

Institute of Nanoscience & Nanotechnology
National Center for Scientific Research "DEMOKRITOS"
Laboratory of Statistical Mechanics and Dynamical Systems

Dr. A. Provata
INN
NCSR "Demokritos"
15310 Athens, Greece

Tel: +30 210 650 3964
E-mail: a.provata@inn.demokritos.gr
Homepage: https://inn.demokritos.gr/person/astero-provata/

Athens, 30/06/2025

FOREIGN SCIENTIFIC SUPERVISOR'S REFERENCE
on the dissertation of PhD candidate Toiganbayeva Nazgul
for the thesis " Development of tools and creation of a database for optical recognition of handwritten text of the Kazakh-Russian language",
presented for the Doctor of Philosophy (PhD) degree in the specialty 6D060200 – Informatics

The dissertation is devoted to a relevant and underexplored topic in Kazakhstani scientific practice – the automatic recognition of handwritten Kazakh text using modern machine learning methods and deep neural networks. The main objective of the research is the creation of the first open dataset of Kazakh handwritten texts – KOHTD (Kazakh Offline Handwritten Text Dataset), as well as the development and comparative analysis of text recognition models based on this dataset.

Within the scope of the study, the author carried out extensive work on the collection, digitization, and annotation of data: the KOHTD dataset includes over 3000 examination sheets, containing more than 140, 335 segmented images and approximately 922,010 characters. A special Telegram bot was developed for annotation, which allowed for the efficient involvement of volunteers and implementation of a crowdsourcing approach.

To segment text at the line and word levels, the dissertation proposes an original method based on a genetic algorithm, which made it possible to achieve high accuracy in separating handwritten text despite its variability.

Following the relevant publication by the candidate, the study includes a large-scale experimental comparison of several state-of-the-art deep learning models, including Bluche, Puigcerver, Flor, and Scheidl. The Flor model demonstrated the best performance in terms of metrics such as CER (Character Error Rate), WER (Word Error Rate), and SER.

The KOHTD dataset and the methods developed in the dissertation and described in their publication have broad applicability in both academic and practical tasks related to natural language processing, analysis of handwritten archives, and the digitization of examination and official documents.

Thus, the dissertation combines scientific novelty, practical relevance, and a high level of technical implementation, making a significant contribution to the development of computer vision, digital humanities, and the Kazakh language in the context of artificial intelligence.

As the foreign scientific supervisor, I am satisfied with the overall outcome of the work. I believe that the dissertation of N. Toiganbayeva on the topic *"Development of tools and creation of a database for optical recognition of handwritten text of the Kazakh-Russian language"*,

Сканировано с CamScanner

submitted for the degree of Doctor of Philosophy (PhD) in the specialty 6D060200 – Informatics, meets the qualification requirements for dissertations in the given specialty.

Overall, the dissertation is a complete research work, performed at a high technical level, and characterized by unity and integrity of the research. The scope of the study and the results obtained fully comply with the criteria established by the Committee for Quality Assurance in Science and Higher Education of the Republic of Kazakhstan for the award of the PhD degree.

The dissertation is recommended for defense.

Foreign Scientific Supervisor
PhD, Institute of Nanoscience and Nanotechnology,
NCSR "Demokritos"



Provata Astero

Сканировано с CamScanner

Наноғылым және нанотехнология институты
«"DEMOKRITOS » ұлттық ғылыми зерттеулер орталығы
Статистикалық механика және динамикалық жүйелер зертханасы

Доктор А. Провата
ННИ
«DEMOKRITOS» ҮҒЗО
15310 Афины, Грешия

Тел: +30 210 650 3964
Электрондық пошта: a.provata@inn.demokritos.gr
Басты бет: https://inn.demokritos.gr/person/astero-provata/

Афины, 30.06.2025

«6D060200 – Информатика» мамандығы бойынша философия докторы (PhD) дәрежесін алу үшін ұсынылған Тойганбаева Назгуль Абеновнаның «Қазақ-орыс тілінің қолжазба мәтінін оптикалық тану үшін құралдар әзірлеу және деректер базасын құру» тақырыбына орындаған диссертациялық жұмысына

ШЕТЕЛДІК ҒЫЛЫМИ КЕҢЕСШІНІҢ ПІКІРІ

Диссертациялық жұмыс заманауи қазақстандық ғылыми кеңістікте өзекті әрі жеткілікті деңгейде зерттелмеген мәселе – заманауи машиналық оқыту әдістері мен терең нейрондық желілер негізінде қазақ қолжазба мәтіндерін автоматты түрде тану проблемасына арналған. Зерттеудің басты мақсаты – қазақ тіліндегі қолжазба мәтіндердің алғашқы ашық деректер қорын (Kazakh Offline Handwritten Text Dataset – KOHTD) құру. сондай-ақ осы деректер қоры негізінде мәтінді тану модельдерін әзірлеу және олардың салыстырмалы талдауын жүргізу.

Зерттеу шеңберінде автор деректерді жинау, цифрландыру және аннотациялау бойынша ауқымды жұмыс атқарды: KOHTD деректер жинағы 3000-нан астам емтихан парағын қамтиды, олардың құрамында 140 335-тен астам сегменттелген кескін және шамамен 922 010 таңба бар. Аннотациялау процесін жүзеге асыру үшін арнайы Telegram-бот әзірленіп, ол еріктілерді тиімді түрде тартуға және краудсорсинг тәсілін іске асыруға мүмкіндік берді.

Мәтінді жол және сөз деңгейінде сегменттеу үшін диссертацияда генетикалық алгоритмге негізделген өзіндік әдіс ұсынылған. Бұл әдіс қолжазба мәтіннің әртүрлілігіне қарамастан, оны жоғары дәлдікпен бөліп алуға мүмкіндік берді.

Үміткердің тақырыпқа қатысты жарияланымына сүйене отырып, зерттеу барысында Bluche, Puigcerver, Flor және Scheidl сияқты заманауи терең оқыту модельдерінің кең ауқымды эксперименттік салыстырмалы талдауы жүргізілді. Нәтижесінде, CER (Character Error Rate), WER (Word Error Rate) және SER көрсеткіштері бойынша ең жоғары тиімділікті Flor моделі көрсеткен.

KOHTD деректер жинағы мен диссертация аясында әзірленген әдістер табиғи тілді өңдеу, қолжазба мұрағаттарды талдау, сондай-ақ емтихан және ресми құжаттарды цифрландыруға қатысты ғылыми-зерттеу және практикалық міндеттерде кеңінен қолданылу әлеуетіне ие.

Осылайша, диссертацияда ғылыми жаңалық, практикалық маңыздылық және техникалық іске асырудың жоғары деңгейі үйлесім тапқан, бұл өз кезегінде компьютерлік көру, цифрлық гуманитаристика және жасанды интеллект контексіндегі қазақ тілінің дамуына елеулі үлес қосады.

Шетелдік ғылыми жетекші ретінде мен атқарылған жұмыстың жалпы нәтижесіне қанағаттанамын. Н. Тойганбаеваның 6D060200 – «Информатика» мамандығы бойынша философия докторы (PhD) дәрежесін алу үшін ұсынылған «Қазақ-орыс тілінің қолжазба мәтінін оптикалық тану үшін құралдар әзірлеу және деректер базасын құру»

Сканировано с CamScanner

тақырыбындағы диссертациялық жұмысы аталған мамандыққа қойылатын біліктілік талаптарына толық сәйкес келеді деп есептеймін.

Жалпы алғанда, диссертациялық жұмыс жоғары техникалық деңгейде орындалған, зерттеудің біртұтастығы мен толықтығын сақтай отырып, толыққанды ғылыми-зерттеу еңбегі болып табылады. Зерттеу көлемі мен алынған нәтижелер Қазақстан Республикасы Ғылым және жоғары білім саласында сапаны қамтамасыз ету комитеті белгілеген PhD дәрежесін тағайындау жөніндегі талаптар мен критерийлерге толық сәйкес келеді.

Диссертацияны қорғауға ұсынуға болады.

Шетелдік ғылыми кеңесші,
PhD, Наноғылым және нанотехнология
Институты «DEMOKRITOS» ҰҒЗО Провато Астеро

/Қолы бар, мөрі бар/

Сканировано с CamScanner