ISSN: 3030-3346

A HYBRID METHOD BASED ON PROBABILISTIC MODELS FOR GRAMMATICAL ERROR CORRECTION IN UZBEK TEXTS

Narkhol Eshkaraeva[1], Farrukh Boymurodov[2]

1Professor of Karshi state university
2PhD in Technical Sciences, Karshi State University
E-mail: farruxboymurodov96@gmail.com
ORCID 0009-0000-9453-5273

KEYWORDS

natural language processing, grammatical error correction, computational linguistics, Bayes theorem, linguo-mathematical model, Uzbek language, hybrid algorithm

ABSTRACT

This paper investigates a hybrid approach based on the integration of probabilistic models and linguistic rules to improve the efficiency of automatic grammatical error detection and correction in Uzbek texts. Taking into account the morphological richness characteristic of agglutinative languages, an improved computational model of Bayes theorem is proposed for ranking candidate words during error correction. Using the developed linguo-mathematical model and algorithm, the text editing process analyzes not only the statistical probability of a word occurrence but also its contextual and morphological correctness. Experimental results demonstrate that the proposed hybrid method achieves higher accuracy compared to traditional purely statistical methods.

INTRODUCTION

Natural Language Processing (NLP) has recently witnessed monumental advancements across various global languages. However, developing robust tools for low-resource and morphologically rich languages, such as Uzbek, remains a significant challenge. Among these challenges, automatic Grammatical Error Correction stands out due to the complex nature of word formation and sentence structure. The Uzbek language belongs to the Turkic family and is characterized by its agglutinative morphology. In Uzbek, a single root word can generate hundreds of distinct word forms through the sequential attachment of various inflectional and derivational suffixes. Consequently, standard dictionary-based or purely rule-based validation methods suffer from severe limitations, such as an inability to handle out-of-vocabulary tokens or a failure to adapt to context-dependent syntax errors.

Conversely, purely statistical or probabilistic approaches, while flexible, frequently fail to capture the strict morphotactic constraints of the language, leading to grammatically invalid suggestions. To overcome these limitations, this paper proposes a hybrid approach that integrates an improved Bayesian computational model with strict linguistic constraints. By leveraging a refined probability calculation, the proposed model efficiently ranks candidate corrections, ensuring both statistical likelihood and grammatical correctness within the specific context.

METHODOLOGY

Natural Language Processing (NLP) technologies play a pivotal role across various domains, particularly in optimizing business processes. These technologies contribute significantly to the development of decision-support systems, enhance the efficiency of information processing, and streamline workflows through automation. One of the most widespread applications of NLP is automatic text classification. In the context of processing and preparing text for further structural analysis, the pipeline relies on several fundamental preprocessing stages [1]:

Tokenization: The process of segmenting a continuous text string into individual words, terms, or specific sequences of characters (tokens).

(cc) BY

Lemmatization: The morphological transformation of words into their canonical, base, or dictionary form (lemma).

Stemming: The computational reduction of words to their base stem or root form by stripping away inflectional and derivational affixes.

Part-of-Speech (POS) Tagging: The automatic assignment of grammatical categories (such as nouns, verbs, adjectives, etc.) to each token based on its syntactic role and context within the sentence.

Through these interconnected preprocessing stages, it becomes feasible to mathematically analyze textual data, determine structural regularities, and process text according to its underlying semantic properties. Consequently, these foundational NLP techniques are extensively applied across a wide range of tasks, including automatic machine translation systems, advanced text analytics, information indexing, and the development of intelligent conversational agents [2].

Let an erroneous word entered by a user be denoted as $\tilde{d}$, and its corresponding correct candidate word form be denoted as *d*. The primary task of the correction system is to identify candidates *d* that maximize the conditional probability $P(d \mid \tilde{d})$. In natural language processing, resolving grammatical and spelling errors focuses on determining the most statistically probable correct variant from the corrupted input string. This process is mathematically modeled using a probabilistic framework based on Bayes' theorem. For a given erroneous word $\tilde{d}$, the optimal correct word form *C* is determined by the following optimization formula:

$$C = \arg\max_{(k,l)\in A} P(d_k \mid \tilde{d}_l).$$

where:

A represents the vector of *(k, l)* pairs obtained from the minimal editing operations within a defined Damerau-Levenshtein distance threshold.

P(d) is the prior probability of the occurrence of the correct word *d* within the language corpus.

$P(d \mid \tilde{d})$ is the error channel probability, which reflects the likelihood that the correct word *d* is misspelled as $\tilde{d}$ by the user.

The primary scientific novelty of the proposed model lies in the integration of specialized linguistic criteria into the mapping process between the erroneous token $\tilde{d}$ and the correct candidate word *d*. In standard text processing frameworks, traditional edit distance algorithms-specifically the classical Damerau-Levenshtein distance-rely exclusively on four conventional editing operations to compute the transformation cost between two strings.

However, real-world digital texts written in the Uzbek language frequently exhibit a unique orthographic anomaly: the unintentional mixing of Cyrillic and Latin alphabets within a single token due to keyboard layout switching errors or visual homoglyph similarities. Since purely statistical models categorize these visual substitutions as standard character errors, they fail to assign appropriate weights to them, thereby distorting the candidate ranking process.

To resolve this limitation, this research enhances the classical metric by establishing a comprehensive framework based on five distinct computational edit operations specifically optimized for the Uzbek language ecosystem:

- Insertion: Captures instances where an extra, unintended character is added to a valid word structure (e.g., maktab => makhtab).
- Deletion: Models cases where a mandatory character is omitted from the word sequence (e.g., maktab => maktb).
- Substitution: Evaluates standard typographical replacement errors where an incorrect character within the same alphabet replaces the intended one (e.g., maktab => maktap).
- Transposition: Occurs during rapid typing when two neighboring letters accidentally swap their positions (e.g., maktab => matkab).
- Cross-alphabet visual character substitution: Introduced as the core architectural expansion in this study. This operation specifically flags and handles instances where characters from the Latin alphabet (Ω) are mistakenly interchanged with visually identical homoglyphs from the Cyrillic alphabet (G), or

(cc) BY

vice versa (e.g., replacing Latin ‘o’, ‘a’, ‘e’ with Cyrillic ‘o’, ‘a’, ‘e’, such as maktab => мaktab).

Once the five-operation distance metrics are established, the next phase of the proposed model involves calculating the error channel probability using a Naive Bayes classifier configuration. For any given observed erroneous token $\tilde{d}$ and a candidate correct word *d*, a sequence of optimal edit operations required to transform $\tilde{d}$ into *d* is determined via the enhanced Damerau-Levenshtein matrix.

This deterministic sequence of editing operations is mathematically designated as *f*. Under this probabilistic formulation, the likelihood of a user accidentally typing the corrupted string $\tilde{d}$ when intending to generate the correct word *c* is directly equivalent to the joint probability of executing the sequence of errors defined by *f*. Consequently, the transformation process between the structural pairs of $\tilde{d}$ and *d* is formalized as a finite set of sequential operations:

$$f = \{f_1, f_2, ..., f_n\},$$

Where each individual operation component f_k is drawn from the unified five-action operational pool optimized for the target language:

$$f_k \in \{qo'shish, o'chirish, almashtirish, o'rinlashtirish, o'xshashlik\}$$

Using this structured set, the conditional error channel probability for the specific sub-indices is formalized as follows:

$$P(\tilde{d}_l \mid d_k) = P(f \mid \tilde{d}_l \to d_k);$$

This mathematical formulation states that the conditional probability of encountering the misspelled string $\tilde{d}$ given the intended target *d* is computed as the probability of committing the sequence of modifications *f*. The operational components map precisely to character insertion, character deletion, character substitution, adjacent transposition, and cross-alphabet visual character substitutions.

To compute error probabilities effectively, any comprehensive lexicon database containing misspelled word forms paired with their ground-truth equivalents is sufficient. However, from a computational perspective, sequentially multiplying numerous conditional probabilities often introduces floating-point underflow problems, as the joint probabilities quickly approach zero. To simplify algorithmic execution and mitigate numerical underflow challenges, the mathematical framework is transformed into log-space.

By applying a logarithmic transformation, the computational operations shift from costly and numerically unstable multiplications to linear additions. Based on this approach, the complete mathematical formulation for the final ranking and selection model is derived as follows:

$$\hat{c} \cong \arg\max_{(k,l)\in A} P(d_k \mid \tilde{d}_l) \cong \arg\max_{(k,l)\in A} \frac{P(\tilde{d}_l \mid d_k)P(d_k)}{P(\tilde{d}_l)} \cong \arg\max_{(k,l)\in A} P(\tilde{d}_l \mid d_k)P(d_k) \cong$$

$$\cong \arg\max_{(k,l)\in A} P(d_k)\prod_{i=1}^{n} P(f_i) \cong \arg\max_{(k,l)\in A} \ln P(\tilde{d}_l) + \sum_{i=1}^{n_{\tilde{d}\to d}} \ln P(f_i^{\tilde{d}\to d}).$$

In this optimization framework, expressing the probabilistic values in a logarithmic format significantly optimizes the system's processing throughput. If a sufficiently large textual corpus is accessible within our database, the prior probabilities $P(d_k)$ of words occurring within the native language structure can be accurately computed.

Furthermore, if a structured training database consisting of known typographical variants is available, the corresponding error channel probabilities can be precisely estimated. Consequently, the combination of these two statistical resources forms a robust foundation for building the enhanced correction model based on the integrated Damerau-Levenshtein distance and Bayesian classification constraints.

DISCUSSION

To evaluate the practical efficacy of the proposed model, an extensive experimental evaluation was conducted. The underlying database developed within the scope of this research comprises approximately 1 million correctly spelled Uzbek words written in the Latin script. For every legitimate word form stored in

(cc) BY

this lexicon, a comprehensive set of potential typographical error variants was systematically generated and indexed. Specifically, the minimum number of simulated spelling fault variations mapped to each individual word is at least equal to the character length of that specific word token.

During real-time text processing, a misspelled word entered by a user can often be mapped to multiple valid dictionary candidates. The proposed system effectively generates a prioritized list of these legitimate alternatives, enabling the user to dynamically select the most contextually appropriate word. In modern computational linguistics, text analysis and automated editing represent foundational pillars closely tied to the evolution of digital word processors. The automated workflow designed in this study ensures that orthographic anomalies are instantly detected, corrected, and presented to the user as optimized suggestions. To achieve a comprehensive and high-quality editing framework, the system is designed to facilitate multi-level validation spanning orthographic, stylistic, grammatical, syntactic, semantic, and paradigmatic text layers.

To test the retrieval and ranking performance of the enhanced Bayesian and five-operation Damerau-Levenshtein metric, a computational query was executed using the corrupted input string “kitib”. The system searched the 1-million-word dictionary for correct candidate words within a maximum edit distance threshold of 2 (distance <= 2).

From a cognitive perspective, when a user types “kitib”, there is a high probability that the intended word was “kitob”, since the character ‘i’ is located directly to the left of ‘o’ on a standard QWERTY keyboard, making it a classic adjacent typographical slip. On average, computing the distance matrix, evaluating the linguistic constraints, and ranking the entire candidate pool for such queries requires less than one second of computational time.The generated alternative candidates, along with their respective log-space values *ln(P)*, raw probabilities *P*, and computed edit distances, are structured in Table 1.

Table 1.

Candidate word generation and probabilistic ranking metrics for the erroneous token “kitib”

Candidate Word	ln(P)	P	Distance
kitob	- 9.789339873	5.60E-05	1
kotib	- 10.77150552	2.10E-05	1
kutib	-13.7570807	1.06E-06	1
kiyib	-14.67258986	4.24E-07	1
kulib	-14.74590114	3.94E-07	2
ketib	-15.0648961	2.87E-07	2
kelib	-15.11324941	2.73E-07	2
qilib	-16.02821899	1.09E-07	2
qotib	-16.04261972	1.08E-07	2
ilib	-16.17396402	1.08E-07	2
isib	-16.45004986	7.18E-08	2
izib	-16.6357114	5.96E-08	2

As demonstrated in Table 1, the candidate “kitob” correctly achieves the highest mathematical priority ($ln(P) = -9.789339873$, $P = 5.6*10^{-5}$) at an edit distance of 1, closely followed by “kotib” and “kutib”. Candidates located at a distance of 2 are penalized accordingly by the error channel model, resulting in significantly lower prior probabilities.

This experimental setup proves that the log-space Bayesian classification framework successfully manages high-volume candidate evaluation without encountering floating-point errors, ensuring precise phonetic and structural ranking for digital Uzbek text inputs.

To assess the practical applicability of the developed hybrid model and algorithms, a dedicated software complex named *UZMORF* was

(cc) BY

successfully developed. The system is architected as an intelligent digital assistant designed to execute multi-level text processing, specifically incorporating functional modules for grammatical, orthographic, and punctuational validation.

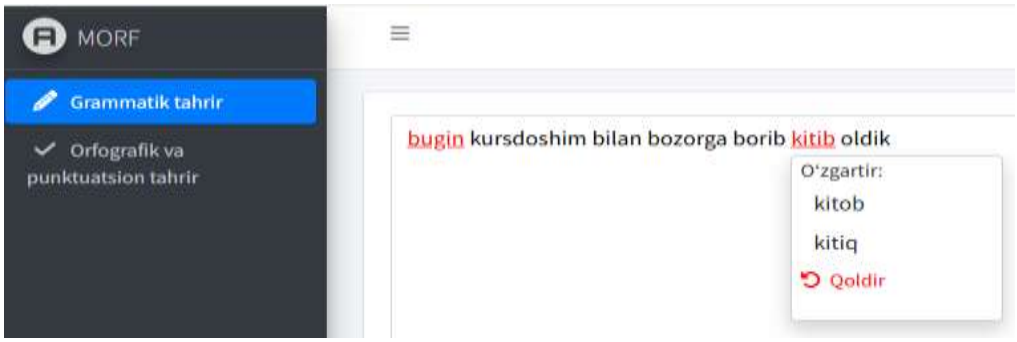


Figure 1. Illustrates the graphical user interface of the "UZMORF" system during a real-time spell-checking and editing session.

As demonstrated in Figure 1, the user inputs a sample conversational sentence: *"bugin kursdoshim bilan bozorga borib kitib oldik"*. The underlying linguistic and probabilistic engines instantly process the input string and flag typographical faults using visual indicators (red underlines). In this specific scenario, two distinct errors are successfully identified:

1. The word "bugin" is flagged due to an orthographic omission the missing apostrophe or vowel mutation typical in fast digital typing for the word "bugun".
2. The token "kitib" is detected as an out-of-vocabulary element that violates morphological rules.

Upon interacting with the flagged token "kitib", the system dynamically invokes the prioritized candidate ranking menu "O'zgartir:". Driven by the enhanced five-operation Damerau-Levenshtein distance matrix and log-space Bayesian scoring analyzed in the sections above, the software generates highly relevant contextual suggestions.

As established in the mathematical evaluation, "kitob" is retrieved and positioned at the top of the recommendation list as the most statistically and morphologically probable correction, followed by structural variants like "kitiq". Additionally, the interface provides a manual override action "Qoldir", allowing users to preserve the original token string if it represents a specialized proper noun or intentional colloquial variant.

This implementation validates that the theoretical linguo-mathematical models can be seamlessly embedded into intuitive, low-latency text-editing applications optimized for the Uzbek language infrastructure.

CONCLUSION

This research addresses a critical gap in computational linguistics for the Uzbek language by developing an enhanced, multi-level text-editing and spell-checking framework. By shifting away from rigid, language-agnostic text processing models, this study successfully integrates language-specific typographical and orthographic realities into a robust probabilistic architecture.

The primary scientific and practical conclusions derived from this research are as follows:

Architectural novelty: The classical Damerau-Levenshtein distance metric was successfully enhanced by expanding it from four to five distinct computational edit operations. The inclusion of a specialized cross-alphabet visual character substitution operation provides a native, low-overhead solution for resolving widespread mixed-script anomalies typical in digitized Uzbek texts.

Mathematical optimization: To resolve computational vulnerabilities associated with sequential probability multiplication - specifically floating-point underflow - the error channel calculations within the Naive Bayes classifier were mapped into a logarithmic space. Experimental testing proved that this transformation

(cc) BY

successfully replaces volatile multiplications with linear additions, ensuring high numerical stability.

Practical validation: The developed models were validated using a structured database comprising approximately 1 million valid Uzbek words alongside systematically generated error pairs. Computational test scenarios demonstrated that for a given erroneous input token, the system can dynamically retrieve, score, and rank complex candidate sets based on keyboard topology and language frequency priors in less than one second.

Software implementation: The theoretical algorithms were seamlessly deployed within a functional graphical application named *UZMORF*. The software interface demonstrates high real-time efficacy in visualizing orthographic, grammatical, and punctuational issues, offering prioritized, highly accurate recommendation menus that maximize user throughput and text-editing precision.

Future work will focus on scaling the *UZMORF* pipeline to handle context-aware semantic validation by integrating n-gram language models, thereby expanding the system's capacity to resolve syntax-level ambiguities and context-dependent homophone errors in continuous Uzbek discourse.

REFERENCE

1. Hutchins J. Machine Translation: History // In Encyclopedia of Language & Linguistics. Oxford: Elsevier, 2006. – pp. 375–383.
2. Jayanthi R., Jeevitha C. An Approach for Effective Text Pre-Processing Using Improved Porters Stemming Algorithm // IJISET. 2015. – Vol. 2, № 7. – pp. 797–807.
3. Эшкараева Н, Боймуродов Ф РАЗРАБОТКА ПРОГРАММУ СЛОВОФОРМ УЗБЕКСКОГО ЯЗЫКА С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМА ПОРТЕРА //International Scientific and Practical Conference on Algorithms and Current Problems of Programming. – 2023. – Т. 1. – №. 01.
4. Nazirova, E. S., Abidova, S. B., Xolmuminov, Y. X., Boymurodov, F. F., Medetbayeva, N. B., & Usmonova, K. A. Mathematical Model and Algorithm for Grammatical Editing of Given Words in Uzbek Language //International Seminar on Mathematics in Industry. – Cham : Springer Nature Switzerland, 2024. – С. 175-182.
5. Nazirova, E., Uzakova, M., Boymurodov, F., & Xolmo‘minov, Y. (2024). Turkiy tillar oilasiga mansub o ‘zbek va qirg ‘iz tillarida elektron tarjimaning lingvo-matematik modelini ishlab chiqish: Turkiy tillar oilasiga mansub o ‘zbek va qirg ‘iz tillarida elektron tarjimaning lingvo-matematik modelini ishlab chiqish //MODERN PROBLEMS AND PROSPECTS OF APPLIED MATHEMATICS. – 2024. – Т. 1. – №. 01.
6. Boymurodov F. Microsoft word uchun o ‘zbek tili so ‘zlarni imloviy tahrirlash plaginini yaratish algoritmlari va dasturi //DIGITAL TRANSFORMATION AND ARTIFICIAL INTELLIGENCE. – 2025. – Т. 3. – №. 1. – С. 39-44.
7. N.Eshqarayeva, Boymurodov F.F. Алгоритмы и программное обеспечение для автоматического редактирования орфографии узбекских слов. Ilim ha’m ja’miyet. Нукус. Нукус давлат педагогика институти. Ilmiy-uslubiy jurnal. Ajiniyoz Qo‘siboy o‘g‘lining 200 yilligiga bag‘ishlanadi. 2024 (№ 6/ qo‘shimcha). https://journal.ndpi.uz/uz/issues/2024-6-extra/read
8. Nazirova E.SH., Abidova SH.B., Uzakova M.A, Boymurodov F.F. Mathematical model for evaluation of trilingual electronic translation. Bulletin of TUIT: Management and Communication Technologies. ISSN:2181-1083 Date:29/12/2024 № 4(40)2024.
9. Nazirova E., Boymurodov F. O ‘ZBEK TILIDAGI DARAK GAPLARNI PUNKTUATSION XATOLARNI ANIQLASH VA TAHRIRLASH LINGVO-MATEMATIK MODELLARI //Techscience. uz-Texnika fanlarining dolzarb masalalari. – 2025. – Т. 3. – №. 8. – С. 28-36.
10. Nazirova, E., Abidova, S., Uzakova, M., Xolmuminov, Y., & Boymurodov, F. (2025, September). Linguistic-mathematical model of Kyrgyz and Karakalpak languages implementing bilingual electronic translation with feedback. In *AIP Conference Proceedings* (Vol. 3356, No. 1, p. 050001). AIP Publishing LLC.

(cc) BY