ABSTRACT

of the PhD thesis by Aitim Aigerim Kairatkyzy on «Models and methods for the automatic processing of unstructured information», submitted for the degree of Doctor of Philosophy (PhD) in the EP 8D06101 – Clever Systems

Relevance of the research is determined by the urgent need to develop intelligent language technologies for the Kazakh language, as part of the broader national strategy for digital transformation, scientific innovation, and cultural modernization in the Republic of Kazakhstan. The work lies at the intersection of artificial intelligence, natural language processing (NLP), and national language policy, addressing both scientific and socio-technical challenges of integrating Kazakh into modern digital systems.

One of the main drivers of this research is the implementation of the "Digital Kazakhstan" government program, launched in 2017, which aims to modernize the economy, public services, and infrastructure through the introduction of digital technologies. A key component of this program is the creation of inclusive digital services that are accessible in the Kazakh language. However, the absence of robust tools for the automatic processing of Kazakh-language texts such as part-of-speech taggers, morphological analyzers, and semantic extractors presents a significant barrier to this goal. This dissertation directly contributes to overcoming this gap by building computational models capable of processing unstructured Kazakh-language texts with high linguistic fidelity.

Moreover, the research supports the goals of the "Rukhani Zhangyru" (Spiritual Modernization) program, which emphasizes the preservation, development, and digital dissemination of Kazakh language and culture. The proposed NLP tools including the morphological analyzer, KazBERT-based taggers, and annotated corpora promote the use of the Kazakh language in modern communication technologies, education platforms, and cultural archives. By enabling linguistic analysis and understanding of large-scale digital content, this work contributes to the linguistic sovereignty and digital identity of Kazakhstan.

In addition, the study aligns with the priorities outlined in the National Science Development Strategy until 2025, which highlights the importance of developing AI technologies, big data methods, and digital linguistic resources for the Kazakh language. This research introduces a scientifically rigorous methodology for building language resources and deep learning models tailored to the Kazakh linguistic system, thereby contributing to the technological advancement of low-resource language processing within the country.

The research also supports the "Educated Nation" (Білімді ұлт) initiative, which focuses on the digitalization of education and the development of innovative tools for personalized learning. The NLP components developed in this dissertation can be integrated into intelligent tutoring systems, Kazakh

language learning applications, and digital textbooks, ensuring high-quality and linguistically informed educational content in the state language.

Taken together, these national programs create a strategic demand for advanced computational models capable of processing and understanding unstructured information in Kazakh. The results of this research including the QNLP software library, linguistic ontology, and annotated datasets - offer practical solutions for embedding the Kazakh language into the core of Kazakhstan's digital infrastructure. This ensures that future technological systems in the country will not only support Kazakh users but also reinforce the role of the Kazakh language in the information age.

The object of the research is the unstructured textual information in the Kazakh language as encountered in real-world domains, including digital news, social media, literature, and official documentation.

The subject of the research is the features and principles of computational models and methods for the automatic processing, analysis, and linguistic annotation of unstructured Kazakh texts, focusing on fundamental NLP tasks such as tokenization, POS tagging, morphological analysis, syntactic parsing, and semantic labeling.

The purpose of the research is to develop an integrated system and a set of linguistic tools that combine machine learning methods with language rules for the automatic and accurate processing of unstructured textual information in the Kazakh language. Special attention is given to the analysis of real-world Kazakh-language texts from news sources, where sentence structures vary and do not conform to formalized patterns, requiring flexible and adaptive solutions in the field of computational linguistics.

Objectives of the research is to develop theoretical foundations, computational models, and software tools for the automatic processing of unstructured textual information in the Kazakh language:

– to conduct a literature review on methods and models used in the automatic processing of unstructured textual data, with a special focus on agglutinative and Turkic languages.

– to identify the limitations and challenges of applying existing NLP architectures to the Kazakh language.

– to collect corpus of unstructured Kazakh texts from online sources such as news portals, blogs, and official documents.

– to design and implement automatic text cleaning, deduplication, language filtering, and pre-annotation pipelines.

– to develop and evaluate part-of-speech tagging models, using both rule-based and data-driven (machine learning, deep learning) approaches.

– to construct models for named entity recognition (NER) and dependency parsing using Kazakh-specific linguistic features and pretrained language models (e.g., KazBERT).

– to implement a hybrid model architecture for sequence labeling tasks.

– to design and develop the QNLP library as a unified, modular software toolkit that includes all developed models and resources.

Research questions:

How can we build NLP models capable of addressing the agglutinative and rich morphological characteristics of the Kazakh language?

What are the most effective methods for building annotated corpora in low-resource settings?

To what extent can transfer learning and pretrained multilingual models like KazBERT be adapted to Kazakh?

How can rule-based and machine learning approaches be effectively combined for morphological and syntactic tasks?

Research hypothesis is postulated that the integration of morphological rule-based models with pretrained Kazakh language transformers, in combination with deep neural architectures, will lead to significant improvements in the accuracy and robustness of automatic text processing systems for the Kazakh language. Furthermore, it is assumed that the development of a carefully constructed corpus and a linguistic ontology will enable the effective modeling of the unique linguistic structures of Kazakh, thereby facilitating cross-task generalization in natural language processing under low-resource conditions.

Scientific novelty.

The thesis work obtained the following main scientific results:

– an original corpus of unstructured Kazakh text has been collected, cleaned, annotated released for reproducible experiments.

– a new developed Python library QNLP for linguistic processing, offering comprehensive tools for morphological analysis, POS tagging, dependency parsing, and NER - all tailored to the specific characteristics of the Kazakh language.

– the first implementation of rule-based and KazBERT + BiLSTM + CRF model trained on an annotated Kazakh corpus for sequence labeling tasks.

Scientific provisions submitted for defense:

– an annotated corpus for major Kazakh NLP tasks, created through a hybrid annotation strategy.

– a hybrid architecture for automatic linguistic analysis of Kazakh unstructured texts, combining traditional linguistic rules and deep learning techniques.

– A Python-based QNLP library has been developed for morphological analysis, POS tagging, dependency parsing, and named entity recognition (NER) of Kazakh texts.

Theoretical significance lies in advancing the theory and methodology of natural language processing for morphologically rich and agglutinative

languages, particularly in defining computational models adapted to Turkic linguistic typology.

Practical significance involves the development of ready-to-use NLP tools and annotated datasets that can be employed in real-world applications such as machine translation, search engines, digital assistants, and linguistic education platforms in Kazakh.

The practical value of the study is reusable open-source NLP library (QNLP) for Kazakh with practical modules for POS tagging, morphological analysis, and NER. A linguistic annotation protocol and annotation toolchain adapted for Kazakh. Contribution to language preservation through digital means by enabling automated processing of Kazakh texts. Tools and resources from this study can be integrated into national digital infrastructure and language technology products.

Research methods. The study employs a combination of qualitative and quantitative methods: collection and preprocessing of large-scale Kazakh text corpora. rule-based morphological modeling, POS tagging, and dependency grammar construction, application of statistical and neural sequence labeling models, accuracy, F1-score metrics used for model performance assessment, use of hybrid manual and automated annotation techniques.

Approbation of work. The main propositions and scientific results of the work were presented and discussed at seminars of the «Information Systems» department at the International Information Technology University and International Conferences:

1. The 6th International Conference on Engineering & MIS 2020, ICEMIS'20 (DTESI'20), September 14–16, 2020, ACM International Conference Proceeding Series, 2020.
2. The 15th International Conference on Emerging Ubiquitous Systems and Pervasive Networks/ EUSPN, Procedia Computer Science, 2024.
3. The 9th International Conference Digital Technologies in Education, Science, and Industry, 2024.

Publications: The main results obtained during the dissertation work have been published in 13 printed works, including 4 articles in publications recommended by the Committee for Control in the Field of Education and Science of the Ministry of Education and Science of the Republic of Kazakhstan, 2 articles in Eastern-European Journal of Enterprise Technologies (Q3) indexed by

Scopus in a high-impact scientific journal with cite score 2.0 and a percentile of 46, and 4 articles in proceedings of international conferences, of which one scientific article with cite score 4.5 and a percentile of 69.

The results obtained on the topic of the dissertation are presented in the following publications:

1. Aitim A.K., Satybaldiyeva R.Zh., Linguistic ontology as means of modeling of a coherent text, Bulletin of Abai KazNPU. Series of Physical and

mathematical sciences. 3 (Sep. 2022), https://doi.org/10.51889/2022-3.1728-7901.18.

2. Aitim A.K., Developing methods for automatic processing systems of Kazakh language, Bulletin of KazATC 133 (4), 2024, ISSN 1609-1817, ISSN Online 2790-5802, https://doi.org/10.52167/1609-1817.

3. Aitim A.K., Satybaldiyeva R.Zh., A systematic review of existing tools to automated processing systems for Kazakh language. Bulletin of Abai KazNPU. Series of Physical and mathematical sciences. 87, 3 (Sep. 2024), 106–122, https://doi.org/10.51889/2959-5894.2024.87.3.009.

4. Aitim A.K., Satybaldiyeva R.Zh., Building methods and models for automatic processing systems of Kazakh language, Bulletin of KazATC №2-137-2025, https://doi.org/10.52167/1609-1817-2025-137-2-346-356

5. Aitim A.K., Satybaldiyeva R.Zh. A comparison of Kazakh language processing models for improving semantic search resultsEastern-European Journal of Enterprise Technologies, 1(2 (133), 66–75, 2025. https://doi.org/10.15587/1729-4061.2025.315954

6. Aitim, A., Sattarkhuzhayeva, D., & Khairullayeva, A. (2025). Development of a hybrid CNN-RNN model for enhanced recognition of dynamic gestures in Kazakh Sign Language. Eastern-European Journal of Enterprise Technologies, 2025, 2(2 (134), 58–67. https://doi.org/10.15587/1729-4061.2025.315834

7. Aitim A.K., Satybaldiyeva R.Zh., Wojcik W.The construction of the Kazakh language thesauri in automatic word processing system the 6th International Conference on Engineering & MIS 2020, ICEMIS'20 (DTESI'20), September 14–16, 2020, ACM International Conference Proceeding Series, 2020, https://doi.org/10.1145/3410352.341078

8. Aitim A.K., Abdulla M.A. Data processing and analyzing techniques in UX research15th International Conference on Emerging Ubiquitous Systems and Pervasive Networks/ EUSPN, Procedia Computer Science, volume 251, 2024, Pages 591-596, https://doi.org/10.1016/j.procs.2024.11.154

9. Aitim A.K., Abdulla M.A., Altayeva A.B. Sentiment Analysis using Natural Language Processing 9th International Conference Digital Technologies in Education, Science and Industry, October 16-17, 2024, Almaty.

10. Aitim A.K., Abdulla M.A., Altayeva A.B. Human-Centric AI: Improving User Experience with Natural Language Interfaces, 9th International Conference Digital Technologies in Education, Science and Industry, october 16-17, 2024, Almaty.

11. Aitim A.K., I. Khlevna. Models of natural language processing for improving semantic search results // International Journal of Information and Communication Technologies. 2022. Vol. 3. Is. 2. Number 10. Pp. 82–91. https://doi.org/10.54309/IJICT.2022.10.2.008.

12. Aitim A.K., Satybaldiyeva R.Zh. Analysis of methods and models for automatic processing systems of speech synthesis. International Journal of

Information and Communication Technologies, 1(2), 2020. https://doi.org/10.54309/IJICT.2020.2.2.019

13. Aitim A.K., Wojcik W. Satybaldiyeva R.Zh. Methods of applying linguistic ontologies in text processing. Journal, News of the scientific and technical society KAKHAK, 2021, Volume-1, Issue - 72, Pages - 132-137.

Thesis Structure. The work consists of an introduction, four sections, a conclusion, a list of references, and applications.

Main content of the thesis.

This work consists of four main chapters.

The first chapter presents a thorough literature review and conceptual foundation for Kazakh language processing. It analyzes the historical development and current state of NLP tools for Kazakh, highlights the complexity of agglutinative morphology, and introduces modern neural architectures such as KazBERT, CRF, and BiLSTM. The chapter also reviews prior linguistic methods, early systems, and the challenges of adapting universal models to the specificities of Kazakh.

The second chapter introduces QCrawler, a custom-built system designed to automatically collect Kazakh-language texts from the web. It outlines the architecture, algorithm, and mathematical model behind the crawler, and compares its efficiency and coverage with baseline scraping methods. The chapter also presents a preliminary application of named entity recognition (NER) using KazBERT+CRF on the newly collected corpus.

The third chapter focuses on the linguistic foundations of QNLP, including the development of a morphological analyzer, a Kazakh thesaurus, and an ontology of Kazakh morphology. It describes how word structure is handled through root-suffix modeling, enabling high-accuracy segmentation and generation of inflected forms. These lexical and rule-based components support deeper linguistic analysis and enhance model interpretability.

The fourth chapter presents the full QNLP toolkit - an end-to-end open-source framework for Kazakh NLP. It details the modular architecture, component interactions, and internal mathematical formulations. The chapter includes experiments evaluating POS tagging, NER, and morphology, an ablation study on KazBERT, BiLSTM, and CRF, and a comparison with existing tools. It concludes with visualizations, accuracy trends, and performance benchmarks showing QNLP’s state-of-the-art effectiveness.