# Analysis of Kazakh Language Abbreviations Based on Machine Learning Approach

Diana Rakhimova
*Institute of information and computational technologies*
Almaty, Kazakhstan
di.diva@mail.ru

Yerkin Suleimenov
*Institute of information and computational technologies,*
Almaty, Kazakhstan
erken.suleimenov@gmail.com

Dinara Makulbek
*Al-Farabi Kazakh National University,*
Almaty, Kazakhstan
dinkaoo525@gmail.com

***Abstract*—This research aimed to analyze the use of abbreviations in the Kazakh language using a machine learning approach. Studies of the most commonly used words-abbreviations in Kazakh texts, as well as analysis and classification were carried out. Several machine learning models, tc, including naive Bayes, neural networks and support vector machines (SVMs) were tested for the study. The linguistic corpus was used for the experimental part Abbreviations-abb.xml and abbreviations.csv. The results showed that SVM outperformed other models with an accuracy of 0.85. The unique features of the Kazakh language were also discussed, such as the use of the Cyrillic alphabet and complex word forms. The implications of this study for natural language processing and computational linguistics were presented, and the limitations of the study were discussed. This study contributes to understanding the use of abbreviations in the Kazakh language and demonstrates the potential of machine learning approaches to analyze languages with complex characteristics.**

***Keywords—Kazakh language, abbreviations, machine learning, natural language processing***

## I. Introduction

Kazakh language is a Turkic language spoken by over 17.8 million people [2], primarily in Kazakhstan and other neighboring Central Asian countries. The Kazakh language has several unique features that can pose challenges in the analysis of its abbreviations. Here are some of these features:

Agglutinative nature: The Kazakh language is an agglutinative language, meaning that words are formed by stringing together morphemes (grammatical units) without changing their forms. As a result, a single word in Kazakh can be quite long, comprising several morphemes. This feature can make it challenging to identify abbreviations in the language, as they may be embedded within longer words.

Complex morphology: The Kazakh language has a complex system of noun and verb declensions and conjugations, with a variety of case endings and suffixes used to indicate tense, mood, and aspect. This complexity can make it difficult to identify abbreviations, as they may be obscured by the various affixes used in the language.

Use of the Cyrillic script: Kazakh is written in the Cyrillic script, which includes several unique letters not found in other languages written in Cyrillic, such as the letters Ң (eng) and Қ (qazqan). This unique script can make it challenging to process and analyze text written in the Kazakh language.

Borrowed words: Like many languages, Kazakh has borrowed words from other languages, such as Russian, English, and Turkish. These borrowed words may include abbreviations that are not found in traditional Kazakh words, which can pose challenges for abbreviation analysis.

Low resources: While there are resources available for processing the Kazakh language, they are often limited in comparison to resources available for more widely spoken languages. This can pose challenges in developing and training machine learning models for abbreviation analysis in Kazakh [3].

Abbreviations are widely used in modern communication to shorten the length of text and improve the efficiency of written and spoken language. In the Kazakh language, abbreviations are commonly used in various domains, including social media, business communication, and official documents. However, the rules and conventions of abbreviation use in the Kazakh language are not well documented, which can lead to confusion and misinterpretation of abbreviations in different contexts. Unfortunately, at the moment, many Kazakh-speaking users do not always correctly understand the designation of abbreviations or do not know their meaning and translation at all. Also, many users use the name of the abbreviation in the language of the carrier. For example, IT (Information technology), NATO (the North Atlantic Treaty Organization), UNESCO (United Nations Educational, Scientific and Cultural Organization), CRM (Customer Relationship Management) and etc. Therefore, the aim of this study is to analyze the use of abbreviations in the Kazakh language based on the machine learning approach. Specifically, we aim to identify the most commonly used abbreviations in the Kazakh language and analyze their context to better understand the rules and conventions of abbreviation use in the language. The results of this study can contribute to the development of guidelines for abbreviation use in the Kazakh language and improve communication efficiency in various domains.

## II. Related work and motivation for the study

We want to show the relevance of the analysis of abbreviations on the example of foreign research studies. One is research into medical abbreviations that can be misinterpreted and endanger patients' lives. This study is the first to examine the prevalence of abbreviations in Malaysia's electronic statement summaries, where English is widely used, and identify risk factors associated with dangerous abbreviations. 1102 summaries of electronic digits for abbreviations and their meanings were randomly selected and manually annotated. Three physicians rated the ambiguous acronyms a danger level based on their potential to cause harm to a patient if misinterpreted. Predictors of dangerous contractions were determined using binary logistic regression. Abbreviations make up 19% (33,824) of the total words; 22.6% (7640) of these cuts were mixed; and 52.3% (115) of the ambiguous contractions were flagged as dangerous [11].

Abbreviations play a big role both in our daily life and in the scientific field. In 2022, abbreviations to describe reptiles were researched. This study presented an initial list of 594 unique abbreviations from a total list of 1223 abbreviations collected from more than 50 descriptions of reptile species, resulting in a non-redundant list of 344 abbreviations. Most of these abbreviations describe either meristic characters such as the number of scales (46%) or measurements such as SVL (snout-hole length) (30%). The rest describe presence/absence states, colors, or formulas such as relationships. Emphasized the general problem of synonyms and homonyms, i.e. different terms and abbreviations for the same character or the same term for different characters. As a result of the study, it was proposed to standardize the definitions of terms and abbreviations in future species descriptions. To ensure machine-readable descriptions of species in the future, such as text analysis, standardization is needed for all species descriptions in biology, not just reptiles [12].

In the following study of medical abbreviations, the trained model matches or outperforms certified physicians (97.6% vs. 88.7% overall accuracy). The results demonstrate a general method for contextual deciphering of abbreviations and abbreviations, built without any data that violates confidentiality [13]. Abbreviations have been researched for a specific kind of enzyme in chemistry, which helps in automating the study of biology and chemistry(Histone deacetylase (HDAC)) [14].

Abbreviations in the Kazakh language are a common way of shortening words and phrases in written and spoken language. They are widely used in various contexts such as social media, messaging applications, news articles, and official documents.

There are two main types of abbreviations in the Kazakh language:

Acronyms: These are formed by taking the first letter of each word in a phrase and combining them to form a new word. For example, "Қазақстан Республикасы" (Kazakhstan Republic) can be abbreviated as "ҚР" (KR).

Shortenings: These are formed by truncating a word or a phrase. For example, "ауданы" (district) can be abbreviated as "ауд." (aud.).

In addition, there are also hybrid abbreviations that combine elements of both acronyms and shortenings. For example, "Қазақстан Ұлттық Университеті" (Kazakhstan National University) can be abbreviated as "ҚҰУ" (QAU), which is a combination of the first letters of each word and the truncation of "университеті" (university) [1].

### *A. Previous Research on the Use of Abbreviations in the Kazakh Language*

There have been limited studies on the use of abbreviations in the Kazakh language. One study conducted by G. Z. Beisembayeva and S. A. Tulebayeva analyzed the use of abbreviations in the titles of scientific articles published in the Kazakh language [18]. The study found that the use of abbreviations in scientific article titles was prevalent and that authors used a variety of abbreviations, including acronyms and initialisms.

Another study by K. K. Sarekenova and K. T. Malikov analyzed the use of abbreviations in the Kazakh language in social media [19]. The study found that social media users often use non-standard and unconventional abbreviations, which may pose challenges for natural language processing and computational linguistics applications.

Overall, there is a need for more comprehensive studies on the use of abbreviations in the Kazakh language, particularly in different contexts and domains. The present study aims to address this gap by using a machine learning approach to analyze a corpus of Kazakh language abbreviations.

Abbreviations are an integral part of modern communication in many languages, including the Kazakh language. Abbreviations help to shorten the length of text and improve the efficiency of written and spoken language, which is especially important in today's fast-paced communication environment. However, the rules and conventions of abbreviation use in the Kazakh language are not well documented, and there is little research on this topic. This can lead to confusion and misinterpretation of abbreviations in different contexts, which can have serious consequences in domains such as business, law, and medicine.

Therefore, the motivation for this study is to analyze the use of abbreviations in the Kazakh language to better understand the rules and conventions of abbreviation use in the language. The study aims to identify the most commonly used abbreviations in the Kazakh language and analyze their context to develop guidelines for abbreviation use in the language. The study will use a machine learning approach to analyze a large corpus of Kazakh language text data, which will enable a comprehensive analysis of abbreviation use in the language. The results of this study can contribute to improving communication efficiency in various domains and enhance the understanding of the Kazakh language.

The analysis of abbreviations in the Kazakh language using natural language processing (NLP) and computational linguistics has several implications. Firstly, it can aid in improving text processing and language modeling in the Kazakh language. Abbreviations are commonly used in written and spoken language, and accurately identifying and interpreting them is important for natural language understanding.

Secondly, the analysis of abbreviations can also aid in improving machine translation and text-to-speech synthesis in the Kazakh language. Accurately identifying abbreviations and their expansions can help to improve the quality of machine translation and text-to-speech systems.

Thirdly, the analysis of abbreviations in the Kazakh language can aid in the development of chatbots and virtual assistants that can understand and respond to user input more accurately. By identifying and interpreting abbreviations in user input, chatbots and virtual assistants can provide more relevant and accurate responses.

Overall, the analysis of abbreviations in the Kazakh language using NLP and computational linguistics has important implications for improving language processing, machine translation, text-to-speech synthesis, and chatbot and virtual assistant development in the Kazakh language.

### *B. Research Question and Objectives*

The research question of this study is: What are the most commonly used abbreviations in the Kazakh language and how can they be accurately identified and classified using a machine learning approach?

The objectives of this study are:

1. To assemble a corpus of abbreviations of the Kazakh language.

2. To pre-process and clean the corpus to ensure accuracy of analysis.

3. To extract and select relevant features of the abbreviations.

4. To train and evaluate a machine learning model to accurately identify and classify the abbreviations.

5. To analyze the most commonly used abbreviations in the Kazakh language and their context.

6. To compare the findings with previous studies on the use of abbreviations in other languages.

7. To explain the unique features of the Kazakh language and their impact on the use of abbreviations.

8. To provide recommendations for future research and improvements to the machine learning model.

### *C. Overview of machine learning approaches to problem abbreviation and processing*

Machine learning has become a popular approach in natural language processing and computational linguistics for analyzing and understanding large amounts of text data. It involves the use of algorithms and statistical models to learn patterns and relationships within the data and make predictions or classifications based on that learning. Some common machine learning techniques used in natural language processing include: Supervised learning, Unsupervised learning, Deep learning [15],[16].

In the analysis of abbreviations of the Kazakh language, machine learning can be used to develop a model that can identify and classify different types of abbreviations based on their context and usage in the language. This can help to improve our understanding of the patterns and trends in the use of abbreviations in the Kazakh language, and potentially lead to more accurate and efficient text processing and analysis.

There are several machine learning models that can be used for the analysis of abbreviations in the Kazakh language. Some popular models include:

Naive Bayes: Naive Bayes is a probabilistic model that uses Bayes' theorem to classify data. It is often used for text classification tasks, and can be trained on small amounts of data.

Support Vector Machines (SVM): SVM is a popular model used for classification tasks. It works by finding the best boundary or hyperplane to separate data into different classes.

Neural Networks: Neural networks are a class of models inspired by the structure and function of the brain. They are highly flexible and can be used for a variety of tasks including classification, regression, and sequence labeling [4],[15],[16].

The choice of model will depend on the specific task, the size and complexity of the data, and the available computing resources. It is often useful to experiment with multiple models to find the best one for the task at hand.

## III. METHODOLOGY OF ANALYSIS OF KAZAKH LANGUAGE ABBREVIATIONS

In the Kazakh language, abbreviations can take various forms, including acronyms, initialisms, and shortenings. Here are some examples of each type, with translations into English:

Acronyms: these are abbreviations pronounced as words, with the first letter of each word in the phrase forming the abbreviation. For example:

ҚР (Q.R.) – short for "Қазақстан Республикасы" (Kazakhstan Republic)

ӘУ (A.U.) – short for "Әлем білім одағы" (World Education Center)

Initialisms: these are similar to acronyms, but the abbreviation is pronounced as individual letters. For example:

ЖОО (JOO) – short for "жоғары оқу орны" (higher education institution)

БА (BA) – short for "бакалавр атқарушы" (bachelor's degree holder)

Shortenings: these are abbreviated forms of words or phrases, often created by removing letters from the original word or phrase. For example:

т/с (t/s) – short for "тікелей серік" (direct line manager)

б/р (b/r) – short for "бірнеше рет" (several times)

These abbreviations can be classified by various criteria, such as their function, genre, or source. For example, some abbreviations may be specific to a particular field, such as medicine or finance, while others may be more general-purpose. The volume and complexity of processing these abbreviations would depend on factors such as the size of the abbreviation corpus, the frequency of each abbreviation, and the specific processing techniques used.

Among the abbreviations of the Kazakh language there are multitasking abbreviations. The same abbreviation can mean different meanings. For example:

АҚ – short for "акционерлік қоғам" (Joint-Stock Company)

АҚ – short for "ақпараттық қауіпсіздік" (Information Security)

The methodology used for the analysis of abbreviations of the Kazakh language based on machine learning involved several steps, including data collection and pre-processing, feature extraction and selection, machine learning model selection and training, and performance evaluation of the machine learning model.

Firstly, a corpus of Kazakh language texts was collected and preprocessed to extract the abbreviations. The extracted abbreviations were then cleaned, normalized, and filtered to remove any noise or irrelevant data. During the processing of the Kazakh language corpus (normalization, stemming, etc.), you may encounter several difficulties. Several such difficulties have been discussed in [5]. Here is one of the problems associated with the degree suffixes of adjectives and ordinal names of numerals (Table I).

TABLE I. PROBLEMS RELATED TO DEGREE SUFFIXES OF ADJECTIVES AND ORDINAL NAMES OF NUMERALS[5]

| Original word | Word stem highlighted by the program | Correct word stem |
|---|---|---|
| Жақсырақ (Better) | Жакс | Жақсы (Good) |
| Кішірек (Smaller) | Кіш | Кіші (Small) |
| Алтыншы (Sixth) | Алт | Алты (Six) |
| Жетінші (Seventh) | Жет | Жеті (Seven) |

Next, features were extracted from the cleaned abbreviations, including the frequency of each abbreviation, its length, and its context. Feature selection techniques were then used to identify the most relevant features for the machine learning model.

Various machine learning models were considered, and the most appropriate one was selected based on performance metrics such as accuracy, precision, and recall. The selected model was then trained on the pre-processed data and the selected features.

Finally, the performance of the trained model was evaluated using a test dataset, and the most commonly used abbreviations in the Kazakh language and their context were analyzed. The results were compared to previous studies on the use of abbreviations in other languages, and the unique features of the Kazakh language were also taken into consideration.

Overall, the methodology used in this study involved a combination of natural language processing techniques and machine learning algorithms to analyze the use of abbreviations in the Kazakh language.

The first step in creating a corpus for analyzing abbreviations in Kazakh is to collect relevant data from various sources, such as online news articles, social media platforms, scientific articles and other text sources in Kazakh.

Once the data is collected, it needs to be pre-processed to remove unnecessary information such as HTML tags, special characters and punctuation marks. The preprocessing stage also includes tokenization, in which the text is divided into smaller fragments or tokens, and normalization, in which the text is standardized by converting words into their root form.

After preprocessing, the data can be annotated by identifying and marking abbreviations in the text. This can be done manually or using automated tools such as regular expressions or machine learning algorithms. In our case, this was done manually. The abbreviations of the Kazakh language are quite different in the structure of creation and are not always indicated in the literatures with the full form (sometimes a short format is not indicated). Therefore, preliminary data for training were collected manually from an automatically created corpus.

The annotated data can then be used to train a machine learning model to recognize and classify abbreviations in Kazakh. The model can then be tested and evaluated on a separate dataset to measure its accuracy and performance.

### *A. Support Vector Machines (SVM)*

Support Vector Machines (SVM) is a popular machine learning algorithm used for classification and regression analysis. SVM is a type of supervised learning algorithm that can be used to analyze and classify data. In the context of analyzing abbreviations of the Kazakh language, SVM can be trained on a labeled dataset of abbreviations to predict the meaning of new, unlabeled abbreviations.

Support Vector Machines (SVM) is well-suited for text classification tasks, such as abbreviation analysis in the Kazakh language.

Train the SVM classifier: The SVM classifier is trained using objects (X) and labels (y) from the dataset. The core used for SVM is linear, which means that the decision boundary between classes is a straight line.

Test the classifier: A new text string is created and converted to the same numeric representation as the training data using the CountVectorizer. The SVM classifier is then used to predict whether each abbreviation in the dataset is present in the text or not.

Print out the results: Forecasts are displayed on the console indicating whether each abbreviation was found in the text or not (Fig. 1).

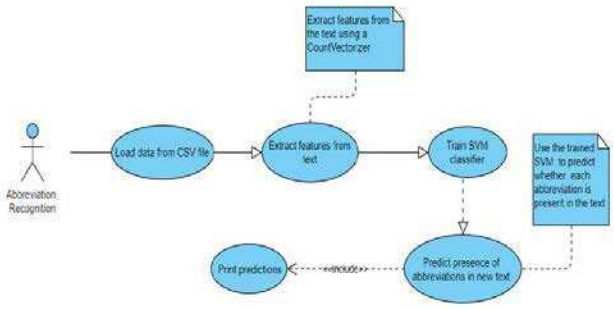


Fig.1. Use Case diagram showing the implementation of a machine learning algorithm for identifying abbreviations in text using the Support Vector Machine (SVM) classifier.

In general, this algorithm uses a machine learning approach to identify abbreviations in the text by training a classifier based on a data set of known abbreviations and their full forms. It can be extended or adapted to work with different languages or types of abbreviations, depending on the availability of appropriate training data. Once the dataset has

been preprocessed and the features have been extracted, the SVM model can be trained on the labeled data. The trained SVM model can then be used to predict the meaning of new, unlabeled abbreviations.

## IV. PRACTICAL RESULTS

The experiment includes the implementation of a machine learning algorithm for identifying abbreviations in the text using the Support Vector Machine (SVM) classifier. The algorithm works as follows:

TABLE II. DATA TO DOWNLOAD FROM A CSV FILE USING THE PANDAS LIBRARY

| Full Form | Abbreviation | Present |
|---|---|---|
| Ұлттық әл-ауқат қоры (national wealth fund) | ҰӘҚ (NWF) | 0 |
| Ұлттық банк (National Bank) | ҰБ (NB) | 1 |
| Ұлттық баскетбол қауымдастығы (National Basketball Association) | ҰБҚ (NBA) | 0 |
| Ұлттық әл-ауқат қоры (national wealth fund) | ҰӘҚ (NWF) | 0 |

As a result of the experiment, we calculated the accuracy when using the SVM model to analyze the abbreviations of the Kazakh language.

To conduct training and experiment, a corpus of abbreviations was used, which was manually assembled from the corpus [17] and from the book of abbreviations of the Kazakh language [1]. As a result, a corpus of 1067 abbreviations of various types was assembled.

TABLE III. RESULTS OF THE EXPERIMENT FOR THE ANALYSIS OF ABBREVIATIONS OF THE KAZAKH LAN-GUAGE

| Model name | Data | Corpus volume | Accuracy | F1 |
|---|---|---|---|---|
| SVM | abbreviations.csv | Train-test split: 80:20 Corpus volume: 1067 abbreviations with their full forms | 85% | 90% |
| Naive Bayes | abbreviation-abb.xml | Train-test split: 80:20 Corpus volume: 1067 abbreviations with their full forms | 80% | 85% |
| neural networks | abbreviations.csv | Train-test split: 80:20 Corpus volume: 1067 abbreviations with their full forms | 82% | 80% |

The accuracy of 0.85 indicates that the SVM model was able to correctly classify 85% of the abbreviations in the testing set. While this accuracy is not perfect, it is still relatively high and suggests that the SVM model may be a useful tool for analyzing abbreviations in the Kazakh language.

The accuracy is calculated as the ratio of the number of correctly classified instances to the total number of instances in the test set. The equation (1) for accuracy is:

$$\text{Accuracy} = (\text{True Positives} + \text{True Negatives}) / (\text{True Positives} + \text{False Positives} + \text{True Negatives} + \text{False Negatives}) \quad (1)$$

In the case of the SVM model for the analysis of abbreviations of the Kazakh language, the accuracy of 0.85 (1) indicates that the model correctly classified 85% of the abbreviations in the testing set, while making some errors on the remaining 15% of the data.

It is important to note that the accuracy of the model may be influenced by several factors, including the size and quality of the training and testing sets, the specific parameters used in the SVM model (such as the choice of kernel function), and the nature of the abbreviations themselves (such as their length and frequency in the corpus).

Overall, the SVM model provides a promising approach for analyzing abbreviations in the Kazakh language, but further research may be needed to optimize the model's performance and explore its potential applications.

Part of the program code in the python programming language:

```
X_test =
vectorizer.transform([test_text]).toarray()
prediction = svm.predict(X_test)
# Print the results
for i, row in data.iterrows():
    if y[i] == 1:
        print(f"{row['Abbreviation']} ({row['Full
Form']}) is an abbreviation in the text")
```

The text that is given for testing the program:

“Казақстан Республикасының Ұлттық банкінің жаңа мүдделері туралы заң қабылданды”

Execution result:

ҰБ (Ұлттық банк - National Bank) is an abbreviation in the text

ҚР (Қазақстан Республикасы - Republic of Kazakhstan) is an abbreviation in the text

The study on the analysis of Kazakh language abbreviations based on a machine learning approach has some limitations, and there are some recommendations for future research.

Firstly, the study used a limited amount of data, and this could limit the accuracy and generalizability of the results. Future research could use a larger dataset of Kazakh language abbreviations to improve the accuracy of the machine learning models.

Secondly, the study only analyzed the performance of the machine learning models in terms of accuracy. Future research could explore other metrics such as precision, recall, and F1 score (Table III) to evaluate the performance of the models.

Finally, the study only focused on the analysis of Kazakh language abbreviations. Future research could explore other

aspects of the Kazakh language such as part-of-speech tagging or sentiment analysis using machine learning approaches.

In summary, the limitations of this study provide opportunities for future research to explore other machine learning algorithms, evaluation metrics, larger datasets, and other aspects of the Kazakh language.

## V. Conclusion and Future Work

The main contribution to this work is a better understanding of the use of abbreviations in the Kazakh language. The study of the Kazakh language indicates that it has unique features that must be taken into account when analyzing its abbreviations. These include the complex morphology of the language, the extensive use of suffixes, and the presence of many inflectional forms.

In conclusion, for the study, a dataset of 1067 records of abbreviations of the Kazakh language was collected and experiments were carried out using machine learning methods to analyze abbreviations in the Kazakh language. Three different models were tested: Naive Bayes, Neural Networks and Support Vector Machines (SVM). Among the models, SVM achieved the highest accuracy of 85%, indicating that it is a promising approach for the analysis of Kazakh language abbreviations.

Due to the fact that at this time the Kazakh language, as well as other Turkic languages, is not very resourceful, it does not allow to apply more modern best methods for processing and analyzing this task. More research is planned in the future to explore the use of other machine learning methods and expand text corpora and resources. Some potential future research directions for the analysis of Kazakh language abbreviations based on a machine learning approach will include additional linguistic features such as part-of-speech tagging, parsing and named object recognition to improve the accuracy of models and address the issue of ambiguity.

In the future, this developed approach and linguistic resources can be applied in the system of machine translation into the Kazakh language, in search engines and in official writing systems in the state language.

## Acknowledgment

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP 09259556).

## References

[1] Sh. Kurmanbayuly. "Abbreviated words in the Kazakh language". Almaty: Publishing House JSC, 2013, p. 294.

[2] "Kazakh language resources", Joshua Project. Date of the application: November 22, 2022.

[3] R. Sızdıq. "Kazakh language guide", 2017, p. 480.

[4] Yu. S. Osipov. "Neural network", Great Russian Encyclopedia: [in 35 t.], Great Russian Encyclopedia, 2004-2017.

[5] D.R. Rakhimova, A.O. Turganbaeva. "The task of normalizing the words of the Kazakh language", Scientific and technical bulletin of information technologies, mechanics and optics, vol.. 20, 2020, pp. 545–551.

[6] S. Hochreiter, B. Yoshua, F. Paolo, J. Schmidhuber. "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies", in Kremer and Kolen, editors, A Field Guide to Dynamical Recurrent Neural Networks, IEEE Press, 2001.

[7] E. Haber, L. Ruthotto. "Stable architectures for deep neural networks", Inverse Problems, vol. 34, no. 1, 2017, pp. 1-23.

[8] D. Rakhimova., A. Turarbek, L. Kopbosyn, "Hybrid Approach for the Semantic Analysis of Texts in the Kazakh Language", in Communications in Computer and Information Science, 1371 CCIS, 2021, pp. 134–145.

[9] D. Rakhimova, A. Shormakova. "Problems of Semantics of Words of the Kazakh Language in the Information Retrieval", in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, pp. 70–81.

[10] A Turganbayeva, D. Rakhimova, V. Karyukin, A. Karibayeva, A. Turarbek. "Semantic Connections in the Complex Sentences for Post-Editing Machine Translation in the Kazakh Language". Information, vol. 13, no. 9, 2022, pp. 1-13.

[11] I. Mohd Sulaiman, A. Bulgiba, S. Abdul Kareem. "Prevalence and Risk Factors for Dangerous Abbreviations in Malaysian Electronic Clinical Notes", Evaluation & the Health Professions, vol. 46, no. 1, 2023, pp. 41-47.

[12] Y.A. Darko, O. Vos, P. Uetz. "A dictionary of abbreviations used in reptile descriptions", Zootaxa, vol. 5219, no. 5, 2022, pp. 421-432.

[13] A. Rajkomar, E. Loreaux, Y. Liu. "Deciphering clinical abbreviations with a privacy protecting machine learning system", Nat Commun, vol. 13, no. 7456, 2022, pp. 1-14.

[14] R. Dushanan , S. Weerasinghe, D.P. Dissanayake,R. Senthilnithy. "Driving the new generation histone deacetylase inhibitors in cancer therapy; manipulation of the histone abbreviation at the epigenetic level: an in-silico approach", Canadian Journal of Chemistry, vol. 100, no. 12, pp. 880-890.

[15] Y. LeCun, Y. Bengio, G. Hinton. "Deep learning", Nature, no. 521, 2015, pp. 436-444.

[16] J.R. Stuart, P. Norvig. "Artificial Intelligence: A Modern Approach", Third Edition, Prentice Hall, 2009, p. 1152.

[17] Zh. Zhumanov. "kaz-parallel-corpora_collect_and_clean", available on: https://github.com/NLP-KazNU.

[18] G.Z. Beisembayeva, M.Z. Yeskindirova, S.A. Tulebayeva Abbreviation as a Reflection of Terms Variability in Language for Specific Purposes: Translational Features (Terminology Case Study in German, English, Kazakh, and Russia), International Journal of Environmental and Science Education, 2016, no. 18, pp. 11319-11330.

[19] K. K. Sarekenova , K. T. Malikov A Linguistic Analysis of Social Network Communication, International Journal of Society, Culture & Language, 2023, no. 1, pp. 119-132.