significantly increased. Experimental studies conducted for the Uzbek language confirm that the BiLSTM model achieves over 90% accuracy in spelling correction. In addition, POS tagging and morphological analysis play a crucial role in determining contextual appropriateness [6:210–215; 8:33].

Proposed Automatic Correction Algorithm



Research findings indicate the following performance levels:

Table 4. For agglutinative languages, a single algorithm is not sufficient; hybrid models yield the best results.

Method	Accuracy
Dictionary-based	60–70%
Levenshtein	75–85%
N-gram	80–88%
BiLSTM	≈90%

Conclusion

In conclusion, this study provides a comprehensive analysis of the automatic detection and correction of spelling errors in the Uzbek language based on modern Natural Language Processing approaches. The problem of automatic spelling correction in Uzbek represents a relevant and important research direction within NLP.

The results demonstrate that due to morphological complexity, simple dictionary-based systems are insufficient. The Levenshtein algorithm proves effective for error detection, while contextual language models significantly improve correction accuracy. A hybrid approach (combining dictionary-based methods, edit distance, and neural models) is shown to be the most optimal solution.

Furthermore, the study scientifically confirms that the agglutinative nature of the Uzbek language where words are formed through numerous suffixes limits the effectiveness of traditional spell-checking algorithms. Therefore, purely dictionary-based systems fail to provide sufficient results in real-world applications.

The analysis also shows that the process of spelling error correction consists of two main stages: error detection and context-based correction. These stages require the integration of various algorithmic approaches. While the Levenshtein distance algorithm demonstrates high efficiency in detecting typographical errors, statistical N-gram models and neural language models significantly improve correction accuracy by incorporating contextual information. In particular, context-aware models enriched with morphological analysis are highly important for adapting to the Uzbek language.

Based on the research findings, the most effective solution is a hybrid algorithmic model that integrates dictionary-based verification, edit distance, and probabilistic language modeling. This model not only detects misspelled words but also selects semantically appropriate corrections. Such an approach expands the possibilities of effectively using the Uzbek language in electronic document processing, educational platforms, search engines, and automatic speech processing systems.

For future research, it is recommended to develop large annotated corpora for the Uzbek language, localize transformer-based language models (such as BERT and GPT families), and design real-time spell-checking systems. These advancements will significantly enhance the functional capabilities of the Uzbek language in the digital environment and contribute to the development of national language technologies.

REFERENCES

[1] L. Bobojonova, A. Akhundjanova, P. Ostheimer, and S. Fellenz, “BERT-based part-of-speech tagging for Uzbek language,” arXiv, 2025.

[2] F. J. Damerau, “A technique for computer detection and correction of spelling errors,” Communications of the ACM, vol. 7, no. 3, pp. 171–176, 1964.

[3] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., 2023.

[4] C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999.

[5] K. Kukich, “Techniques for automatically correcting words in text,” ACM Computing Surveys, vol. 24, no. 4, pp. 377–439, 1992.

[6] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” Soviet Physics Doklady, vol. 10, no. 8, pp. 707–710, 1966.

[7] M. M. Ochilov, O. O. Narzullayev, and O. A. Xolmatov, “Mashinali o‘qitish algoritmlari asosida o‘zbek tili matnlaridagi imlo xatolarini aniqlash va tuzatish,” 2025.

[8] B. Elov and M. Ahmedova, “Development of a spell correction system based on N-grams,” 2025.

[9] U. Salaev, “UzMorphAnalyser: Morphological analysis model for Uzbek language,” arXiv, 2024.

[10] M. Minin, “Norvig and SymSpell spelling correction algorithms,” 2024.

[11] R. S. Madatovich and S. D. Maxamadiyevna, “The Role Of The System Of Education And Family Education In Forming Youth's World View,” European Journal of Humanities and Educational Advancements, vol. 4, no. 4, pp. 128–130.

[12] R. Madatovich, “The role of civic responsibility in educating youth in a healthy spiritual environment in an information society,” Pubmedia Social Sciences and Humanities, vol. 3, no. 1, pp. 6, 2025.

[13] R. S. Madatovich, "The role of preschool education and family education in the raising of a healthy balanced generation," For Teachers, vol. 57, no. 4, pp. 520–523, 2024.
[14] R. O'. SirojmuRODOV, "Yoshlarda sog'lom turmush tarzi rivojlanishida milliy va dunyoviy qadriyatlarni uyg'unlashtirishning ijtimoiy-falsafiy tahlil," ACTA NUUz, vol. 1, no. 1.10.1, pp. 184–186, 2024.
[15] S. Ruzimurodov and Sh. Artikov, "Anakharsis–velikiy filosof iz Centralnoy Azii," Innovatsii v tekhnologiyakh i obrazovanii, pp. 340–342, 2016.