IDENTIFICATION OF SPEECH ERRORS IN THE UZBEK LANGUAGE USING ARTIFICIAL INTELLIGENCE AND THEIR LINGUISTIC CLASSIFICATION

Dushamova Nodirabegim – Independent Researcher

Abstract. This article analyzes the possibilities of automatically identifying, linguistically classifying, and correcting speech errors in the Uzbek language using artificial intelligence technologies. The agglutinative nature of Uzbek, the diversity of morphological forms, flexible word order, lexical-semantic relations, and pragmatic factors associated with speech situations are examined as key factors influencing the error detection process. The article also discusses the application of rule-based algorithms, statistical models, and neural networks to the identification of spelling, morphological, syntactic, lexical-semantic, stylistic, and pragmatic errors. Within the scope of the study, an approach based on linguistic criteria is proposed for the systematic classification and automatic analysis of speech errors in Uzbek.

Keywords: artificial intelligence, Uzbek language, speech errors, computational linguistics, natural language processing, automatic analysis, morphological error, syntactic error, semantic analysis, pragmatics.

Today, the development of artificial intelligence technologies is expanding the possibilities of language learning, text editing, machine translation, and human speech analysis. Natural language processing technologies play an important role in detecting grammatical and spelling errors in texts, analyzing sentence structures, and evaluating textual meaning. At the same time, the distinctive grammatical and semantic characteristics of the Uzbek language require a specialized approach to developing automatic error detection systems. The concept of a speech error encompasses phenomena in which linguistic units fail to conform to literary language norms, grammatical rules, or the requirements of a communicative situation. In written texts, such errors may appear as spelling and grammatical deviations, whereas in spoken language, they may manifest as

pronunciation problems, inappropriate word choice, incorrect grammatical forms, or difficulties in expressing ideas.

Identifying speech errors using artificial intelligence involves more than simply detecting misspelled words. A system must also determine the linguistic level to which an error belongs, identify the factors contributing to its occurrence, and establish how it can be corrected in context. For example, the sentence “Men maktabga bordim kecha” (“I went to school yesterday”) may be grammatically acceptable in certain contexts. However, in neutral written discourse, the word order “Men kecha maktabga bordim” (“I went to school yesterday”) may sound more natural stylistically. Therefore, it is incorrect to classify every nonstandard expression as an error.

Uzbek belongs to the Turkic language family and is distinguished by its agglutinative structure. Grammatical and new lexical meanings are formed by attaching various suffixes to word stems in a specific sequence. Consequently, correctly identifying word stems, suffix order, and grammatical functions is essential for automatic linguistic analysis. For example, the word “kitoblarimizdan” (“from our books”) combines the meanings of a root, plurality, possession, and case. The incorrect use of suffixes within this structure may result in a morphological error. However, a system should not automatically classify a rarely occurring word form as incorrect, since it may be a correctly formed word that appears infrequently in the corpus.

Another factor complicating error detection in Uzbek is the parallel use of the Latin and Cyrillic scripts. In addition, variations in the use of apostrophes and other diacritical marks, alternative spellings of borrowed words, dialectal forms, and informal online communication may affect the accuracy of automatic checking systems. When analyzing speech errors, it is necessary to consider standard language rules, discourse style, and the communicative situation simultaneously. This is because certain forms that are natural in everyday conversation may be inappropriate in official documents or academic texts.

The linguistic classification of speech errors by linguistic level is essential for developing an automatic error detection system for Uzbek. The following classification is proposed for practical analysis.

Spelling errors occur when established spelling conventions are violated. These include omitting letters, inserting unnecessary letters, substituting one letter for another, and violating rules governing compound and separate word spelling.

Examples:

“kitobhon” → “kitobxon” (correct spelling of “book lover” or “reader”);

“ma’sul” → “mas’ul” (correct spelling of “responsible”);

Ensuring that the words in the phrase “ma’lumotlar bazasi” (“database”) are written according to spelling conventions and separated appropriately.

Artificial intelligence can identify spelling errors using dictionaries, character-sequence comparison, edit-distance algorithms, and contextual language models. In this process, both the form of a word and its meaning within the sentence must be taken into account.

Morphological errors are associated with the incorrect formation or use of grammatical word forms. They may occur in case, possession, plurality, tense, person-number agreement, and other grammatical categories.

Examples:

“Men kitobni o‘qidi” → “Men kitobni o‘qidim” (“I read the book”);

“Biz maktabga bordim” → “Biz maktabga bordik” (“We went to school”).

To identify such errors, a system must check person-number agreement between the subject and predicate, verb forms, and the grammatical functions of suffixes. Given the agglutinative nature of Uzbek, morphological analysis should not rely exclusively on pre-

existing dictionary forms. Segmenting word structures and identifying grammatical features can improve the reliability of error detection.

Syntactic errors arise when grammatical relationships between sentence constituents are violated or when sentence structure interferes with the intended meaning.

Example:

“Men kitob o‘qiyapti” → “Men kitob o‘qiyapman” (“I am reading a book”).

In the first sentence, person-number agreement between the subject and predicate is incorrect. An automatic system can identify this error by combining syntactic analysis with the comparison of morphological features.

Nevertheless, word order in Uzbek may vary according to communicative purpose. Therefore, a sentence should not automatically be classified as syntactically incorrect merely because its word order differs from the usual pattern. A system must distinguish grammatically acceptable alternatives from stylistic preferences.

Lexical-semantic errors occur when a word is selected incorrectly, semantically incompatible units are combined, or an expression inappropriate to the context is used. For example, in the sentence “U baland ovozda kitobni ko‘rdi” (“He/she saw the book in a loud voice”), the phrase “baland ovozda” (“in a loud voice”) is semantically incompatible with the verb “ko‘rdi” (“saw”). Depending on the author's intended meaning, alternatives such as “U kitobni diqqat bilan ko‘rdi” (“He/she looked at the book carefully”) or “U kitobni baland ovozda o‘qidi” (“He/she read the book aloud”) may be more appropriate.

Identifying such errors is more complex than detecting spelling errors because every word in a sentence may be correctly spelled when considered individually. The system must analyze semantic relationships between words, the semantic requirements of verbs, and the overall context.

Stylistic errors are associated with the inappropriate use of linguistic units in relation to the text type, communicative purpose, or requirements of a particular literary style. For

example, excessively colloquial expressions in an academic article or vague, emotionally charged language in an official document may negatively affect text quality. Artificial intelligence can help address these issues by identifying text styles, evaluating word choice, and suggesting alternative expressions. However, stylistic recommendations should not be presented as absolute grammatical rules.

Pragmatic errors are characterized by the inappropriate use of linguistic forms in relation to a particular situation, the relationship between interlocutors, or the communicative purpose. For example, in a formal situation, using the informal Uzbek pronoun “sen” instead of the respectful “Siz” may violate communicative conventions. However, the use of “sen” may be entirely natural in a conversation between friends. Therefore, a system must consider conversational context, the participants in the interaction, and the purpose of communication when identifying pragmatic errors. Such errors cannot be fully identified through grammatical rules alone.

Several technological approaches can be used to automatically identify speech errors in Uzbek.

The first approach involves rule-based systems. These systems operate using spelling dictionaries, grammatical rules, and predefined patterns. They can be effective in detecting clear grammatical errors but have limitations when analyzing complex contexts and stylistic variations.

The second approach involves statistical and probabilistic models. This method considers the frequency of words and grammatical forms in a corpus, as well as their relationships with the surrounding context. However, there is a risk of classifying rare but correct forms as errors.

The third approach involves neural networks and contextual language models. These models can learn the overall meaning of sentences, relationships between words, and long-distance grammatical dependencies. When trained on large amounts of high-quality data, they can improve the detection of complex errors.

The fourth approach involves hybrid systems. These systems combine linguistic rules, morphological analysis, dictionaries, and neural models. This approach is promising for Uzbek because it incorporates grammatical knowledge while retaining the benefits of contextual analysis.

The identification and linguistic classification of speech errors in Uzbek using artificial intelligence constitute an important area of contemporary computational linguistics. Classifying speech errors into spelling, morphological, syntactic, lexical-semantic, stylistic, and pragmatic categories helps define the functions of automatic analysis systems more precisely. The agglutinative structure of Uzbek, the diversity of grammatical forms, variations in writing systems, and context-dependent semantic features require specialized approaches to error detection. Therefore, combining rule-based methods with contextual neural models and linguistic knowledge is advisable.

The practical significance of this study lies in the potential application of the proposed classification and methodology to the development of systems for automatic proofreading, editing, and linguistic analysis of Uzbek texts. In the future, creating high-quality annotated corpora, accounting for dialectal and stylistic variations, and testing models in real communicative situations will contribute to the development of this field.

References

1. Hojiyev, A. *Tilshunoslik terminlarining izohli lug‘ati* [Explanatory Dictionary of Linguistic Terms]. Tashkent: O‘zbekiston Milliy Ensiklopediyasi, 2002.
2. Ne’matov, H., and Rasulov, R. *O‘zbek tili sistem leksikologiyasi asoslari* [Fundamentals of Systemic Lexicology of the Uzbek Language]. Tashkent: O‘qituvchi, 1995.
3. Sayfullayeva, R., et al. *Hozirgi o‘zbek adabiy tili* [Modern Uzbek Literary Language]. Tashkent: Fan va texnologiya, 2009.
4. Abdujalilova, Sh. A. *O‘zbek tili kompyuter lingvistikasi asoslari* [Fundamentals of Uzbek Computational Linguistics]. Tashkent: Noshir, 2021.