

RECEPȚIONAT

Agenția Națională pentru Cercetare
și Dezvoltare _____

____ decembrie 2025

AVIZAT

Secția AȘM _____

____ 2025

RAPORT ȘTIINȚIFIC ANUAL

(pentru etapa 2025)

privind implementarea proiectului din cadrul concursului
Stimularea excelenței cercetărilor științifice

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

(titlul proiectului)

Cifrul proiectului 25.80012.0807.50SE

Prioritatea strategică Provocări societale

Rectorul USM

Șarov Igor

(numele, prenumele)

(semnătura)

Președintele Consiliul științific/Senat

Titchev Inga

(numele, prenumele)

(semnătura)

Conducătorul proiectului

Petic Mircea

(numele, prenumele)



Chișinău, 2025

CUPRINS:

1. Scopul etapei 2025 conform proiectului depus la concurs.....	3
2. Obiectivele etapei 2025.....	3
3. Acțiunile planificate pentru realizarea scopului și obiectivelor etapei 2025.....	3
4. Acțiunile realizate pentru atingerea scopului și obiectivelor etapei 2025.....	3
5. Rezultatele obținute	5
6. Diseminarea rezultatelor la foruri științifice.....	9
7. Impactul științific, social și/sau economic al rezultatelor științifice obținute în cadrul proiectului 2025.....	10
8. Colaborare la nivel național în cadrul implementării proiectului 2025.....	10
9. Colaborare la nivel internațional în cadrul implementării proiectului 2025.....	10
10. Dificultăți în realizarea proiectului: financiare, organizatorice, legate de resursele um....	10
11. Recomandări, propuneri.....	11
12. Rezumatul activității și a rezultatelor obținute în proiect 2025 în limba română și în limba engleză (Anexa 1).....	12
13. Lista lucrărilor științifice, publicate în anul 2025 (Anexa 2).....	14
14. Executarea devizului de cheltuieli din contractul de finanțare pentru anul 2025 (Anexa 3).....	16
15. Componenta echipei conform contractului de finanțare pentru anul 2025 (Anexa 4).....	17

1. Scopul proiectului conform proiectului depus la concurs.

Scopul proiectului propus este elaborarea unui corpus unic în felul său dedicat folclorului basarabean din sec. XIX-XX, astfel contribuind la preservarea patrimoniului cultural, prin utilizarea tehnologiilor avansate în procesul de digitizare, precum și utilizarea algoritmilor de AI în generarea imaginilor care reflecta sensul expresiilor idiomatice cât și transliterarea în scopul extinderii platformei HeDy.

2. Obiectivele etapei 2025

- Analiza cantitativă și calitativă a textelor folclorice din patrimoniul cultural basarabean din secolele XIX- XX pentru a identifica trăsăturile distinctive și caracteristicile acestora (gen. fonturi, stilurile folclorice, structura sintactică etc.);
- Crearea modelelor de recunoaștere a textelor cu caractere chirilice din secolul XIX-XX;
- Aplicarea platformei HeDy pentru digitizarea textelor folclorice din patrimoniul cultural basarabean (scanarea, pre-procesarea, OCR-izarea, post-procesarea etc.)
- Validarea (manuală și automatizată) a procesului de digitizare, identificarea tipurilor de erori și soluții de îmbunătățire a procesului de digitizare.
- Pregătirea resurselor digitale pentru crearea corpusului diacronic-paralel

3. Acțiunile planificate pentru realizarea scopului și obiectivelor etapei 2025

Activitatea 1 Identificarea resurselor textuale din bibliotecile din Republica Moldova (Biblioteca Națională, Biblioteca Universității de Stat din Moldova, Biblioteca Academiei de Științe a Moldovei „Andrei Lupan”, Biblioteca Științifică a Universității de Stat „Alec Russo” din Bălți) și întocmirea catalogului de publicații relevante.

Activitatea 2 Digitalizarea primară a resurselor existente pentru obținerea variantei scanate a textelor folclorice specifice secolelor XIX- XX

Activitatea 3 Analiza instrumentelor existente și identificarea posibilităților de adaptare a acestora pentru specificul textelor folclorice din secolele XIX- XX;

Activitatea 4 Testarea posibilităților platformei HeDy, proiectarea anumitor ajustări pentru crearea modelelor specifice;

Activitatea 5 Elaborarea publicațiilor științifice pentru a prezenta rezultatele proiectului în cadrul conferințelor

Activitatea 6 Elaborarea raportului pentru anul 2025

Activitatea 7 Seminar organizat pentru studenți la Universitatea de Stat din Moldova

4. Acțiunile realizate pentru atingerea scopului și obiectivelor etapei 2025

Activitatea 1 În cadrul conlucrării cu reprezentanții bibliotecilor au fost selectate mai multe cărți de valoare pentru cercetările din proiect și elaborată lista statistică a operelor literare folclorice din patrimoniul cultural basarabean din secolele XIX- XX. În colaborare cu cercetătorii de la Institutul de Filologie română “B.P.Hasdeu” a fost stabilită lista trăsăturilor distinctive și caracteristicile

acestora pentru textele folclorice din patrimoniul cultural basarabean din secolele XIX-XX. Aceasta ne-a permis să elaborăm structura corpusului și să identificăm metadatele, care urmează să însoțească înscrisurile în corpus.

Activitatea 2 A fost efectuată digitalizarea primară a resurselor existente pentru obținerea variantei scanate a textelor folclorice specifice secolelelor XIX- XX. Au fost selectate sursele literare care să cuprindă mai multe tipuri de texte folclorice. În colaborare cu bibliotecile, dar și cu ajutorul instrumentelor disponibile în cadrul Institutului de Matematică și Informatică „Vladimir Andrunachievici”, USM, au fost scanate integral 8 cărți ceea ce ar constitui peste 2500 pagini. Lucrul asupra scanării se efectuează în continuare, inclusiv și după prezentarea acestui raport pentru a atinge parametrii stabiliți.

Activitatea 3 Analiza instrumentelor existente ne-a permis să concludem, că cea mai potrivită scopurilor noastre este platforma HeDy, care deja a fost antrenată și conține anumite modele ce pot fi adaptate pentru specificul textelor folclorice din secolele XIX- XX;

Activitatea 4 Testarea posibilităților platformei HeDy, proiectarea anumitor ajustări pentru crearea modelelor specifice;

Platforma HeDy a fost antrenată pe diferite seturi de modele și dicționare. Astfel, pentru secolul al XVII-lea, există două modele OCR antrenate pe peste 10.000 de exemple de antrenament și un dicționar de peste 4.500 de cuvinte; pentru secolele al XVIII-lea și al XIX-lea, există 10 modele OCR cu un dicționar de peste 10.000 de cuvinte; pentru secolul al XX-lea, este suficient un singur model OCR, care are și cea mai bogată resursă de antrenament, cu peste 450.000 de cuvinte. Având în vedere că proiectul FolkAI este axat în principal pe lucrul cu texte din secolul al XX-lea, există o șansă mai mare de a obține un rezultat OCR cu un nivel destul de bun de precizie. Cu toate acestea, precizia nu este de 100%, așa că este necesară intervenția manuală pentru corectarea erorilor, care se poate face fie pe textul chirilic recunoscut, fie pe textul transliterat. Suportul pentru corectarea manuală a erorilor este asigurat de HeDy. Pe de altă parte, textele folclorice au, de regulă, un vocabular specific, care nu este reflectat integral în modelul existent. Aceasta duce la necesitatea de creare a noi modele pentru a mări acuratețea recunoașterii.

Activitatea 5 Elaborarea publicațiilor științifice pentru a prezenta rezultatele proiectului în cadrul conferințelor

Au fost elaborate 2 articole și un rezumat care au fost prezentate în cadrul Workshop on Intelligent Information Systems (WIIS2025 – 16-18 octombrie 2025 la Chișinău), The 20th edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing (ConSLR-2025 – 8-10 octombrie 2025, București) și Scientific Conference with international participation dedicated to the 35 anniversary of the Faculty of Mathematics and Informatics from St. Cyril and St. Methodius University of Veliko Tarnovo (27-28 Noiembrie 2025, Veliko Tarnovo, Bulgaria).

Activitatea 6 Elaborarea raportului pentru anul 2025

Este elaborat.

Activitatea 7 Seminar organizat pentru studenți la Universitatea de Stat din Moldova

În ziua de 12 noiembrie 2025 a fost organizat la Universitatea de Stat din Moldova, Facultatea de Matematică și Informatică seminarul „Digitizarea și crearea corpusului FolkAI cu texte ale folclorului basarabean din secolele XIX–XX” cu prezentarea rezultatelor științifice din cadrul proiectului. La seminar au participat circa 20 de studenți. În cadrul evenimentului au fost prezentate rezultatele și direcțiile proiectului FolkAI, axat pe digitizarea și analiza textelor populare din Basarabia, în scopul creării unui corpus digital valoros, accesibil cercetării moderne prin instrumente de inteligență artificială și procesare a limbajului natural. Studenții au descoperit cum tehnologiile actuale pot sprijini conservarea tradițiilor și studiul limbii române, explorând potențialul aplicării AI în domeniile umaniste.

5. Rezultatele obținute (descriere narativă 3-5 pagini)

Au fost identificate sursele literare care vor constitui baza corpusului folcloric diacronic. Din ele fac parte următoarele:

- a. Grigore Botezatu. ș.a. 1973. Folclor din părțile Codrilor, Chisinau, 362 p.
- b. Grigore Botezatu. ș.a. 1981. Creația populară moldovenească: proverbe și zicător, Chisinau, 309 p.
- c. Grigore Botezatu ș.a. 1978. Strigături, amintiri și scrisori versificate, Chișinău, Editura Știința, 347 p.
- d. Grigore Botezatu ș.a. 1986. Folclor din stepa bălților, Chișinău, Editura Știința, 302 p.
- e. Băeșu N. M. ș.a. 1966. Folclor moldovenesc. Hrestomație, Chișinău, 314 p.
- f. Grigore Botezatu 1967. Povestea lui Aliman – feciorul lui Verde Împărat, Chișinău, 26 p.
- g. Grigore Botezatu ș.a. 1957. Poezie populară moldovenească, Chișinău, Editura de Stat a Moldovei, 470 p.
- h. Grigore Botezatu ș.a. 2001. Dicționar de proverbe și zicători. Chișinău, Editura Litera, p.272
- i. Grigore Botezatu ș.a. 1983. Folclor din nordul Moldovei. Chișinău, Editura Știința, 320 p.
- j. Grigore Botezatu ș.a. 1988. Povești populare moldovenești. Chișinău, Literatura artistică, 184 p.

A fost stabilită lista trăsăturilor distinctive și caracteristicile acestora pentru textele folclorice din patrimoniul cultural basarabean din secolele XIX-XX. Din ele fac parte următoarele:

- Componentă identitară și memorie colectivă

Folclorul basarabean funcționează ca reper identitar, conservând valori, credințe și experiențe istorice comune.

- Interferență între tradiția românească și influențe regionale

Deși se înscrie în matricea folclorului românesc, folclorul din Basarabia prezintă influențe slave, ucrainene, rusești și sud-est europene datorită poziției geografice și istoriei complexe.

- Stratificare istorică și ideologică

Textele folclorice, în special cele colectate în a doua jumătate a secolului XX, includ adesea distorsiuni ideologice sovietice, inserate în colecții altfel autentice. Această stratificare reclamă reconstituire critică.

- *Transmitere orală, caracter memorabil și formule stabile*

Creațiile sunt transmise pe cale orală și utilizează repetiții, ritm, paralelisme și formule fixe care facilitează memorarea și reproducerea.

- *Varietate tipologică și genologică extrem de amplă*

Include un spectru larg: proverbe, zicători, strigături; legende, basme, povestiri; cântece (epice, lirice, ritualice); ghicitori și formule de leac; texte ritualice și cu funcție magică.

- *Grafii multiple și tranziții alfabetice*

Coexistența alfabetului *chirilic* (în principalele colecții din sec. XIX și început de XX) și trecerea progresivă la alfabetul *latin*, reflectând transformările socio-politice.

- *Arhaisme, regionalisme și particularități dialectale*

Lexicul abundă în forme vechi, neologisme populare, împrumuturi regionale și structuri sintactice arhaice, ceea ce face corpusul valoros pentru studiul evoluției limbii.

- *Symbolism și metafore tradiționale*

Textele sunt puternic metaforice, folosind simboluri naturale, cosmice, zoomorfe și ritualice pentru a transmite sensuri culturale și morale.

- *Funcție normativă și educativă*

Proverbele și zicătorile codifică norme morale, reguli de comportament, principii comunitare și modele de înțelepciune populară.

- *Potențial vizual și narativ ridicat*

Bogăția imagistică, scenariile mitice și firul narativ expresiv se pretează la ilustrare vizuală și la augmentări AI.

- *Valori documentare și etnografice*

Textele reflectă modul de viață rural, ocupațiile tradiționale, credințele și comunitățile multi-etnice, fiind surse primare pentru etnologie și antropologie.

- *Vulnerabilitate la degradare și pierdere*

O parte din creațiile autentice sunt fragmentare, incomplete sau afectate de trecerea timpului, ceea ce face necesară digitizarea lor sistematică.

- *Importanță pentru studiul diacronic al limbii române*

Textele permit urmărirea evoluției lexicului, a morfologiei și a stilului între secolele XIX–XX, cu o valoare unică pentru corpusurile istorice.

- *Necesitatea unei abordări critice moderne*

Datorită intervențiilor istorice, colectivei și ideologice, multe texte cer: verificare sursă, normalizare, recontextualizare, reconstituire comparată cu alte regiuni românești.

- *Relevanță pentru aplicații moderne (AI și educație)*

Textele au potențial ridicat pentru: corpusuri digitale, transliterare automată, ilustrare AI, instrumente educaționale interactive, patrimoniu cultural digital deschis.

În cazul nostru am realizat colectarea sistematică a textelor folclorice, provenite în principal din cărți. Pentru a asigura reprezentativitatea și diversitatea culturală, cărțile au fost atent selectate și ne-am extins în mod deliberat atenția pentru a include variații regionale - cum ar fi narațiunile nordice ale Moldovei, folclorul din stepa Bălți, regiunea Codrului și zona Soroca. O atenție egală a fost acordată diversității genurilor folclorice, rezultând în includerea legendelor, basmelor, cântecelor tradiționale, anecdotelor umoristice, doinelor, pildelor, proverbelor și zicătorilor, bocetelor și exclamațiilor ritualice. Această abordare multiregională și multigen oferă o bază cuprinzătoare pentru construirea unui corpus echilibrat de folclor basarabean. Colecția de texte folclorice specifice secolelor XIX- XX a fost digitizată (prin scanare).

Pentru realizarea acțiunilor următoare au fost efectuate cercetări riguroase în scopul selectării instrumentelor potrivite sarcinilor puse în acest proiect. În urma analizei ne-am oprit la utilizarea posibilităților oferite de platforma HeDy. Această platformă a fost dezvoltată în cadrul Institutului de Matematică și Informatică „V. Andrunachievici” (Universitatea de Stat din Moldova, Chișinău) și există în versiuni desktop și web. O versiune demo (cu funcționalități destul de bogate) este disponibilă pentru acces liber pe site-ul institutului. Platforma oferă instrumente și resurse pentru digitalizarea documentelor românești scrise în alfabet chirilic, tipărite între secolele XVII-XX. Aceasta include module de procesare a imaginilor, modele de recunoaștere optică a caracterelor, dicționare utilizate pentru recunoaștere, module pentru transliterare din alfabetul chirilic în alfabetul latin și dicționare de excepții utilizate pentru corectarea textului transliterat.

Primul pas l-a constituit preprocesarea imaginilor, necesară în special atunci când se lucrează cu texte vechi, afectate de timp sau de condiții necorespunzătoare de depozitare și utilizare. Pentru preprocesarea imaginilor în cadrul platformei HeDy utilizatorul poate alege între ABBY FineReader, OpenCV sau Scan Tailor, ultima opțiune fiind aplicabilă doar în versiunea desktop. Platforma HeDy oferă cea mai extinsă gamă de funcții pentru preprocesarea imaginilor prin FineReader, inclusiv următoarele opțiuni: corectarea rezoluției imaginii, corectarea orientării paginii, conversia imaginii în alb-negru, reducerea zgomotului și îndreptarea liniilor.

Pentru recunoașterea textelor scanate a fost necesară ajustarea modelelor de recunoaștere a textelor chirilice din sec. XIX-XX din platforma HeDy în scopul procesării textelor folclorice, astfel creându-se modele noi și îmbogățindu-se dicționarele. Pentru aceeași perioadă istorică, pot exista mai multe modele, în special în cazul alfabetelor de tranziție, iar utilizatorul îl selectează pe cel mai potrivit. Uneori poate fi util să încerce mai multe modele pentru a-l găsi pe cel mai apropiat. Platforma HeDy a fost antrenată pe diferite seturi de modele și dicționare.

Cu toate acestea, precizia sistemului Hedy nu este de 100%, așa că este necesară intervenția manuală pentru corectarea erorilor, care se poate face fie pe textul chirilic recunoscut, fie pe textul transliterat. Asistența pentru corectarea manuală a erorilor este asigurată de HeDy.

A fost elaborată structura corpusului diacronic-parallel al folclorului basarabean, evidențiind modul în care textele în grafie chirilică pot fi digitizate, recunoscute și transliterate în alfabet latin prin intermediul platformei HeDy. Vom menționa, că paralelismul în corpus nu se referă la bilingvism, ci la coexistența și alinierea celor două sisteme grafice. A fost propus un model de metadata specializat pentru organizarea materialelor folclorice, bazat pe experiența unor arhive internaționale (AILLA, ELAR, SKS) și pe „pașapoartele de creație” folosite în tradiția cercetării moldovenești. Modelul dezvoltat a structurat metadatale în patru categorii: (1) metadata textuale

(gen, motive ATU/Thompson, cuvinte-cheie, rezumate), (2) metadata despre performer (vârstă, gen, rol cultural, dialect), (3) metadata despre colectare (culegător, loc, dată, context performativ) și (4) metadata tehnice (format, nivelul transcrierii, licențiere). Această structură permite atât analiza lingvistică și etnografică aprofundată, cât și integrarea materialelor în aplicații AI, precum căutarea semantică sau educația digitală.

Au fost testate posibilitățile de aplicare a diferitelor instrumente de inteligență (ChatGPT, NightCafe, NanoBanana, Canva, Freepik) pentru ilustrarea subiectelor folclorice.

Concomitent, s-a lucrat la construirea **FolkAI Treebank**, o resursă lingvistică aliniată la standardul Universal Dependencies, dar adaptată particularităților folclorului basarabean, stabilindu-se modul în care textele digitizate pot fi transliterate, tokenizate, normalizate și adnotate în format CoNLL-U. Primele rezultate au inclus două lucrări procesate, însoțite de avansuri semnificative în transliterare și adnotare sintactică.

Astfel, rezultatele obținute rezumă în următoarele:

1. A fost selectat și digitizat prin scanare un set de cărți cu conținut folcloric, reflectând texte literare folclorice de diferit gen, acestea fiind culese din diferite regiuni basarabene.
2. O parte din aceste cărți a fost procesată cu ajutorul platformei HeDy, parcurgându-se etapele de procesare a imaginilor, recunoaștere a caracterelor, transliterare.
3. Specificul textelor folclorice a necesitat ajustarea platformei, în special lucrul de completare a dicționarilor cu vocabularul respectiv.
4. A fost elaborată structura corpusului diacronic cu stabilirea metadatelor, care reflectă atât genul creației folclorice, cât și date ce corespund pașaportului de creație, utilizat de cercetătorii din domeniu.
5. Au fost testate posibilitățile de aplicare a diferitelor instrumente de inteligență (ChatGPT, NightCafe, NanoBanana, Canva, Freepik) pentru ilustrarea subiectelor folclorice.

6. **Diseminarea rezultatelor** obținute în proiect în formă de publicații (obligatoriu) și în formă de prezentări la foruri științifice (comunicări, postere – pentru cazurile când nu au fost publicate în materialele conferințelor). Lista lucrărilor publicate este prezentată în Anexă.

Participări la manifestări științifice internaționale

1. The 20th International Conference on Linguistic Resources and Tools for NLP, Bucharest, 08-10 October 2025, Romania
2. The Workshop on Intelligent Information Systems WIIS2025, October 16-18, 2025, Chisinau, Republic of Moldova

Participări la manifestări științifice naționale cu participare internațională

1. Scientific Conference with International participation dedicated to the 35-th anniversary of the Faculty of Mathematics and Informatics. November 27-28, 2025, Veliko Tarnovo

Participări la manifestări științifice naționale

1. "INFOINVENT". Expoziția Internațională Specializată. Ediția a XIX-a. 3-5 decembrie 2025. CIE Moldexpo SA

Altele – Seminare în cadrul universităților

1. Seminarul „Digitizarea și crearea corpusului FolkAI cu texte ale folclorului basarabean din secolele XIX–XX”, Universitatea de Stat din Moldova, Facultatea de Matematică și Informatică, Chișinău. 12 noiembrie 2025

Lista publicațiilor din anul 2025 în care se reflectă doar rezultatele obținute în proiect, perfectată conform cerințelor față de lista publicațiilor (a se vedea Anexa 2)

Notă: Lista va include și brevetele de invenții și alte obiecte de proprietate intelectuală, materiale la saloanele de invenții în cazul în care sunt (conform Anexei 2)

1. Caftanator O., Cojocaru S., Malahov L., Petic M. FolkAI: a diachronic corpus of Bessarabian folklore texts. In: Proceedings of the 20th International Conference on Linguistic Resources and Tools for NLP, Bucharest, 08-10 October 2025, to be published.
2. Petic M., Caftanator O., Cojocaru S., Malahov L., and Atul Kr. Ojha. 2025. Digitizing of Bessarabian Folklore: From Textual Heritage to Digital Corpus. In: Proceedings of the Workshop on Intelligent Information Systems WIIS2025, October 16-18, 2025, Chisinau, Republic of Moldova, 319-332.
3. Caftanator O., Cojocaru S., Malahov L., Petic M. FolkAI: Metadata Design and Corpus Structuring for the Preservation of Moldavian Folk Heritage. In: Book of Abstracts of Scientific Conference with International participation dedicated to the 35-th anniversary of the Faculty of Mathematics and Informatics. November 27-28, 2025, Veliko Tarnovo, to be published.
4. Petic M., Cojocaru S., Caftanator O., Malahov L. FolkAI: Corpus diacronic paralel cu texte ale folclorului basarabean / FolkAI: Parallel diachronic corpus with texts of bessarabian

folklore. In: INFOINVENT 2025. Editia XIX. CATALOG OFICIAL. Expoziția Internațională Specializată, 3-5 decembrie, II.B.78, pag.215.

7. Impactul științific, social și/sau economic al rezultatelor științifice obținute în cadrul proiectului

Proiectul **FolkAI** contribuie semnificativ la conservarea și valorificarea patrimoniului cultural imaterial prin digitalizarea, transliterarea și vizualizarea inteligentă a textelor folclorice basarabene. Rezultatele au impact socio-economic prin promovarea identității culturale, stimularea turismului cultural și integrarea resurselor în educația digitală. Din perspectivă tehnică, proiectul dezvoltă un flux inovativ de procesare a limbajului natural pentru limbi cu resurse reduse, combinând recunoașterea optică, transliterarea automată și generarea vizuală bazată pe inteligență artificială — un model transferabil pentru alte domenii ale patrimoniului cultural.

Folclorul reprezintă o parte esențială a identității culturale și a memoriei colective a unui popor, servind și drept o modalitate de reflectare a istoriei, tradițiilor sociale și economice din trecut. Totodată, activitățile bazate pe valorificarea folclorului autohton pot servi drept un catalizator de dezvoltare a turismului, aducându-și astfel contribuția la dezvoltarea economică a comunităților.

8. Colaborare la nivel național în cadrul implementării proiectului

În cadrul proiectului e firesc să colaborăm pentru atingerea rezultatelor cu Biblioteca Națională, Biblioteca Universității de Stat din Moldova, Biblioteca Academiei de Științe a Moldovei „Andrei Lupan”, Biblioteca Științifică a Universității de Stat „Alec Russo” din Bălți.

Totodată este firească colaborarea și cu Facultatea de Matematică și Informatică USM, Facultatea de Litere USM, dar și Institutul de Filologie Română „Bogdan Petriceicu- Hașdeu” al USM.

9. Colaborare la nivel internațional în cadrul implementării proiectului Suntem deschiși pentru toți care sunt interesați de cercetările care sunt realizate în cadrul proiectului: Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu” (București), Facultatea de Informatică de la UAIC (Iași).

10. Dificultățile în realizarea proiectului de natură financiară, organizatorică, legate de resursele umane etc. (obligatoriu).

- Procedurile birocratice la nivelul bugetului, dar și a procedurilor de raportare.
- Lipsa flexibilității în gestionarea surselor financiare.
- Volumul mare de date - obiectivul de 5.000+ pagini, 100.000+ tokeni și aproximativ 1.000 de ilustrații presupune efort tehnic și uman substanțial pentru digitizare, validare, adnotare și curățare.

11. Recomandări, propuneri (opțional).

- Flexibilizarea procedurilor de gestionarea proiectului la nivel de buget și gestionarea surselor financiare.

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

Data: 04.12.2025

LȘ

Rezumatul activității și a rezultatelor obținute în proiect în anul 2025

Cifra 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-paralel 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-paralel, 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.

**Lista lucrărilor științifice, științifico-metodice și didactice
publicate în anul 2025 în cadrul proiectului**

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

1. Monografii (recomandate spre editare de consiliul științific/senatul organizației din domeniile cercetării și inovării)

1.1. monografii internaționale

1.2. monografii naționale

2. Capitole în monografii naționale/internaționale

3. Editor culegere de articole, materiale ale conferințelor naționale/internaționale

4. Articole în reviste științifice

4.1. în reviste din bazele de date Web of Science și SCOPUS (cu indicarea factorului de impact IF)

4.2. în alte reviste din străinătate recunoscute

4.3. în reviste din Registrul National al revistelor de profil, cu indicarea categoriei

4.4. în alte reviste naționale

5. Articole în culegeri științifice naționale/internaționale

5.1. culegeri de lucrări științifice editate peste hotare

CAFTANATOV, O.; COJOCARU, S.; MALAHOV, L.; PETIC, M., FolkAI: a diachronic corpus of Bessarabian folklore texts. In: Proceedings of the 20th International Conference on Linguistic Resources and Tools for NLP, Bucharest, 08-10 October 2025, to be published.

5.2 culegeri de lucrări științifice editate în Republica Moldova

6. Articole în materiale ale conferințelor științifice

6.1. în lucrările conferințelor științifice internaționale (peste hotare)

6.2. în lucrările conferințelor științifice internaționale (Republica Moldova)

PETIC, M.; CAFTANATOV, O.; COJOCARU, S.; MALAHOV, L., and ATUL Kr. Ojha, Digitizing of Bessarabian Folklore: From Textual Heritage to Digital Corpus. In: Proceedings of the Workshop on Intelligent Information Systems WIIS2025, October 16-18, 2025, Chisinau, Republic of Moldova, 319-332.

6.3. în lucrările conferințelor științifice naționale cu participare internațională

6.4. în lucrările conferințelor științifice naționale

7. Teze ale conferințelor științifice

7.1. în lucrările conferințelor științifice internaționale (peste hotare)

CAFTANATOV, O.; COJOCARU, S.; MALAHOV, L.; PETIC, M., FolkAI: Metadata Design and Corpus Structuring for the Preservation of Moldavian Folk Heritage. In: Book of Abstracts of Scientific Conference with International participation dedicated to the 35-th anniversary of the Faculty of Mathematics and Informatics. November 27-28, 2025, Veliko Tarnovo, to be published.

7.2. în lucrările conferințelor științifice internaționale (Republica Moldova)

7.3. în lucrările conferințelor științifice naționale cu participare internațională

7.4. în lucrările conferințelor științifice naționale

Notă: vor fi considerate teze și nu articole materialele care au un volum de până la 0,25 c.a.

8. Alte lucrări științifice (recomandate spre editare de o instituție acreditată în domeniu)

8.1. cărți (cu caracter informativ)

8.2. enciclopedii, dicționare

8.3. atlase, hărți, albume, cataloage, tabele etc. (ca produse ale cercetării științifice)

9. Brevete de invenții și alte obiecte de proprietate intelectuală, materiale la saloanele de invenții

PETIC, M.; COJOCARU, S.; CAFTANATOV, O.; MALAHOV, L., FolkAI: Corpus diacronic paralel cu texte ale folclorului basarabean / FolkAI: Parallel diachronic corpus with texts of bessarabian folklore. In: INFOINVENT 2025. Editia XIX. CATALOG OFICIAL. Expoziția Internațională Specializată, 3-5 decembrie, II.B.78, pag.215.

10. Lucrări științifico-metodice și didactice

10.1. manuale pentru învățământul preuniversitar (aprobate de ministerul de resort)

10.2. manuale pentru învățământul universitar (aprobate de consiliul științific /senatul instituției)

10.3. alte lucrări științifico-metodice și didactice

11. Recomandări, propuneri.

Nu sunt.

**Executarea devizului de cheltuieli,
conform anexei nr. 2.3 din contractul de finanțare pentru anul 2025**

Cifrul proiectului: 25.80012.0807.50SE

Cheltuieli				
Denumirea	Cod		Anul de gestiune	
	Eco (k6)	Aprobat	Modificat +/-	Precizat
Deplasări în interes de serviciu peste hotare	222720	35640		35640
Servicii medicale	222810			
Servicii editare	222910			
Servicii de cercetări științifice	222930	180420,0		180420,0
Servicii neatribuite altor aliniate	222999	74400		74400
Cheltuieli curente neatribuite la alte categorii	281900	8000		8000
Procurarea materialelor de uz gospodăresc și rechizite de birou	316110			
Procurarea activelor nemateriale	317110			
Procurarea materialelor pentru scopuri didactice, științifice și alte scopuri	335110			
Procurarea materialelor de uz gospodăresc și rechizitelor de birou	336110	1540		1540
Total		300000		300000

Conducătorul organizației Șarov Igor (numele, prenumele, semnătura)

Contabil șef Cojocaru Liliana / (numele, prenumele, semnătura)

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



Componența echipei conform contractului de finanțare 2025

Cifrul proiectului 25.80012.0807.50SE

Echipa proiectului conform contractului de finanțare (la semnarea contractului) pentru 2025						
Nr	Nume, prenume, (conform contractului de finanțare)	Anul nașterii	Titlul științific	Norma de muncă sau nr. de ore conform contractului	Data angajării	Data eliberării
1.	Petic Mircea	1983	Dr.	24	01.08.2025	31.12.2025
2.	Malahov Ludmila	1947		24	01.08.2025	31.12.2025
3.	Caftanator Olesea	1986		25	01.08.2025	31.12.2025
4.	Cojocaru Svetlana	1952	Dr. Hab. , acad.	24	01.08.2025	31.12.2025

Conducătorul organizației Șarov Igor / (numele, prenumele, semnătura)Contabil șef Cojocaru Liliana / (numele, prenumele, semnătura)Conducătorul de proiect Petic Mircea / (numele, prenumele, semnătura)