

Rezumatul activității și a rezultatelor obținute în proiect în anul 2025

Cifrul proiectului 25.80012.0807.50SE

Denumirea Proiectului **„Digitizarea și crearea corpusului FolkAI cu texte ale folclorului basarabean din sec. XIX-XX”**

Proiectul a urmărit dezvoltarea unui corpus diacronic-parallel unic dedicat folclorului basarabean din secolele XIX–XX, prin integrarea digitizării avansate, transliterării automate și generării de imagini cu algoritmi AI, contribuind astfel la conservarea și valorificarea patrimoniului cultural. În etapa 2025, activitățile s-au axat pe analiza trăsăturilor distinctive ale textelor folclorice, crearea modelelor OCR pentru grafia chirilică, digitizarea primară a surselor și pregătirea materialelor pentru corpus.

Au fost identificate și selectate resurse folclorice relevante din bibliotecile din Republica Moldova, iar în colaborare cu Institutul de Filologie Română „B.P. Hașdeu” a fost definit setul de caracteristici lingvistice și culturale specifice folclorului basarabean. Au fost scanate integral opt cărți (peste 2.500 de pagini), urmând să fie procesate în continuare pentru atingerea obiectivelor proiectului.

Analiza instrumentelor a confirmat că platforma HeDy este cea mai adecvată pentru digitizarea textelor chirilice istorice. Platforma a fost testată și adaptată pentru corpus, fiind utilizate modele OCR preexistente și dezvoltate noi modele pentru vocabularul specific folcloric. Recunoașterea automată a textelor necesită corectări manuale, însă HeDy oferă suport integrat pentru transliterare și post-procesare.

A fost elaborată structura corpusului diacronic-parallel, bazată pe alinierea textelor chirilice și latine, precum și un model extins de metadata, inspirat de arhive internaționale (AILLA, ELAR, SKS) și adaptat practicilor locale, cuprinzând patru categorii: textuale, legate de performer, informații de colectare și date tehnice. Totodată, s-a avansat în construcția pe formatul **Treebank**, conform Universal Dependencies, realizându-se transliterarea, tokenizarea și primele adnotări sintactice.

Proiectul a generat rezultate științifice semnificative, materializate în două articole, un rezumat și participări la conferințe internaționale (ConsILR 2025 București, WIIS 2025 Chișinău, conferința de la Veliko Tarnovo), precum și la expoziția INFOINVENT. De asemenea, a fost organizat un seminar la Universitatea de Stat din Moldova, unde au fost prezentate studenților rezultatele și direcțiile de dezvoltare ale corpusului.

Impactul proiectului este unul cultural, educațional și tehnico-științific: contribuie la conservarea folclorului basarabean, facilitează studiul limbii române, sprijină dezvoltarea instrumentelor digitale educaționale și poate stimula turismul cultural. Printre dificultățile întâmpinate se numără procedurile birocratice, rigiditatea gestionării financiare și volumul foarte mare al datelor ce necesită digitizare și adnotare.

Summary of activities and results achieved in the project in 2025

Project code **25.80012.0807.50SE**

Project name: "**Digitization and creation of the FolkAI corpus with texts of Bessarabian folklore from the 19th-20th centuries.**"

The project aimed to develop a unique diachronic-parallel corpus dedicated to Bessarabian folklore from the 19th to the 20th centuries by integrating advanced digitization, automatic transliteration, and AI-based image generation, thus contributing to the preservation and valorization of cultural heritage. During the 2025 stage, the activities focused on analyzing the distinctive features of folklore texts, creating OCR models for Cyrillic script, conducting primary digitization of the sources, and preparing materials for the corpus.

Relevant folklore resources were identified and selected from libraries in the Republic of Moldova, and in collaboration with the Institute of Romanian Philology "B.P. Hasdeu", a set of linguistic and cultural characteristics specific to Bessarabian folklore was established. Eight books (over 2,500 pages) were fully scanned and will continue to be processed to meet the project objectives.

The analysis of available tools confirmed that the HeDy platform is the most suitable for digitizing historical Cyrillic texts. The platform was tested and adapted for the corpus, using pre-existing OCR models and developing new models to accommodate the specific folklore vocabulary. Automatic text recognition requires manual corrections, but HeDy provides integrated support for transliteration and post-processing.

The structure of the diachronic-parallel corpus was developed, based on aligning Cyrillic and Latin texts, along with an extensive metadata model inspired by international archives (AILLA, ELAR, SKS) and adapted to local practices. The metadata covers four categories: textual, performer-related, collection-related, and technical. In parallel, progress was made on building the Treebank in accordance with Universal Dependencies, achieving transliteration, tokenization, and the first syntactic annotations.

The project generated significant scientific results, including two articles, an abstract, and participation in international conferences (ConsILR 2025 Bucharest, WIIS 2025 Chişinău, the conference in Veliko Tarnovo), as well as at the INFOINVENT exhibition. A seminar was also organized at the Moldova State University, where the project's outcomes and future development directions of the corpus were presented to students.

The impact of the project is cultural, educational, and technical-scientific: it contributes to the preservation of Bessarabian folklore, facilitates Romanian language studies, supports the development of digital educational tools, and may stimulate cultural tourism. The challenges encountered include bureaucratic procedures, rigid financial management, and the very large volume of data requiring digitization and annotation.

Conducătorul de proiect Petic Mircea (numele, prenumele, semnătura)

Data: 04.12.2025

L.Ş.