

## **О Т З Ы В**

на диссертационную работу **Әйтiм Әйгерiм Қайратқызы** «Модели и методы автоматической обработки неструктурированной информации», представленную на соискание ученой степени доктора философии (PhD) по ОП: 8D06101 – «Интеллектуальные системы»

Диссертация, выполненная **Әйтiм Әйгерiм Қайратқызы**, представляет собой всестороннее исследование, посвящённое одной из наиболее актуальных и наукоёмких задач современной лингвистической информатики - разработке комплексной архитектуры автоматизированной обработки неструктурированных текстов на казахском языке с использованием современных методов искусственного интеллекта и машинного обучения. В условиях цифровой трансформации Казахстана и стремительного роста объема текстовой информации на казахском языке возрастает необходимость в полноценных NLP-инструментах, адаптированных под особенности казахской морфологии, синтаксиса и семантики. На сегодняшний день, несмотря на отдельные инициативы, отсутствуют решения, способные в полной мере охватить задачи: морфологического анализа, частеречной разметки (POS), распознавания именованных сущностей (NER), синтаксико-семантического анализа, порождения словоформ, вопросно-ответных систем, основанных на спецификах агглютинативного строя казахского языка. Работа **Әйтiм Ә.Қ.** отвечает на вызовы времени, предлагая архитектурно и лингвистически выверенную многоуровневую систему на стыке компьютерной лингвистики, нейросетевых технологий и национальной языковой политики.

Целью диссертационной работы является построение комплексной интеллектуальной системы автоматической обработки неструктурированной текстовой информации на казахском языке на основе современных методов глубинного обучения и лингвистического моделирования. Для достижения цели были решены следующие задачи: создание масштабного, представительного корпуса казахских текстов; автоматизированная очистка, нормализация и структурирование данных; построение POS-модели с учетом морфологических особенностей; реализация собственного морфологического анализатора и генератора словоформ, обучение моделей NER, объединение всех компонентов в единую библиотеку QNLP. Структура диссертации включает логически взаимосвязанные главы, переходящие от теоретических основ к прикладной реализации: постановка задачи и обоснование темы; анализ состояния области; создание и обработка корпуса; реализация морфологического анализатора; построение и обучение нейросетевых моделей; интеграция в библиотеку и экспериментальная проверка.

Научная новизна исследования заключается в построении первой комплексной архитектуры NLP-библиотеки QNLP с поддержкой полного цикла обработки казахских текстов. Разработан оригинальный морфоанализатор и генератор словоформ, реализован KazBERT + BiLSTM + CRF для POS и NER, создан вручную размеченный корпус, формализована структура казахской морфологии и предложены алгоритмы лингвистического анализа.

Практическая значимость работы подтверждается тем, что библиотека QNLP уже применяется в научных, образовательных и технологических проектах, включая анализ социальных сетей, автоматическую проверку письменных работ, распознавание тем и сущностей в новостях, генерацию казахских ответов в чат-ботах. Все модули протестированы, сопровождаются документацией и открытым исходным кодом, что подтверждает высокий уровень реализации и пригодность к тиражированию.

Диссертация написана на высоком научном уровне, логично структурирована, снабжена наглядными диаграммами, таблицами, графиками, лингвистическими примерами и блоками кода. Автор демонстрирует глубокое понимание как лингвистических, так и технических аспектов задач, использует широкий инструментарий анализа, соблюдает академический стиль изложения. Результаты исследования представлены в публикациях в рецензируемых журналах и докладах на международных конференциях, что свидетельствует о признании их научной ценности.

Учитывая вышесказанное, считаю, что докторская диссертация «Модели и методы автоматической обработки неструктурированной информации» является законченной самостоятельной научно-исследовательской работой, отвечающей всем требованиям предъявляемым к диссертационным работам по данной специальности, а ее автор **Әйтiм Әйгерiм Қайратқызы** заслуживает присуждения ученой степени доктора философии (PhD) по ОП: 8D06101 – «Интеллектуальные системы»

Научный консультант,  
к.т.н, ассоциированный профессор  
кафедры «Кибербезопасности, обработки и  
хранения информации», Satbayev University



Сатыбалдиева Р.Ж.

## **REVIEW**

on the PhD thesis by Aitim Aigerim Kairatkyzy entitled "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

The thesis presented by Aitim Aigerim Kairatkyzy is a comprehensive research study focused on one of the most relevant and technologically demanding problems in modern computational linguistics - the development of an integrated architecture for the automated processing of unstructured Kazakh-language texts using advanced methods of artificial intelligence and machine learning.

In the context of Kazakhstan's ongoing digital transformation and the rapid growth in the volume of Kazakh-language textual data (media, blogs, governmental services, educational and legal documents), there is a pressing need for robust NLP tools that are tailored to the morphological, syntactic, and semantic features of the Kazakh language. To date, despite several partial initiatives, no complete solutions exist that adequately address the full range of tasks including morphological analysis, part-of-speech (POS) tagging, named entity recognition (NER), syntactic-semantic parsing, word form generation, and question answering - especially considering the agglutinative nature of the Kazakh language.

The work of Aitim A.K. responds to this urgent demand by proposing a linguistically and architecturally sound multi-level system situated at the intersection of computational linguistics, neural network technologies, and national language development strategies.

The aim of the thesis is to construct a comprehensive intelligent system for the automatic processing of unstructured Kazakh-language textual information based on modern deep learning methods and linguistic modeling. To achieve this, the following objectives were addressed: the creation of a large, representative corpus of Kazakh texts; automated cleaning, normalization, and structuring of data; development of a POS tagging model accounting for Kazakh morphological features; implementation of a custom morphological analyzer and word form generator; training of models for NER, semantic role labeling (SRL), and coreference resolution; and the integration of all components into a unified library called QNLP. The thesis is structured into logically connected chapters progressing from theoretical foundations to practical implementation: problem formulation and motivation; review of existing approaches; corpus construction and preprocessing; development of a morphological analyzer; design and training of neural models; and finally, integration into the QNLP library with extensive experimental evaluation.

The scientific novelty of the research lies in the development of the first comprehensive NLP library for Kazakh - QNLP, supporting the full pipeline of text processing. The work introduces an original morphological analyzer and word form generator, adapts KazBERT + BiLSTM + CRF for POS and NER tasks, and builds a manually annotated corpus for training and evaluation. Additionally, the author formalized the structure of Kazakh morphology and proposed linguistically informed analytical algorithms.

The practical significance of the work is demonstrated by the fact that QNLP is already being used in scientific, educational, and technological projects, including social media analysis, automated evaluation of student writing, topic and entity recognition in news media, and Kazakh-language chatbot systems. All modules have been tested, include proper documentation, and are distributed as open-source software - indicating a high level of implementation readiness and potential for wide-scale deployment.

The thesis is written to a high scientific standard, logically organized, and supported with clear diagrams, tables, charts, linguistic examples, and code snippets. The author demonstrates deep understanding of both the linguistic and technical aspects of the problem, applies a wide range of analytical tools, and maintains a formal academic style. The research results have been published in peer-reviewed journals and presented at international conferences, confirming their recognition and value to the scientific community.

Based on the above, I conclude that the dissertation "Models and methods for the automatic processing of unstructured information" is a completed independent research work that fully meets the standards for dissertations in this specialty. The author, Aitim Aigerim Kairatkyzy, fully deserves to be awarded the degree of Doctor of Philosophy (PhD) in the EP 8D06101 – Clever Systems. -

**Scientific consultant**  
**Candidate of Technical Sciences,**  
**Associate Professor of the**  
**Department of Cybersecurity, Information**  
**Processing and Storage, Satbayev University**



**Satybaldiyeva R.Zh.**