DIACRONIA
Impavidi progrediamur!

ARTICOL

doi:10.17684/i14A196ro
ISSN: 2393-1140
www.diacronia.ro

Digitalizarea scrierilor cu alfabet chirilic (românesc) prin utilizarea platformei Transkribus: noi perspective

Constanța Burlacu [1]*, Achim Rabus [2]

[1] *Merton College, Universitatea din Oxford, Merton St., OX1 4JD Oxford, Marea Britanie*
[2] *Departmentul de Lingvistică Slavă, Universitatea „Albert Ludwig", Werthmannstr. 14, 79085 Freiburg, Germania*

Despre articol

Istoric:
Primit 17 septembrie 2021
Acceptat 26 septembrie 2021
Publicat 12 decembrie 2021

Cuvinte-cheie:
Transkribus
recunoașterea textului din manuscrise
chirilica românească
științe umaniste digitale

Rezumat

Prezentul articol aduce în discuție aplicarea platformei software Transkribus (*transkribus.eu*), un instrument de inteligență artificială pentru recunoașterea scrisului de mînă (HTR), la manuscrise românești din secolul al XVI-lea și la surse tipărite redactate cu alfabet chirilic. După o trecere în revistă a funcționalității de bază a tehnologiei HTR și a programului Transkribus, discutăm despre sursele românești și cele bilingve slavo-române pe care le-am utilizat, detaliind modelele de instruire specifice și generice, precum și pe cele inteligente (respectiv transliterarea din sistemul de scriere chirilic în cel latin), evaluînd performanța acestora și discutînd apoi implicațiile HTR asupra cercetării filologice în epoca digitală. În concluzii ne referim perspectivele viitoare de cercetare.

1. Introducere: ce este Transkribus și cum funcționează?

Transkribus (*transkribus.eu*, Muehlberger *et al.*, 2019) este un instrument de inteligență artificială care poate fi antrenat pentru transcrierea automată a manuscriselor și a tipăriturilor vechi, redactate într-o varietate de limbi, stiluri și sisteme de scriere. Mai mult decît atît, Transkribus servește ca platformă pentru lucrul colaborativ pe surse diverse. Tehnologia de recunoaștere a scrisului de mînă (HTR) pe care o utilizează este semnificativ mai avansată decît tehnologia OCR (recunoașterea optică a caracterelor)[1], din perspectiva faptului că procesul de recunoaștere nu se axează numai pe caractere individuale, ci funcționează la nivelul rîndului, luînd în considerare grafemele învecinate și chiar cuvinte întregi pentru a determina transcrierea cea mai potrivită.

Abordarea bazată pe inteligența artificială pe care se bazează tehnologia HTR aparține așa-numitei învățări supervizate. Acest lucru înseamnă că algoritmul HTR are nevoie de o anumită cantitate de imagini digitale de înaltă calitate a surselor supuse analizei, precum și de transcrierile corectate manual ale acestora. Pe parcursul mai multor etape („epochs")[2], modelul învață trăsăturile lingvistice și paleografice ale sursei analizate. Pentru sursele ușor descifrabile și pentru cele obișnuite, se poate iniția formarea modelului de transcriere pornind de la doar 2000 de cuvinte transcrise. Rezultate satisfăcătoare se obțin de regulă cu aproximativ 10 000 de cuvinte, respectiv modele cu capacități generice, în sensul că sînt capabile să transcrie diferite mîini sau chiar stiluri de scriere de mînă; diverse stiluri de scriere au fost testate pe mai mult de 100 000, uneori chiar pe milioane de cuvinte. În cadrul Transkribus, modelele HTR pot fi partajate colegilor sau pot fi puse la dispoziția publicului.

Măsura principală a calității modelelor HTR este reprezentată de rata de eroare a caracterelor la recunoașterea caracterelor/ Character Error Rate (CER). Modelele bune pentru o anumită sursă ajung la

* Adresă de corespondență: *constanta.burlacu@merton.ox.ac.uk*.

[1] Felix Dietrich compară distanța de la OCR la HTR cu cea de la un algoritm de forță brută, precum cel implementat în Deep Blue, computerul care l-a învins pentru prima oară pe campionul mondial la șah în anul 1997, la mult mai sofisticatul AlphaGo care a învins campionul la Go în 2016 (*readcoop.eu*).

[2] Potrivit *fon.hum.uva.nl*, conceptul de „epoch" reprezintă "o prezentare completă a datelor setate pentru învățare pentru un program de învățare automată".

o valoare CER mai mică de 5%, ceea ce înseamnă că mai puțin de un caracter din 20 (incluzînd aici și semnele de punctuație) este transcris în mod eronat. Modelele utilizabile care permit corectarea manuală a erorilor produse într-un interval de timp mai mic decît cel necesar pentru transcrierea manuală a întregului material de la zero au o valoare a CER sub 10%.

Curba de învățare tipică pentru un model HTR este prezentată în Fig. 1. După cum se poate observa, pe durata primelor aproximativ cinci epoci, rata de eroare a transcrierii caracterelor scade drastic, în timp ce îmbunătățirea pe durata celor rămase (în acest caz aproximativ 50) este redusă.

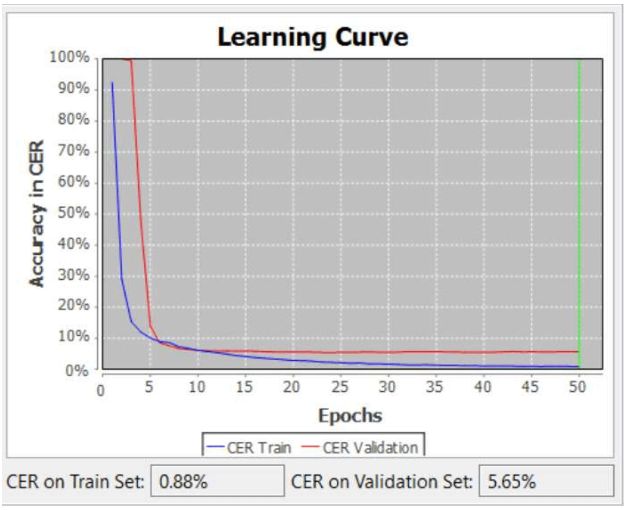

Figura 1: Exemplu de curbă de învățare HTR în Transkribus

Pentru scrierile în sistem chirilic pre-modern, au fost publicate modelele Transkribus publice atît pentru slavonul *ustav*, cît și pentru *poluustav* (Rabus, 2019). Acestea au fost antrenate pe sute de mii de token-uri de cuvinte, ceea ce a condus la capacități generice ale modelelor. Acest lucru înseamnă că, într-o anumită măsură, aceste modele pot transcrie surse redactate în diverse stiluri de scriere, din diverse regiuni și aparținînd unor perioade istorice diferite. Figurile 2–4 exemplifică performanța modelelor generice disponibile public pentru slavona bisericească.

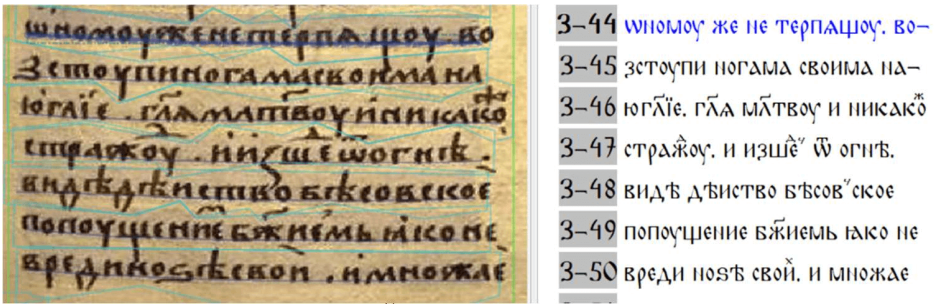

Figura 2: VMČ, SIN 993, aprilie, secolul al XVI-lea

Așa cum se poate observa, în pofida unor diferențe ce țin de calitatea transcrierii, per ansamblu, performanța modelelor publice pentru slavonă este destul de convingătoare. Prin intermediul platformei Transkribus, aceste modele sînt accesibile gratuit oricărui utilizator[3].

Cu toate acestea, datorită faptului că, așa cum am menționat mai sus, modelele HTR învață nu numai trăsături paleografice, ci și lingvistice (de exemplu probabilități de combinare a unor grafeme sau chiar a

[3]Transkribus și toate modelele publice pot fi utilizate gratuit. La înregistrare, utilizatorul primește 500 de credite, suficiente pentru transcrierea a aproximativ 400 de pagini. Utilizatorii care doresc să transcrie un număr mai mare de pagini pot obține credite suplimentare accesând *readcoop.eu*.

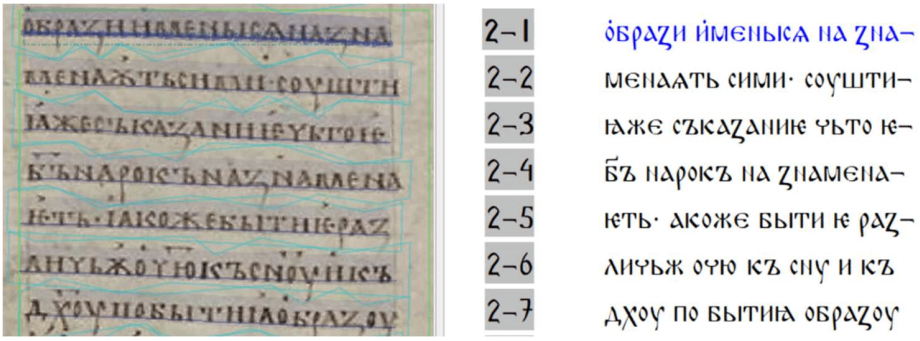

Figura 3: Izbornik Svjatoslava, 1073

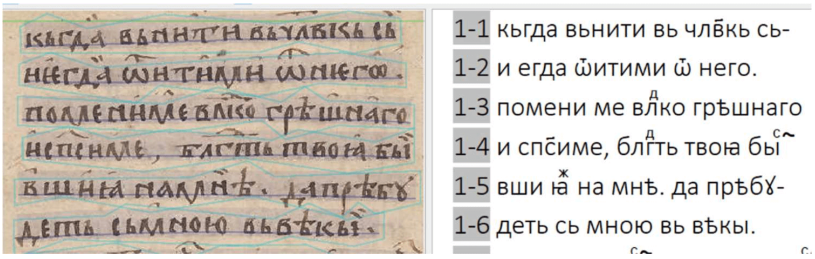

Figura 4: Bdinski Zbornik, secolul al XIV-lea

unor cuvinte), performanța acestor modele pentru scrierea chirilică românească este mediocră sau chiar slabă. Cu alte cuvinte, aceste modele încearcă să detecteze cuvinte sau combinații de grafeme slave în texte românești, iar rezultatele sînt departe de a fi satisfăcătoare. Din acest motiv se evidențiază necesitatea de a antrena modele HTR specifice pentru scrierea chirilică românească (și pentru sursele bilingve).

2. Formarea și evaluarea modelelor pentru scrierea chirilică românească

Modelele pe care le-am dezvoltat pînă în prezent pentru scrierea chirilică românească au fost testate pe manuscrise și materiale tipărite datînd din secolul al XVI-lea, în principal texte din *Faptele Apostolilor* și *Psaltire*. Inițial, ne-am propus să dezvoltăm modele specifice pentru anumite surse, astfel încît primul manuscris analizat a fost *Codicele Voronețean*, pentru care am dezvoltat ulterior două modele (Romanian_Cyrillic_0.01 și 0.02, v. Fig. 5).

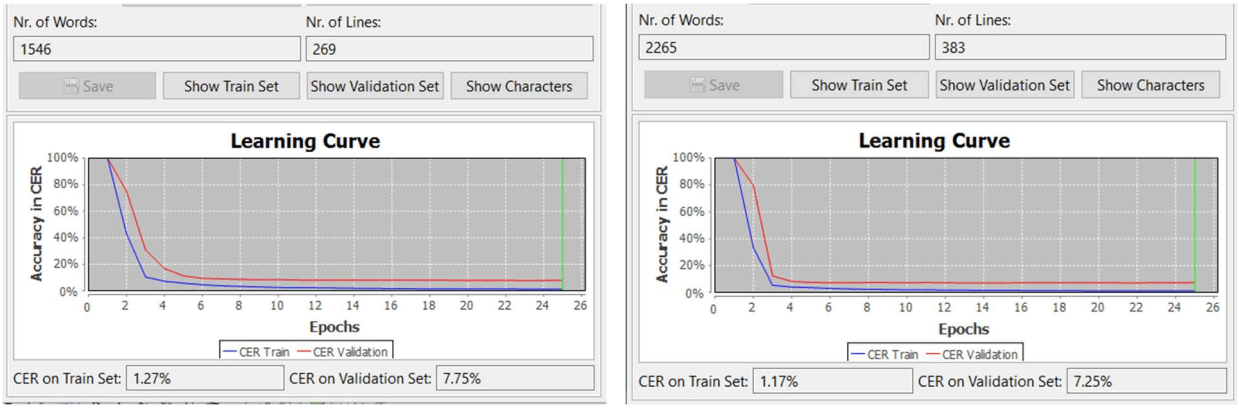

Figura 5: Curbe de învățare pentru Romanian_Cyrillic_0.01 (stînga) și 0.02 (dreapta)

Cu această ocazie au fost transcrise aproximativ 30 de pagini, care se adaugă celor peste 2000 de token-uri de cuvinte care pot fi folosite ca standard de bază pentru generarea unui nou model. Așa cum se observă,

rata de eroare este de 7,25%, care, luînd în considerare volumul redus de date furnizate, indică faptul că modelul poate fi deja utilizat pentru transcriere. Exemplul din Fig. 6 ilustrează eficacitatea modelului.

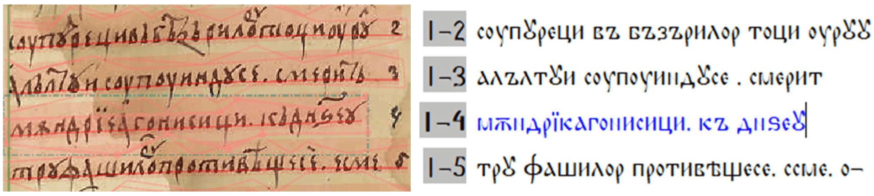

Figura 6: Transcriere din *Codicele Voronețean*, f. 82r, cu Romanian_Cyrillic_0.02

După cum se poate vedea, acest prim model are dificultăți în identificarea suprascrierilor (**соуп8ф̲рецивъ**, **оур8л̲**), a ligaturilor (**бът̲р̲ърилор**), a literelor precum **к** și **є** în **мжндрїє̲** și **є** **смє-**. În plus, există diverse erori în ceea ce privește spațierea, ca de exemplu primul verb transcris **соуп8рецивъ** sau **мжндрїє агонисици** de pe rîndul 4. Cu toate acestea, modelul poate fi utilizat pentru transcrierea altor pagini din aceeași sursă și astfel pentru a îmbunătăți modelul inițial, ceea ce am făcut prin crearea celui de-al treilea model (Romanian_Cyrillic_0.03) adăugînd aproximativ 5000 de de cuvinte, rezultatul fiind o rată de eroare de 5,85%. În această nouă transcriere literele sînt identificate mai bine decît în cadrul modelului anterior, la fel ca și ligatura **тр** și suprascrierile[4], deși modelul nu recunoaște slovele suprascrise de pe ultimul rînd și consideră primele două cuvinte ca fiind unul singur: **тр8фашилѡ҄ противѣщєсє** (Fig. 7).

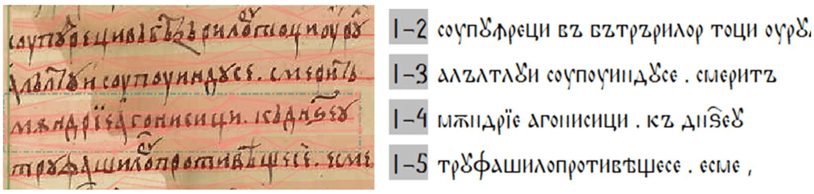

Figura 7: Transcriere din *Codicele Voronețean*, f. 82r, cu Romanian_Cyrillic_0.03

Deși dezvoltarea modelelor de transcriere pentru manuscrise necesită o cantitate mare de date analizate și corectate manual de experți (Ground Truth data), modelele pentru materiale tipărite necesită considerabil mai puține date. La crearea modelului pentru *Faptele Apostolilor*, tipărit de Coresi în 1563, am obținut o rată de eroare de doar 4,46% folosind mai puțin de 4000 de cuvinte pentru antrenare. Transcrierea unei pagini care nu mai fusese procesată anterior (p. 70) din *Faptele Apostolilor* a indicat rezultatele din Fig. 8.

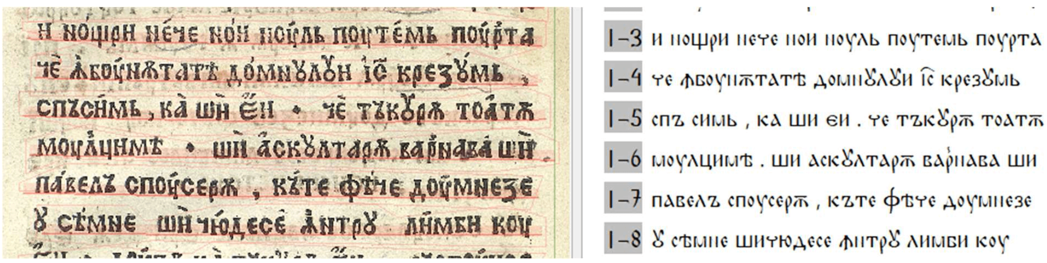

Figura 8: Transcriere din *Faptele Apostolilor* din 1563, p. 70, cu BRV12_Romanian Printing 16th c

Toate literele sînt recunoscute corect, existînd doar erori de spațiere în rîndurile 4 (**ꙟвоунжтатѣ**), 5 (**спъсимь**) și 8 (**шичюдесе**). Particularitatea acestui text tipărit este faptul că el conține foarte puține suprascrieri și abrevieri și respectă regulile moderne de editare în privința spațiilor. Datorită acestor aspecte și

[4]În aceste prime modele am decis să aducem suprascrierile la nivelul rândului, deși nu se întâmplă același lucru și în modelele pe care le-am dezvoltat ulterior. Redarea fidelă a slovelor suprascrise, așa cum se regăsesc în textul original, a fost facilitată de Daniel Bunčić, care ne-a furnizat driverele de tastatură pentru tastarea în slavona bisericească. A se vedea *obshtezhitie.net*.

faptului că, de regulă, literele tipărite au un aspect mult mai regulat decît cele scrise de mînă, procesul HTR pentru acest tip de material este mult mai facil decît în cazul manuscriselor. Dacă încercăm analizarea unui manuscris cu acest model, rezultatul este inutilizabil (Fig. 9).

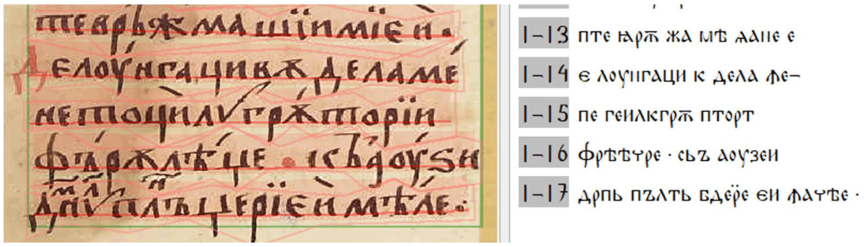

Figura 9: Transcriere din *Psaltirea Scheiană*, f. 7^{r}, cu BRV12_Romanian Printing 16th c

Cu toate acestea, dacă unim cele două modele menționate mai sus, respectiv Romanian_Cyrillic_0.03 și BRV12_Romanian Printing 16th c, putem obține un model combinat care furnizează o transcriere relativ bună, atît pentru manuscrise, cît și pentru tipărituri. Aceeași pagină din *Psaltirea Scheiană* este transcrisă după cum se arată în Fig. 10.

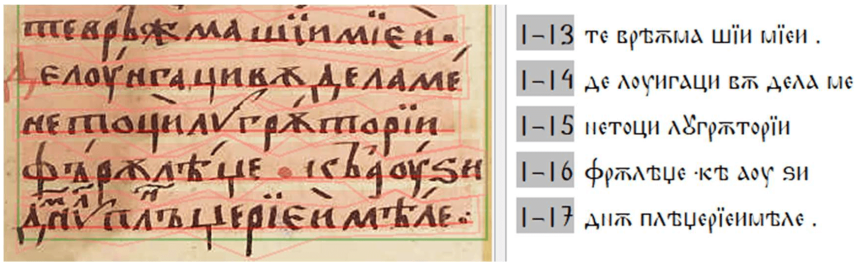

Figura 10: Transcriere din *Psaltirea Scheiană*, f. 7^{r}, cu Combined Romanian Cyrillic

Transcrierea este departe de a fi perfectă, deoarece necesită corectare manuală atît în ceea ce privește identificarea literelor, cît și în cazul spațierii. Cu toate acestea, dacă luăm în considerare faptul că sursa este complet nouă pentru model și că dimensiunea acestuia din urmă este relativ redusă, transcrierea obținută este un punct de plecare bun. În mod similar, dacă aplicăm modelul combinat la o sursă tipărită, rezultatul este o transcriere funcțională (Fig. 11).

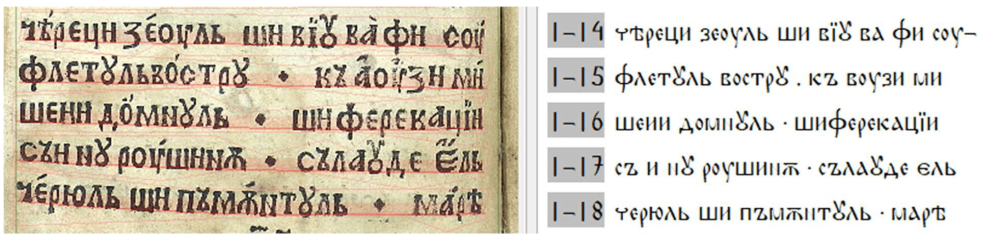

Figura 11: Transcriere din *Psaltirea* din 1570, f. 74^{v}, with Combined Romanian Cyrillic

Deși un model comun, care să poată fi utilizat pe o varietate de surse neprocesate anterior, este de dorit, pînă în prezent modelul Combined Romanian Cyrillic model este prea redus ca dimensiune ca să poată funcționa optim pe scrieri pentru care nu a fost antrenat. Pentru a dovedi acest aspect, am transcris o pagină din *Psaltirea Hurmuzaki* (f. 30^{r}) cu modelul Combined (CER 5,65%) și un al doilea model, dezvoltat special pentru textul psaltirii, respectiv pe transcrieri provenite din textul în discuție (CER 10,11%). La compararea performanței celor două modele se poate observa cum modelul specializat, deși are o valoare CER mai mare, furnizează un rezultat mai bun decît modelul generic. Redăm în Fig. 12 performanța modelului Combined Romanian Cyrillic, urmată de cea a modelului pentru Psaltirea Hurmuzaki 2.

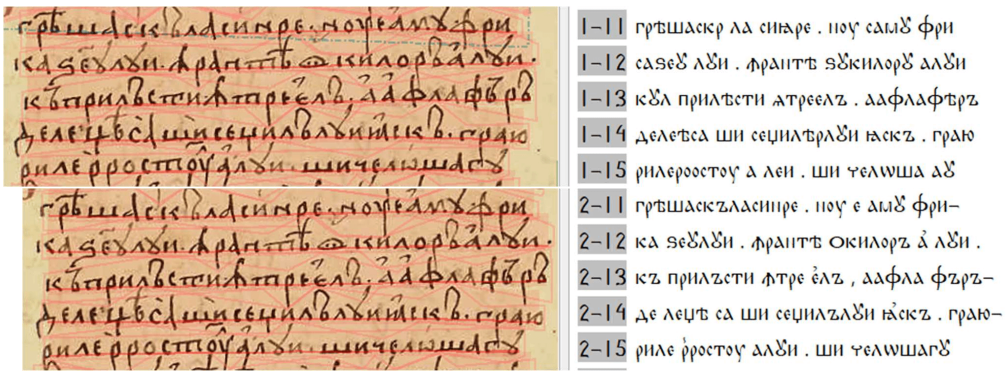

Figura 12: Modele Combined Romanian Cyrillic și Psaltirea Hurmuzaki 2 applicate *Psaltirii Hurmuzaki*, f. 30[r]

Discrepanța cea mai evidentă în ceea ce privește performanța celor două modele este faptul că modelul Hurmuzaki recunoaște mai bine forma literelor, deși există multe erori de spațiere și recunoaștere a cuvintelor, în timp ce rezultatul modelului Combined este practic imposibil de utilizat în acest caz. Motivul este faptul că modelul Combined a fost expus la și verificat pe o cantitate limitată de date, provenite doar din două surse, niciuna dintre acestea aparținînd *Psaltirii Hurmuzaki*. În momentul în care modelului Combined i se adaugă datele de antrenare utilizate pentru modelul Hurmuzaki, transcrierea automată se îmbunătățește în mod evident. La aplicarea noului model Combined aceleiași secțiuni a *Psaltirii Hurmuzaki*, obținem rezultatul din Fig. 13.

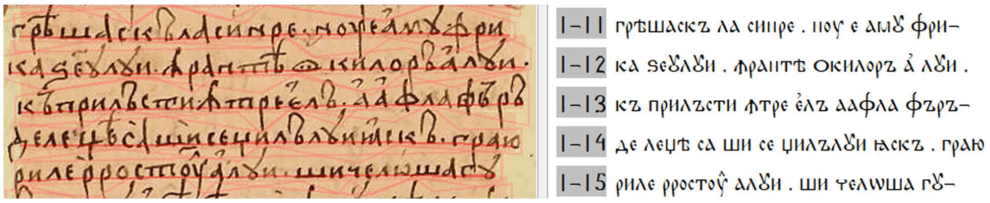

Figura 13: Modelul Combined_Romanian_Hurmuzaki aplicate *Psaltirii Hurmuzaki*, f. 30[r]

Toate literele sînt identificate corect, singurele erori fiind cauzate de separarea cuvintelor din rîndurile 13 **а҃ афла**, unde marca de infinitiv *a* nu este separată de verb, 14 **џилълѧиѭскъ**, unde cuvîntul nu este recunoscut ca atare, și 15 **л лѧи . ши челшшагшѕ-**, unde primele două cuvinte sînt considerate ca fiind unul singur, în timp ce ultimele 6 litere sînt parte a unui cuvînt care continuă pe rîndul următor **челшшагѕлъ**. Modelul prezintă o îmbunătățire dacă îl comparăm cu cele două modele anterioare, respectiv recunoașterea și reproducerea suprascrisului **л** pe rîndul 15. Așa cum este de așteptat, dacă modelul de transcriere primește mai multe date de formare și învață particularitățile unei scrieri specifice, performanța acestuia o va depăși pe cea a modelelor mai reduse ca dimensiune. Mergînd pe acest raționament, am decis să combinăm toate modelele generale dezvoltate pînă în prezent (inclusiv cele bilingve discutate mai jos) și să observăm dacă un model general cu date de antrenare însumînd pînă la 30 900 de cuvinte și cu o rată de eroare de 8,31% ar funcționa mai bine pentru transcrierea *Psaltirii Hurmuzaki*. Deși diferența de performanță pe această pagină nu este semnificativă, mai ales în ceea ce privește spațierea, este interesant să observăm că modelul Combined_mono-bilingual_Romanian identifică cuvîntul **џилълѧиѭскъ** din rîndul 14 ca atare (Fig. 14). Acest lucru se întîmplă cel mai probabil datorită activării caracteristicii *language model* (model lingvistic) incluse în Transkribus, care pune accentul pe aspectul lingvistic mai degrabă decît pe cel paleografic în ceea ce privește datele de formare utilizate în construirea modelului[5].

Un alt aspect important al producției de texte de pe teritoriile românești din secolul al XVI-lea este caracterul lor bilingv—multe dintre cărțile religioase au fost scrise sau tipărite atît în slavona bisericească, cît și în română. Deși forma literelor nu diferă de la o limbă la alta, este totuși necesară dezvoltarea unui

[5]Informații suplimentare referitoare la această funcție pot fi accesate la *readcoop.eu*.

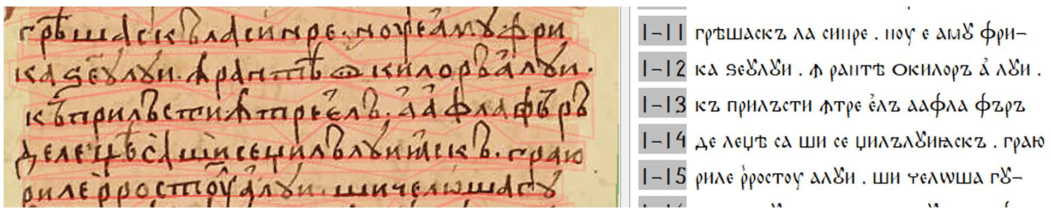

Figura 14: Modelul Combined_mono-bilingual_Romanian aplicat *Psaltirii Hurmuzaki*, f. 30[r]

nou model HTR pentru aceste surse. De fapt, așa cum am menționat anterior, în comparație cu tehnologiile OCR (recunoaștere optică a caracterelor) care se axează pe recunoașterea literelor individuale, HTR prezintă o formă de „inteligență lingvistică", analizînd proximitatea literelor în relație cu poziția lor în cadrul rîndului și încercînd să determine distribuția cea mai probabilă pe baza aspectelor pe care le-a învățat despre limbă din datele de antrenament. În consecință, slavona bisericească și româna fiind două limbi diferite, tehnologiile HTR ar trebui să învețe unele trăsături lingvistice aparținînd ambelor limbi. La fel ca în cazul abordării la care am recurs pentru modelele anterioare, am decis să reutilizăm munca depusă pînă la acel moment și să construim pe baza rezultatelor anterioare. Prima încercare a fost de a utiliza modelul Combined Romanian Cyrillic pe *Codicele Bratu*, un Apostol bilingv de la jumătatea secolului al XVI-lea, pentru a obține astfel o transcriere inițială, care a fost apoi verificată manual și adusă la nivelul de date de antrenament. Din păcate, imaginea din sursă este de foarte slabă calitate, astfel că un volum ceva mai mare de token-uri de cuvinte (11 500) nu a fost suficient pentru a obține o valoare satisfăcătoare a ratei de eroare, care în prezent se ridică la 11,12%. De fapt, antrenarea unui nou model de transcriere pentru *Psaltirea Ciobanu*, o altă sursă bilingvă disponibilă în format digital, cu imagini de foarte bună calitate, a avut rezultate mult mai bune, deși modelul a fost antrenat pe mai puține date (CER 7,97% cu 7800 de cuvinte). Combinarea celor două modele ne-a condus la o valoare CER de 10,26%, care, aplicat la o nouă sursă, are rezultatele arătate în Fig. 15.

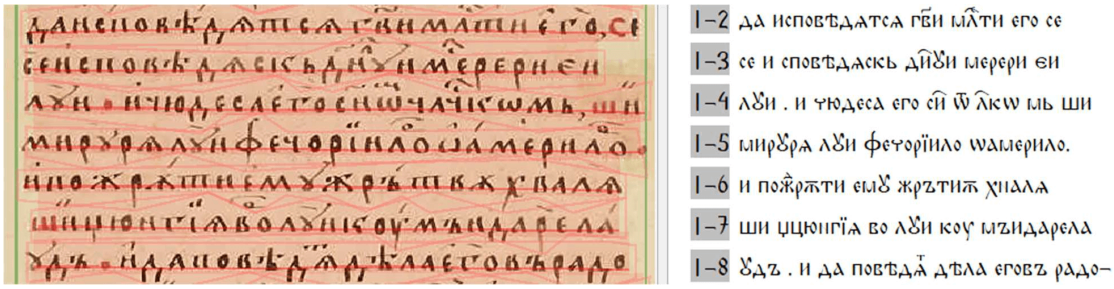

Figura 15: Modelul Bilingual_Romanian_Slavic_Cyrillic_0.01 applicat *Psaltirii Voronețene*, f. 19[r]

Modelul recunoaște cu dificultate suprascrierile, atît pentru cuvintele românești, cît și pentru cele slavone (a se vedea, spre exemplu, rîndurile 5 **фечорїило ѡамерило** și 4 **и чюдеса его сн҃ѿ чл҃кѡмь**), ceea ce reprezintă una din problemele specifice pentru modelele Transkribus HTR. În plus, există cîteva erori la identificarea anumitor litere – **н** este redat ca **и** în **дн҃ѣи** pe rîndul 3 și **коумънъдаре** pe rîndul 7, iar **в** este redat fie ca **и**, fie ca **н** în **жрътвж хвалм** pe rîndul 6. Deși alternarea între cele două limbi este marcată prin semne de punctuație și spațiere, textul în ansamblul său este scris în *scriptura continua*, astfel că separarea cuvintelor constituie o altă sursă de erori pentru acest model. Cu toate acestea, combinația între imagini de înaltă calitate și un volum mai mare de date de antrenament ar putea conduce la modele de transcriere eficiente pentru surse bilingve, utile pentru orice lucrare filologică axată pe limba română veche.

Ultimul aspect pe care l-am investigat în cadrul utilizării platformei Transkribus pe texte românești vechi a fost transliterarea scrierii chirilice în scriere latină. Încă de la jumătatea secolului al XX-lea, cercetătorii români au început să încline tot mai mult în favoarea transliterării mai degrabă decît a transcrierii în edițiile critice ale textelor vechi (Fischer, 1962; Avram, 1964), astfel că dezvoltarea modelelor de translite-

rare ar putea fi utilă. Pentru a construi un set de date de antrenament pentru modelul de transliterare, am ales un tabel standard de transliterare din scriere chirilică românească în alfabet latin, apoi am utilizat un convertor pentru a translitera datele de antrenament, am reîncărcat datele și am antrenat un nou model inteligent cu abilitatea de a translitera din alfabet chirilic în alfabet latin[6]. Sîntem pe deplin de acord cu faptul că principiile transliterării aplicate de către acest model sînt, într-o oarecare măsură, controversate. În orice caz, obiectivul nostru principal a fost să dovedim că modelele de transliterare HTR pentru scrierea chirilică românească funcționează și că nu este necesară corespondența biunivocă între imaginea vizuală a sursei și caracterul transcris. Pe viitor, comunitatea științifică ar trebui să decidă asupra unui sistem de transliterare general acceptat care să fie utilizat pentru modelele HTR de transliterare pentru care regulile de transcriere interpretativă fonetică produc rezultate nesatisfăcătoare. Performanța modelului este redată în Fig. 16 (Coresi, *Faptele Apostolilor*, 1563, p. 37).

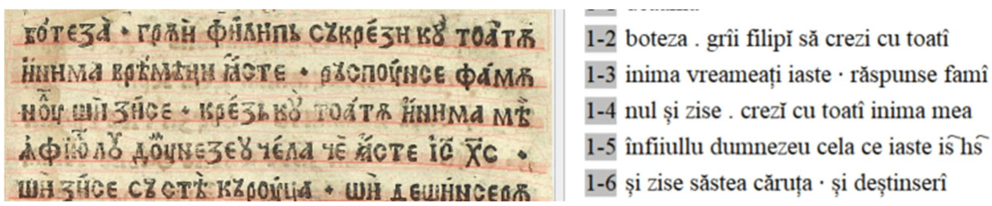

Figura 16: BRV12_Transliterated Romanian aplicat la *Faptele Apostolilor*, 1563, p. 37

Așa cum se întîmplă și în cazul transcrierii, pentru transliterare este important ca sursa originală să fie redată în manieră diplomatică, demers care, la rîndul său, conferă consistență datelor de antrenament utilizate pentru antrenarea unui model HTR. În acest caz, modelul (CER 5,32%) recunoaște cu exactitate toate slovele și spațierea (cu excepția rîndurilor 5 ꙟ фїїѡ лꙋ/*în fiiul lu* și 6 pentru съ стѣ/*să stea*) în conformitate cu regulile de transliterare aplicate inițial, astfel că ѫ este redat în mod constant prin *î*, ъ prin *ă*, ь prin *ĭ* etc. În plus, literele suprascrise sînt identificate corect și aduse pe rînd, textul transliterat fiind relativ ușor de citit.

3. Implicații

Am arătat că, prin utilizarea tehnologiei HTR și a programului Transkribus, un număr însemnat de manuscrise sau cărți tipărite din perioada timpurie pot fi pre-transcrise în mod eficient. Deși rata de eroare a modelului computerizat este mai mare decît cea a unui specialist în domeniu, costurile sînt incomparabil mai mici. La fel ca în cazul proiectelor editoriale tradiționale (manuale) unde există cel puțin o etapă de corectare după prima transcriere, efectuată de regulă de revizorul principal al proiectului editorial, și după transcrierea HTR este necesară cel puțin o etapă de corectare, care poate fi consumatoare de timp. Cu toate acestea, așa cum indică Rabus (in press), costul total al unui proiect editorial în care se utilizează Transkribus reprezintă sub o zecime din costul total al unui proiect tradițional bazat exclusiv pe munca manuală.

Luînd în considerare acest factor, sîntem îndreptățiți să afirmăm că tehnologia HTR poate schimba cursul științelor umaniste în epoca digitală. Pentru prima dată în istorie această tehnologie permite digitalizarea în masă a unui volum uriaș de surse needidate anterior într-un interval de timp redus și cu implicarea unor resurse financiare cu mult mai reduse decît în cazul proiectelor tradiționale.

Mai mult decît atît, se conturează și perspectiva de a utiliza tehnologia HTR pentru digitalizarea unor întregi arhive, ceea ce ar permite identificarea de texte din diverse manuscrise (v., spre exemplu, *transkribus.eu*). În plus, din perspectivă cantitativă, se pot dezvolta modalități de cercetare noi și interesante, care să nu se bazeze pe post-corectarea manuală a rezultatelor HTR și care să deschidă perspective complet noi asupra datelor textuale istorice (de ex., Camps *et al.*, 2020).

[6]Tabelul de corespondențe a fost preluat de pe *Wikipedia*, iar instrumentul de conversie este *Protea*.

4. Concluzii și perspective

În cadrul acestui studiu am prezentat primele noastre experimente cu tehnologia de recunoaștere a scrisului de mînă (HTR) pentru scrierea chirilică românească prin utilizarea platformei Transkribus. Deși modelele curente nu sînt încă perfecte și uneori produc transcrieri eronate, avem speranța că am reușit să arătăm potențialul acestei tehnologii pentru filologia românească în era digitală. Pe viitor, o provocare interesantă ar fi să antrenăm modele pentru diverse stiluri de scriere, precum scrierea chirilică sau latină cursivă.

Datorită faptului că versatilitatea și calitatea modelelor HTR Transkribus depind în mod crucial de volumul de date de antrenare disponibile, este deosebit de important să creăm modele mult mai mari decît cele discutate în prezentul studiu. Această sarcină poate fi asumată doar prin prisma colaborării, de aceea lansăm un apel tuturor cercetătorilor interesați de scrierea chirilică românească să-și unească forțele, să recicleze transcrierile/transliterările pe care le-au realizat inițial în alt scop (de exemplu pentru elaborarea de ediții tipărite sau online) și astfel să creeze date de antrenament prin colaborare, în manieră rapidă și eficientă. Încurajăm toți cercetătorii interesați de HTR asistată de inteligența artificială să ne contacteze pentru a explora în echipă noi posibilități de a crea și aplica noile modele HTR pentru scrierea (chirilică) românească. Credem că tehnologia HTR reprezintă un punct de referință pentru filologia digitală modernă și avem speranța că întreaga comunitate științifică românească se va angaja în promovarea acestei cauze, în beneficiul tuturor celor implicați.

Bibliografie

A. Izvoare

A.1. Manuscrise

Codicele Bratu, ediție critică de Al. Gafton, disponibilă online.
Psaltirea Ciobanu, psaltire slavo-română, Moldova, 1573–1585, ms. rom. 3465 BAR, disponibilă online.
Psaltirea Hurmuzaki, secolul al XVI-lea, ms. rom. 3077 BAR, disponibilă online.
Psaltirea Scheiană, secolul al XVI-lea, ms. rom. 449 BAR, disponibilă online.
Codicele Voronețean, secolul al XVI-lea, ms. rom. 448, disponibilă online.
Psaltirea Voronețeană, secolul al XVI-lea, ms. rom. 693, disponibilă online.

A.2. Ediții

Psaltire: [Brașov, diaconul Coresi, 1570], accessibil online, CRV 16.
Faptele Apostolilor: [Brașov, diaconul Coresi, 1563], accessibil online, CRV 12.

B. Literatură secundară

Avram, A. (1964). *Contribuții la interpretarea grafiei chirilice a primelor texte românești*, în „Studii și cercetări lingvistice", **XV** (1–5).

Camps, J-B., Thibault Clérice, T., & Ariane Pinche, A. (2020). *Stylometry for Noisy Medieval Data: Evaluating Paul Meyer's Hagiographic Hypothesis*, [online].

Fischer, I. (1962). *Principii de transcriere a textelor românești*, în „Limba română", **IX** (5), p. 577–581.

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto S. *et al.* (2019). *Transforming Scholarship in the Archives Through Handwritten Text Recognition*, în „Journal of Documentation", **75** (5), p. 954–976, Crossref.

Rabus, A. (2019). *Recognizing Handwritten Text in Slavic Manuscripts: A Neural-Network Approach Using Transkribus*, în „Scripta & e-Scripta", **XIX**, p. 9–32.

Rabus, A. (in press). *Automatische computergestützte Transkription paläoslavistischer Quellen und ihre Folgen für Korpuslinguistik und Editionsphilologie*, în „Proceedings Humboldt Kolleg Venice 2020".

C. Modele Transkribus pentru scrierea chirilică românească

Numele modelului	**Surse**	**Cuvinte**	**CER**
BRV12_Romanian Printing 16th c	1563 Apostolos, printed	3951	4,46%
Romanian_Cyrillic_0.01/0.02/0.03	Voroneț Codex	5013	5,84%
Combined Romanian Cyrillic_manuscript	1563 Apostolos + Voroneț Codex	8964	5,65%
Psaltirea Hurmuzaki 1/2	Hurmuzaki Psalter	11 439	10,11%
Combined_Romanian_Hurmuzaki	Hurmuzaki Psalter + 1563 Apostolos + Voroneț Codex	20 310	6,78%
CB_Bilingual Romanian-Slavonic	Bratu Codex	11 553	11,12%
PCb_Bilingual Romanian-Slavonic	Ciobanu Psalter	7865	7,97%
Bilingual_Romanian_Slavic_Cyrillic_0.01	Bratu Codex + Ciobanu Psalter	19 418	10,26%
Combined_mono-and_bilingual_Romanian_0.01	Hurmuzaki Psalter + 1563 Apostolos + Voroneț Codex + Bratu Codex + Ciobanu Psalter	39 728	8,31%