

Raport privind prezentarea ipotezelor cercetării și a rezultatelor obținute. Justificarea încadrării rezultatelor la nivelul TRL 3

Titlu proiect: *De la arhivă la canon: o lectură de la distanță a romanului românesc (1845-1947)*

Tipul proiectului: Proiect de Cercetare Exploratorie – PCE 2020

Codul proiectului: PN-III-P4-ID-PCE-2020-2690

Perioada de derulare a proiectului: 4 ianuarie 2021 – 29 decembrie 2023

Programul care a acordat finanțarea: PNCDOI III (prin CNCS-UEFISCDI)

Prezentul document are rolul de a justifica încadrarea *modulului de identificare a microgenurilor literare, adaptat pentru a măsura nivelul rezilienței climatice în texte narative ficționale* (MICROLIT) realizat în cadrul proiectului ARCAN la **nivelul de maturitate tehnologică TRL 3**, în conformitate cu definițiile și criteriile oficiale aplicabile.

1. Contextul proiectului ARCAN și poziționarea rezultatului analizat

Proiectul ARCAN [*De la arhivă la canon: o lectură de la distanță a romanului românesc (1845-1947)*] a investigat relația dintre dimensiunea cantitativă a producției romanului românesc și calitatea estetică a acestuia, prin integrarea instrumentelor calitative ale teoriei genurilor literare cu metode computaționale bazate pe tehnologii de procesare a limbajului natural (NLP). Metodologia ARCAN s-a bazat pe abordarea de tip *distant reading*, care permite analiza sistematică a unui volum foarte mare de texte, imposibil de parcurs prin metode tradiționale de lectură.

Originalitatea proiectului constă în dimensiunea sa interdisciplinară, care articulează cercetarea umanistă cu informatica, în amploarea fără precedent a corpusului analizat și în nivelul ridicat de granularitate al categoriilor utilizate, materializat prin sistemul de microgenuri literare. ARCAN reprezintă prima inițiativă de acest tip în spațiul românesc și una dintre puținele la nivel european.

Prin rezultatele sale, printre care se numără și *Modulul de identificare a microgenurilor literare, adaptat pentru a măsura nivelul rezilienței climatice în texte narative ficționale*, proiectul contribuie la dezvoltarea unor modele analitice inovatoare pentru studiul istoriei literare, la consolidarea infrastructurii digitale pentru cercetarea umanistă și la generarea unui impact semnificativ la nivel național și internațional, cu efecte culturale, educaționale și economice care depășesc cadrul strict academic.

2. Identificarea și descrierea rezultatului exploatabil

Rezultatul analizat este definit ca: *Modul de identificare a microgenurilor literare, adaptat pentru a măsura nivelul rezilienței climatice în texte narative ficționale* (MICROLIT)

MICROLIT este un sistem computațional dedicat analizei automate a diversității stilistice și a microgenurilor literare în proza românească. Sistemul abordează o problemă fundamentală în studiile literare computaționale: depășirea clasificării simplificate a operelor literare în genuri unice și explorarea naturii eterogene și hibride a textelor literare.

Modulul răspunde unui decalaj semnificativ în domeniul analizei literare computaționale, având în vedere faptul că majoritatea cercetărilor se concentrează pe limbi pentru care există o infrastructură digitală și tehnologică mult mai dezvoltată, precum engleza, franceza sau germana. Simpla preluare

mecanică a unor modele analitice și computaționale dezvoltate pentru clasificarea acestor limbi consacrate în cercetarea internațională nu reprezintă o soluție viabilă, întrucât o asemenea abordare nu este suficient de sensibilă la particularitățile morfo-sintactice, stilistice și culturale ale limbii române. Literatura română, cu specificul său istoric și estetic, necesită instrumente adaptate care să țină cont de particularitățile lingvistice și de contextul cultural al dezvoltării sale.

3. Prezentarea ipotezelor cercetării și a rezultatelor obținute

Inițiativa a pornit de la construirea și curățarea unui corpus extins de texte literare românești din perioada în care genurile narrative moderne prindeau contur. Corpusul reunește 175 de romane scrise de 106 autori între 1845 și 1920, clasificate la nivel de operă în 17 genuri literare distincte și preprocesate la nivel de capitol și paragraf (în total 3583 de capitole). Metodologic, cercetarea combină două abordări care se completează reciproc. Pe de o parte, extragerea automată a peste 100 de indici de complexitate textuală, meniți să surprindă trăsăturile lingvistice ale textelor. Pe de altă parte, utilizarea modelelor lingvistice mari (LLM) pentru predicția microgenurilor la un nivel granular de detaliu. Rezultatele sunt validate prin teste statistice non-parametrice și prin compararea încrucișată a celor două metode. Ceea ce iese la suprafață din această analiză este tocmai complexitatea internă a literaturii române din epocă: hibriditatea genurilor, coexistența mai multor microgenuri în interiorul aceleiași opere și o distribuție care variază semnificativ de la capitol la capitol, de la paragraf la paragraf.

Pornind de la constatările privind hibriditatea microgenurilor literare, metodologia MICROLIT este extinsă pentru a identifica tipare narrative asociate schimbărilor climatice în literatura română modernă și contemporană. Modelele lingvistice mari open-source permit analiza granulară a textelor la nivel de capitol și paragraf, evidențiind situații ecologice, exploatare de resurse naturale sau comportamente responsabile și iresponsabile față de mediu. Această abordare recunoaște că, asemenea microgenurilor, temele climatice nu se manifestă uniform într-o operă, ci se integrează adesea în mod hibrid în structura narativă, intersectând elemente sociale, istorice sau psihologice.

Sistemul poate genera prezentări coerente și structurate ale operelor și autorilor, sintetizând informații despre contextul istoric, caracteristicile stilistice și relațiile narrative dintre personaje și mediu. Analiza combină trăsături lingvistice cuantificabile cu predicții ale LLM-urilor la nivel de paragraf, permițând identificarea frecvențelor și tipologiilor situațiilor descrise, precum și modul în care acestea contribuie la percepția sensibilității ecologice în text.

Prin această metodologie, cercetătorii pot explora modul în care literatura română reflectă preocupările față de mediu, înțelegând nu doar prezența explicită a temelor climatice, ci și formele subtile prin care acestea sunt integrate în narativ. În plus, instrumentul digital rezultat oferă o bază solidă pentru analiza comparativă a operelor, valorificând hibriditatea stilistică pentru a dezvălui tipare emergente care altfel ar rămâne dificil de detectat prin metode tradiționale de lectură critică.

Pe lângă analiza microgenurilor, MICROLIT integrează un modul dedicat **căutărilor tematiche**, prin care sistemul detectează automat și cartografiază distribuția unui set de **teme predefinite legate de mediu și schimbări climatice** în cadrul corpusului. Teme precum exploatarea resurselor naturale, relația omului cu pământul, dezastrele naturale, defrișările, seceta sau transformările peisajului rural sunt identificate la nivel de paragraf și capitol, oferind o imagine de ansamblu asupra modului în care sensibilitatea ecologică a evoluat în literatura română de-a lungul timpului. Această funcționalitate oferă cercetătorilor un punct de intrare rapid și structurat în corpus și deschide totodată noi perspective

pentru înțelegerea modului în care literatura română a reflectat tensiunile dintre om și mediul său natural.

4. Validarea rezultatelor

Rezultatul a fost validat formal prin aprobarea Raportului final de activitate al proiectului ARCAN.

MICROLIT îndeplinește criteriile specifice TRL 3 prin validare experimentală realizată în laborator. Toate componentele au fost testate într-un mediu controlat, pe un set de date curățat și structurat, demonstrând fezabilitatea metodei în identificarea și caracterizarea microgenurilor literare în limba română.

Analiza cantitativă riguroasă a oferit măsurători precise ale performanței sistemului și semnificației rezultatelor obținute. Experimentele au fost documentate și puse la dispoziția comunității științifice prin cod open-source și date publice, asigurând astfel reproductibilitatea studiului. Setul de date poate fi găsit accesând link-ul https://huggingface.co/datasets/upb-nlp/lumro_175_novels/, iar codul poate fi accesat la <https://github.com/upb-nlp/LUMRO>.

Validarea conceptuală a demonstrat că trăsăturile lingvistice extrase în mod automat pot discrimina între genuri literare cu semnificație statistică, iar modelele LLM pot detecta microgenuri la scară mică fără a necesita fine-tuning specific. Combinarea celor două abordări oferă o perspectivă complementară asupra diversității stilistice, iar analiza confirmă că hibriditatea genurilor literare este o caracteristică intrinsecă a operei, nu o excepție.

5. Justificare TRL 3

RCE-ul MICROLIT se află în prezent la nivelul TRL 3, demonstrând fezabilitatea conceptului și validarea experimentală inițială a metodologiei propuse. Construirea și validarea corpusului de date a inclus asamblarea a 175 de romane românești publicate între 1845 și 1920, preprocesate prin normalizare ortografică, segmentare automată în capitole cu validare manuală și filtrarea conținutului non-narativ. Textul a fost structurat la nivel multi-nivel, cu 2.959 de capitole segmentate în paragrafe, pentru a permite analiza granulară a microgenurilor literare.

Componenta de analiză lingvistică a fost testată prin extragerea automată a peste 100 de indici de complexitate textuală folosind framework-ul ReaderBench (indici referitori la sintaxa frazei, precum numărul de complemente de cauză, numărul vocativelor; indici referitori la morfologie, precum media substantivelor per frază, numărul maxim de pronume personale; indici referitori la complexitatea cuvintelor, precum numărul de diferențe între leme, sau referitori la lungimea frazei). Datele au fost prelucrate statistic pentru eliminarea atributelor neinfluențe și reducerea multicolinearității, iar testele non-parametrice au identificat trăsături lingvistice semnificative atât la nivel de document, cât și la nivel de capitol.

Componenta de clasificare cu LLM a avut la bază un protocol de prompting structurat în limba română și o schemă JSON pentru consistența răspunsurilor modelelor. Modelele Llama 3.3 70B și DeepSeek-R1 70B au fost comparate prin validare încrucișată, obținând acuratețe top-5 de 80.57% și 89.71%, respectiv, pentru predicția genului dominant al romanului. Modelele au demonstrat, de asemenea, capacitatea de a detecta microgenuri secundare și de a evidenția hibriditatea genurilor literare. Analiza distribuției microgenurilor la nivel de paragraf a confirmat prezența simultană a mai multor genuri în toate romanele studiate și a permis cartografierea variației acestora pe parcursul operelor, evidențiind

heterogenitatea narativă și complementaritatea dintre analiza lingvistică globală și detectarea detaliată cu LLM.

6. Concluzie

Modulul de identificare a microgenurilor literare, adaptat pentru a măsura nivelul rezilienței climatice în texte narative ficționale este un rezultat exploatabil al proiectului ARCAN, cu grad de maturitate tehnologică TRL 3 – concept demonstrat analitic, conform definițiilor agreate la nivel național și internațional.

10.12.2023

Director proiect,
Prof. univ. dr. Andrei Terian