

3-бөлім

Раздел 3

Section 3



Информатика

Информатика

Computer  
science

IRSTI 20.53.19

DOI: <https://doi.org/10.26577/JMMCS2024-v124-i4-a4>

M.E. Mansurova\* , D.R. Rakhimova   
Al-Farabi Kazakh National University, Almaty, Kazakhstan  
\*e-mail: mansurova.madina@gmail.com

## MORPHOLOGICAL PARSING OF KAZAKH TEXTS WITH DEEP LEARNING APPROACHES

Morphological analysis is a crucial task in Natural Language Processing (NLP) that greatly contributes to enhancing the performance of large language models (LLMs). Although NLP technologies have seen rapid advancements in recent years, the creation of efficient morphological analysis algorithms for morphologically complex languages, such as Kazakh, continues to be a significant challenge. This research focuses on designing a morphological analysis algorithm for the Kazakh language, specifically optimized for integration with LLMs.

The study will address the following tasks: data corpus collection and processing, selection and adaptation of suitable algorithms, and model training and evaluation. This paper delivers a detailed exploration of using deep learning models for the morphological analysis of the Kazakh language, specifically highlighting Recurrent Neural Networks (RNN) and Transformer models. Because of Kazakh is an agglutinative language, where word formation is achieved by attaching multiple suffixes and prefixes, the task of morphological analysis poses 25 unique challenges for computational models.

The performance of Recurrent Neural Networks (RNNs) is evaluated, including those with LSTM and GRU enhancements, in comparison with Transformer models, focusing on their capability to analyze the complex morphology of Kazakh. The results outline the benefits and limitations of each approach for processing agglutinative languages, indicating that RNNs are often more effective for Kazakh morphological analysis, whereas Transformer models may require additional fine-tuning to achieve optimal results with such languages.

**Key words:** Kazakh language, morphological analysis, RNN, Transformer.

М.Е. Мансурова\*, Д.Р. Рахимова  
әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан  
\*e-mail: mansurova.madina@gmail.com

### Терең оқыту модельдері негізінде қазақ тілінің морфологиялық талдауы

Морфологиялық талдау – табиғи тілді өңдеудегі (NLP) саласындағы іргелі есеп, үлкен тілдік модельдердің (LLM) өнімділігін арттыруда шешуші рөл атқарады. Соңғы жылдары NLP технологиялары қарқынды дамып келеді, алайда бай морфологиясы бар қазақ тілі үшін морфологиялық талдаудың тиімді алгоритмдерін әзірлеу өзекті есеп болып қала береді. Бұл зерттеудің мақсаты LLMs-ді пайдалануға арнайы бейімделген қазақ тілі үшін морфологиялық талдау алгоритмін әзірлеу болып табылады.

Зерттеу келесі міндеттерді қарастырады: мәліметтер корпусын жинау және өңдеу, сәйкес алгоритмдерді таңдау және бейімдеу, модельдерді оқыту және бағалау жасау. Бұл мақалада қазақ тілінің морфологиялық талдауы үшін терең оқыту модельдерін қолдану, атап айтқанда, рекурентті (қайталанатын) нейрондық желілер (RNN) және трансформаторлық модельдер туралы егжей-тегжейлі зерттеу жасалған. Қазақ тілі агглютинативті тіл болғандықтан, сөз-жасамға бірнеше жұрнақтар мен префикстерді қосу арқылы қол жеткізіледі, морфологиялық талдау міндеті есептеу модельдері үшін ерекше қиындық туғызады.

Рекуррентті (қайталанатын) нейрондық желілердің (RNNs), соның ішінде LSTM және GRU жақсартулары бар желілердің өнімділігі трансформаторлық модельдермен салыстырғанда бағаланады, олардың қазақ тілінің күрделі морфологиясын талдау қабілетіне баса назар аударылады. Нәтижелер агглютинативті тілдерді өңдеудің әрбір тәсілінің артықшылықтары мен шектеулерін сипаттайды, Бұл RNN көбінесе қазақ морфологиялық талдауы үшін тиімдірек екенін көрсетеді, Ал трансформер модельдер мұндай тілдерде оңтайлы нәтижелерге қол жеткізу үшін қосымша баптауларды талап етеді.

Біздің нәтижелеріміз агглютинативті тілдік тапсырмалар үшін әрбір модельдің күшті жақтары мен шектеулерін көрсетеді, бұл RNN қазақ тілін морфологиялық талдау үшін қолайлырақ екенін, ал Transformer модельдері осындай тілдерді одан әрі оңтайландыру арқылы ұтымды болуын көрсетеді.

**Түйін сөздер:** Қазақ тілі, морфологиялық талдау, RNN, Transformer.

М.Е. Мансурова\*, Д.Р. Рахимова

Казахский национальный университет имени аль-Фараби, г. Алматы, Казахстан

\*e-mail: mansurova.madina@gmail.com

### **Морфологический анализ казахского языка с использованием моделей глубокого обучения**

Морфологический анализ представляет собой центральную задачу в области обработки естественного языка (NLP), существенно влияя на улучшение эффективности больших языковых моделей (LLM). Несмотря на значительный прогресс в технологиях NLP за последние годы, разработка эффективных алгоритмов для морфологического анализа морфологически сложных языков, таких как казахский, остаётся актуальной и сложной задачей. В данном исследовании рассматривается разработка алгоритма морфологического анализа, специально адаптированного для казахского языка и интегрированного с LLM.

В рамках исследования будут решены следующие ключевые задачи: сбор и обработка корпуса данных, выбор и адаптация алгоритмов, а также обучение и оценка моделей. Статья детализирует применение моделей глубокого обучения для морфологического анализа казахского языка, особенно сосредоточив внимание на рекуррентных нейронных сетях (RNN) и трансформаторных моделях. Поскольку казахский язык является агглютинативным, где морфологические изменения достигаются путём добавления множества суффиксов и префиксов, задача морфологического анализа предъявляет специфические требования к вычислительным моделям.

Оценка производительности RNN, включая улучшенные версии с LSTM и GRU, по сравнению с трансформаторными моделями позволяет выявить их способности к анализу сложных морфологических структур казахского языка. Результаты показывают как преимущества, так и ограничения каждого подхода для обработки агглютинативных языков, указывая на то, что RNN часто более эффективны для морфологического анализа казахского языка, в то время как трансформаторные модели могут требовать дополнительной настройки для достижения оптимальных результатов.

**Ключевые слова:** Казахский язык, морфологический анализ, RNN, Transformer.

## **1 Introduction**

Morphological analysis is a central task in language processing, where the input is a word, and the output reveals various morphological components, providing its morphological representation. It is often the first step in various types of text analysis in any language. A morphological analyzer is used in speech synthesis, speech recognition, segmentation, lemmatization, search engines, and machine translation. The Kazakh language, characterized by its agglutinative nature, features a highly intricate and elaborate morphological structure. Grammatical functions are conveyed through the attachment of numerous suffixes to a base

root. Generally, a single word can incorporate a minimum of two or three affixes (suffixes and endings).

The application of machine learning (ML) methods for the morphological analysis of the Kazakh language involves the automated identification of word structures, including roots, suffixes, prefixes, and grammatical features. As an agglutinative language, Kazakh presents a highly intricate morphology where grammatical nuances are conveyed through extensive affixation:

1. Morphological complexity of agglutinative languages. Unlike languages with analytical or inflectional structures, Kazakh words can contain multiple affixes that convey various grammatical elements such as person, number, tense, and mood. This complexity poses challenges for traditional analysis methods and necessitates the use of more flexible ML approaches.

2. Processing data with low resources. The scarcity of labeled data for Kazakh, which complicates the application of standard machine learning methods. In such conditions, transfer learning

3. Morphology modeling. Automated morphological analysis can utilize both supervised and unsupervised models. Supervised models depend on pre-labeled data, while unsupervised methods, such as clustering, can uncover morphological patterns from unlabeled data.

## 2 Related works

A range of approaches has been employed to create morphological analyzers for diverse languages:

- Rule-based approaches implemented using finite-state transducers;
- Rule-based approaches implemented using dictionaries;
- Statistical methods;
- Machine learning and neural networks.

Research over the past decade, publicly accessible resources for the morphological parsing of Kazakh language texts (including words and sentences) predominantly utilize rule-based methods such as dictionaries, tables, and finite automata.

In [4], a morphological structure of Kazakh language texts (words, sentences) based on dictionaries is presented. In [5,6], the same research group proposes universal methodologies for stemming, segmentation, and morphological analysis of Turkic languages (including Kazakh), based on the "Complete Suffix Enumeration" (CSE) model of Turkic morphologies. The CSE model is based on four types of suffixes: plural, case, personal, and possessive. A special relational data model-solution table is created for morphological analysis, and morphological tagging of the text is performed using this table.

In [7-10], ontological models of nouns and adjectives for Kazakh and Turkish languages are explored. The research results in an ontological model of Kazakh morphological rules. Based on this model, new word forms can be generated.

In [11], an automatic morphological analyzer is developed to identify parts of speech and extract lemmas. Lemma extraction utilizes Porter's algorithm, adapted to the grammatical rules of the Kazakh language, while detailed word analysis is carried out using a word-form dictionary. Different word forms are produced for each lemma and recorded in a database alongside associated metadata, including morphological properties.

In [12], a method of morphological analysis and disambiguation for the Kazakh language is proposed, considering both inflectional and derivational morphology. The method is data-driven and does not require manually generated rules. Transition chains help discard false segmentations while retaining correct ones, with ambiguity resolution using the standard HMM approach.

In [13], foreign researchers present a comprehensive two-level morphological analysis of modern Kazakh using the Nuve Framework. The root dictionary contains 24,000 roots, and the suffix dictionary has 150 suffixes.

In recent years, machine and deep learning-based methods have become popular for text processing. A key challenge with these methods is the lack or absence of structured, clean datasets for model training. The quality of these technologies depends on the size and content of the training dataset. Machine and deep learning are novel applications for Kazakh morphological analysis.

### 3 Developing a morphological analysis model for kazakh based on deep learning models

#### 3.1 The construction of a model for the morphological analysis of texts Kazakh using Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are advanced computational frameworks frequently employed in natural language processing (NLP) applications. RNNs process one word at a time, retaining memory of previously seen words, allowing different words to be processed based on their position in a sentence. This property of RNNs makes them applicable to our task of morphological analysis for the Kazakh language.

RNNs have loops that allow information to be passed between neurons when reading input data.

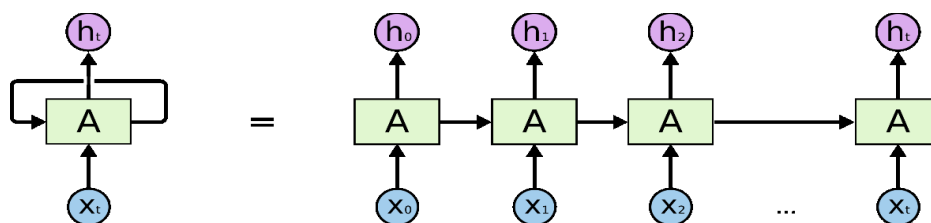


Figure 1: Architecture of the RNN model.

The idea behind using RNN for morphological analysis is leveraging the data sequence. However, RNNs are limited by their ability to look back only a short distance. Therefore, Long Short-Term Memory (LSTM) units are used to achieve better results. Long Short-Term Memory (LSTM) networks incorporate memory cells that retain contextual information from the start of the input sequence. For instance, when predicting the 11th word in a sequence of 10 words, an RNN processes all 10 words, with LSTM units preserving the weights at each step. LSTM cells, functioning alongside the hidden layer, maintain information that may not be directly relevant to immediate predictions. These memory components enable RNNs to generate more precise outcomes.

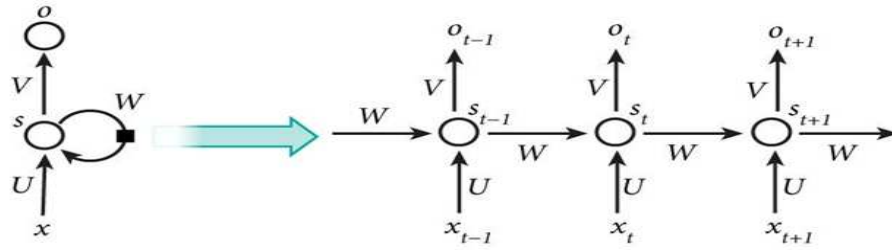


Figure 2: Unfolded RNN.

Figure 2 illustrates the process by which a Recurrent Neural Network (RNN) is unfolded into a fully connected network. This unfolding involves extending the network to encompass the entire sequence, such as representing a sentence with six words as a six-layer neural network.

The computations within an RNN are defined as follows:

$a_t$  – represents the input at time step  $t$ , which can be a vector corresponding to the second word in a sentence.

$b_t$  – denotes the hidden state at time step  $tt$ , computed based on the previous hidden state and the current input:  $b_t = f(Ua_t + Wb_{t-1})$ , where,  $f$  is typically a non-linear function like like tanh or ReLU, and  $b_{t-1}$ , s initialized to zeros.

$c_t$  – is the output at step  $t$ , such as a probability vector over the vocabulary:  $c_t = \text{softmax}(Vb_t)$ .

A detailed description of the problem setup is as follows: The training dataset comprises a large collection of sequences of characters (words in sentences), denoted as:

$$V_{train} = \{a^1, \dots, a^y\}$$

$$\text{Where, } a = [a_1, \dots, x_n].$$

For example

$$a = [\text{мынау}, \text{көл}, \text{мөлдір}, \text{STOP}]$$

$$\phi = [\text{this}, \text{lake}, \text{clear}, \text{STOP}]$$

The objective is to develop a model that estimates:  $P(a)$ ,  $\forall a \in V^{Maxn}$  where  $V$  – is the vocabulary,  $V^{MaxN}$  – represents all possible sentences.

The probability is expressed as:

$$(a) = (a_1, \dots, a_n)$$

$$P(a_1) = P(a_2, \dots, a_n | a_1)$$

$$P(a_1) = P(a_2 | a_1), \dots, P(a_n | a_1)$$

So, probabilities as a logistic regression can be written as:

$$P(a_n = k | a_{n-1} \dots a_1) = \frac{\exp(w_k \cdot \varphi(a_{n-1} \dots a_1))}{\sum_{k'=1 \text{ to } V} \exp(w_{k'} \cdot \varphi(a_{n-1} \dots a_1))} \quad (1)$$

$w_k - t$  represents the weights for word  $k$ ,  $\phi(a_{n-1}, \dots, a_1)$  – denotes features extracted from preceding words  $(a_{n-1} \dots a_1)$ .

The conditional probability  $P(a_n | a_{n-1}, \dots, a_1)$  is computed as:

$$P(a_n|a_{n-1}, \dots, a_1) = \text{softmax}(W\phi(a_{n-1}, \dots, a_1))$$

Where,  $W \in R^{|\text{vocab}| \times |\text{contexts}|}$  contains weights for each word and context, and  $b_{n-1} \in R^d$  represents the context memory.

The probability distribution for word  $a_n$  is given by:

$P(a_n|a_{n-1}, \dots, a_1) = \text{softmax}(Vb_{n-1})$ , where  $V \in R^{|\text{vocab}| \times d}$  represents the context in the probability distribution. To obtain  $b_{n-1}$ , the RNN is used:

$b_n = \sigma(Wb_{n-1} + Ua_n)$  with  $b_{n-1} \in R^d$  retaining the context of  $a_{n-1}$  and  $a_{n-1}$ , and  $U \in R^{|\text{vocab}| \times d}$  – containing word vectors for all words. To determine the probability distribution for word  $a_n$ :  $O_{n-1} = P(a_n|a_{n-1}, \dots, a_1) = \text{softmax}(Vb_{n-1})$

Training requires learning the word vectors  $U$ , as well as the parameters of the hidden layer  $W$  and output layer  $V$ . Since standard backpropagation is ineffective due to parameter sharing in the hidden layer, Backpropagation Through Time (BPTT) [14] is employed. BPTT involves expanding the network graph over  $N$  time steps and aggregating the gradient contributions to update the parameters.

### 3.2 Transformers model

Similar to Recurrent Neural Networks (RNNs), Transformers are engineered to handle sequential data. However, unlike RNNs, Transformers do not require sequential data to be processed in order. For example, if the input is a natural language sentence, the Transformer does not need to process it from start to finish. Because of this feature, Transformers allow for much greater parallelism than RNNs, and therefore reduce training time. Most competing neural sequence transformation models utilize an encoder-decoder architecture. In this framework, the encoder transforms an input sequence of symbols  $(a_1, \dots, a_n)$  into a sequence of continuous representations  $(z_1, \dots, z_n)$ . Using these continuous representations  $z$ , the decoder generates an output sequence of symbols  $(y_1, \dots, y_m)$ , producing one symbol at a time. The model operates in an autoregressive manner, incorporating previously generated symbols as additional input to produce the subsequent symbol.

Figure 3 illustrates the use of the Transformer model for morphological analysis of a Kazakh word, exemplified by the term "Кітабымды" (book + my + accusative case).

Model Description:

1. Input Word: The model is fed the word "Кітабымды" which is a word with several affixes.

2. Embedding Layer: This stage transforms the input word into numeric vectors (embeddings), which reflect the semantic meaning of the word and its morphemes.

3. Self-Attention: This mechanism allows the model to take into account the dependencies between different parts of the word, for example, the relationship between the root ("кітап") and the affixes ("ым ды"). This is a key component that helps the transformer capture long-term relationships.

4. Feed Forward: This stage further processes the data to clarify the relationship between morphemes and their grammatical function.

5. Output Layer: The model produces a segmented representation of the word, such as "Кітап+ым+ды," where the root and affixes are distinctly identified.

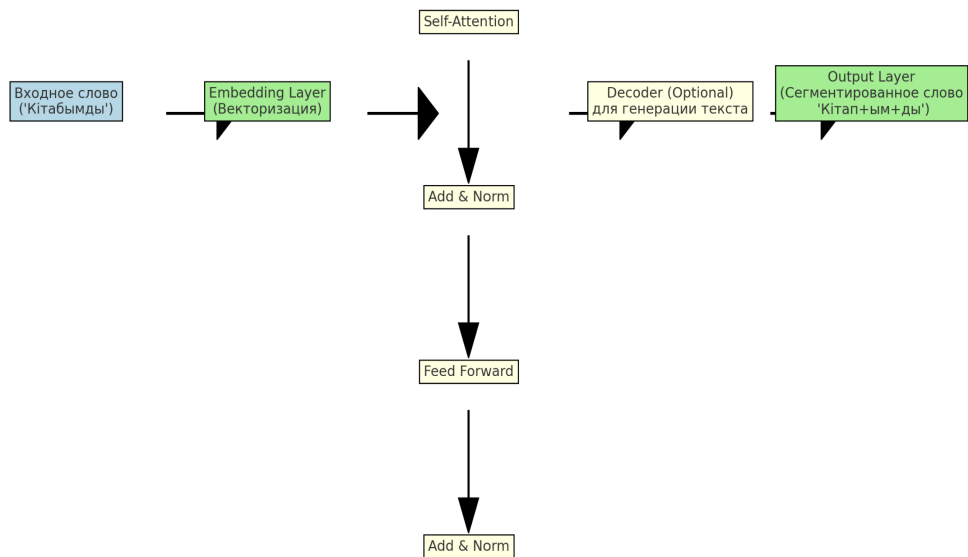


Figure 3: An example of morphological analysis of the Kazakh language with the identification of key blocks and connections between them in the Transformer model.

#### 4 Experimental work

RNN and Transformer architectures were applied to build the morphological analysis model for Kazakh texts.

Table 1. Training dataset size

Dataset	Count
Training corpus (Kazakh texts with parallel morphological tagging)	20,000 sentences
Testing	
- test 1	13 000 sentences
- test 2	7 000 sentences
Dictionary	
- or untagged corpus	15 000 words
- for tagged corpus	12 000 words

Table 2. Example of results of the morphological analyzer

Models	Kazakh text	Morphological analysis (trained models)	Reference morphological analysis
Trained RNN model	Сондықтан да, біз мұны әрдайым сақтай отырып, үнемі есте ұстауға тиіспіз. [Therefore, we must always keep this in mind, always keeping it in mind.]	^сондықтан<cnjadv>+ да<postadv>\$ ^,<cm>\$ ^біз<prn> <pers> <p1> <pl> <nom>\$ ^бұл<prn><dem> <acc>\$ ^әрдайым<adv>\$ ^сақта<v><tv> <prc_impf>\$ ^отыр<vaux> <gna_perf>\$ ^, <cm>\$ ^үнемі<adv>\$ есте ұстауға ^тиіс<adj>+ e<cop><aor> <p1> <pl>\$ ^.<sent>\$	^сондықтан <cnjadv>+ да <postadv> \$ ^,<cm> \$ ^біз <prn> <pers> <p1> <pl> <nom>\$ ^бұл<prn> <dem> <acc>\$ ^әрдайым<adv>\$ ^сақта<v> <tv> <prc_impf>\$ ^отыр<vaux> <gna_perf>\$ ^,<cm>\$ ^үнемі<adv>\$ ^ec<n> <loc>\$ ^ұста<v> <tv> <ger> <dat>\$ ^тиіс<adj>+ e<cop> <aor> <p1> <pl>\$ ^.<sent>\$
Transformer	Бұлар – Қазақстанның болашағына кілттер	^бұл<prn> <dem><pl> <nom>\$ - ^Қазақстан<np> <top> <gen>\$ ^бола- шақ<n> <px3sp> <dat>\$ кілттер ^.<sent>\$	^бұл<prn> <dem> <pl> <nom> \$ ^-<guio> \$ ^Қаза- қстан<np> <top> <gen>\$ ^бо- лашақ<n> <px3sp> <dat>\$ ^кілт<n> <pl> <nom>+ e<cop> <aor> <p3> <pl>\$ ^.<sent>\$

To evaluate the quality of the trained morphological analysis model, the following metrics were used: BLEU, WER, TER. (show table 3)

**BLEU (Bilingual Doubler Assessment)** is an algorithm used to assess the quality of machine-translated text from one natural language to another. In this context, the output is evaluated based on the marked-up case.

**WER (Word error rate)** measures the normalized distance between a candidate translation and multiple reference translations. It represents the edit distance, or the number of insertions, deletions, and substitutions required to transform the candidate translation ( $t$ ) into the reference translation ( $r$ ).

**TER (Translation error rate)** takes into account the number of corrections required to change the result so that it semantically corresponds to the correct (reference) translation.

Table 3. Indicators of learning models for morphological analysis of texts (words, sentences) in the Kazakh language

Trained morphological analysis models	BLEU	WER	TER
RNN	46.93	39	39
Transformer	41.78	46	46



From the obtained estimates in table 3 through experiments, we found that while Transformer models leverage the self-attention mechanism to capture long-range dependencies and handle larger contexts efficiently, they faced challenges with the intricacies of Kazakh morphology, especially in recognizing subtle morphological patterns within shorter contexts.

In contrast, RNN-based models, particularly those enhanced with LSTM or GRU units, demonstrated slightly better performance in terms of accuracy. The sequential nature of RNNs allowed them to more effectively capture the step-by-step morphological transformations in Kazakh words, making them more adept at identifying the hierarchical structure of morphemes. As a result, the RNN models outperformed Transformer models by a few percentage points in key metrics such as segmentation accuracy and morpho-syntactic parsing.

## 5 Conclusion

Morphological analysis of the Kazakh language is challenging due to its agglutinative nature and rich morphological structure. This study explored the application of deep learning models, particularly RNNs, for solving this task. Research showed that RNN models have significant potential to improve the accuracy of morphological analysis by capturing long-term dependencies in word sequences.

The use of RNN models has significantly improved the analysis results compared to traditional rule-based methods. This is due to the fact that deep learning models are able to automatically extract complex dependencies and language features from large amounts of data, which is especially relevant for agglutinative languages such as Kazakh. In addition, the paper discusses the importance of pre-training and adapting models to the specifics of the Kazakh language. The experimental results show that while Transformer models have potential due to their scalability and efficiency in tasks with large amounts of data, RNNs may be more suitable for specific language tasks with low resources, such as morphological analysis of the Kazakh language. Future work can focus on hybrid approaches that combine the strengths of both architectures or on exploring further tuning of Transformer-based models to better fit agglutinative language structures. The experimental results showed that transfer learning and additional training of models on data specific to the Kazakh language can significantly improve their efficiency and accuracy. Thus, the implementation of deep learning in the task of morphological analysis of the Kazakh language opens up new prospects for creating more advanced and accurate natural language processing systems. In the future, research can be expanded by integrating RNN models with other deep learning architectures to further improve the quality of morphological analysis.

## Acknowledgments

This research was funded by project BR24993001 with support from the Ministry of Science and Higher Education of the Republic of Kazakhstan.

## References

- [1] Start learning Kazakh now! <https://www.soyle.kz/>. (Date of access: 03/18/2024).
- [2] Learn Kazakh with TIL-QURAL. <https://tilqural.kz/>. (Date of access: 02/20/2024).
- [3] Declension according to the rules of the Kazakh language. <http://morpho.kz/>. (Date of access: 02/25/2024).
- [4] Tukeyev U., Sundetova A., Abduali B., Akhmadiyeva Z., Zhanbussunov N., "Inferring of the morphological chunk transfer rules on the base of complete set of Kazakh endings", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, (2016): 563–574.
- [5] Tukeyev U., Karibayeva A., Zhumanov Z.H., "Morphological segmentation method for Turkic language neural machine translation", *Cogent Engineering*, (2020): 7.
- [6] Tukeyev U., Karibayeva A., Turganbayeva A., Amirova D., "Universal Programs for Stemming, Segmentation, Morphological Analysis of Turkic Words", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, (2021): 643–654.
- [7] Mukanova A., Yergesh B., Bekmanova G., Razakhova B., Sharipbay A.A., "Formal models of nouns in the Kazakh language", *Leonardo Electronic Journal of Practices and Technologies*, 13(25) (2014): 264–273.
- [8] Yergesh B., Mukanova A., Sharipbay A. A., Bekmanova G., Razakhova B., "Semantic Hyper-graph Based Representation of Nouns in the Kazakh Language", *Computación y Sistemas*, 18(3) (2014): 627–635.
- [9] Zhetkenbay L., Sharipbay A. A., Bekmanova G., Kamanur U., "Ontological modeling of morphological rules for the adjectives in kazakh and Turkish languages", *Journal of Theoretical and Applied Information Technology*, 91(2) (2016): 257–263.
- [10] Zhetkenbay L., Bekmanova G., Sharipbay A.A., Altenbek G.A., "Uniform Morphological Analyzer for the Kazakh and Turkish Languages", *Conference: Sixth International conference on Analysis of Images, Social Networks, and Texts*, (2017): 1–11.
- [11] Akhmed-Zaki D., Mansurova M., Madiyeva G., Kadyrbek N., Kyrgyzbayeva M., "Development of the information system for the Kazakh language preprocessing", *Cogent Engineering*, 8(1) (2021): 1–15.
- [12] Makhambetov O., Makazhanov A., Sabyrgaliyev I., Yessenbayev Zh., "Data-driven morphological analysis and disambiguation for Kazakh", *Computational Linguistics and Intelligent Text Processing*, (2015): 151–163.
- [13] Yiner Z., Kurt A., "Two Level Kazakh Morphology\*", *Acta Infologica*, 5(1) (2021): 79–98.
- [14] Mesnil G., Dauphin Y., Yao K., "Using recurrent neural networks for slot filling in spoken language understanding", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3) (2015): 530–539.
- [15] Lee W., Jung B., Shin J., Lee J-H., "Adaptation of Back-translation to Automatic Post-Editing for Synthetic Data Generation", *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, (2021): 3685–3691.
- [16] Rubino R., Marie B., Dabre R., "Extremely low-resource neural machine translation for Asian languages", *Machine Translation*, 34 (2020): 347–382.
- [17] Imankulova A., Sato T., Komachi M., "Improving low-resource neural machine translation with filtered pseudo-parallel corpus", *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, (2017): 70–78.
- [18] Sennrich R., Haddow B., Birch A., "Improving Neural Machine Translation Models with Monolingual Data", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1 (2016): 86–96.
- [19] Lopes A., Farajian A., Correia G., Trénous, J., Martins A., "Unbabel's Submission to the WMT2019 APE Shared Task: BERT-Based Encoder-Decoder for Automatic Post-Editing", *Proceedings of the Fourth Conference on Machine Translation*, 3 (2019): 118–123.
- [20] Zhumanov Zh., Madiyeva A., Rakhimova D., "New Kazakh Parallel Text Corpora with On-line Access", *Conference on Computational Collective Intelligence Technologies and Applications*, (2017): 501–508.
- [21] Sapin A.S., "Building neural network models for morphological and morpheme analysis of texts", *Trudy ISP RAN/Proc. ISP RAS*, 33(4) (2021): 117–130 (in Russian). DOI: 10.15514/ISPRAS2021-33(4)-9.

**Information about authors:**

*Madina Mansurova (corresponding author) – Candidate of Physical and Mathematical Sciences, Professor of the Department of Artificial Intelligence and Big Data at the Information Technology Faculty of Al-Farabi Kazakh National University (Almaty, Kazakhstan, email: mansurova.madina@gmail.com);*

*Diana Rakhimova – PhD, Associate Professor of the Department of Information Systems at the Information Technology Faculty of Al-Farabi Kazakh National University (Almaty, Kazakhstan, email: di.diva@mail.ru);*

**Авторлар туралы мәлімет:**

*Мадина Мансурова (корреспондент автор) - физика-математика ғылымдарының кандидаты, әл-Фараби атындағы Қазақ ұлттық университетінің ақпараттық технологиялар факультетінің жасанды интеллект және үлкен деректер кафедрасының профессоры (Алматы қ., Қазақстан, email: mansurova.madina@gmail.com);*

*Диана Рахимова – PhD, әл-Фараби атындағы Қазақ ұлттық университетінің ақпараттық технологиялар факультетінің ақпараттық жүйелер кафедрасының қауымдастырылған профессор м.а. (Алматы қ., Қазақстан, электрондық пошта: di.diva@mail.ru).*

*Received: September 23, 2024*

*Accepted: November 15, 2024*