ISBN 978-989-9121-45-4
58th ISTANBUL International Conference on Research in "Science, Engineering and Technology" (IRSET-24)
Dec. 19-20, 2024 Istanbul (Turkiye)

Identification of Errors in the Kazakh Language: Approaches to Processing Semi-Structured Texts

S.A. Tussupova[1], L.M. Baitenova[1], D. Rakhimova[1,2] , K.A. Tussupov[1] and A. Turarbek[1,2]

[1]"Turan" University, Almaty, Kazakhstan
[2]Al Farabi Kazakh National University, Almaty, Kazakhstan

Abstract: *The rapid expansion of digital content on the Internet and social networks has underscored the significance of computational linguistics for processing, analyzing, and validating natural language data. This domain supports key technologies like information retrieval, dialogue systems, and machine translation, all of which depend on precise text processing. Among the diverse tasks in natural language processing (NLP), error detection and correction—especially in identifying and rectifying incorrect words—is essential for ensuring language accuracy.*

The study offers an overview of semi-structured data, methods, and technologies for identifying incorrect words in natural languages. It also compares text-checking and correction systems used in social networks with technologies designed for detecting incorrect words. Additionally, the paper presents an approach for identifying incorrect Kazakh words, and analyzes the features and capabilities of this approach.

Keywords*: Kazakh Language, Semi-Structured Data, Text Processing, Error Detection, Natural Language Processing and Social Networks*

1. Introduction

The immense flow of information on the Internet and social networks has significantly contributed to the rapid development of natural language processing and computational linguistics. Currently, various research mechanisms support projects aimed at information exchange, machine translation, email verification, and the development of question-answering systems among users [1]. Error detection and correction in texts and words remains one of the primary tasks in natural language processing. Over the past half-century, this topic has maintained its relevance, with new methods emerging and applications expanding.

Applications such as Instagram, VKontakte, Facebook, and other social networks are highly appealing for processing and analyzing information due to the real-time and dynamic nature of the content generated on these platforms [2]. However, texts on the Internet often deviate from conventional language norms, leading to various types of errors caused by intentional word distortions [3]. These errors complicate the readability and processing of texts. Natural language processing requires standard word patterns, as spelling errors or digitization issues reduce the informational value of texts. For instance, a spelling error in medical records can hinder diagnostic processes, while errors in online communication may negatively affect research or organizational activities [4].

As the Kazakh language belongs to the group of low-resource languages, it faces a lack of translation systems, dictionaries, corpora (multilingual or bilingual), and tools for detecting and correcting errors. Thus, developing programs and systems to identify orthographic errors in resource-constrained languages like Kazakh has become crucial.

Semi-structured data refers to data that does not adhere to the strict structure of relational database models, such as tables and relationships. Information on the Internet is not always specific to a particular domain. Consequently, many organizations and researchers are developing specific algorithms to construct text structures unrelated to the educational field [5].

Semi-structured data has become a significant object of study as it serves as a linking format (e.g., JSON, XML) between full-text documents and databases, essential for Internet development. Examples of systems containing semi-structured data include user comments, posts, and texts found on websites and social networks

https://doi.org/10.17758/DIRPUB16.DiR1224135

[6]. Data extracted from such systems is of great interest for research and applications, enabling real-time sentiment analysis and contributing to the spread of information. Additionally, it helps reshape public perspectives on business, politics, and social systems. Each type of data has unique features that must be considered during data collection, preparation, preprocessing, and object description.

This study utilized information from the Internet and social networks, which, as described above, is semi-structuredIand was applied practically during the research.

The challenges of detecting and correcting orthographic errors in texts date back to the 1960s and continue to this day. Efforts to enhance quality and productivity and expand potential applications provide strong justification for ongoing research in this area. Although system-level programs (e.g., processors) have become more complex, they still fail to assist users in correcting numerous evident spelling errors in input data sources [7]. Over 50 years of addressing the issue of error detection and correction, researchers have tested various methods, ranging from character codes, n-gram recognition tables, and direct application of the Damerau-Levenshtein distance to incorporating phonetic information and machine learning approaches in error detection systems. However, constructing error detection and correction systems still faces fundamental challenges, including compact dictionary storage, efficient morphological and syntactic analysis, and developing scientific editor systems for technical and literary works [8].

Popular text correction systems for English include Grammarly, Grammarchecker, and ReversoSpeller, while Russian systems include Orfogramma, Advego, ORFO, and LINAR. For agglutinative languages such as Turkish or Kyrgyz, systems like the MS Word spell checker are available. Unfortunately, these systems are unsuitable for Kazakh. Moreover, publicly available analogs for Kazakh are nonexistent.

During the research and analysis, texts of various styles, including content from the Internet and social networks, were considered. Additionally, text checking and correction systems for English and Russian were analyzed to identify their strengths and weaknesses. Table I summarizes the comparative characteristics of these systems.

An analysis of widely-used text correction systems revealed a significant limitation: these systems are unsuitable for the Kazakh language due to its agglutinative nature and complex morphological and lexical structures [1].

To create an effective system for identifying and correcting errors in the Kazakh language, it is essential to consider its unique characteristics. As an agglutinative language, Kazakh features complex morphological and syntactic rules, with sentence semantics playing a central role.

As a result of this research, an electronic dictionary for the Kazakh language and a system capable of checking the accuracy of Kazakh texts at an industrial level were developed. However, systems for evaluating the accuracy of semi-structured texts in Kazakh remain inaccessible, and even commercial software for this purpose is challenging to find online [9].

Orthographic errors generally fall into two categories: *typographical* and *cognitive*. Cognitive errors occur when a word is not in the dictionary, often due to phonetic or orthographic similarity between words, and typically arise when a person is unsure of the correct spelling. Typographical errors, on the other hand, result from keyboard input issues, such as pressing adjacent letter keys by mistake.

Beyond these, additional types of errors have been identified, especially in semi-structured data, highlighting the complexity of error types in modern text analysis [10, 11].

A variety of error types exists in semi-structured data, particularly in Kazakh, where common errors include:

- Typographical errors: kitap becomes kiap
- Spelling mistakes: muhit becomes mýhit
- Intentional word distortions: algaaa
- Grammatical and punctuation errors
- Non-standard alphabets: using Russian or Latin scripts
- Abbreviations and slang expressions

TABLE I. Comparative Characteristics of Text Checking and Correction Systems for English and Russian Languages

Text Checking and Correction Systems	*Disadvantages*	*Advantages*	*Price*
Advego	*Does not check punctuation*	*Spelling checks. Detects missing or extra letters, spaces. Advanced SEO features (stop words, readability, word/character count).* *Supports large volumes (up to 100,000 characters). Available in 20 languages.*	*Free*
LanguageTool	*Does not check punctuation*	*Detects grammatical and stylistic errors. Checks punctuation. Integrates with text editors and browsers. Supports up to 30 languages. Provides correction suggestions.*	*Free/Paid.*
Istio	*Does not check punctuation*	*Spelling checks. Inspects web pages. Unlimited text checking. SEO text analysis. Suggests replacements.*	*Free*
Orfogramka - https://orfogramka.ru/	*Only available as a paid version.*	*Checks spelling, punctuation, style, typography, and semantics. Identifies tautologies and cacophony. Corrects errors with explanations. Offers SEO analysis. Regularly updated dictionary.*	*Paid*
Text.ru	*For unregistered users, there is a long waiting time for verification. The free version limits articles to a maximum of 15,000 characters. The dictionary is incomplete.*	*Checks spelling and grammar. Analyzes text uniqueness. Identifies incorrect usage of case, parentheses, spaces, and repetitions.* *Provides SEO parameters (e.g., readability, spam levels, character count).* *Offers error replacement suggestions.*	*Free/Paid.*
Orfograf	*No correction suggestions.* *Minimal functionality.* *Does not check punctuation.* *Limited dictionary.*	*Checks spelling.* *Identifies errors in web page content.* *Customizable marker design.*	*Free.*
ORFO - https://online.orfo.ru/	*Performs well with individual words but struggles with large texts, missing some errors. Does not check punctuation.*	*Checks spelling. Supports 50 languages.* *Analyzes uniqueness and other SEO metrics.* *Capable of checking individual web pages and entire websites.*	*Free/Paid.*

2. Research Methodology

Detecting orthographic errors is closely related to identifying inaccuracies in words or sentences. The primary method involves dictionary-based spelling checks, which compare each word in a text against entries in a dictionary. Words not found in the dictionary are flagged as potentially erroneous. Common methods for this include: *n-gram algorithms:* identify patterns within sequences of letters; *morphological analysis:* break down words into roots and affixes; *machine learning algorithms:* use trained models to predict and identify errors.

Hybrid methods combining multiple approaches are frequently employed for greater accuracy [12].

A dictionary typically consists of all the words in the Kazakh language, listed alphabetically, with each word on a new line. This method is among the most widely used for detecting errors. If all the letters in a word match an entry in the dictionary, the word is deemed correct; otherwise, it is flagged as incorrect [13]. Rule-based dictionaries also verify that words comply with specific language rules.

Another approach involves detecting errors without dictionaries, focusing instead on: **capitalization rules:** ensuring that letters following a period are capitalized; **repetition checks:** flagging exact matches between repeated words as potential errors; **n-gram analysis:** analyzes letter sequences to detect unusual or missing patterns, marking words with such anomalies as erroneous.

Unlike dictionary-based methods, n-gram analysis is language-independent and does not require specific linguistic knowledge [14, 15].

Further research into error-detection models has led to the development of methods tailored to identifying incorrect words in the Kazakh language. Figure 1 illustrates the convergence approach used for error detection.

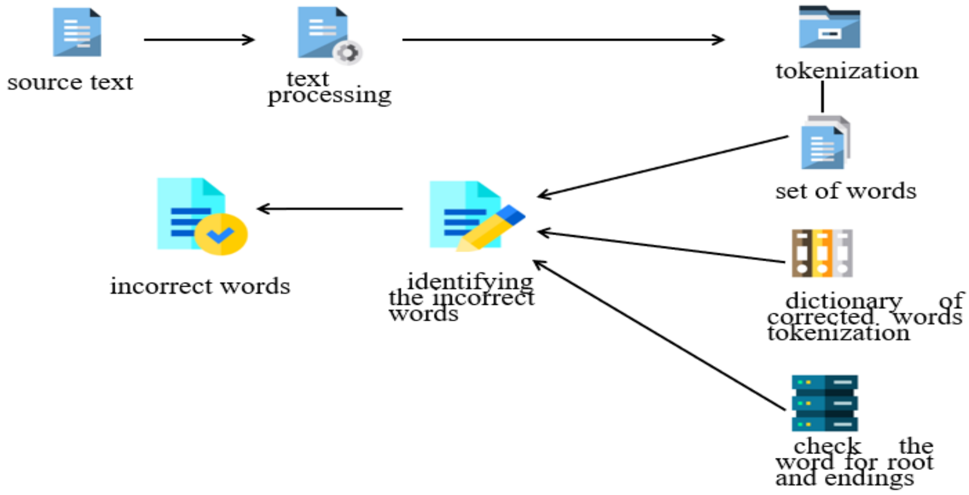

source text
text processing
tokenization
set of words
incorrect words
identifying the incorrect words
dictionary of corrected words tokenization
check the word for root and endings

Fig. 1. Diagram of the Methodology for Identifying Incorrect Words

In the described method, text is first collected from semi-structured data sources and then preprocessed. The text is divided into sentences, which are further split into individual words, creating a set of words for analysis.

Apertium is a free, open-source rule-based platform for machine translation [10, 18]. It comprises program modules (aligned with specific processing stages), lexical dictionaries, and rule-based dictionaries.

Program Modules (Aligned with Processing Stages):

1. **De-formatter:** Separates the translatable text from formatting elements.
2. **Morphological Analyzer:** Splits text into lexemes and identifies the lexical forms (dictionary forms) for each lexeme.
3. **POS Tagger (Part-of-Speech Identification):** Resolves morphological ambiguities and identifies the correct part of speech. For words with multiple lexical forms, the most appropriate is selected using a combination of handwritten rules and Hidden Markov Models.
4. **Lexical Transfer:** Maps each lexical form in the source language to its corresponding form in the target language using a finite state machine built from bilingual dictionaries.
5. **Lexical Selection:** For words with multiple possible translations, the most contextually appropriate one is chosen using rule-based finite state machines.
6. **Structural Transfer:** Groups lexical forms into segments based on sentence patterns and applies pattern-action rules. Processing is performed from left to right, using the longest applicable rule first.

6.1 **Chunker:** Segments parts of the sentence into manageable "chunks."

6.2 **Interchunk:** Adjusts or reorders these chunks as necessary.

6.3 **Post-chunk:** Processes modified chunks and arranges words in the final output format accepted by the generator.

7. **Morphological Generator:** Produces the correct morphological form for each lexical unit in the translated sentence using a finite state machine derived from a morphological dictionary.
8. **Postgenerator:** Executes orthographic operations, such as: contractions (merging adjacent vowels into a single vowel or diphthong); elisions (omitting sounds to ease pronunciation); inserting apostrophes where needed.
9. **Reformatter:** Restores formatting removed during the initial de-formatting step.

Morphological Dictionary for the Kazakh Language (File name: *apertium-kaz.kaz.lexc*): The lexical dictionary for Kazakh is located within the Apertium project repository (*apertium-kaz*). This dictionary includes Kazakh words with syntactic annotations corresponding to their roles in a sentence [10].

N1-ABBR Lexicon:

%<n%>%<attr%>: # ;

%<n%>:% – POSSESSIVES ;

%<n%>:% – CASES-ETC ;

%<n%>: CASES-ETC ; ! Director/LR

%<n%> is a noun, and %<attr%> represents the abbreviation of its attributes. For example:

Kazu

ABBR ; ! "Al-Farabi Kazakh National University"

N1 Lexicon:

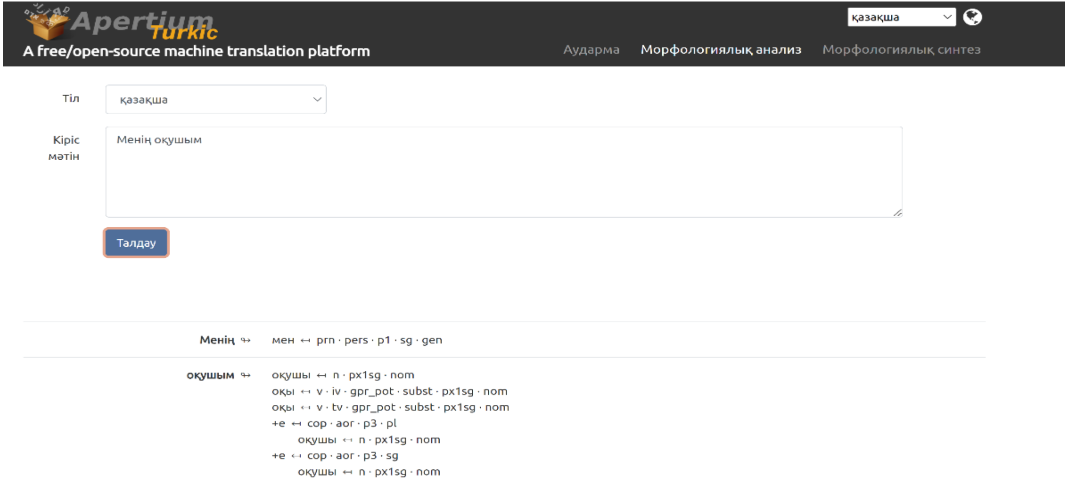
%<n%>%<attr%>: # ;

%<n%>: FULL-NOMINAL-INFLECTION ;

Here, %<attr%> refers to an attribute, which is a distinct feature, inherent property, or integral part. For example:

ағаш:ағаш N1 ; ! " ағаш "

аға: аға N1 ; ! " аға "



Тіл: қазақша

Кіріс мәтін: Менің оқушым

```
Менің ↬   мен ↤ prn · pers · p1 · sg · gen

оқушым ↬  оқушы ↤ n · px1sg · nom
          оқы ↤ v · iv · gpr_pot · subst · px1sg · nom
          оқы ↤ v · tv · gpr_pot · subst · px1sg · nom
          +е ↤ cop · aor · p3 · pl
              оқушы ↤ n · px1sg · nom
          +е ↤ cop · aor · p3 · sg
              оқушы ↤ n · px1sg · nom
```

Fig. 2. Example of Morphological Analysis of Kazakh Text Using the Apertium Platform

2.1. Research Results

During the research, a program for constructing a corpus in the Kazakh language was developed. Using Python code implemented in the Google Colab environment, a corpus containing 120,198 sentences in Kazakh was created. Python libraries were utilized to collect the corpus from websites. The collected dataset was analyzed using scripts based on the HTML structure of the websites.

Statistical calculations were conducted on the compiled corpus, individual words were analyzed, and a dictionary of correct words was created to supplement the data. Table 2 outlines the input data sources used to implement the program.

TABLE II: Input Data Sources for Program Implementation

Corpus Sources	*Number of Sentences*
akorda.kz	*93847*
nur.kz	*26351*

2.2. Challenges in Corpus Construction

During the construction of the corpus, one of the main challenges was the need for preprocessing the data. This was necessary due to the varied formats of information available on the websites. Tasks included removing unnecessary symbols, correcting sentences, and splitting them into individual words. These steps were time-consuming but essential for ensuring data consistency.

For experimentation, a publicly available program described in [10, 16] was used. Additionally, linguistic resources for the Kazakh language, such as a stop-word dictionary and a complete set of affixes, were employed [17]. The program underwent multiple tests, with results shown in Table 3.

TALE III: Experimental Results of the Algorithm

Number of Words Checked	*Accuracy (%)*
190	*89*
687	*92*
1299	*90*
3438	*93*
998	*94*
1431	*96*

2.3. Analysis of Experimental Results

After running the program, the experimental results were analyzed. Using the learning algorithm, the roots and affixes of words were extracted. During the annotation of input data, the following errors were observed:

- Affixes were not correctly identified.
- The algorithm failed because some affixes were missing from the affix database.

Solutions:

- Expanding the list of root words and completing the set of affixes.
- Addressing the complexity of collecting root words: root words were added with each experiment, and the results from previous tests were incorporated into subsequent training. This iterative approach required significant time and effort, as each test necessitated compiling universal and unique root words.

To resolve this, root words gathered from each experiment were compared with the program's main word list. Additionally, texts in various styles of the Kazakh language were compiled, and experiments were conducted to collect new root words.

During program operation, essential resources such as root words, stop words, and Kazakh language affixes were collected. Table 4 summarizes the linguistic resources gathered during the experiments.

TABLE IV: Linguistic Resources Compiled During the Experiment

Linguistic Resource	*Number of Elements*
dictionary of stop words	*596*
dictionary of stem words	*69878*
dictionary of Kazakh language endings	*4983*
dictionary of correct words	*69997*

2.4. The Correct Words Dictionary in Error Detection

To identify incorrect words in semi-structured data in Kazakh, a dictionary of correct words was used. This dictionary was created with the help of the corpus and took into account the morphological features of the Kazakh language. Experiments were conducted to test the dictionary-based approach and methods for checking the roots and affixes of words. Comparative analyses were performed on approximately 3,000 messages and texts. Figure 3 illustrates the experimental results.

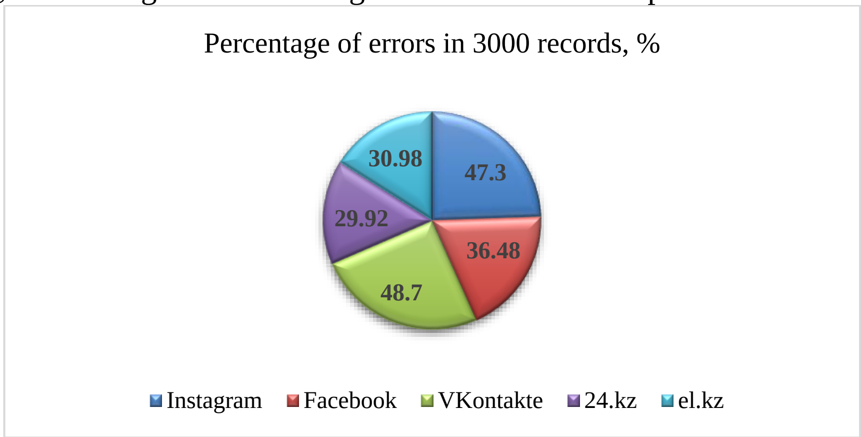


Fig. 3. Frequency of Incorrect Words on Social Networks

During the study, data were collected from websites and social networks, both automatically (via software) and manually. The error detection process was implemented using a developed algorithm. The error rate was calculated as the ratio of incorrect words to the total number of words in the messages of a given object. Percentages of different error types were similarly calculated. Linguistic specialists were involved in error classification and analysis, with experiments conducted as shown in Figure 4. Comparative analysis considered posts from websites and social networks, with 600 messages collected from each of five objects, amounting to a total of 3,000 messages. This approach enabled the identification of incorrect words and their types in semi-structured Kazakh data.

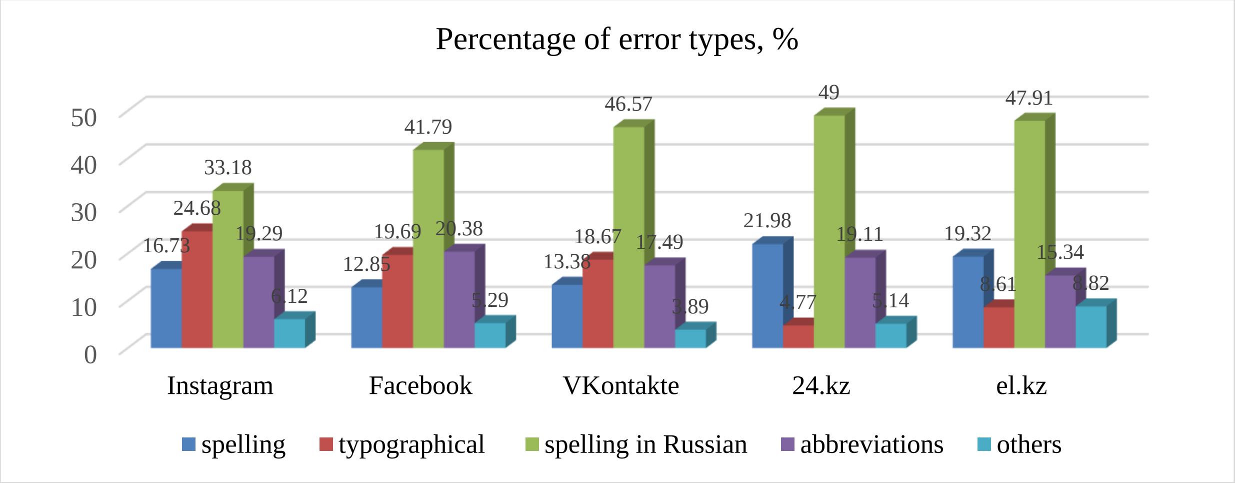


Fig. 4. Types and Frequency of Errors on Social Networks

Many errors in semi-structured text on social networks are not only introduced by users but also by editors. These errors can lead to the spread of misinformation and complicate decision-making processes.

3. Conclusion

The study reviewed works by various researchers and analyzed methods and technologies for detecting incorrect words in natural languages. A comparative analysis of text correction systems was also conducted. The findings revealed that these systems cannot be applied to Kazakh due to its complex morphological and lexical structures as an agglutinative language. Consequently, methods and systems for detecting errors in Kazakh texts were developed, and a dictionary of correct words for the language was created.

A corpus tailored to the features of the Kazakh language was built using specialized code. The developed method for detecting errors in semi-structured data involved morphological analysis through the Apertium platform. Post-analysis, the developed algorithm identified common error types.

Experiments were conducted using data from social networks and news portals. Detecting and correcting errors improved site reputation, supported business growth, attracted target audiences, and enabled keyword identification for search engines through user opinion analysis. Overall, the accuracy of error detection exceeded 90%.

The primary contribution of this work is the development of an error-detection method for Kazakh texts that considers linguistic characteristics. Programs for collecting and analyzing linguistic resources and semi-structured data were also created. Future work will focus on refining error correction, expanding the corpus, and enhancing the methods for enriching data.

4. Acknowledgments

This study was conducted and funded under the grant project AP23487753, “Innovative technologies for automated correction of texts in Kazakh: machine learning and morphological analysis,” supported by the Ministry of Science and Higher Education of the Republic of Kazakhstan.

5. References

[1] Computational processing of the Kazakh language: collection of scientific works (materials) / edited by Rakhimova D.R. –Almaty: Qazaq Universiteti, 2020. -146 p.

[2] Han B., Baldwin T.: Lexical normalisation of short text messages: Makn sens a# twitter. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. – Volume 1. – Association for Computational Linguistics, 2011. – P. 368-378.

[3] Farra N. et al.: Generalized Character-Level Spelling Error Correction. ACL (2). – 2014. – P. 161-167. https://doi.org/10.3115/v1/P14-2027

[4] Hladek, Daniel, et al.: Survey of Automatic Spelling Correction. Electronics [Basel], vol. 9, no. 10, 2020, p. 1dj+. Accessed 30 Apr. 2021.
https://doi.org/10.3390/electronics9101670

[5] Peter Buneman.: Semistructured data. In: Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. May 11-15, 1997, Tucson, Arizona, United States. – P. 117-121.
https://doi.org/10.1145/263661.263675

[6] Brill E., Moore R.C.: An improved error model for noisy channel spelling correction. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. – Association for Computational Linguistics, 2000. – P. 286-293.
https://doi.org/10.3115/1075218.1075255

[7] Farag Ahmed, Ernesto William De Luca, and Andreas Nürnberger.: Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness. Polibits (40) 2009.

[8] Kaufmann M., Kalita J.: Syntactic normalization of twitter messages. International conference on natural language processing, Kharagpur, India. – 2010.

[9] Rakhimova D.R.: Research of models and methods of semantics of machine translation from Russian into Kazakh language. Dissertation. – Almaty, 2014.

[10] Tukeyev U.A., Rakhimova D.R., Zhumanov Zh.M., Sundetova A.M. Machine translation of the Kazakh language into English and Russian (and vice versa) based on the Apertium platform: monograph / – Almaty: Kazakh University, 2017. – 286 p. ISBN 978-601-04-2926-0.

[11] Rakhimova D., Assem S. Problems of Semantics of Words of the Kazakh Language in the Information Retrieval. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, 11684 LNAI, pp. 70–81.
https://doi.org/10.1007/978-3-030-28374-2_7

[12] Shaalan K., Aref R., Fahmy A.: An approach for analyzing and correcting spelling errors for non-native Arabic learners. Published 2010, Computer Science, 2010 The 7th International Conference on Informatics and Systems (INFOS).

[13] Taktashkin D.V., Mokrousova Ye.A.: Methods and algorithms for checking the spelling of test documents (Paper in Russian). Electronic scientific & practical journal «Modern scientific researches and innovations». 2017. №5. URL: https://web.snauka.ru/issues/2017/05/72892 - (date of access: 12.04.2021)

[14] Rakesh Kumar, Minu Bala, and Kumar Sourabh.: 2018. A study of spell checking techniques for indian languages. JK Research Journal in Mathematics and Computer Sciences, 1(1).

[15] Rakhimova D., Turganbayeva A. Approach to Extract Keywords and Keyphrases of Text Resources and Documents in the Kazakh Language. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020, 12496 LNAI, pp. 719–729.
https://doi.org/10.1007/978-3-030-63007-2_56

[16] Tukeyev U.A., Turganbaeva A.O.: Lexicon-free stemming for the Kazakh language. In: Materials of the International Scientific Conference "Computer science and Applied Mathematics" dedicated to the 25th anniversary of the Independence of the Republic of Kazakhstan and the 25th anniversary of the Institute of Information and Computational Technologies, Part II, Almaty, September 21-24, 2016.

[17] Ualsher Tukeyev, Aliya Turganbayeva, Aidana Karibayeva, Dina Amirova, and Balzhan Abduali.: Language_Resources_for_Kazakh_language.https://github.com/NLPKazNU/Language_Resources_for_Kazakh_language

[18] Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages
https://turkic.apertium.org/index.kaz.html?choice=kaz&qA=%D0%9C%D0%B5%D0%BD%D1%96%D2%A3%20%D0%BE%D2%9B%D1%83%D1%88%D1%8B%D0%BC%20#analyzation