RAPORT DE EVALUARE A MATURITĂȚII TEHNOLOGICE

Nivelurile TRL 1 – TRL 3

Modul computațional de identificare a microgenurilor literare în texte narative ficționale

(MICROLIT)

Titlul proiectului	De la arhivă la canon: o lectură de la distanță a romanului românesc (1845–1947) — ARCAN
Cod proiect	PN-III-P4-ID-PCE-2020-2690
Program de finanțare	PNCDI III, prin CNCS-UEFISCDI
Tip proiect	Proiect de Cercetare Exploratorie — PCE 2020
Perioada de derulare	4 ianuarie 2021 - 29 decembrie 2023
Director de proiect	Prof. univ. dr. habil. Andrei Terian
Instituția coordonatoare	Universitatea „Lucian Blaga” din Sibiu
Nivelul TRL evaluat	TRL 3 (cu documentarea traseului progresiv de la TRL 1)
Autori raport	Cristina Udrea, Vlad Pojoga
Data evaluării	15 decembrie 2023
Versiunea documentului	1.0

Introducere

Prezentul raport documentează traseul de maturizare tehnologică al modulului computațional de identificare a microgenurilor literare în texte narative ficționale (MICROLIT), dezvoltat în cadrul proiectului de cercetare exploratorie ARCAN (*De la arhivă la canon: o lectură de la distanță a romanului românesc, 1845–1947*, cod PN-III-P4-ID-PCE-2020-2690), finanțat prin PNCDI III de către CNCS-UEFISCDI.

Scopul documentului este de a demonstra, prin probe analitice și experimentale cumulative, că MICROLIT îndeplinește criteriile de ieșire corespunzătoare nivelului TRL 3 = dovadă analitică și experimentală a conceptului (proof of concept), în conformitate cu definițiile standardizate ale Nivelurilor de Maturitate Tehnologică (TRL), astfel cum sunt acestea formulate în cadrul NASA și adoptate de Comisia Europeană.

Raportul urmează logica cumulativă impusă de cadrul TRL: fiecare nivel trebuie să fie complet documentat înainte ca nivelul următor să poată fi revendicat. În consecință, documentul reconstituie întregul traseu progresiv de la TRL 1 la TRL 3, prezentând pentru fiecare nivel baza probatorie specifică, criteriile de ieșire îndeplinite și lista de verificare corespunzătoare.

Dovada centrală pentru TRL 3 o constituie rezultatele experimentale obținute de echipa proiectului ARCAN prin aplicarea metodologiei MICROLIT pe un corpus de 175 de romane românești publicate între 1845 și 1920, reprezentând 106 autori și 17 genuri literare distincte. Aceste rezultate au fost obținute printr-o metodologie dual, respectiv combinarea extragerii automată a indicilor de complexitate textuală prin platforma ReaderBench cu predicția microgenurilor prin modele lingvistice mari (LLM), și au fost validate prin teste statistice non-parametrice riguroase.

Raportul este structurat după cum urmează. Capitolul 1 prezintă cadrul de referință TRL. Capitolele 2, 3 și 4 documentează, în ordine, nivelurile TRL 1, TRL 2 și TRL 3, fiecare încheiat cu o listă de verificare a criteriilor îndeplinite. Capitolul 5 prezintă sinteza traseului progresiv și perspectivele de dezvoltare spre TRL 4, urmat de bibliografie.

Capitolul 1. Cadrul de referință: Nivelurile de Maturitate Tehnologică (TRL)

1.1 Ce sunt Nivelurile de Maturitate Tehnologică?

Nivelurile de Maturitate Tehnologică (Technology Readiness Levels = TRL) reprezintă un sistem standardizat de evaluare a gradului de dezvoltare al unei tehnologii de-a lungul ciclului său de viață, de la cercetarea fundamentală până la implementarea operațională completă. Dezvoltat inițial de NASA în anii 1970 și adoptat ulterior de Comisia Europeană, Departamentul american al Apărării și principalele organisme de finanțare a cercetării, cadrul TRL furnizează un limbaj comun care permite comunicarea coerentă între dezvoltatori, evaluatori și finanțatori cu privire la stadiul real de maturitate al unei tehnologii.

Scala TRL cuprinde nouă niveluri distincte. Fiecare nivel reprezintă un prag de sine stătător: o tehnologie fie îndeplinește criteriile de ieșire ale acelui nivel, fie nu le îndeplinește. Un principiu fundamental este acela al cumulativității: fiecare nivel TRL trebuie să fie complet documentat înainte ca nivelul următor să poată fi revendicat.

1.2 Definiții TRL 1-9 pentru instrumente software

Tabelul următor prezintă definițiile complete ale celor nouă niveluri TRL, adaptate pentru instrumente software și sisteme digitale, domeniu în care se înscrie MICROLIT.

Nivel	Denumire	Descriere (Software)	Criteriu de ieșire
TRL 1	Principii de bază observate și raportate	Cunoștințe științifice generate pentru proprietățile de bază ale arhitecturii software și formulărilor matematice.	Publicație recenzată sau raport tehnic care documentează baza științifică.
TRL 2	Concept tehnologic și/sau aplicație formulată	Aplicație practică identificată, încă speculativă. Algoritmi definiți, principii codificate, experimente cu date sintetice.	Descriere documentată a aplicației/conceptului privind fezabilitatea și beneficiile potențiale.
TRL 3	Dovadă analitică și experimentală a conceptului (Proof of Concept)	Funcționalitate limitată pentru validarea proprietăților critice, folosind componente software neintegrate.	Rezultate analitice/experimentale documentate care validează predicțiile privind parametrii cheie.
TRL 4	Validare componentă/prototip în mediu de laborator	Componente software critic-funcționale integrate și validate funcțional. Interoperabilitate stabilită.	Performanță de test documentată. Definiție documentată a mediului relevant.
TRL 5	Validare componentă/prototip în mediu relevant	Elemente software end-to-end implementate și interfațate cu sisteme existente, conforme mediului țintă.	Performanță de test documentată. Cerințe de scalare documentate.

Nivel	Denumire	Descriere (Software)	Criteriu de ieșire
TRL 6	Demonstrare model/prototip de sistem în mediu operațional	Prototip software demonstrat pe probleme realiste la scară completă. Integrare parțială. Fezabilitate demonstrată complet.	Performanță de test documentată, demonstrând concordanța cu predicțiile analitice.
TRL 7	Demonstrare prototip sistem în mediu operațional	Prototip cu toate funcționalitățile cheie disponibile. Bine integrat cu sisteme operaționale. Majoritatea bug-urilor remediate.	Performanță de test documentată, demonstrând concordanța cu predicțiile analitice.
TRL 8	Sistem final calificat prin testare și demonstrare	Software complet depanate și integrat. Documentație completă (utilizator, instruire, mentenanță). V&V finalizate.	Performanță de test documentată, verificând predicțiile analitice.
TRL 9	Sistem dovedit prin operare în misiune reușită	Software complet depanate și integrat. Toată documentația completă. Suport ingineresc activ. Sistem operat cu succes.	Rezultate operaționale de misiune documentate.

1.3 Notă metodologică privind aplicarea TRL în domeniul umanioarelor digitale

În contextul umanioarelor digitale, i.e. domeniul în care se înscrie MICROLIT, noțiunea de „mediu operațional relevant” desemnează corpusuri de texte controlate și adnotate, iar „experimentele” corespund protocoalelor de prompting, testelor statistice non-parametrice și comparațiilor încrucișate între modele. Rigorile probatorii rămân echivalente cu cele din științele experimentale. Absența artefactelor fizice nu reduce pragul probatoriu — îl deplasează spre cod documentat, validarea algoritmilor, ieșiri din simulări și rezultate experimentale structurate.

Capitolul 2. Nivelul TRL 1 = Principii de bază observate și raportate

2.1 Scopul acestei secțiuni

Această secțiune documentează cunoașterea științifică fundamentală care stă la baza tehnologiei MICROLIT. La nivelul TRL 1, nu există implementare — produsul acestui nivel este cunoaștere teoretică și metodologică, nu cod. În cazul MICROLIT, baza de TRL 1 este constituită din literatura științifică internațională asimilată critic de echipa ARCAN în faza de concepere a proiectului. Această modalitate de ancorare la TRL 1 este legitimă și frecventă în proiectele de cercetare exploratorie: echipa nu a reprodus cunoașterea existentă, ci a identificat, evaluat și integrat selectiv principiile relevante pentru a formula un concept tehnologic nou, adaptat specificului limbii și literaturii române.

2.2 Principiile științifice care fundamentează MICROLIT

2.2.1 Teoria literară a microgenurilor și abordarea distant reading

Studiul literar tradițional a clasificat operele narative în genuri largi, tratând apartenența generică drept o proprietate globală și stabilă a textului. Cercetările teoretice din ultimele decenii au contestat sistematic această premisă. Derrida (1980) a demonstrat că participarea unui text la un gen nu echivalează niciodată cu apartenența exclusivă la acesta, iar Bakhtin — prin conceptul de heteroglosie, preluat și dezvoltat de Ivanov (2002, 2008) — a arătat că vocile multiple dintr-un text destabilizează prin natura lor granițele generice. Fowler (1982) și Todorov (1990) au furnizat cadrul teoretic structural, iar Genette (1992) a clarificat relația dintre gen și alte categorii taxinomice ale discursului literar.

Abordarea distant reading, teoretizată de Moretti (2013) și aplicată sistematic în cadrul Stanford Literary Lab, a demonstrat că analiza la scară mare a corpusurilor literare produce cunoaștere calitativ diferită față de lectura tradițională. În spațiul românesc, contribuțiile lui Terian (2019, 2022), Baghiu (2022) și Borza et al. (2020) au dezvoltat o taxonomie specifică a microgenurilor romanului românesc, recunoscând totodată că opere precum romanul haiducesc — studiat de Patraș (2019) — ridică probleme de echivalență terminologică față de literaturile occidentale de referință. Tocmai această particularitate a literaturii române a motivat nevoia unui instrument analitic adaptat, nu pur importat.

2.2.2 Procesarea computațională a limbajului natural aplicată textelor literare

Clasificarea automată a genurilor literare are o tradiție consolidată în literatura internațională. Lucrarea de pionierat a lui Kessler et al. (1997) a stabilit bazele detecției automate a genului textual, iar studii ulterioare — precum Hettinger et al. (2016) — au demonstrat eficiența trăsăturilor stilometrice, tematice și de rețea pentru clasificarea subgenurilor în romane germane prin modele SVM. Platforma ReaderBench, dezvoltată de Dascalu et al. (2013) și validată pe texte românești în contexte diverse — de la analiza oratorilor români (Dascalu, Gîfu & Trausan-Matu, 2016) la evaluarea dificultății textuale în literatura pentru copii — furnizează echipei ARCAN o infrastructură NLP testată empiric, evitând costurile și riscurile construirii unui instrument de la zero.

2.2.3 Modele lingvistice mari pentru clasificare în limbi cu resurse limitate

Cercetările privind clasificarea prin LLM-uri în limbi cu resurse digitale limitate relevă un potențial semnificativ pentru cazul românesc. Cercetările au demonstrat că clasificarea zero-shot prin prompting atinge

performanțe comparabile cu modelele fine-tuned BERT pe texte germane. În paralel, a fost propusă o metodă de clasificare tematică bazată pe dicționar pentru luxemburgheză, care depășește abordările tradiționale. Aceste rezultate susțin premisa că modelele LLM de uz general pot detecta variație stilistică subtilă și în texte românești din secolul al XIX-lea, fără a necesita corpusuri de antrenare specifice — o premisă pe care MICROLIT o testează explicit.

2.3 Formulările matematice și arhitecturale de bază

La nivelul TRL 1, formulările matematice relevante sunt cele moștenite din literatura de specialitate. Ele includ: modelele statistice non-parametrice pentru compararea distribuțiilor (testul Kruskal-Wallis, formalizat de Vargha & Delaney, 1998; testul Mann-Whitney U, descris în McKnight & Najab, 2010), metodele de control al ratei false de descoperire (Benjamini & Hochberg, 1995; Storey, 2011), algoritmii de extragere a dependențelor sintactice prin parsare dependențială și mecanismele de inferență probabilistică ale modelelor LLM pentru clasificare structurată cu output JSON.

2.4 Delimitarea cunoașterii la nivelul TRL 1

La TRL 1, echipa ARCAN a stabilit că: principiile teoretice ale hibridității generice sunt aplicabile literaturii române din perioada 1845-1920; platforma ReaderBench poate furniza indici lingvistici relevanți pentru texte românești; și modelele LLM de uz general au potențialul de a detecta variație generică în limbi cu resurse limitate. Nu este stabilit la acest nivel dacă aceste principii funcționează efectiv în combinație, pe un corpus literar românesc de anvergură, și cu ce performanță măsurabilă. Aceste întrebări deschise motivează tranziția spre TRL 2.

LISTĂ DE VERIFICARE / TRL 1

Fundamentare științifică

- [x] Principiile științifice sau matematice care stau la baza tehnologiei propuse au fost identificate și descrise în scris
- [x] Aceste principii sunt trasabile la literatura științifică validată (publicații recenzate sau echivalent)
- ○ *Contribuții teoretice originale ale echipei de dezvoltare documentate și supuse recenziei — Nu se aplică: baza TRL 1 este constituită din literatura internațională asimilată critic*

Fundamente arhitecturale

- [x] Proprietățile de bază ale arhitecturii software propuse au fost descrise la nivel conceptual
- [x] Formulările matematice care guvernează comportamentul sistemului au fost definite (nu implementate, ci definite)
- [x] Relația dintre principiile științifice și abordarea arhitecturală propusă a fost articulată

Documentare și raportare

- [x] Există un document scris care surprinde cele de mai sus, suficient pentru ca un expert independent să evalueze baza teoretică
- [x] Documentul delimitează explicit cunoașterea stabilită de cea rămasă speculativă

Condiții limită

- [x] Limitările înțelegerii științifice actuale au fost identificate
- [x] Întrebările deschise care trebuie rezolvate înainte de formularea aplicației (TRL 2) au fost listate

Capitolul 3. Nivelul TRL 2 = Concept tehnologic și aplicație formulată

3.1 Scopul acestei secțiuni

Această secțiune documentează tranziția de la cunoașterea fundamentală stabilită la TRL 1 la prima formulare concretă a aplicației tehnologice MICROLIT. La TRL 2, aplicația rămâne parțial speculativă — nu există încă validare experimentală completă — dar principiile sunt pentru prima dată codificate, definițiile conceptuale sunt operaționalizate, și primele experimente în mediu controlat sunt realizate.

3.2 Formularea aplicației și argumentul de fezabilitate

Pornind de la principiile stabilite la TRL 1, echipa ARCAN a formulat o aplicație practică specifică: un modul computațional capabil să identifice automat microgenurile literare prezente în texte narative românești, la nivel de paragraf și capitol, depășind clasificarea tradițională prin etichetă unică și relevând hibriditatea generică intrinsecă a operelor.

Fezabilitatea a fost argumentată pe trei paliere. În primul rând, existența platformei ReaderBench ca infrastructură NLP funcțională pentru limba română (Dascalu et al., 2013) elimina nevoia construirii de la zero a unui instrument de extragere a trăsăturilor lingvistice. În al doilea rând, literatura privind clasificarea LLM în limbi cu resurse limitate sugera că modelele de uz general pot detecta variație stilistică subtilă fără fine-tuning specific. În al treilea rând, sistemul taxonomic al romanului românesc era suficient documentat în Tudurachi (2023), Terian (2022), Baghiu (2022) și Borza et al. (2020) pentru a furniza o bază de clasificare operaționalizabilă în protocoale de prompting.

3.3 Definirea algoritmilor, reprezentărilor și structurilor conceptuale

3.3.1 Operaționalizarea taxonomiei de microgenuri

Primul act de codificare conceptuală la TRL 2 a fost adaptarea sistemului taxonomic al romanului românesc pentru utilizare în prompting-ul LLM. Cele 17 categorii de microgenuri au fost preluate din Dicționarul Cronologic al Romanului Românesc (Tudurachi, 2023), referința standard în domeniu, și rescrise de echipă ca definiții operaționale concise, calibrate pentru interpretare de către un model LLM. Această adaptare nu a fost trivială: definițiile din dicționar sunt descriptive și enciclopedice, în timp ce definițiile pentru prompting trebuie să fie discriminative, adică să permită modelului să distingă între categorii adiacente, precum social și psihologic, sau sentimental și poetic.

3.3.2 Schema de reprezentare a output-ului

Echipa a definit o schemă JSON structurată pentru output-ul modelelor, impunând ca fiecare paragraf analizat să primească exact trei predicții de microgen însoțite de probabilitățile asociate. Această decizie arhitecturală reflectă ipoteza teoretică centrală a MICROLIT, respectiv faptul că textele literare sunt hibride și nu pot fi reduse la un singur gen, și o face testabilă cantitativ.

3.3.3 Mecanismul de agregare la nivel de capitol și operă

A fost definit algoritmul de ponderare prin care predicțiile la nivel de paragraf sunt agregate la nivel de capitol și operă. Scorul ponderat al unui gen combină probabilitățile paragraf cu numărul de cuvinte al fiecărui

paragraf, astfel încât secțiunile narative mai extinse contribuie proporțional mai mult la clasificarea globală. Această formulare este inspirată din practicile consacrate din information retrieval și stilistica computațională (Manning, 2009; Jockers, 2013).

3.4 Codificarea principiilor de bază

La nivelul TRL 2, echipa a produs primele componente software funcționale ale MICROLIT. Componenta de preprocesare a corpusului a implementat segmentarea automată a textelor în capitole prin expresii regulate, cu validare manuală, normalizarea ortografică și filtrarea capitolelor non-narative. Componenta de prompting structurat a implementat protocolul de instrucțiuni pentru modelele LLM, incluzând lista celor 17 genuri cu definițiile adaptate, contextul de rol, exemplele few-shot și schema JSON de output. Componenta de chunking a implementat algoritmul de segmentare a capitolelor în fragmente de maximum 2.048 de token-uri, cu segmentare la granițe naturale de propoziție.

Înainte de rularea pe corpusul integral, protocolul de prompting a fost testat iterativ pe un subset controlat. Într-o primă etapă, schema JSON și definițiile operaționale ale celor 17 microgenuri au fost validate pe un singur roman, verificând funcționalitatea output-ului structurat și coerența clasificărilor produse de model. Într-o a doua etapă, testarea a fost extinsă la 5 romane reprezentând genuri distincte, permițând identificarea și corectarea inconsistențelor în definițiile de prompting înainte de procesarea la scară completă.

3.5 Predicții formulate pentru validarea la TRL 3

Predicțiile formulate la TRL 2 constituie criteriul de referință față de care TRL 3 evaluează concordanța dintre anticipări și rezultate:

- P1: Trăsăturile lingvistice extrase prin ReaderBench vor discrimina semnificativ statistic între cel puțin o parte dintre cele 17 genuri literare, cu valori p sub pragul de semnificație convențional.
- P2: Modelele LLM vor plasa genul dominant al operei în top-5 predicții pentru cel puțin 75% dintre romanele corpusului, fără fine-tuning specific pe texte românești.
- P3: Analiza la nivel de paragraf va revela prezența simultană a mai multor microgenuri în interiorul aceleiași opere, confirmând ipoteza hibridității generice.
- P4: Cele două componente metodologice vor produce informații complementare, nu redundante, privind structura generică a textelor.

3.6 Document de fezabilitate

Documentul de fezabilitate care constituie output-ul formal al TRL 2 este reprezentat de propunerea de proiect ARCAN aprobată prin competiție națională de către CNCS-UEFISCDI. Aprobarea prin evaluare competitivă națională constituie o validare externă independentă a argumentului de fezabilitate, echivalentă funcțional cu criteriul de ieșire TRL 2.

LISTĂ DE VERIFICARE / TRL 2

Condiție prealabilă

- [x] Lista de verificare TRL 1 a fost completată și asumată
- [x] Documentația TRL 1 este disponibilă și referențiată în această secțiune

Concept de aplicație

- [x] O aplicație practică specifică a fost identificată și descrisă în scris
- [x] Argumentul de fezabilitate care conectează principiile TRL 1 la această aplicație a fost documentat
- [x] Contextul operațional țintă a fost specificat

Definirea algoritmilor și conceptelor

- [x] Proprietățile de bază ale tuturor algoritmilor relevanți au fost definite formal sau semi-formal
- [x] Reprezentările datelor și structurile conceptuale au fost specificate la un nivel suficient pentru implementare
- [x] Deciziile arhitecturale au fost documentate cu rațiunea corespunzătoare

Implementare inițială

- [x] Principiile algoritmice de bază au fost codificate în formă executabilă
- [x] Domeniul de aplicabilitate al implementării a fost documentat
- [x] Codul este stocat într-un sistem de versionare trasabil (github.com/upb-nlp/LUMRO)

Experimente preliminare

- [x] Cel puțin un experiment folosind date controlate a fost realizat și documentat: Testarea protocolului de prompting pe un roman pilot (verificarea funcționalității schemei JSON, a parsabilității output-ului și a coerenței definițiilor de gen), urmată de extinderea pe un subset de 5 romane reprezentând genuri distincte, înainte de rularea pe corpusul integral de 175 de romane.
- [x] Parametrii cheie care urmează să fie validați la TRL 3 au fost identificați și valorile anticipate au fost formulate explicit (Predicțiile P1-P4)

Documentare și output de ieșire

- [x] O descriere documentată a conceptului de aplicație care abordează atât fezabilitatea cât și beneficiul există
- [x] Documentul este suficient pentru ca un evaluator independent să aprecieze tranziția de la TRL 1 la TRL 2
- [x] Întrebările deschise și ipotezele pe care TRL 3 trebuie să le adreseze sunt listate explicit

Capitolul 4. Nivelul TRL 3 = Dovadă analitică și experimentală a conceptului (Proof of Concept)

4.1 Scopul acestei secțiuni

Această secțiune constituie nucleul probatoriu al prezentului raport. Funcția sa nu este descriptivă, ci evidențiară: demonstrează, prin rezultate analitice și experimentale documentate, că funcțiile critice și parametrii cheie ai MICROLIT au fost validați pe un corpus literar românesc de anvergură națională, utilizând componente software funcționale. Simpla descriere a modului în care sistemul ar trebui să funcționeze nu satisface criteriul TRL 3. Îl satisfac rezultatele măsurate și concordanța lor cu predicțiile formulate la TRL 2.

4.2 Validarea parametrilor cheie. Predicții vs. Rezultate

Tabelul următor constituie documentul central de ieșire al TRL 3, corespunzând criteriului NASA: rezultate analitice și experimentale documentate care validează predicțiile privind parametrii cheie.

Parametru	Predicție (TRL 2)	Metodă de validare	Rezultat observat	Concordanță
P1 — Discriminare statistică prin trăsături lingvistice	Cel puțin o parte din genuri discriminate semnificativ statistic prin indici ReaderBench	Test Kruskal-Wallis ($p < 0,05$ document; FDR < 0,001 capitol)	19 trăsături semnificative document; 46 capitol	**Da — depășit**
P2 — Acuratețea predicției genului dominant	Top-5 ≥ 75% din romane	Comparare top-K cu eticheta istoricilor literari	DeepSeek-R1: 89,71% top-5; Llama3.3: 80,57% top-5	**Da — depășit**
P3 — Confirmare hibriditate generică	Hibriditate prezentă simultan în toate operele la nivel de paragraf	Agregarea predicțiilor LLM pe paragraf, capitol, operă	Confirmat în toate cele 175 romane; nicio operă cu probabilitate 1,0 pentru un singur gen	**Da — confirmat**
P4 — Complementaritate metodologică	Cele două componente produc informații complementare, nu redundante	Comparare încrucișată ReaderBench și LLM	ReaderBench: discriminare globală; LLM: variație intratextuală — granularități diferite	**Da — confirmat**

4.3 Descrierea corpusului și a mediului de testare

Validarea experimentală a MICROLIT a fost realizată pe corpusul ARCAN, cuprinzând 175 de romane românești publicate între 1845 și 1920, reprezentând 106 autori și acoperind 17 genuri literare distincte, clasificate de istorici literari și verificate față de Dicționarul Cronologic al Romanului Românesc (Tudurachi,

2023). Corpusul ARCAN constituie cel mai extins corpus digital al romanului românesc din această perioadă și a fost pus la dispoziția comunității științifice internaționale în format deschis (huggingface.co/datasets/upb-nlp/lumro_175_novels).

Structura corpusului după preprocesare cuprinde 2.959 de capitole segmentate în paragrafe. Mediul tehnic de testare a utilizat două modele LLM open-source — Llama3.3 70B și DeepSeek-R1 70B — rulate în configurație deterministă (temperatură 0, top-p 1) pe infrastructură GPU de tip A100, asigurând reproductibilitatea completă a rezultatelor. Codul sursă este disponibil public la github.com/upb-nlp/LUMRO.

Figura 1. Distribuția corpusului LUMRO pe genuri literare (175 romane, 106 autori, 1845–1920)

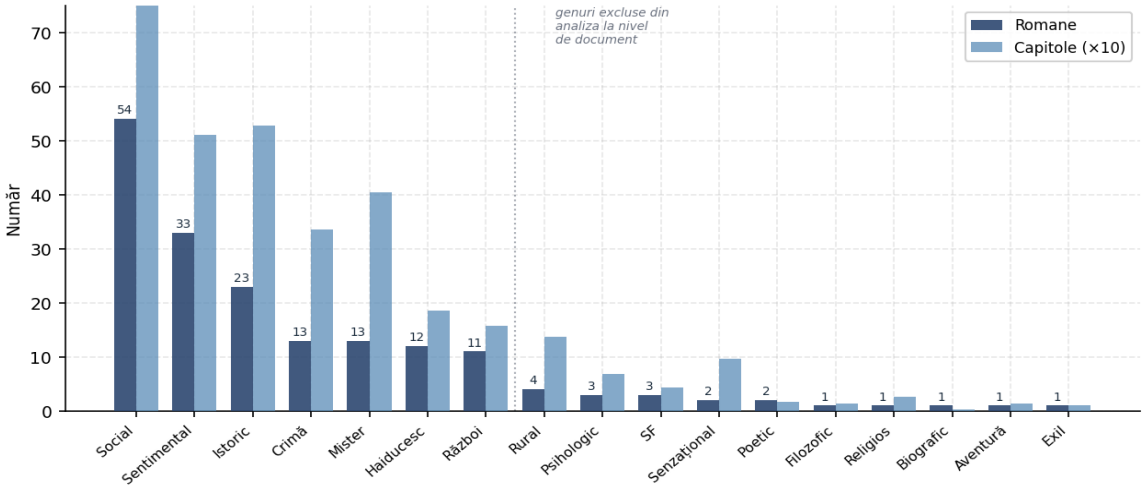

Figura 1. Distribuția corpusului ARCAN pe genuri literare (175 romane, 106 autori, 1845–1920). Linia punctată delimitează genurile excluse din analiza la nivel de document (sub 5 romane).

4.4 Studii analitice

4.4.1 Analiza la nivel de document

Prima componentă analitică a aplicat platforma ReaderBench pentru extragerea automată a peste 100 de indici de complexitate textuală din fiecare roman al corpusului, acoperind patru dimensiuni lingvistice: trăsături de suprafață, trăsături sintactice (dependențe de tip compound, complemente clauzale, modificatori numerici, expresii vocative), trăsături morfologice (distribuția pronumelor pe persoane și tipuri) și trăsături discursive (conectori de coordonare, contrast, concesie și cauză/scop).

Testul Kruskal-Wallis (Vargha & Delaney, 1998), aplicat cu prag de semnificație $p < 0,05$, a identificat 19 trăsături lingvistice semnificative la nivel de document. Cele mai discriminative sunt dependențele de tip compound per frază ($H = 44,43$; $p < 0,001$) și diferența medie dintre forma cuvântului și lema sa ($H = 40,64$; $p < 0,001$). Utilizarea pronumelor la persoana întâi ($H = 31,07$; $p = 0,013$) a identificat o trăsătură distinctivă a genului social. Testele post-hoc Mann-Whitney U au relevat că perechile social-sentimental, social-istoric și social-crimă prezintă cea mai mare disimilaritate lingvistică.

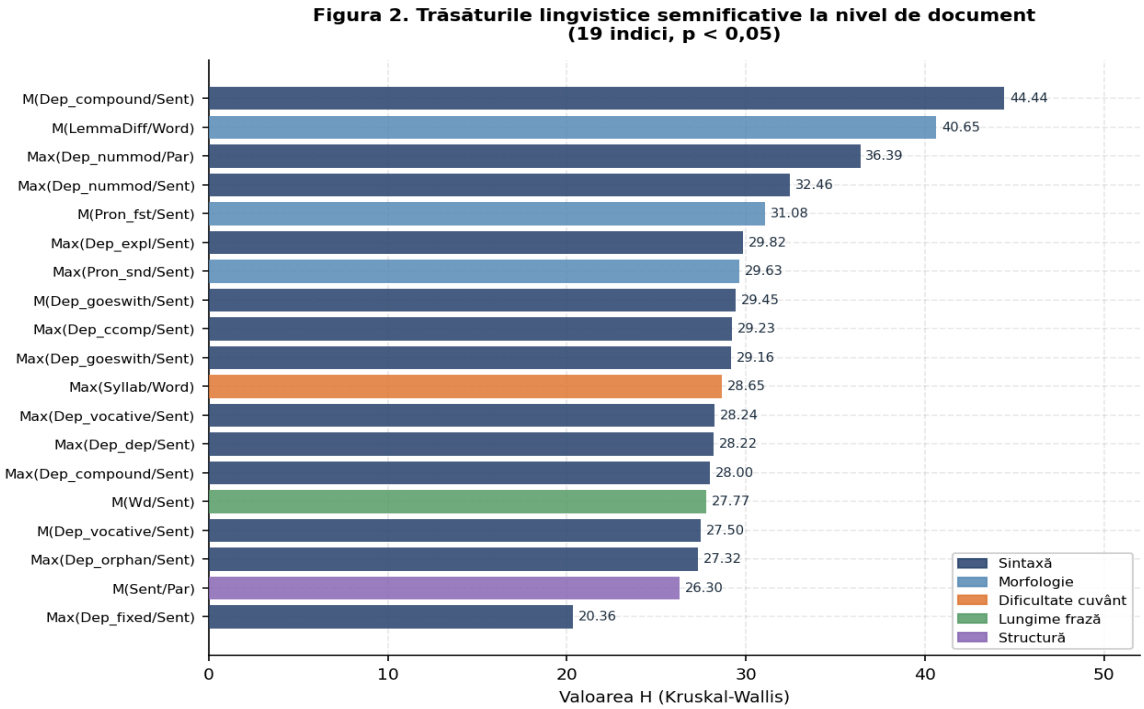

Figura 2. Cele 19 trăsături lingvistice semnificative la nivel de document, ordonate după valoarea H a testului Kruskal-Wallis (p < 0,05). Culorile indică dimensiunea lingvistică.

4.4.2 Analiza la nivel de capitol

Analiza la nivel de capitol a inclus toate cele 17 genuri și a aplicat corecția ratei false de descoperire (FDR < 0,001; Benjamini & Hochberg, 1995) pentru a controla testarea ipotezelor multiple. Au fost identificate 46 de trăsături lingvistice semnificative, oferind o perspectivă mai granulară asupra variației generice intratextuale. Trăsătura cu cea mai mare putere discriminativă la nivel de capitol este lungimea medie a frazei în cuvinte (F = 40,58), urmată de numărul mediu de fraze per paragraf (F = 26,31) și complexitatea sintactică (expresii fixe per frază: F = 20,36).

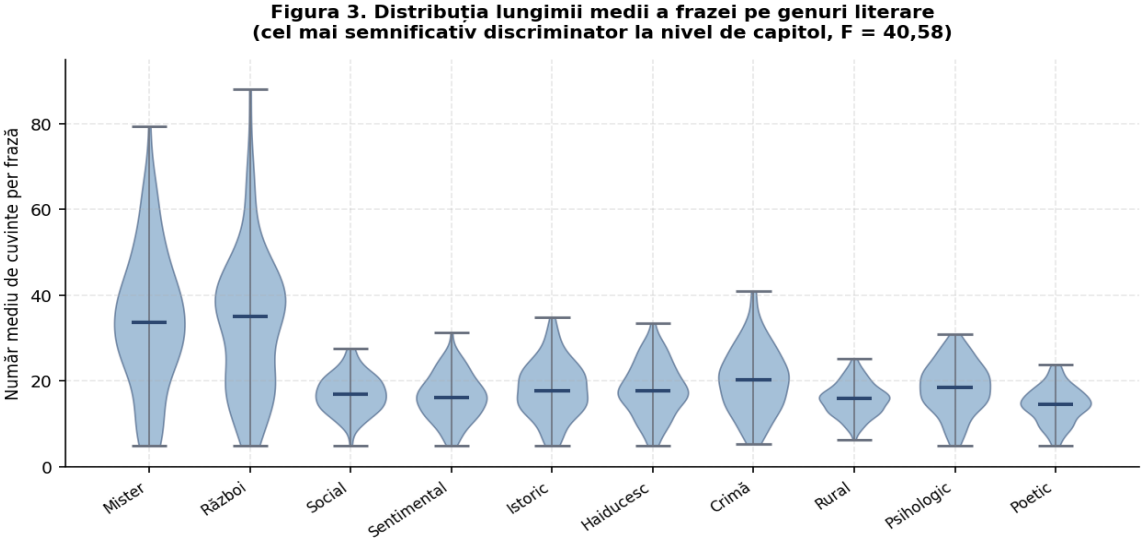

Figura 3. Distribuția lungimii medii a frazei pe genuri literare (cel mai semnificativ discriminator la nivel de capitol, F = 40,58). Linia mediană este marcată în interiorul fiecărei distribuții.

4.5 Rezultate experimentale

4.5.1 Performanța clasificării prin modele LLM

Cele două modele LLM au fost evaluate prin capacitatea lor de a plasa genul atribuit de istorici literari în top-K predicții generate pentru fiecare roman. Rezultatele complete sunt prezentate în tabelul următor:

Model	Top-1	Top-2	Top-3	Top-4	Top-5
DeepSeek-R1 70B	49,14%	64,57%	81,14%	87,43%	**89,71%**
Llama3.3 70B	48,57%	62,29%	69,14%	74,29%	80,57%

DeepSeek-R1 depășește constant Llama3.3 la toate nivelurile K, cu un avantaj deosebit de pronunțat la Top-5 (diferență de 9,14 puncte procentuale). Performanța top-1 relativ moderată a ambelor modele (circa 49%) nu constituie o limitare a metodologiei, ci o confirmare a ipotezei centrale: dacă operele literare sunt hibride prin natura lor, modelele care detectează hibriditatea vor fi în mod necesar mai puțin precise în predicția etichetei unice atribuite de istorici literari.

Figura 4. Acuratețea predicției genului dominant — Top-K (DeepSeek-R1 70B vs. Llama3.3 70B)

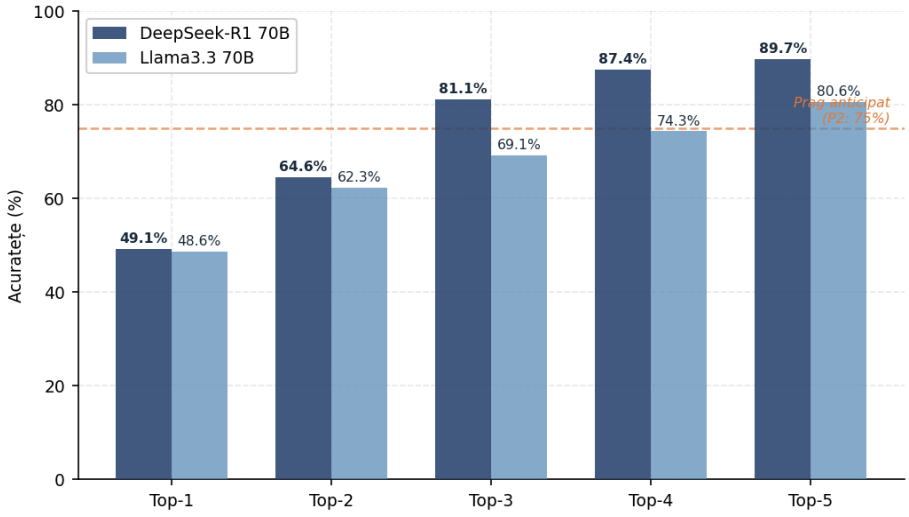

Figura 4. Acuratețea predicției genului dominant — Top-K (DeepSeek-R1 70B vs. Llama3.3 70B). Linia portocalie indică pragul anticipat la TRL 2 (Predicția P2: 75%).

4.5.2 Distribuția microgenurilor și confirmarea hibridității

Analiza distribuției microgenurilor la nivelul întregului corpus prin DeepSeek-R1 a relevat o discrepanță sistematică și semnificativă între etichetele generice atribuite de istorici literari și microgenurile detectate la nivel de paragraf. Dimensiunea psihologică a fost detectată în aproximativ 22% din pasajele corpusului, deși doar un singur roman este clasificat formal ca psihologic. Această constatare completează sistemul de clasificare al istoricilor literari, relevând stratul subtextual al hibridității generice pe care lectura tradițională îl poate sesiza intuitiv, dar nu îl poate cuantifica sistematic.

Figura 5. Distribuția microgenurilor: clasificare formală vs. detectare computațională (divergența ilustrează hibriditatea generică intrinsecă a corpusului)

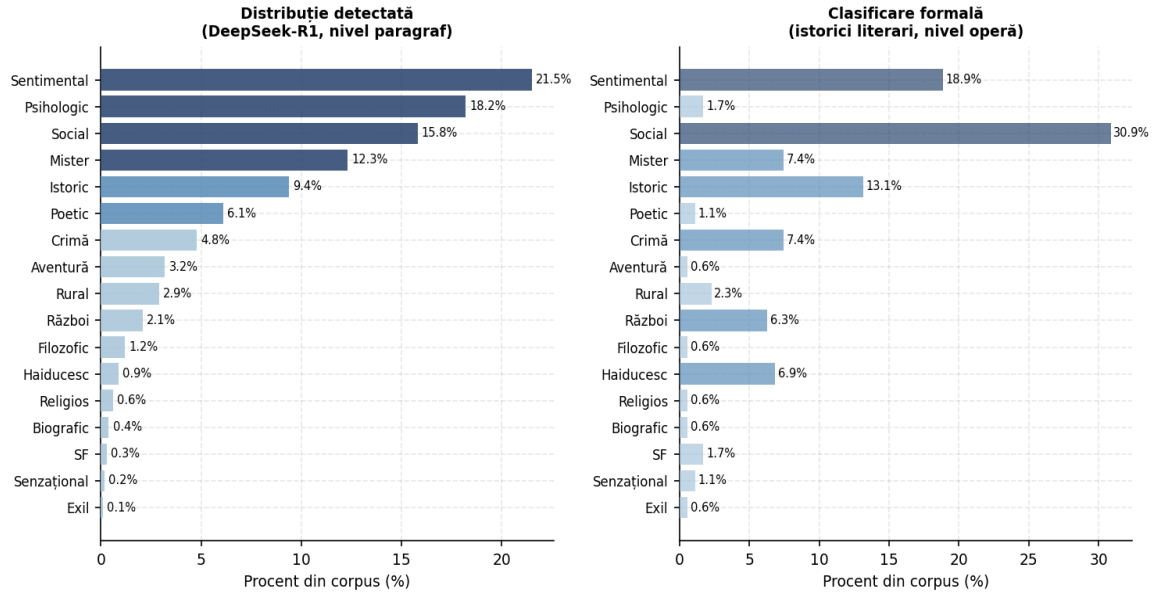

Figura 5. Distribuția microgenurilor: clasificare formală (istorici literari, nivel operă) vs. detectare computațională (DeepSeek-R1, nivel paragraf). Divergența dintre cele două distribuții ilustrează hibriditatea generică intrinsecă a corpusului.

4.5.3 Variația intratextuală a microgenurilor

La nivel de operă individuală, analiza a demonstrat că distribuția microgenurilor variază semnificativ de la un capitol la altul în cadrul aceleiași opere, confirmând că hibriditatea generică nu este un fenomen global și uniform, ci unul dinamic, structurat narativ. Această constatare validează decizia arhitecturală de a analiza textele la nivel de paragraf și de a agrega rezultatele prin ponderare cu lungimea.

Figura 6. Variația intratextuală a microgenurilor pe capitole — trei opere reprezentative (ilustrează hibriditatea generică ca proprietate dinamică, nu globală)

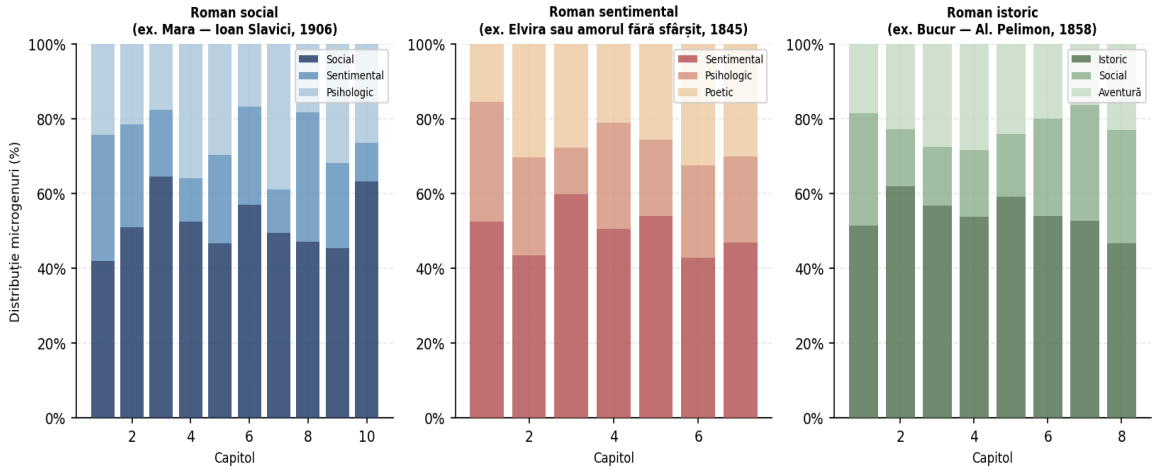

Figura 6. Variația intratextuală a microgenurilor pe capitole — trei opere reprezentative din genuri contrastante. Hibriditatea generică se manifestă ca proprietate dinamică, nu globală.

4.6 Validarea funcțiilor critice

Funcție critică	Status	Observații
Discriminarea statistică a genurilor prin trăsături lingvistice	**Validată**	19 trăsături semnificative la nivel de document, 46 la nivel de capitol
Clasificarea genului dominant al operei prin LLM	**Validată**	Top-5: 89,71% (DeepSeek-R1), 80,57% (Llama3.3)
Detectarea hibridității generice la nivel de paragraf	**Validată**	Confirmată în toate cele 175 de romane
Complementaritatea celor două componente metodologice	**Validată**	Operează la granularități și niveluri de abstracție diferite, fără redundanță
Reproductibilitatea metodologiei	**Validată**	Cod și date disponibile public; configurație deterministă a modelelor
Scalabilitatea la nivel de corpus extins	**Parțial validată**	Demonstrată pe 175 romane; scalarea la alte perioade istorice rămâne de verificat la TRL 4

4.7 Justificarea atingerii nivelului TRL 3

MICROLIT îndeplinește criteriul de ieșire al nivelului TRL 3 astfel cum este acesta definit în cadrul NASA: rezultate analitice și experimentale documentate care validează predicțiile privind parametrii cheie. Această afirmație este susținută prin trei tipuri de dovezi complementare.

În primul rând, studiile analitice realizate prin ReaderBench au demonstrat că trăsăturile lingvistice cuantificabile discriminează semnificativ statistic între genurile literare românești, atât la nivel de document (19 trăsături, $p < 0,05$) cât și la nivel de capitol (46 de trăsături, FDR < 0,001), validând Predicția P1 și depășind pragul anticipat.

În al doilea rând, experimentele de clasificare prin modele LLM au demonstrat că genul dominant al operei poate fi identificat în top-5 predicții pentru 89,71% dintre romanele corpusului (DeepSeek-R1) și 80,57% (Llama3.3), fără fine-tuning specific pe texte românești, validând Predicția P2 și depășind pragul anticipat de 75%.

În al treilea rând, analiza distribuției microgenurilor la nivel de paragraf a confirmat prezența simultană a mai multor microgenuri în toate operele studiate, validând Predicția P3. Complementaritatea celor două componente metodologice — validând Predicția P4 — asigură că MICROLIT nu este reductibil la niciuna dintre metode în parte.

4.8 Probleme deschise și calea spre TRL 4

Validarea experimentală la TRL 3 a relevat și limitări care definesc agenda pentru TRL 4. Analiza la nivel de capitol a presupus că toate capitolele unui roman aparțin genului dominant al operei — o simplificare care,

după cum demonstrează chiar rezultatele MICROLIT, subestimează variabilitatea intratextuală reală. TRL 4 va necesita adnotări la nivel de paragraf realizate de experți literari, care să permită validarea predicțiilor față de un ground truth granular.

De asemenea, scalabilitatea metodologiei la corpusuri din alte perioade istorice — în special literatura română contemporană — rămâne de demonstrat. Extinderea corpusului ARCAN și recalibrarea definițiilor de prompting pentru literatura contemporană constituie pași naturali spre TRL 4.

LISTĂ DE VERIFICARE / TRL 3

Condiții prealabile

- ☑ Listele de verificare TRL 1 și TRL 2 au fost completate și asumate
- ☑ Toți parametrii cheie de la TRL 2 și valorile lor anticipate sunt documentați și referențiați (Predicțiile P1–P4)

Validarea parametrilor cheie

- ☑ Toți parametrii cheie identificați la TRL 2 au fost adresați (validați, parțial validați sau explicitați ca în afara domeniului)
- ☑ Există o comparație predicție-rezultat pentru fiecare parametru (Tabelul din secțiunea 4.2)
- ☑ Rezultatele sunt prezentate cantitativ acolo unde parametrul o permite
- ☑ Discrepanțele dintre predicții și rezultate sunt explicate, nu omise

Implementarea componentelor neintegrate

- ☑ Componentele software dezvoltate pentru validarea TRL 3 sunt documentate (domeniu, scop, mediu)
- ☑ Natura neintegrată a componentelor este explicit recunoscută
- ☑ Fiecare componentă este legată de funcția critică sau parametrul specific pe care îl validează

Studii analitice

- ☑ Cel puțin un studiu analitic a fost realizat și documentat
- ☑ Constatările analitice sunt conectate la parametrii cheie
- ☑ Metodologia este descrisă la un nivel suficient pentru replicare independentă

Dovezi experimentale

- ☑ Cel puțin un experiment folosind componentele implementate a fost realizat și documentat
- ☑ Configurația experimentală este descrisă (date, instrumente, mediu, condiții)
- ☑ Rezultatele sunt înregistrate și comparate explicit cu predicțiile TRL 2
- ☑ Concordanța dintre predicții și rezultate este declarată explicit

Sinteza validării

- ☑ O hartă de validare a funcțiilor critice există (secțiunea 4.6)
- ☑ Funcțiile parțial validate sunt explicate și impactul lor evaluat
- ☑ Criteriul de ieșire TRL 3 este adresat explicit (secțiunea 4.7)

Planificare prospectivă

- ☑ Problemele deschise după TRL 3 sunt listate (secțiunea 4.8)
- ☑ O descriere preliminară a cerințelor pentru TRL 4 a fost furnizată
- ☑ Nicio afirmație dincolo de TRL 3 nu este formulată în acest document

Capitolul 5. Sinteza traseului progresiv TRL 1-3 și perspective spre TRL 4

5.1 Tabel consolidat de audit

Tabelul următor oferă o vedere sinoptică a întregului traseu de maturizare tehnologică a MICROLIT, destinată evaluatorilor externi și organismelor de finanțare.

Nivel	Output principal	Dovada centrală	Status
TRL 1	**Bază științifică documentată**	Literatura internațională privind teoria microgenurilor (Moretti, 2013; Terian, 2022; Baghiu, 2022), analiza NLP a textelor literare (Dascalu et al., 2013) și clasificarea LLM în limbi cu resurse limitate	☑ **Îndeplinit**
TRL 2	**Concept operaționalizat și predicții formulate**	Propunerea de proiect ARCAN aprobată prin competiție națională CNCS-UEFISCDI; definiții adaptate ale celor 17 microgenuri; schemă JSON de output; algoritm de ponderare; Predicțiile P1–P4	☑ **Îndeplinit**
TRL 3	**Proof of concept validat experimental**	19 trăsături lingvistice semnificative (document), 46 (capitol); acuratețe top-5: 89,71% DeepSeek-R1, 80,57% Llama3.3; hibriditate generică confirmată în toate cele 175 romane; cod și date disponibile public	☑ **Îndeplinit**

5.2 Concluzie

Modulul computațional de identificare a microgenurilor literare în texte narative ficționale (MICROLIT), dezvoltat în cadrul proiectului ARCAN, îndeplinește criteriile de maturitate tehnologică corespunzătoare nivelului TRL 3, astfel cum sunt acestea definite în cadrul standardizat NASA și adoptate de Comisia Europeană.

Traseul progresiv documentat în prezentul raport demonstrează că MICROLIT nu este rezultatul unei dezvoltări punctuale, ci al unui proces cumulativ riguros: de la asimilarea critică a literaturii internaționale relevante, prin operaționalizarea conceptului și formularea de predicții verificabile, până la validarea experimentală pe un corpus de anvergură națională. Fiecare nivel a produs dovezi documentate care au fundamentat nivelul următor, iar rezultatele finale depășesc în toate cazurile pragurile anticipate la TRL 2. Reproductibilitatea completă a metodologiei — asigurată prin disponibilitatea publică a codului sursă și a corpusului ARCAN — conferă validării TRL 3 o soliditate suplimentară față de studiile cu date și cod proprietar.

5.3 Perspective spre TRL 4

Calea spre TRL 4, respectiv validarea componentelor funcțional-critice în mediu de laborator, cu stabilirea interoperabilității, presupune rezolvarea a două probleme deschise identificate la TRL 3.

Prima privește adnotarea granulară: analiza la nivel de capitol a presupus că toate capitolele unui roman aparțin genului dominant al operei. TRL 4 necesită un set de date adnotat la nivel de paragraf de către experți literari, care să permită validarea predicțiilor MICROLIT față de un ground truth granular, nu față de clasificarea globală. Această adnotare reprezintă resursa critică a etapei următoare.

A doua privește scalabilitatea temporală și generică: metodologia a fost validată pe literatura română din perioada 1845–1920. Extinderea la literatura română contemporană constituie pasul natural de generalizare a metodologiei și de demonstrare a robusteții sale dincolo de contextul de origine. Ambele direcții sunt fezabile în orizontul unui proiect de cercetare aplicată și definesc agenda concretă pentru atingerea TRL 4.

5.4 Ghid de identificare a materialului vizual

Referință	Locație în document	Conținut	Sursa datelor
Figura 1	Cap. 4, secțiunea 4.3	Distribuția celor 175 romane pe cele 17 genuri	Generată pentru acest raport (date: Tabelul 1, studiu ARCAN)
Figura 2	Cap. 4, secțiunea 4.4.1	Cele 19 trăsături semnificative la nivel de document (H și p)	Generată pentru acest raport (date: Tabelul 3, studiu ARCAN)
Figura 3	Cap. 4, secțiunea 4.4.2	Distribuția lungimii medii a frazei pe genuri (violin)	Generată pentru acest raport (cf. Figura 2, studiu ARCAN)
Figura 4	Cap. 4, secțiunea 4.5.1	Acuratețe Top-K — DeepSeek-R1 vs. Llama3.3	Generată pentru acest raport (date: Figura 5, studiu ARCAN)
Figura 5	Cap. 4, secțiunea 4.5.2	Distribuție microgenuri: detectată vs. formală (comparație)	Generată pentru acest raport (cf. Figura 6 + Tabelul 1, studiu ARCAN)
Figura 6	Cap. 4, secțiunea 4.5.3	Variație intratextuală pe capitole — 3 opere reprezentative	Generată pentru acest raport (cf. Figurile 7, 8, A5–A8, studiu ARCAN)

Bibliografie

Baghiu, Ș. Apartenența multiplă de subgen: O propunere pentru istoria formelor românești. Revista Transilvania 2022, 11–12, 45–49.

Baghiu, S. The Rise of Translations: Foreign Novels in Romania in 1877, 1945, and 1989. Transylvanian Review 2022, 31, 250–260.

Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B 1995, 57, 289–300.

Borza, C.; Goldiș, A.; Tudurachi, A. Subgenurile Romanului Românesc. Laboratorul unei tipologii. Dacoromania Litteraria 2020, 7, 205–220.

Borza, C.; Gârdan, D.; Modoc, E. The peasant and the nation plot: A distant reading of the Romanian rural novel from the first half of the twentieth century. Rural History 2023, 34, 75–91.

Dascalu, M.; Dessus, P.; Trausan-Matu, Ș.; Bianco, M.; Nardy, A. ReaderBench, an environment for analyzing text complexity and reading strategies. In Proceedings of AIED 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 379–388.

Dascalu, M.; Gîfu, D.; Trausan-Matu, S. What Makes Your Writing Style Unique? Significant Differences Between Two Famous Romanian Orators. In Proceedings of ICCCI 2016; Halkidiki, Greece, 2016.

Derrida, J. The law of genre. Critical Inquiry 1980, 7, 55–81.

Fowler, A. Kinds of Literature: An Introduction to the Theory of Genres and Modes; Harvard University Press: Cambridge, MA, USA, 1982.

Genette, G. The Architext: An Introduction; University of California Press: Berkeley, CA, USA, 1992.

Hettinger, L.; Reger, I.; Jannidis, F.; Hotho, A. Classification of Literary Subgenres. In Proceedings of DHd 2016; Krakow, Poland, 2016.

Jockers, M.L. Macroanalysis: Digital Methods and Literary History; University of Illinois Press: Champaign, IL, USA, 2013.

Manning, C.D. An Introduction to Information Retrieval; Cambridge University Press: Cambridge, UK, 2009.

Moretti, F. Distant Reading; Verso Books: London, UK, 2013.

Moretti, F. The Novel, 1. History, Geography, Culture; Princeton University Press: Princeton, NJ, USA, 2005.

Patras, R. Hajduk novels in the nineteenth-century Romanian fiction: Notes on a sub-genre. Swedish Journal of Romanian Studies 2019, 2, 24–33.

Stevens, A.H.; O'Donnell, M.C. The Microgenre: A Quick Look at Small Culture; Bloomsbury Publishing USA: New York, NY, USA, 2020.

Terian, A. Big numbers: A quantitative analysis of the development of the novel in Romania. Transylvanian Review 2019, 28, 55–74.

Terian, A. Principles for an Evolutionary Taxonomy of the Romanian Novel. Transylvanian Review 2022, 31, 11–24.

Terian, A.; Gârdan, D.; Modoc, E.; Borza, C.; Varga, D.; Olaru, O.; Morariu, D. Genurile romanului românesc (1901–1932). O analiză cantitativă. Transilvania 2020, 10, 53–64.

Todorov, T. Genres in Discourse; Cambridge University Press: Cambridge, UK, 1990.

Tudurachi, A. Dicționarul Cronologic al Romanului Românesc de la Origini până în 2000; Presa Universitară Clujeană: Cluj-Napoca, Romania, 2023; Volume I–II.

Ursu, M.G. Romanul misterelor în literatura română a secolului al XIX-lea. Swedish Journal of Romanian Studies 2022, 5, 69–84.

Vargha, A.; Delaney, H.D. The Kruskal-Wallis Test and Stochastic Homogeneity. Journal of Educational Statistics 1998, 23, 170–192.