UDK 681.142.2:411(045)

DEVELOPMENT OF ALGORITHMS AND A SOFTWARE TOOL FOR PROPER NOUN RECOGNITION AND CORRECTION IN UZBEK LANGUAGE

Nazirova Elmira Shodmonovna
Tashkent University of information technologies named after Muhammad al-Khwarizmi, DcS, professor
E-mail: elmira_nazirova@mail.ru
Eshqarayeva Narxol G‘uzarovna
Karshi State University, PhD, associate professor
E-mail: 1970narxol@gmail.com
Boymurodov Farrux Farxod o‘g‘li
Karshi State University, doctoral student
E-mail: farruxboymurodov96@gmail.com

Abstract. In the era of rapid development of globalization and digital technologies, there is a growing need for tools that serve to automatically analyze, detect errors and edit texts created in the Uzbek language. In the field of natural language processing (NLP), one of the important issues is the correct identification of proper nouns and their editing. This article considers the development of algorithms and software designed to identify and automatically edit proper nouns in Uzbek texts. The proposed algorithms analyze lexical units in the text morphologically and syntactically and identify them based on grammatical features important for proper nouns. The article also discusses the capabilities of the created software tool, its operating principle, the methods used and the results.

Keywords. NLP, NER, spelling correction, grammar correction, orthographic editing, UZMORF.

Annotatsiya. Hozirgi kunda kompyuter lingvistikasi sohasida tabiiy tillarni raqamlashtirish jarayonlari jadal rivojlanayotgan davrda o‘zbek tilidagi matnlarni avtomatik tahlil qilish, xatolarni aniqlash va ularni tahrirlashga xizmat qiluvchi vositalarga ehtiyoj ortib bormoqda. Tabiiy tilni qayta ishlash (NLP) sohasida atoqli otlarni to‘g‘ri aniqlash va ularni tahrirlash muhim masalalardan biri hisoblanadi. Ushbu maqola NLP sohasi uchun muhim bo'lgan NER (Nomed Entity Recognition) xususiyatlari va aniqlash usullaridan o‘zbek tili matnlarida atoqli otlarni aniqlash va ularni avtomatik tahrirlashga mo‘ljallangan algoritmlar va dasturiy ta’minot ishlab chiqish masalasi ko‘rib chiqiladi. Taklif etilgan algoritmlar matndagi leksik birliklarni morfologik va sintaktik jihatdan tahlil qilib, atoqli otlar uchun muhim bo‘lgan grammatik belgilarga asoslangan holda ularni aniqlaydi.

ISSN 3030-3907

Shuningdek, maqolada yaratilgan dasturiy vosita imkoniyatlari, uning ishlash prinsipi, foydalanilgan metodlar hamda natijalari yoritiladi.

Kalit so‘zlar. NLP, NER, imlo tahrir, gramatik tahrir, orfografik tahrir, UZMORF.

Аннотация. В эпоху стремительного развития глобализации и цифровых технологий растет потребность в инструментах, которые служат для автоматического анализа, обнаружения ошибок и редактирования текстов, созданных на узбекском языке. В области обработки естественного языка (NLP) одним из важных вопросов является правильная идентификация имен собственных и их редактирование. В данной статье рассматривается разработка алгоритмов и программного обеспечения, предназначенных для идентификации и автоматического редактирования имен собственных в узбекских текстах. Предложенные алгоритмы анализируют лексические единицы в тексте морфологически и синтаксически и идентифицируют их на основе грамматических признаков, важных для имен собственных. В статье также рассматриваются возможности созданного программного инструмента, принцип его работы, используемые методы и полученные результаты.

Ключевые слова. NLP, NER, исправление орфографии, исправление грамматики, орфографическое редактирование, UZMORF.

Introduction. In the context of global digital transformation, Natural Language Processing (NLP) technologies are enabling in-depth study, automation, and effective use of the Uzbek language. Such technologies include automatic text analysis, error detection, translation, and voice and text interfaces. In particular, the correct spelling of proper nouns in official documents, scientific works, or media texts is one of the important criteria for determining the accuracy and reliability of the text [1].

Any geographical object that surrounds us has a name, and this “name” serves to distinguish objects from each other. For this reason, many words in the language have become names (proper nouns). The words “library”, “school”, “kindergarten” in our everyday life are also words denoting objects, but they are common nouns (appellative - Latin “proper noun”)). The concepts of common and proper nouns differ from each other according to the “naming” nature of the object they denote. That is, a named and unnamed object are distinguished by the expansion of content and the quality of renaming [4].

Due to the complex and rich morphological structure of the Uzbek language, automatic detection and editing of proper nouns is a rather complex task. Proper nouns differ from other words syntactically in their place in the text, and semantically in their content. Also, since they are usually written with a capital

ISSN 3030-3907

letter, and sometimes need to be determined from the context, their precise separation requires additional analysis.

This article considers the development of algorithms and software tools for automatic detection and editing of proper nouns in Uzbek texts. The software system is developed as a plug-in for the Microsoft Word environment, processes the text entered by the user in real time, detects proper nouns and provides recommendations for correcting errors in their writing. The research describes in detail the principle of operation of the algorithms, main modules, analysis methods, and software architecture.

Methodology. In this study, an algorithmic approach was developed to automatically detect and correct proper nouns in Uzbek texts. The system operates on the basis of a plug-in developed for the Microsoft Word environment and analyzes the text entered by the user in real time. The analysis process consists of several stages and is carried out as follows:

NE (Named Entity) – a named entity within the noun phrase – represents a "named noun". The process of automatically identifying named entities – NER (Named Entity Recognition) – is the recognition of named entities, identifying all named entities in the text. NER is used in many areas of Natural Language Processing (NLP). One of the NER objects is place names (toponyms).

Toponyms. The previously mentioned examples of common nouns such as *"kutubxona", "maktab",* and *"bog'cha"* represent objects associated with places. Upon hearing these words, one typically envisions physical entities categorized as "buildings." However, when such place-designating objects are renamed with specific titles (e.g., *"Amir Temur muzeyi", "Abdulla Oripov maktabi"* or *"Erkatoy bog'chasi"*) they are transformed into proper nouns and referred to as **toponyms**.

Toponyms, also known as **geographical names**, denote named locations or institutions and are classified as a subcategory of proper nouns. The academic study of toponyms falls under the discipline of **toponymy**. The term *toponymy* is derived from the Greek words *topos* meaning "place" and *onoma (or onima)* meaning "name."

Toponyms in Linguistic and Onomastic Studies. Toponyms constitute an integral part of a language's lexicon and function in accordance with the grammatical and lexical norms of the language system. In linguistics, toponyms are studied within the branch of **onomastics**, derived from the Greek *onomastike*, meaning "the art of naming."

Within the domain of toponymy, it is important to distinguish between three closely related terms:

ISSN 3030-3907

Toponymy (*Toponimika*) — the scientific discipline that studies place names, their structure, and the socio-historical processes through which geographical names emerge and evolve.

Toponymic index (*Toponimiya*) — a collected list or register of place names in a specific area (e.g., the Toponymy of Bukhara refers to the catalog of place names associated with Bukhara).

Toponym (*Toponim*) — an individual proper noun referring to a specific geographical location.

Classification of Toponyms

In the field of toponymy, place names are categorized based on the type of geographic or cultural object they represent. The main types of toponyms include:

- **Oikonyms** – place names referring to **settlements or inhabited localities**, such as cities, towns, or villages.
- **Hydronyms** – names of **bodies of water**, including rivers, lakes, seas, and other aquatic formations.
- **Oronyms** – place names denoting **landforms and surface reliefs**, such as mountains, hills, or valleys.
- **Ethnotoponyms** – names of places that originate from or include the names of **ethnic groups or nationalities**.
- **Anthrotoponyms** – place names derived from **personal names**, typically those of individuals or historical figures.
- **Dromonyms** – names of **roads, paths, and transport routes**.

Many scientists have been involved in the process of identifying NER objects. In particular, Rowen Brown mentioned in his articles about the methods of determining NER and anthology [9]. In his scientific works, Yujian Tang has given recommendations on the detection of NER objects through the Python programming language and the use of the SpaCy library [10].

The UZMORF program provides automatic analysis of proper nouns in texts, as well as detection and correction of grammatical and spelling errors. This plug-in, integrated for the Microsoft Word environment, provides users with efficiency and convenience in editing texts in the Uzbek language. Several basic methods are used in grammatical editing of Uzbek words. These methods help to ensure the correct formation, content, and compliance with the rules of the word during the editing process:

1. Orthography: The main thing in the editing of words is the orthography rules of the Uzbek language. This includes:
   - the correct placement and combination of letters.
   - the correct use of additional signs (apostrophes, dashes) and vowels.

ISSN 3030-3907

- capitalization of letters at the beginning of words.
- rules for correct hyphenation and addition of words.

2. Morphology: In grammatical editing, the morphological structure of the word is given special attention. This method includes:

- word formation, suffixes, and their compatibility.
- grammatical compatibility of the word form with other words in the sentence.
- word class (verb, noun, adjective, adverb, etc.) revisions.

3. Lexicology: The content and semantic aspects of words are also examined during the revision process. This method includes:

- determining the meaning of a word in context and translating it into a meaningful form.
- logical and aesthetic correctness of words or phrases.
- expressing the meaning clearly through the use of synonyms, antonyms, or synonyms.

The combined use of these methods ensures the grammatical and stylistic correctness of texts. Systems developed for editing Uzbek texts, including the UZMORF program, based on these methods, provide the ability to automatically analyze and correct words [11] .

The system database has so far included about 20,000 words belonging to proper nouns in the Uzbek Latin script. For each of these words, possible spelling error variants have been identified and added to the system. The probability of a misspelling for each word is considered to be at least equal to the number of letters in its composition, and all possible incorrect spelling forms have been added to the system. The text entered by the user is analyzed by the system. Spelling errors in the text are identified in accordance with the database entered into the system, and their correct forms are suggested [5]. In particular, there may be several correct variants of a word. The system, taking into account the context of the text, provides the user with the most suitable options. This methodology creates the opportunity to effectively identify and correct spelling errors. This approach provides a user-friendly and highly accurate editing system by deeply analyzing the features of the Uzbek language.

The user enters the text into the system for checking, and the system edits the text based on the following steps:

✓ the entered text is divided into words, the type of each word is determined, and suffixes are extracted from the words. This process is carried out to determine the morphological structure of words and make them easier to analyze.

ISSN 3030-3907

✓ each word is checked against the list of word roots available in the database. If the word is available in the database, it is left as is. Otherwise, matching options for this word are searched for among the possible spelling errors included in the database. If words matching the error are identified, their correct options are presented to the user in the form of a list.

✓ words with spelling errors are highlighted to the user. The user corrects these errors in accordance with his text, that is, choosing from among the correct options offered by the system and making the necessary changes.

This process allows the user to quickly identify and correct spelling errors in the text. These stages of the system are aimed at automating the text processing process and take into account the complex morphological structure of the Uzbek language.

A complex mathematical model is proposed based on all stages of morphological, syntactic and statistical analysis of the text, comparison with base words and automatic correction of misspelled proper nouns. This model covers the processes of identifying and editing proper nouns from Uzbek texts: for each word $w_i \in W$, apply a morphological parser:

$$\Theta(w_i) = (r_i, \sigma_i, \tau_i)$$

where:

- r_i: root of the word,
- σ_i: sequence of suffixes,
- τ_i: grammatical category (e.g., noun, verb, etc.).

$$P(y_i = 1 | w_i, r_i, \tau_i) = f_{ML}(x_i)$$

- $y_i \in \{0,1\}$: label indicating whether w_i is a proper noun,
- x_i: feature vector for w_i, including capitalization, position, context embeddings, morphological tags, and frequency.

To detect and correct errors in words that are not in the dictionary or that have just appeared, the program uses an artificial intelligence algorithm - "*spell checker*". This algorithm works on the following principles when detecting and editing word errors in texts:

1. *Analysis based on dictionary and grammatical rules*. Each word is checked according to the dictionaries and grammatical rules available in the program database. At this stage, the presence of the word itself or its morphological forms is determined.

2. *Detection of possible errors*. If the entered word is not found in the dictionary database, the algorithm determines possible error variants of the word. In this case, the following types of errors based on the letters of the word are considered:

- replacement of letters.

ISSN 3030-3907

- addition or omission of unnecessary letters.
- use of incorrect letters.

3. *Suggesting alternatives*. The algorithm suggests the correct word variants to the user based on the possible errors found. The suggestions are sorted based on the proximity of the words in the dictionary to the text, their compliance with the rules, and the level of probability.

4. *Creating convenience for the user*. The user is shown a list of suggested alternatives. The user can select the word that best fits his text and edit the errors. This process saves the user from retyping the text and increases efficiency.

The artificial intelligence algorithm, based on modern approaches to text processing, takes into account the linguistic features of the Uzbek language when implementing these processes. This makes the text analysis and editing processes more efficient and faster.

```
def base_check(word):
    con = sqlite3.connect('atoqli_otlar.db')
    cur = con.cursor()
    cur.execute("SELECT turi FROM atoqli_ot WHERE ot=?", (soz,))
    natija = cur.fetchone()
    con.close()
    return natija[0] if natija else None
```

This approach complements user input with correct words, reducing errors and increasing the accuracy of text analysis. The "*nltk scikit-learn*" Python package is designed to make this process efficient and convenient.

Results and discussion. Automatic editing and parsing is an important area of computational linguistics, and is closely related to the development of text editors, including Microsoft Word. Automatic editing and parsing analyzes text, identifies spelling and grammatical errors, automatically corrects them, and suggests alternatives to the user.

The effectiveness of an automatic text editing system depends on its ability to perform perfectly and with high quality. In this case, texts should be analyzed orthographically, stylistically and grammatically, and syntactic, semantic and paradigmatic approaches should be used in the analysis process. The process of automatic grammatical editing is based on the following stages:

- the entered text is divided into words, the types of words are determined and suffixes are extracted from the words. These suffixes are analyzed sequentially attached to the root of the word.
- each word is compared with the list of root words in the database. If the word is in the database, it is left as is. Otherwise, this word is compared with the list of

ISSN 3030-3907

grammatical errors, the corresponding errors are identified and the correct grammatical form of the word is suggested based on them.

- the user is shown words with grammatical errors. The system provides the user with the corrected text, and the user can also review the text further and make any necessary corrections.

The automatic editing system created as a result of this approach, when integrated with text editors, in particular with the Microsoft Word environment, significantly reduces the time users spend on correcting texts and increases the quality of texts. The algorithms used to detect and correct grammatical errors ultimately increase the accuracy of the system and provide effective service to the user through a user-friendly interface. This technology allows for further improvement and widespread use of text editors working in the Uzbek language.

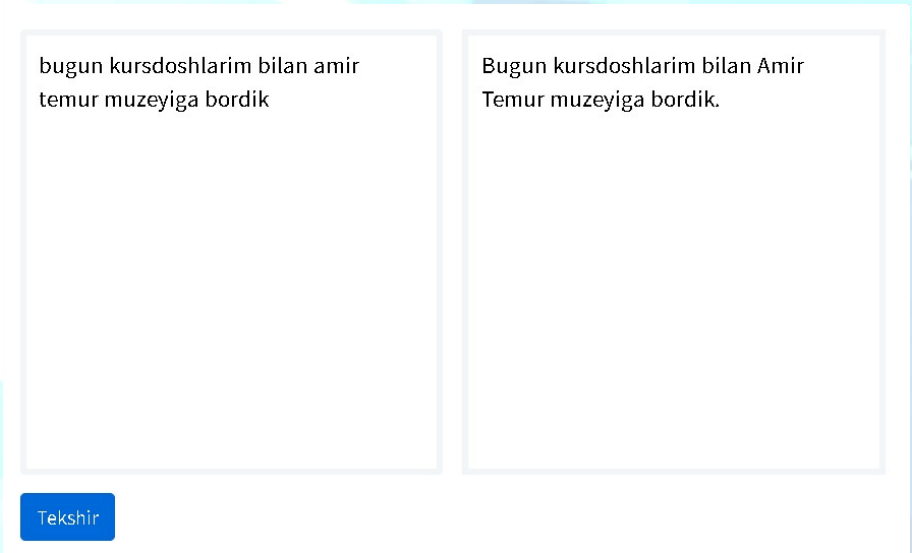


Figure 1. UZMORF program's detection and correction of a proper noun in grammatical editing.

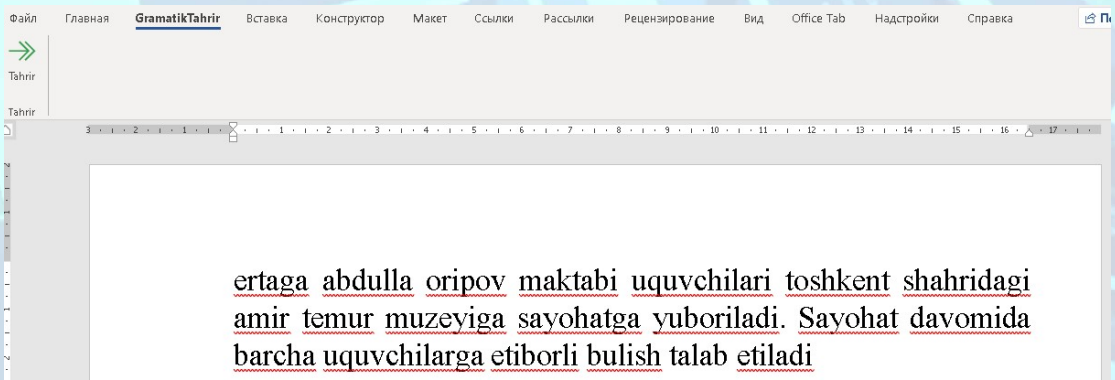


Figure 2. UZMORF program for identifying and correcting a proper noun in Microsoft Word.

ISSN 3030-3907

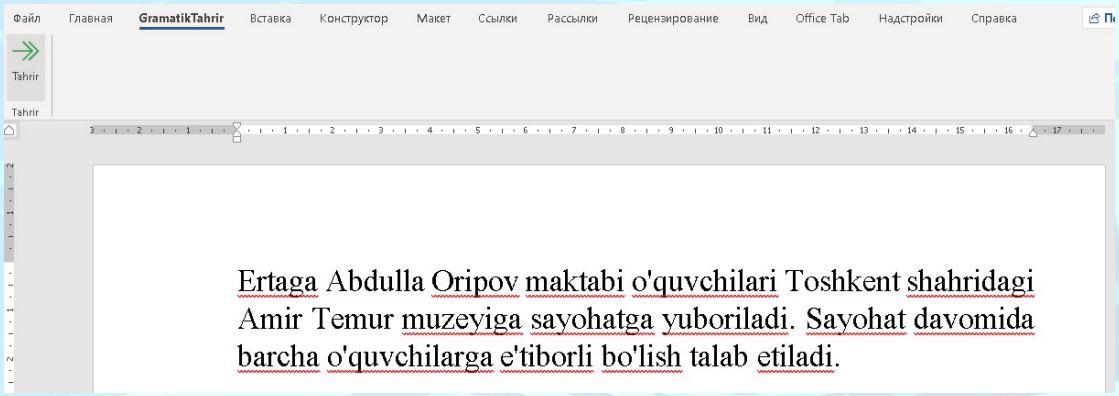


Figure 3. UZMORF program for identifying and correcting a proper noun in Microsoft Word.

This article analyzes the importance and capabilities of algorithms developed for automatic spelling editing of Uzbek words. The results of the study showed that the specific grammatical structure and word formation rules of the Uzbek language require a special approach for this language. The created algorithm and software, integrated with the Microsoft Word environment, allow you to automate the process of grammatical editing of Uzbek texts. The plug-in developed for Microsoft Word was implemented using *"macro"* technology to ensure convenient and efficient operation of the software. This program is fully compatible with the Word environment and allows users to analyze Uzbek texts, identify grammatical errors and show suggestions for their correction.

The following suggestions are made for further development and expansion of the software:

- expand platform integration: the program is not limited to the Microsoft Word environment, but can also be integrated with other text editors and social media platforms.
- improve macro technology: It is recommended to further expand macro-based functionality, for example, add quick analysis or auto-correction capabilities.
- add new features: In addition to spell checking, it is useful to add semantic analysis of the text and the ability to provide contextual suggestions to the user to the program.
- update the dictionary and algorithm: It is necessary to increase the accuracy of the system by regularly updating the program and adding new information on the dictionary and language rules.

Since the created program is fully compatible with the Microsoft Word environment, it provides maximum convenience for users and significantly simplifies the process of editing texts in the Uzbek language. The expansion and

ISSN 3030-3907

development of such projects will ensure the widespread use of the Uzbek language in the field of digital technologies.

REFERENCES

1. 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.
2. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / — М.: МИЭМ, 2011
3. Эшкараева Н. Г., Боймуродов Ф. Ф. Разработка и реализация алгоритма стемматизации и генерации словоформ для глаголов узбекского языка //ҚарДУ ХАБ. – С. 4.
4. Nazirova E, Uzakova M, Boymurodov F 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.
5. Nazirova E, Boymurodov F, Uzakova M O ‘zbek tili so ‘zlarni imloviy tahrirlash va qayta ishlash algoritmi: O ‘zbek tili so ‘zlarni imloviy tahrirlash va qayta ishlash algoritmi //MODERN PROBLEMS AND PROSPECTS OF APPLIED MATHEMATICS. – 2024. – Т. 1. – №. 01.
6. Boymurodov.F.F O ‘ZBEK TILI SO ‘ZLARNI HOSIL QILISH VA AVTOMATIK QAYTA ISHLASH TIZIMLARI TAHLILI: https://doi. org/10.5281/zenodo. 11198320 //PEDAGOGICAL REFORMS AND THEIR SOLUTIONS. – 2024. – Т. 1. – №. 1. – С. 495-498.
7. Boymurodov.F.F DEVELOPMENT OF AN ALGORITHM FOR GENERATING WORD FORMS FOR THE UZBEK LANGUAGE: https://doi. org/10.5281/zenodo. 11198320 //PEDAGOGICAL REFORMS AND THEIR SOLUTIONS. – 2024. – Т. 1. – №. 1. – С. 493-495.Боймуродов Ф. Ф. Разработка морфологического анализатора узбекского языка. – 2023.
8. Boymurodov F. ALGORITHM DEVELOPMENT AND IMPLEMENTATION STEMMATIZATION AND GENERATION OF WORD FORMS OF UZBEK LANGUAGE FOR AUTOMATIC PROCESSING SYSTEMS TEXTS //Конференции. – 2020.
9. Roger Brown, Cogito Tech LLC, Medium. How Does Named Entity Recognition Work: NER Methods.

ISSN 3030-3907

10. Yujian Tang, Dev Genius. The Best Way to do Named Entity Recognition (NER).
11. Boymurodov.F. MICROSOFT WORD UCHUN O ‘ZBEK TILI SO ‘ZLARNI IMLOVIY TAHRIRLASH PLAGININI YARATISH ALGORITMLARI VA DASTURI //DIGITAL TRANSFORMATION AND ARTIFICIAL INTELLIGENCE. – 2025. – T. 3. – №. 1. – C. 39-44.
12. Nazirova E., Abidova S., Uzakova, M. (2024, November). Mathematical models of complex word formation in Uzbek, Karakalpak and Kyrgyz languages belonging to the Turkic language family. In AIP Conference Proceedings (Vol. 3244, No. 1). AIP Publishing.

ISSN 3030-3907