# ANNOTATION

## to the dissertation work of the PhD student Karibayeva Aidana Seilgazykyzy in the specialty "Information systems" 6D070300 on the topic: " Development and research models and methods of morphological segmentation of Kazakh texts for neural machine translation"

**The relevance of the research topic.** Machine translation is one of the central tasks of artificial intelligence, and neural machine translation is the most modern approach, based on which the best results in machine translation and image recognition are obtained.

Many problems of machine translation are not fully understood and require detailed consideration depending on the specifics of the language. Machine translation systems cannot always solve the problem with traditional methods. In rule-based machine translation, all the rules may not be taken into account; in statistical translation, the correct translation is not always determined by context. Today, the use of neural networks has become popular in many subject areas, and machine translation is no exception.

There are different approaches to solving machine translation problems, such as the approach based on the grammar rules of languages; a statistical machine translation approach based on a statistical approach to finding a probabilistic phrase table of translated languages; a neural machine translation approach based on training neural networks of translated languages. Each of these approaches has advantages and disadvantages. Recently, the best results of machine translation have been shown by an approach based on neural networks, neural machine translation. Since the problem of machine translation has not yet been solved at a sufficiently high level, close to professional translation, the problem of machine translation is very urgent. It should be noted that solving the problem of machine translation paves the way for solving other very important problems of artificial intelligence, such as understanding natural language.

Neural machine translation is based on the mechanism of recurrent neural networks based on matrix calculations, which makes it possible to create more complex probabilistic models than statistical machine translators. The direction of neural machine translation is a hot topic in natural language processing, because neural machine translation outperforms rule-based and statistical machine translation.

There are several topical problems in natural language processing problems. One of the most widespread and used tasks for improving translation quality is: segmentation. Segmentation for natural language is an actual research topic in computational linguistics and remains an open problem today. Most methods suggest frequency-based segmentation that does not consider the morphological features of the language. This method includes the BPE (byte-pair encoding) method. BPE-based segmentation does not give good results for agglutinative languages. The paper proposes a segmentation method based on the morphological features of the Kazakh language based on the CSE (complete set of endings) -model.

Dictionaries play an important role in neural machine translation (NMT). However, a large dictionary requires a significant amount of memory, which limits the application of NMT and can cause a memory error. This limitation can be solved by segmenting each word into morphemes in parallel source corpora. Therefore, this study introduces a new morphological segmentation approach for Turkic languages based on the complete set of endings (CSE), which reduces the vocabulary volume of the source corpora. When training NMT, the volume of the corresponding NMT dictionary rapidly increases; therefore, it requires excessive computer memory resources.

There are several preprocessing steps in machine translation. Segmentation of the text is one of the preparatory stages for machine translation to reduce the volume of the vocabulary. The segmentation problem has been investigated for analytical languages in many ways, while for agglutinative languages, namely for Turkic-speaking, the amount of research is small. Neural networks usually create a large dictionary to translate most of the words in the target language. In neural machine translation, the segmentation problem appears when training neural networks to reduce the volume of the vocabulary, when the volume of the vocabulary requires more memory, as well as for solving problems of unknown and rare words. Segmenting for a given text is one such solution. Therefore, the problem of the **relevance** of text segmentation in neural machine translation is increasing.

**The purpose of the dissertation work.** Development of models, algorithms, and software for improving neural machine translation of the Kazakh language based on linguistic features.

**Research objectives.** To achieve the objectives of the study, the following tasks are considered:

1) Improvement of the language model of the morphology of the Kazakh language based on the Complete Set of Endings (CSE - Complete Set of Endings) (expanding the list of possible endings of the Kazakh language);

2) Creation of a morphological segmentation model and algorithm based on the CSE-model of the morphology of the Kazakh language;

3) Conducting and creating experiments on defined tasks on the neural machine translation platform.

**The object of the study.** Kazakh language.

**The subject of the study.** Neural machine translation for the Kazakh language

**Research methods.** Numerical methods of combinatorial analysis, machine learning, deep learning and neural networks were used as research methods.

**Scientific novelty of the work:**

1. An improved computational model of the morphology of the Kazakh language has been developed based on the enumeration of possible endings, which is distinguished by the creation of a complete set of endings of the language.

2. A method and algorithm of morphological segmentation based on an improved computational model of the morphology of the Kazakh language have been developed, which differ from the known ones in the formation of a complete set of language endings in the form of a decision table and provide a reduction in the volume of the neural machine translation dictionary. This allows you to train neural machine translation on a much larger amount of initial data (a larger volume of vocabulary).

**Theoretical and practical significance of the work.** The theoretical importance of this work lies in the creation of a universal new method of morphological segmentation, taking into account the linguistic features of the Kazakh language. The morphological segmentation method based on the created CSE-model can be applied to other Turkic languages as well.

**The practical significance of the work** is that learning neural machine translation based on segmented text continuously reduces the amount of memory and avoids errors with memory.

**The basic concept for defense.** New model and algorithm of morphological segmentation of words in the Kazakh language, the results of experiments on neural machine translation of the Kazakh language, confirming the effectiveness of the proposed model and algorithm of segmentation of words in the Kazakh language.

**The degree of reliability and the results of testing.** The reliability and validity of the results of the research is ensured by the reasonable responsibility of setting the tasks, the analysis of the criteria and the state of research in the given field, the conducted experiments, as well as the improvements of their results in the neural machine translation of the Kazakh language. The results of the dissertation were published in the following publications below.

*Journal article in the Scopus database:*

1) Tukeyev U., Karibayeva A., Zhumanov Zh. Morphological segmentation method for Turkic language neural machine translation. *Cogent Engineering*, 2020, 1 том, номер №1. (Scopus:Q2; CiteScore-2.5; Percentile- 73%))

2) Turgangayeva A., Rakhimova D., Karyukin V., Karibayeva A, Turarbek A. Semantic Connections in the Complex Sentences for Post-Editing Machine Translation in the Kazakh Language. *Information* 2022, *13*(9), 411; https://doi.org/10.3390/info13090411(Scopus: Q2; CiteScore 4.2; Percentile-64%)

3) Rakhimova D., Karibayeva A. Aligning and extending technologies of parallel corpora for the Kazakh languge. *Eastern-European Journal of Enterprise Technologies*, 2022, 4(2-118), стр. 32–39 (Scopus: Q3; Citescore: 2.0; Percentile: 37%)

*In journals submitted by the Committee of Control in the field of education and science:*

1) Karibayeva A., Rakhimova D., Abduali B., Amirova. Analysis of machine translation of the Kazakh language. Bulletin of KazNITU. No. 3 (127), KazNITU, 2018, 90 - 96 p.

2) Рахимова Д., Тұрарбек А., Карюкин В., Карибаева А., Тұрғанбаева Ә. Қазақ тіліне арнлаған заманауи машиналық аударма технологияларына шолу. Вестник КазНИТУ, №5 (141) 2020. -стр. 103-110.

3) Абдуали Б.А., Әмірова Д.Т., Рахимова Д.Р., Кәрібаева А.С. Қазақ тіліндегі мәтінді ресурстар мен құжаттарды аналитикалық өңдеу. Вестник КазНИТУ, №2(132), 2019, стр. 356-362. (CCES)

4) Karibayeva A., Karyukin V., A. Turgynbayeva, A. Turarbek. The translation quality problems of machine translation systems for the Kazakh language. Journal of Mathematics, Mechanics and Computer Science, [S.l.], v. 111, n. 3, p. 132-140, oct. 2021. ISSN 2617-4871. (CCES)

*Conferences on Web Science and Scopus:*

1) Tukeyev U., Amirova D., Karibayeva A., Sundetova A., Abduali B. Combined technology of lexical selection in rule-based machine translation. Computational Collective Intelligence: 9th International Conference, ICCCI 2017, Nicosia, Cyprus, September 27-29, 2017, Proceedings, Part II (Lecture Notes in Computer Science) 1st ed. 2017 Edition, Springer, p. 491-500 (**Q3, SJR=0.25, CS=1.8, Percentile-50%).**

2) Tukeyev U., Karibayeva A., Abduali B. Neural machine translation system based on synthetic corpora. CMES-2018, Poland, Kazimeirz Dolny, 2018, MATEC Web of Conferences. 252. 03006. 10.1051/matecconf/201925203006 (**Web of Science**).

3) Tukeyev U., Turganbayeva A., Abduali B., Rakhimova D., Amirova D., Karibayeva A. Lexicon-free stemming for Kazakh language information retrieval. DOI:10.1109/ICAICT.2018.8747021.AICT-2018, Kazakhstan, Almaty (**Scopus)**.

4) Tukeyev U., Karibayeva A. Inferring the Complete Set of Kazakh Endings as a Language Resource. Proceedings of International Conference on Computational Collective Intelligence, 2020, p. 741-751 **(Q4, SJR=0.209, CS=0.9, Percentile – 16%)**

5) Tukeyev U., Karibayeva A., Turganbayeva A., Amirova D. Universal Programs for Stemming, Segmentation, Morphological Analysis of Turkic Words. Computational Collective Intelligence. ICCCI 2021. Lecture Notes in Computer Science, vol 12876. Springer, Cham. https://doi.org/10.1007/978-3-030-88081-1_48 -p. 643–654 (**Q3, SJR=0.25, CS=1.8, Percentile-50%**).

6) Rakhimova D., Karyukin V., Karibayeva A., Turarbek A., Turganbayeva A. The Development of the Light Post-editing Module for English-Kazakh Translation. DOI:

https://doi.org/10.1145/3492547.3492651. ICEMIS'21: The 7th International Conference on Engineering & MIS 2021, Almaty, Kazakhstan, October 2021 (**Scopus**)

*In the international conferences:*

1. Tukeyev U., Sundetova A., Abduali B., Karibayeva A., Amirova D. Technology of the structural machine translation rules generation, based on the complete set of Kazakh endings // Информатика және қолданбалы информатика: Халықаралық ғылыми конференция материалдары (27-30 қыркүйек 2017 ж). 2-бөлім. - Алматы, 2017, - 38 б.
2. Tukeyev U., Zhumanov Zh., Karibayeva A., Amirova D., Sundetova A., Abduali B. Формирование двуязычного словаря многозначных слов для машинного перевода казахского языка” The Vth International Conference on Computer Processing of Turkic Languages “TurkLang 2017”, 18-21 October, Kazan, Tatarstan.
3. Tukeyev U., Zhumanov Zh., Sundetova A., Abduali B., Karibayeva A., Amirova D. Technology of the structural machine translation rules genearation, based on the complete set of Kazakh endings. The II International Conference “Computer Science and Applied Mathematics”, 2017, Part II, Almaty, Kazakhstan.
4. Tukeyev U., Zhumanov Zh., Rakhimova D., Karibayvea A, Amirova D. Complex technology of machine translation resources extension for the Kazakh language. Varia Informatica 2017 №1, ISBN 978-83-936692-3-3, Lublin, 14 стр
5. Karibayeva A., Abduali B., Amirova D. Formation of the synthetic corpora for Kazakh on the base of endings complete system. Turklang-2018, Uzbekistan, Tashkent, pp. 153 – 161.
6. Кәрібаева А.С, Абдуали Б.А, Тукеев У. А. Разработка программы морфологической сегментации текста казахского языка на основе полной системы окончаний. «Фараби әдемі» атты студенттер мен жас ғалымдардың халықаралық ғылыми конференция", Казахстан, Алматы, 2020, - 53 стр.
7. Әмірова Д.Т., Кәрібаева А.С. Исследование технологии машинного перевода казахско-английской пары языков и обратно на основе трансферной модели нейронной сети. «Фараби әлемі» атты студенттер мен жас ғалымдардың халықаралық ғылыми конференция", Қазақстан, Алматы, 2020, -45 стр.
8. Рахимова Д.Р., Турарбек А., Карибаева А., Карюкин В. Технологий машинного перевода и постредактирования казахского языка. Глава в коллективной монографии «Современные методы и подходы обработки казахского языка» КГТУ, Бишкек 2021

**Personal contribution of the applicant.** The applicant solved the tasks of the dissertation work. A model and method of morphological segmentation of the text was developed in neural machine translation of the Kazakh language. A corpus of parallel texts in the Kazakh language was compiled for training and testing in the neural machine translation system. Experiments were conducted to determine the effectiveness of the developed model and method. A complete list of endings was created for the Kazakh language based on the CSE (complete set of endings) model.

**Connection of the topic of the dissertation with the plans of scientific research.** The thesis was carried out within the framework of the grant research project of the Ministry of Education and Science of the Republic of Kazakhstan on "Development and research of neural machine translation of the Kazakh language" (2017-2020).

**Structure and scope of work.** The dissertation consists of an introduction, 4 sections and a conclusion. The total volume of the dissertation consists of 172 pages, 7 pictures, 54 tables, and 79 references.

The relevance of the work was determined in the **introduction**, and the problems related to the topic were indicated. The idea of the work, the purpose and objectives of the research, the scientific novelty and practical value of the research, research methods are shown.

**The first section** describes the research and analysis of existing technologies for improving neural machine translation. An analytical review of research in the field of machine translation segmentation and language morphology modeling is made. The main advantages and disadvantages of the methods are identified.

**The second section** analyzes the models used in the description of language morphology. Taking into account the morphological models, the work of creating a language model of the Kazakh language morphology based on the complete system of conjunctions is described.

**In the third section**, work was carried out to create a model and algorithm of morphological segmentation using the model of the complete system of connections (CSE). A step-by-step algorithm of morphological segmentation was created. A dictionary of special words was created to avoid wrong segmentation.

**In the fourth chapter**, software is selected for training the system of neural machine translation of the Kazakh language, the program for segmentation of texts in the Kazakh language is described. The main neural network models used in training neural machine translation systems are described. The learning process was implemented with the seq2seq model based on recurrent neural networks in the Tensorflow library. Experiments with a constructed method (CSE) and another method (BPE) are described. The results of the experiment are analyzed. To determine the quality of neural machine translation, experimental work was carried out on the comparison of the segmentation method, in particular, with BPE, quality results were obtained in the BLEU metric.

**The conclusion** presents the main results and conclusions of the dissertation.

The obtained scientific results are confirmed by experiments with different training configurations. Validity and reliability of the study correspond to the results of the developed method.