

REVIEW OF THE OFFICIAL REVIEWER
on the dissertation work of Aitim Aigerim Kairatkyzy
on the topic «Models and methods for the automatic processing of unstructured information»,
provided for the degree of Doctor of Philosophy (PhD) in the educational program: 8D06101 – «Clever Systems»

№	Criteria	Eligibility (please select one answer option)	Justification of the official reviewer's position
1	The topic of the thesis (as of the date of its approval) corresponds to the directions of scientific development and/or government programs	1.1 Compliance with priority areas of scientific development or government programs:	The dissertation <u>corresponds</u> to the objectives of the following state programs: 1. The National Project “Digital Kazakhstan”, approved by the Resolution of the Government of the Republic of Kazakhstan dated December 12, 2017, No. 827. 2. The Concept for the Development of Language Policy of the Republic of Kazakhstan for 2023–2029, approved by the Decree of the President of the Republic of Kazakhstan dated May 5, 2023, No. 1. 3. The Strategic Development Plan of the Republic of Kazakhstan until 2025, approved by the Decree of the President of the Republic of Kazakhstan dated February 15, 2018, No. 636. 4. The Concept for the Development of Science in the Republic of Kazakhstan for 2022–2026, approved by the Resolution of the Government of the Republic of Kazakhstan dated October 19, 2021, No. 736. The “Digital Government” program within the framework of “Digital Kazakhstan” further defines the objectives for the implementation of AI and NLP technologies in public service delivery.
		1. The dissertation was completed within the project or target program financed from the state budget (indicate the name and number of the project or program) 2. The dissertation was completed within the other state program (indicate the name of the program) 3. <u>The thesis corresponds to the priority direction of scientific development, approved by the Higher Scientific and Technical Commission under the</u>	The dissertation topic <u>corresponds to the priority area</u> of scientific development “Information and Communication Technologies” approved by the Higher Scientific and Technical Commission under the Government of the Republic of Kazakhstan.

		<u>Government of the Republic of Kazakhstan (indicate direction)</u>	
2	Importance for science	The work <u>makes/does not make</u> a significant contribution to science and its importance is <u>well-disclosed</u> / not disclosed	The contribution of this work to science is significant, and its importance is <u>well-disclosed</u>. The dissertation makes a significant contribution to the development of domestic and international science in the field of natural language processing, especially in the context of agglutinative and underrepresented languages, such as Kazakh. For the first time, a complete QNLP library has been implemented, covering all the key modules of automatic processing of the Kazakh language - morphological analysis, lemmatization, POS tagging, NER, syntactic parsing, etc. Adapted architectures of modern neural network models (KazBERT + BiLSTM + CRF) are proposed, taking into account the specifics of the Kazakh language.
3	The principle of self-reliance	Self-reliance level: 1) <u>high</u>; 2) average; 3) low; 4) no independence	Level of independence – <u>high</u>. The dissertation work was completed by the author independently at all stages of scientific research.
4	The principle of internal unity	4.1 Justification of the relevance of the thesis: 1) <u>justified</u>; 2) partially justified; 3) not justified	The relevance of the dissertation is <u>justified</u>. In the context of the implementation of the national programs of the Republic of Kazakhstan - "Digital Kazakhstan", "Language Policy Concept for 2023-2029" and "Science Development Concept for 2022-2026" - the need to develop our own intelligent technologies for the Kazakh language is becoming a strategically important task. Without high-quality models of morphological analysis, lemmatization, syntactic and semantic analysis, full-fledged digitalization of document flow, electronic government services, educational and media platforms in the Kazakh language is impossible. Additional relevance is given by the linguistic specificity of the Kazakh language: agglutinativity, vowel harmony, free word order - all this makes it impossible to simply adapt existing Western solutions without considering the national language structure. Consequently, it is necessary to create specialized models, architectures and algorithms that consider the morphological and syntactic features of the Kazakh language. Thus, the work meets the current challenges of the time and contributes to the development of national digital infrastructure, scientific linguistics, and artificial intelligence in the Republic of Kazakhstan.

		4.2 The content of the thesis reflects its topic: 1) <u>reflects</u>; 2) partially reflects; 3) does not reflect	The content of the dissertation <u>reflects</u> the topic of the dissertation on the development of models and methods for automatic processing of unstructured text information in the Kazakh language. All chapters are logically structured in accordance with the goal, objectives and main areas of the study, ensuring theoretical validity and applied significance of the results obtained.
		4.3 The purpose and objectives correspond to the topic of the thesis: 1) <u>correspond</u>; 2) partially correspond; 3) do not correspond	The goal and objectives of the study <u>correspond</u> to the topic of the dissertation, as they are aimed at developing effective models and methods for automatic processing of unstructured information in the Kazakh language. The tasks defined in the work logically follow from the formulated goal and cover all key aspects of the topic - from theoretical analysis of linguistic features to the implementation of applied software solutions, which ensures the scientific integrity and completeness of the study.
		4.4 All sections and provisions of the dissertation are logically interconnected: 1) <u>completely interconnected</u>; 2) the relationship is partial; 3) there is no relationship	All sections and provisions of the dissertation are logically <u>completely interconnected</u>. All sections, chapters and key provisions of the dissertation research are structured in accordance with the internal logic of scientific analysis and are aimed at consistently achieving the stated goal. The study is based on the principle "from problem analysis to solution construction and evaluation", ensuring the integrity of the scientific presentation.
		4.5 The new solutions (principles, methods) proposed by the author are argued and evaluated in comparison with known solutions: 1) <u>there is a critical analysis</u>; 2) partial analysis; 3) the analysis does not represent one's own opinions, but quotes from other authors	The dissertation proposes new methods and architectures aimed at automatic processing of unstructured information in the Kazakh language, and their validity is confirmed theoretically and experimentally, <u>there is a critical analysis</u>. The developed solutions were systematically compared with existing approaches in the field of processing agglutinative languages. The innovations proposed by the author include the combined architecture KazBERT + BiLSTM + CRF, specially adapted to the features of the Kazakh language: agglutination, vowel harmony, free word order and morphological homonymy. This allows to significantly increase the accuracy of morphological and syntactic markup compared to basic BERT or LSTM models. Development of a comprehensive QNLP framework integrating morphoanalysis, POS, NER, syntax in the Kazakh language into a single system. The proposed solutions were experimentally tested on our own corpus and compared with widely used NLP tools, including Stanza, UDPipe, trankit, and spaCy. The results demonstrated: - 5–12% increase in POS tagging accuracy compared to baseline models; - Significant improvement in NER quality for Kazakh named entities; - More accurate morphemic division compared to existing morphoanalyzers. Thus, the author's solutions are not only new and original, but also have practical applicability, which is confirmed by quantitative metrics (Accuracy, F1) and analytical comparison with existing approaches.

5	The principle of scientific novelty	5.1 Are the scientific results and provisions new? 1) <u>completely new</u>; 2) partially new (25–75% are new); 3) not new (less than 25% are new)	All the main provisions and scientific results presented in the dissertation are <u>completely new</u> and original, reflect the author's contribution to the development of natural language processing methods for the Kazakh language and correspond to the current level of research in the field of artificial intelligence and computational linguistics. The new scientific results obtained by the author include: The original architecture KazBERT + BiLSTM + CRF for the tasks of morphological analysis, POS tagging and named entity recognition (NER) was developed and implemented, taking into account the specifics of the Kazakh language. The QNLP framework was developed - the first comprehensive open-source system combining modules of morphoanalysis, lemmatization, POS tagging, NER, syntactic analysis for the Kazakh language. A comparative analysis of modern NLP systems (Stanza, UDPipe, trankit, spaCy) was carried out in relation to the Kazakh language, with experimental validation and justification of the advantages of the author's solutions. The obtained scientific results have been tested in the form of 13 publications, conference reports, 2 implementation certificates, 2 author's certificates, which confirms their novelty, scientific validity and practical value.
		5.2 Are the conclusions of the dissertation new? 1) <u>completely new</u>; 2) partially new (25–75% are new); 3) not new (less than 25% are new)	The conclusions of the dissertation are <u>completely new</u> and logically substantiated. The findings obtained during the study reflect the results of the development of original models and methods for automatic processing of Kazakh-language texts, previously not presented in the scientific literature. The novelty lies in the integration of the adapted architecture of KazBERT + BiLSTM + CRF, the formalization of word formation rules, as well as the creation and annotation of our own text corpus. These findings have both theoretical and practical significance and contribute to the development of natural language processing for agglutinative languages, which is confirmed by publications and testing on international and national platforms.
		5.3 Technical, technological, economic, or managerial decisions are 1) <u>new and justified</u>; 2) completely new; 3) partially new (25–75% are new); 4) not new (less than 25% are new)	The technical and technological solutions presented in the dissertation are <u>new and justified</u>, which is confirmed by the development and implementation of a comprehensive system for automatic processing of unstructured information in the Kazakh language. As part of the study, an original QNLP library was created, including modules for morphological analysis, part-of-speech and named markup, syntactic and semantic analysis, specially adapted to the features of the agglutinative nature of the Kazakh language. The developed ontology of Kazakh morphology and lexicographic base that supports automatic generation of word forms and lemmatization, as well as the formation of our own marked corpus of Kazakh texts confirm the high level of scientific and applied novelty of the work performed, which has no direct analogues in existing domestic and foreign studies.

6	Validity of the main conclusions	All key findings <u>are based</u> / are not based on scientifically sound evidence or are reasonably well funded (for qualitative research and areas of training in the arts and humanities)	The conclusions formulated in the dissertation are <u>scientifically substantiated</u> , since they are obtained based on the author's research, developed models, algorithms and empirical data collected and processed within the framework of this work. They represent the result of a comprehensive analysis of linguistic and computational aspects of automatic processing of the Kazakh language and do not reproduce known solutions. Thus, each of the conclusions in the dissertation is based on its own scientific developments, is an original contribution to science and opens new directions for subsequent research and practical application in the field of digitalization of the Kazakh language.
7	Main provisions submitted for defense	The following questions must be answered for each provision separately: 7.1 Has the position been proven? 1) <u>proven</u>; 2) rather proven; 3) rather not proven; 4) not proven 7.2 Is it trivial? 1) yes; 2) <u>no</u> 7.3 Is it new? 1) <u>yes</u>; 2) no 7.4 Application level: 1) narrow; 2) medium; 3) <u>wide</u> 7.5 Is it proven in the article? 1) <u>yes</u>; 2) no	7.1 The main provisions submitted for defense are <u>fully proven</u>. According to the first provision, the work describes in detail the process of corpus construction: web scraping of Kazakh-language news sites, filtering by language, cleaning HTML noise, normalization, tokenization and saving in JSONL format. Quantitative characteristics are given. POS tagging, morphological markup and NER models were trained and tested on the corpus, which confirms its applicability and reproducibility. According to the second provision, a full-fledged QNLP library in Python with morphological analysis, POS tagging, syntactic analysis and NER modules was implemented. The architecture, testing results on Kazakh texts and examples of practical application are presented. According to the third provision, the architecture of the KazBERT + BiLSTM + CRF model is proposed. KazBERT extracts contextual embeddings, BiLSTM processes bidirectional context, CRF provides global prediction. Preprocessing and filtering improve accuracy. Experiments with the corpus demonstrate high F1 scores for POS and NER. 7.2 The main provisions submitted for defense contain new theoretical knowledge, solutions and implementation methods and are <u>not trivial</u>. According to the first provision, creating a corpus taking into account agglutination, vowel harmony and other features of the Kazakh language required specific solutions - language filtering, noise removal, correct encoding and thematic balance. According to the second provision, developing a library for a low-resource agglutinative language required deep elaboration: creating a lexicon, morpheme analysis, our own tokenizer and visualizer, integrating ML models and language resources. According to the third provision, the emphasis is on deep language adaptation, which goes beyond typical solutions for NLP. 7.3 The provisions and scientific achievements are <u>new</u>, which is confirmed by published articles and experiments.

			The created corpus is the first open, representative and machine-readable resource for the Kazakh language, covering a wide range of sources. QNLP is the first Python library with a full stack of NLP tools for the Kazakh language, including morphology, syntax, NER, visualization and a user-friendly interface. The KazBERT + BiLSTM + CRF model is adapted for the first time for the Kazakh language, taking into account its morphology and filtering in preprocessing, which has not been demonstrated before. 7.4 The models and software solutions developed in the dissertation have a <u>wide</u> level of applicability, covering both scientific and applied areas. In the scientific field, the results are useful for further research on agglutinative languages, formalization of morphology and construction of ontologies. In applied - algorithms and the QNLP library are applicable in intelligent text analysis systems, educational platforms, search engines, voice interfaces and digital assistants in the Kazakh language. Due to adaptation to the grammar of the Kazakh language and the possibility of scaling, the solutions can be used for other Turkic languages. 7.5 <u>Proved</u> in the article The provisions of the dissertation research are proven by the conducted research. The main provisions submitted for defense are presented in 13 publications of the applicant: The main results obtained in the course of the dissertation work were published in thirteen printed works, including 4 articles in publications recommended by the Committee for Control in the Sphere of Education and Science of the Ministry of Education and Science of the Republic of Kazakhstan, as well as 2 articles in the journal Eastern-European Journal of Enterprise Technologies (Q3), indexed in the Scopus database, with a cite score of 2.0 and a percentile of 46. In addition, 4 articles were published in the materials of international conferences, including one scientific article with a cite score of 4.5 and a percentile of 69 (Q2) and 3 articles in other republican publications.
8	The principle of reliability. Reliability of sources and information provided	8.1 The choice of methodology – is the methodology justified or is it described in sufficient detail? 1) <u>yes</u>; 2) no	The research methodology is <u>justified and described in detail</u>: the choice of architectures, preprocessing methods, annotation and model evaluation is determined by the linguistic features of the Kazakh language and the objectives of the dissertation. All stages - from corpus construction to experimental validation - are presented consistently and are accompanied by arguments in favor of the approaches used, which ensures the reproducibility and scientific validity of the results.

		8.2 The results of the thesis were obtained using modern methods of scientific research and techniques for processing and interpreting data using computer technologies: 1) <u>yes</u>; 2) no	The results of the dissertation work were <u>obtained</u> using modern methods of scientific research, including methods of machine learning, neural network architectures and natural language processing. In developing models and analyzing data, reliable statistical methods for quality assessment were used, as well as tools for processing, annotating and visualizing large volumes of texts in the Kazakh language. All stages of the study were implemented using modern computer technologies, which ensured high accuracy, reproducibility and applied value of the results obtained.
		8.3 Theoretical conclusions, models, identified relationships and patterns are proven and confirmed by experimental research (for areas of training in pedagogical sciences, the results are proven based on a pedagogical experiment): 1) <u>yes</u>; 2) no	The theoretical conclusions, developed models, as well as the identified relationships and patterns within the dissertation research have been <u>proven and confirmed</u> by the results of experimental studies. All proposed solutions - from morphological analysis to a comprehensive QNLP framework - have been tested in practice on a representative corpus of Kazakh text. The results have been validated using generally accepted quality metrics and compared with existing solutions, which confirms the reliability and scientific validity of the data obtained.
		8.4 Important statements are <u>supported</u> / partially supported / not supported by references to current and reliable scientific literature	Important claims <u>are supported</u> by references to relevant and reliable scientific literature.
		8.5 The literature sources used are <u>sufficient</u> / are not sufficient for the literature review	The literature sources used <u>are sufficient</u> and cover a wide range of modern scientific works in the field of natural language processing, machine learning, agglutinative linguistics and digital processing of the Kazakh language. The literature review presents both classical approaches and the latest research, including works published in high-ranking journals and at leading scientific conferences, which ensures the theoretical validity and relevance of the dissertation research.
9	Principle of practical value	9.1 The dissertation has theoretical significance: 1) <u>yes</u>; 2) no	The dissertation <u>has theoretical significance</u>, as it develops and substantiates new approaches to processing agglutinative languages using the Kazakh language as an example, including original models of morphological analysis, POS tagging and semantic markup. The proposed theoretical provisions expand modern ideas about the structure and formalization of the Kazakh language in computational linguistics, and also lay the foundation for further research in the field of automatic text processing and the development of language technologies for low-resource languages.

		9.2 The thesis has practical significance and there is a high probability of applying the obtained results in practice: 1) <u>yes</u>; 2) no	The dissertation <u>has practical significance</u>, since within its framework a full-fledged QNLP software framework for automatic processing of texts in the Kazakh language was developed, including modules of morphoanalysis, POS tagging, NER and syntactic analysis. The obtained results can be directly applied in electronic document management systems, digital educational platforms, voice assistants and other IT solutions focused on the Kazakh language. The high degree of adaptability of the models to the linguistic features of the language, open architecture and the possibility of scaling confirm the real applicability of developments in the framework of state and commercial projects.
		9.3 Are the suggestions for practice new? 1) <u>completely new</u>; 2) partially new (25–75% are new); 3) not new	The proposals formulated in the dissertation for practical application are <u>completely new</u>, since they are based on the author's developments and adapted to the specifics of the Kazakh language. A comprehensive approach to building a QNLP language framework is proposed, allowing integrating morphoanalysis, syntactic and semantic processing into a single system.
10.	Quality of writing and design	Quality of academic writing: 1) <u>high</u>; 2) average; 3) below average; 4) low.	The quality of academic writing in the dissertation is <u>high</u> and meets the requirements for scientific research: the presentation of the material is logical, consistent and precise in its formulations. A scientific style of presentation is used with the correct use of terminology in the field of linguistics, computer science and artificial intelligence. All provisions are substantiated, conclusions are justified, and the structure of the text contributes to a clear disclosure of the research topic. Thus, the academic quality of the text ensures a high degree of scientific persuasiveness and perceivability of the dissertation material.
11.	Comments and suggestions on the dissertation		<i>There are duplicate works in the list of references.</i>
12.	The scientific level of the doctoral candidate's articles on the research topic (in the case of defending the dissertation in the form of a series of articles, the official reviewers will comment on the scientific level of each article by the doctoral		The scientific level of the applicant's publications fully complies with the requirements for PhD dissertation.

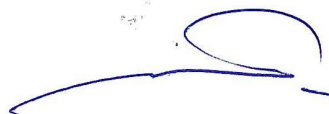
	candidate on the research topic).		
13.	The decision of the official reviewer (according to paragraph 28 of the current Model Regulation)		To petition the Committee for the award of the degree of Doctor of Philosophy (PhD)

Conclusion

The presented dissertation for the degree of Doctor of Philosophy (PhD) by **Aitim Aigerim Kairatkyzy** on the topic «**Models and methods for the automatic processing of unstructured information**», meets the requirements of the Rules, the author deserves to petition the Committee for the award of the degree of Doctor of Philosophy (PhD).

Official reviewer:

doct. of tech. sc., professor of the
Department of Information Systems,
Al Farabi Kazakh National University



Tukeyev Ualsher Anuarbekovich



 ел-Фараби атындағы ҚазҰУ Ғылыми кадрларды даярлау және аттестаттау басқармасының басшысы
 ЗАБЕРЯЮ
 Начальник управления подготовки и аттестации научных кадров КазНУ им. аль-Фараби
Сергей Сергеевич
 «29» 20 21 ж.г.

Copies of reviews from official reviewers are given to the doctoral student no later than 5 (five) working days before the defense of the dissertation.

ОТЗЫВ ОФИЦИАЛЬНОГО РЕЦЕНЗЕНТА

на диссертационную работу Әйтiм Әйгерiм Қайратқызы на тему «Models and methods for the automatic processing of unstructured information» («Модели и методы автоматической обработки неструктурированной информации»), предоставленную на соискание степени доктора философии (PhD) по образовательной программе 8D06101 – «Интеллектуальные системы»

п/п	Критерии	Соответствие критериям (подчеркнуть один из вариантов ответа)	Обоснование позиции официального рецензента (замечания выделить курсивом)
1	Тема диссертации (на дату ее утверждения) соответствует направлениям развития науки и/или государственным программам	1.2 Соответствие приоритетным направлениям развития науки или государственным программам:	Диссертационная работа <u>соответствует</u> задачам государственных программ: 1. Национальный проект «Цифровой Казахстан», утвержден постановлением Правительства Республики Казахстан от 12 декабря 2017 года № 827 2. Концепция развития языковой политики Республики Казахстан на 2023–2029 годы, утверждена Указом Президента Республики Казахстан от 5 мая 2023 года № 1 3. Стратегический план развития Республики Казахстан до 2025 года, утвержден Указом Президента Республики Казахстан от 15 февраля 2018 года № 636 4. Концепция развития науки Республики Казахстан на 2022–2026 годы утверждена постановлением Правительства РК от 19 октября 2021 года № 736 Программа цифровизации государственного управления «Цифровое правительство» в рамках «Цифрового Казахстана», конкретизирует задачи внедрения ИИ и NLP в сфере госуслуг
		1) Диссертация выполнена в рамках проекта или целевой программы, финансируемого(ой) из Государственного бюджета (указать название и номер проекта или программы) 2) Диссертация выполнена в рамках другой государственной программы (указать название программы)	Тема диссертации <u>соответствует</u> приоритетному направлению развития науки «Информационные, коммуникационные технологии», утвержденному высшей научно-технической комиссией при Правительстве РК.

		3) <u>Диссертация соответствует приоритетному направлению развития науки, утвержденному Высшей научно-технической комиссией при Правительстве Республики Казахстан (указать направление)</u>	
2	Важность для науки	Работа <u>вносит/не вносит</u> существенный вклад в науку, а ее важность <u>хорошо раскрыта/не раскрыта</u>	Вклад работы в науку является существенным, а важность работы <u>раскрыта хорошо</u> . Диссертационная работа вносит существенный вклад в развитие отечественной и международной науки в области обработки естественного языка, особенно в контексте агглютинативных и недостаточно представленных языков, таких как казахский. Впервые реализован целостная библиотека QNLP, охватывающий все ключевые модули автоматической обработки казахского языка - морфологический анализ, лемматизацию, POS-теггинг, NER, синтаксический разбор итп. Предложены адаптированные архитектуры современных нейросетевых моделей (KazBERT + BiLSTM + CRF) с учетом специфики казахского языка.
3	Принцип самостоятельности	Уровень самостоятельности: 1) <u>Высокий</u> ; 2) Средний; 3) Низкий; Самостоятельности нет	Уровень самостоятельности – <u>высокий</u> . Диссертационная работа выполнена автором самостоятельно на всех этапах научного исследования.
4	Принцип внутреннего единства	4.1 Обоснование актуальности диссертации: 1) <u>Обоснована</u> ; 2) Частично обоснована; 3) Не обоснована.	<u>Обоснована</u> актуальность диссертации. В условиях реализации национальных программ Республики Казахстан - «Цифровой Казахстан», «Концепция языковой политики на 2023–2029 годы» и «Концепция развития науки на 2022–2026 годы» - необходимость разработки собственных интеллектуальных технологий для казахского языка становится стратегически важной задачей. Без качественных моделей морфологического анализа, лемматизации, синтаксического и семантического анализа, невозможна полноценная цифровизация документооборота, электронных госуслуг, образовательных и медиаплатформ на казахском языке. Дополнительную актуальность придаёт лингвистическая специфика казахского языка: агглютинативность, сингармонизм, свободный порядок слов - всё это делает невозможным простую адаптацию существующих западных решений без учёта национальной языковой структуры. Следовательно, требуется создание специализированных моделей,

		архитектур и алгоритмов, учитывающих морфологические и синтаксические особенности казахского языка. Таким образом, работа отвечает актуальным вызовам времени и вносит вклад в развитие национальной цифровой инфраструктуры, научной лингвистики и искусственного интеллекта в Республике Казахстан.
	4.2 Содержание диссертации отражает тему диссертации: 1) <u>Отражает</u> ; 2) Частично отражает; 3) Не отражает	Содержание диссертации <u>отражает</u> тему диссертации по разработке моделей и методов автоматической обработки неструктурированной текстовой информации на казахском языке. Все главы логично выстроены в соответствии с целью, задачами и основными направлениями исследования, обеспечивая теоретическую обоснованность и прикладную значимость полученных результатов.
	4.3. Цель и задачи соответствуют теме диссертации: 1) <u>соответствуют</u> ; 2) частично соответствуют; 3) не соответствуют	Цель и задачи исследования <u>соответствуют</u> теме диссертации, так как направлены на разработку эффективных моделей и методов автоматической обработки неструктурированной информации на казахском языке. Определённые в работе задачи логично вытекают из сформулированной цели и охватывают все ключевые аспекты темы - от теоретического анализа лингвистических особенностей до реализации прикладных программных решений, что обеспечивает научную целостность и полноту исследования.
	4.4. Все разделы и положения диссертации логически взаимосвязаны: 1) <u>полностью взаимосвязаны</u> ; 2) взаимосвязь частичная; взаимосвязь отсутствует	Все разделы и положения диссертации <u>логически полностью взаимосвязаны</u> . Все разделы, главы и ключевые положения диссертационного исследования структурированы в соответствии с внутренней логикой научного анализа и направлены на последовательное достижение поставленной цели. Исследование строится по принципу «от анализа проблемы - к построению и оценке решения», обеспечивая целостность научного изложения.
	4.5 Предложенные автором новые решения (принципы, методы) аргументированы и оценены по сравнению с известными решениями: 1) <u>критический анализ есть</u> ; 2) анализ частичный; анализ представляет собой не собственные мнения, а цитаты других авторов	В диссертации предложены новые методы и архитектуры, направленные на автоматическую обработку неструктурированной информации на казахском языке, и их аргументированность подтверждена теоретически и экспериментально, <u>есть критический анализ</u> . Разработанные решения были систематически сопоставлены с существующими подходами в области обработки агглютинативных языков. К числу новшеств, предложенных автором, относятся комбинированная архитектура KazBERT + BiLSTM + CRF, специально адаптированная под особенности казахского языка: агглютинацию, сингармонизм, свободный порядок слов и морфологическую омонимию. Это позволяет значительно повысить точность морфологической и синтаксической разметки по сравнению с базовыми BERT или LSTM-моделями. Разработка комплексного фреймворка QNLP, интегрирующего в единую систему морфоанализ, POS, NER, синтаксис на казахском языке.

			Предложенные решения были экспериментально протестированы на собственном корпусе и сравнивались с широко используемыми NLP-инструментами, включая Stanza, UDPipe, trankit и spaCy. Полученные результаты продемонстрировали: - Повышение точности POS-теггинга на 5–12% в сравнении с baseline-моделями; - Существенное улучшение качества NER для казахских именованных сущностей; - Более точное морфемное членение по сравнению с существующими морфоанализаторами. Таким образом, авторские решения не только новы и оригинальны, но и обладают практической применимостью, что подтверждено количественными метриками (Assurasy, F1) и аналитическим сравнением с существующими подходами.
5	Принцип научной новизны	5.1 Научные результаты и положения являются новыми? 1) <u>полностью новые</u>; 2) частично новые (новыми являются 25-75%); 3) не новые (новыми являются менее 25%)	Все основные положения и научные результаты, представленные в диссертационной работе, являются <u>полностью новыми</u> и оригинальными, отражают авторский вклад в развитие методов обработки естественного языка для казахского языка и соответствуют современному уровню исследований в области искусственного интеллекта и вычислительной лингвистики. К числу новых научных результатов, полученных автором, относятся: Разработана и реализована оригинальная архитектура KazBERT + BiLSTM + CRF для задач морфологического анализа, POS-теггинга и распознавания именованных сущностей (NER) с учётом специфики казахского языка. Разработан фреймворк QNLP - первая комплексная open-source система, объединяющая модули морфоанализа, лемматизации, POS-теггинга, NER, синтаксического анализа для казахского языка. Проведён сравнительный анализ современных NLP-систем (Stanza, UDPipe, trankit, spaCy) применительно к казахскому языку, с экспериментальной валидацией и обоснованием преимуществ авторских решений. Полученные научные результаты прошли апробацию в виде 13 публикаций, докладов на конференциях, 2 актов внедрения, 2 авторских свидетельств, что подтверждает их новизну, научную состоятельность и практическую ценность.
		5.2 Выводы диссертации являются новыми? 1) <u>полностью новые</u>; 2) частично новые (новыми	Выводы диссертации являются <u>полностью новыми</u> и логически обоснованными. Полученные в ходе исследования выводы отражают результаты разработки оригинальных моделей и методов автоматической обработки казахскоязычных текстов, ранее не представленных в научной

		являются 25-75%); 3) не новые (новыми являются менее 25%)	литературе. Новизна заключается в интеграции адаптированной архитектуры KazBERT + BiLSTM + CRF, формализации правил словообразования, а также создании и аннотировании собственного корпуса текстов. Эти выводы обладают как теоретической, так и практической значимостью и вносят вклад в развитие обработки естественного языка для агглютинативных языков, что подтверждается публикациями и апробацией на международных и национальных платформах.
		5.3 Технические, технологические, экономические или управленческие решения являются новыми и обоснованными: 1) <u>полностью новые</u> ; 2) частично новые (новыми являются 25-75%); 3) не новые (новыми являются менее 25%)	Представленные в диссертационной работе технические и технологические решения являются <u>полностью новыми</u> , что подтверждается разработкой и реализацией комплексной системы автоматической обработки неструктурированной информации на казахском языке. В рамках исследования создана оригинальная библиотека QNLP, включающая модули морфологического анализа, частеречной и именованной разметки, синтаксического и семантического анализа, специально адаптированные под особенности агглютинативной природы казахского языка. Разработанная онтология казахской морфологии и лексикографическая база, поддерживающая автоматическую генерацию словоформ и лемматизацию, а также формирование собственного размеченного корпуса казахских текстов подтверждают высокий уровень научной и прикладной новизны выполненной работы, не имеющей прямых аналогов в существующих отечественных и зарубежных исследованиях.
6	Обоснованность основных выводов	Все основные <u>выводы</u> основаны/не основаны на весомых с научной точки зрения доказательствах либо достаточно хорошо обоснованы	Выводы, сформулированные в диссертационной работе, являются <u>научно обоснованными</u> , поскольку получены на основе авторских исследований, разработанных моделей, алгоритмов и эмпирических данных, собранных и обработанных в рамках данной работы. Они представляют собой результат комплексного анализа лингвистических и вычислительных аспектов автоматической обработки казахского языка и не воспроизводят известных решений. Таким образом, каждый из выводов в диссертации опирается на собственные научные разработки, является оригинальным вкладом в науку и открывает новые направления для последующих исследований и практического применения в сфере цифровизации казахского языка.
7	Основные положения, выносимые на защиту	Необходимо ответить на следующие вопросы по каждому положению в отдельности: 7.1 Доказано ли положение? 1) <u>доказано</u> ;	7.1 Основные положения выносимые на защиту <u>полностью доказаны</u> . По-первому положению в работе подробно описан процесс построения корпуса: веб-скрапинг казахскоязычных новостных сайтов, фильтрация по языку, очистка HTML-шумов, нормализация, токенизация и сохранение в формате JSONL. Приведены количественные характеристики. На корпусе обучены и протестированы модели POS-теггирования, морфологической

	2) скорее доказано; 3) скорее не доказано; не доказано 7.2 Является ли тривиальным? 1) <u>да</u>; 2) <u>нет</u> 7.3 Является ли новым? 1) <u>да</u>; 2) <u>нет</u> 7.4 Уровень для применения: 1) узкий; 2) средний; 3) <u>широкий</u> 7.5 Доказано ли в статье? 1) <u>да</u>; 2) <u>нет</u>	разметки и NER, что подтверждает его применимость и воспроизводимость. По-второму положению реализована полноценная библиотека QNLP на Python с модулями морфологического анализа, POS-теггирования, синтаксического анализа и NER. Представлены архитектура, результаты тестирования на казахских текстах и примеры практического применения. По-третьему положению предложена архитектура модели KazBERT + BiLSTM + CRF. KazBERT извлекает контекстуальные эмбединги, BiLSTM обрабатывает двунаправленный контекст, CRF обеспечивает глобальное предсказание. Предобработка и фильтрация повышают точность. Эксперименты с корпусом демонстрируют высокие F1-оценки по POS и NER. 7.2 Основные положения, выносимые на защиту, содержат новые теоретические знания, способы решения и реализации и <u>не являются тривиальными</u>. По-первому положению, создание корпуса с учётом агглютинации, сингармонизма и других особенностей казахского языка потребовало специфических решений - языковая фильтрация, устранение шумов, корректная кодировка и тематический баланс. По-второму положению, разработка библиотеки для низкоресурсного агглютинативного языка потребовала глубокой проработки: создание лексикона, морфемный анализ, собственный токенизатор и визуализатор, интеграция ML-моделей и языковых ресурсов. По-третьему положению, акцент на глубокой языковой адаптации, которая выходит за рамки типовых решений для NLP. 7.3 Положения и научные достижения являются <u>новыми</u>, что подтверждается опубликованными статьями и экспериментами. Созданный корпус является первым открытым, репрезентативным и машиночитаемым ресурсом для казахского языка, охватывающим широкий спектр источников. QNLP - первая Python-библиотека с полным стеком NLP-инструментов для казахского языка, включая морфологию, синтаксис, NER, визуализацию и удобный интерфейс. Модель KazBERT + BiLSTM + CRF впервые адаптирована для казахского языка с учётом его морфологии и фильтрации в предобработке, чего ранее не демонстрировалось. 7.4 Разработанные в диссертационной работе модели и программные решения обладают <u>широким</u> уровнем применимости, охватывая как научные, так и прикладные области. В научной области результаты полезны для дальнейших исследований по агглютинативным языкам, формализации
--	---	---

			морфологии и построению онтологий. В прикладной - алгоритмы и библиотека QNLP применимы в интеллектуальных системах анализа текста, образовательных платформах, поисковых системах, голосовых интерфейсах и цифровых ассистентах на казахском языке. Благодаря адаптации к грамматике казахского языка и возможности масштабирования, решения могут быть использованы и для других тюркских языков. 7.5 Доказаны в <u>статье</u> Положения диссертационного исследования доказаны проведенными исследованиями. Основные положение, выносимые на защиту, представлены в 13 публикациях соискателя: Основные результаты, полученные в ходе выполнения диссертационной работы, были опубликованы в тринадцати печатных работах, включая 4 статьи в изданиях, рекомендованных Комитетом по контролю в сфере образования и науки Министерства образования и науки Республики Казахстан, а также 2 статьи в журнале Eastern-European Journal of Enterprise Technologies (Q3), индексируемом в базе Scopus, с cite score 2.0 и перцентилем 46. Кроме того, опубликовано 4 статей в материалах международных конференций, среди которых одна научная статья с cite score 4.5 и перцентилем 69 (Q2) и 3 статьи в других республиканских изданиях.
8	Принцип достоверности Достоверность источников и предоставляемой информации	8.1 Выбор методологии - обоснован или методология достаточно подробно описана 1) <u>да</u> ; 2) нет	Методология исследования <u>обоснована и подробно описана</u>: выбор архитектур, методов предобработки, аннотации и оценки моделей обусловлен лингвистическими особенностями казахского языка и задачами диссертации. Все этапы - от построения корпуса до экспериментальной валидации - изложены последовательно и сопровождаются аргументацией в пользу применённых подходов, что обеспечивает воспроизводимость и научную состоятельность результатов.
		8.2 Результаты диссертационной работы получены с использованием современных методов научных исследований и методик обработки и интерпретации данных с применением компьютерных технологий: 1) <u>да</u> ; 2) нет	Результаты диссертационной работы <u>получены</u> с использованием современных методов научных исследований, включая методы машинного обучения, нейросетевых архитектур и обработки естественного языка. При разработке моделей и анализе данных применялись надёжные статистические методы оценки качества, а также инструментальные средства для обработки, аннотации и визуализации больших объёмов текстов на казахском языке. Все этапы исследования реализованы с использованием современных компьютерных технологий, что обеспечило высокую точность, воспроизводимость и прикладную ценность полученных результатов.

		8.3 Теоретические выводы, модели, выявленные взаимосвязи и закономерности доказаны и подтверждены экспериментальным исследованием (для направлений подготовки по педагогическим наукам результаты доказаны на основе педагогического эксперимента): 1) <u>да</u> ; 2) нет	Теоретические выводы, разработанные модели, а также выявленные взаимосвязи и закономерности в рамках диссертационного исследования <u>доказаны и подтверждены</u> результатами экспериментальных исследований. Все предложенные решения - от морфологического анализа до комплексного фреймворка QNLP - прошли практическую проверку на репрезентативном корпусе казахского текста. Результаты валидированы с применением общепринятых метрик качества и сравнены с существующими решениями, что подтверждает достоверность и научную обоснованность полученных данных.
		8.4 <u>Важные утверждения подтверждены/частично подтверждены/не подтверждены</u> ссылками на актуальную и достоверную научную литературу	Важные утверждения <u>подтверждены</u> ссылками на актуальную и достоверную научную литературу
		8.5 Используемые источники литературы <u>достаточны/не достаточны</u> для литературного обзора	Используемые источники литературы являются <u>достаточными</u> и охватывают широкий спектр современных научных трудов в области обработки естественного языка, машинного обучения, агглютинативной лингвистики и цифровой обработки казахского языка. В литературном обзоре представлены как классические подходы, так и новейшие исследования, включая работы, опубликованные в высокорейтинговых журналах и на ведущих научных конференциях, что обеспечивает теоретическую обоснованность и актуальность диссертационного исследования.
9	Принцип практической ценности	9.1 Диссертация имеет теоретическое значение: 1) <u>да</u> ; 2) нет	Диссертация имеет <u>теоретическое значение</u> , так как в ней разработаны и обоснованы новые подходы к обработке агглютинативных языков на примере казахского, включая оригинальные модели морфологического анализа, POS-теггинга и семантической разметки. Предложенные теоретические положения расширяют современные представления о структуре и формализации казахского языка в вычислительной лингвистике, а также закладывают основу для дальнейших исследований в области автоматической обработки текстов и разработки языковых технологий для малоресурсных языков.

		9.2 Диссертация имеет практическое значение и существует высокая вероятность применения полученных результатов на практике: 1) <u>да</u>; 2) нет	Диссертация имеет <u>практическое значение</u>, поскольку в её рамках разработан полноценный программный фреймворк QNLP для автоматической обработки текстов на казахском языке, включающий модули морфоанализа, POS-теггинга, NER и синтаксического анализа. Полученные результаты могут быть непосредственно применены в системах электронного документооборота, цифровых образовательных платформах, голосовых помощниках и других ИТ-решениях, ориентированных на казахский язык. Высокая степень адаптированности моделей к лингвистическим особенностям языка, открытая архитектура и возможность масштабирования подтверждают реальную применимость разработок в рамках государственных и коммерческих проектов.
		9.3 Предложения для практики являются новыми? 1) <u>полностью новые</u>; 2) частично новые (новыми являются 25-75%); 3) не новые (новыми являются менее 25%)	Предложения, сформулированные в диссертационной работе для практического применения, являются <u>полностью новыми</u>, поскольку основаны на авторских разработках и адаптированы под специфику казахского языка. Предложен комплексный подход к построению языкового фреймворка QNLP, позволяющего интегрировать морфоанализ, синтаксическую и семантическую обработку в единую систему.
10.	Качество написания и оформления	Качество академического письма: 1) <u>высокое</u>; 2) среднее; 3) ниже среднего; 4) низкое.	Качество академического письма в диссертации является <u>высоким</u>, соответствует требованиям, предъявляемым к научным исследованиям: изложение материала отличается логичностью, последовательностью и точностью формулировок. Используется научный стиль изложения с корректным употреблением терминологии в области лингвистики, компьютерных наук и искусственного интеллекта. Все положения аргументированы, выводы обоснованы, а структура текста способствует ясному раскрытию темы исследования. Таким образом, академическое качество текста обеспечивает высокую степень научной убедительности и воспринимаемости диссертационного материала.
11.	Замечания к диссертации		<i>В списке литературы есть дублирования работ.</i>
12.	Научный уровень статей докторанта по теме исследования (в случае защиты диссертации в форме серии статей официальные рецензенты)		Научный уровень публикаций соискателя полностью соответствует предъявляемым требованиям к диссертационным работам уровня PhD.

	комментируют научный уровень каждой статьи докторанта по теме исследования)		
13.	Решение официального рецензента (согласно пункту 28 настоящего Типового положения)		Ходатайствовать перед Комитетом для присуждения докторанту степени доктора философии (PhD)

Заключение

Представленная диссертация на соискание степени доктора философии (PhD) **Әйтiм Әйгерiм Қайратқызы** на тему **«Models and methods for the automatic processing of unstructured information»** («Модели и методы автоматической обработки неструктурированной информации»), является законченной самостоятельной научно-исследовательской работой, которая отвечает требованиям «Правил присуждения степеней», а её автор заслуживает ходатайствовать перед Комитетом о присуждении степени доктора философии (PhD).

Официальный рецензент:

д.т.н., профессор кафедры «Информационные системы»,
КазНУ им. Аль-Фараби

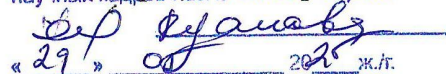


Тукеев Уалшер Ануарбекович

РАСТАЙМЫН
өл-Фараби атындағы ҚазМУ Ғылыми кадрларды
даярлау және аттестаттау басқармасының басшысы
ЗАВЕРЯЮ
Начальник управления подготовки и аттестации
научных кадров КазНУ им. А.Фараби



Копии отзывов официальных рецензентов вручаются докторанту не позднее, чем за 5 (пять) рабочих дней до защиты диссертации.


«29» 2021 ж.г.