Analiza automată a materialelor şcolare franceze din prisma complexităţii textuale

Mihai Dascălu, Lucia Larise, Ștefan Trăușan-Matu, Philippe Dessus, Maryse Bianco

Analiza automată a materialelor şcolare franceze din prisma complexităţii textuale

Mihai Dascălu, Larise Lucia Stavarache, Stefan Trausan-Matu

Universitatea Politehnica din Bucureşti

313 Splaiul Independenței, România

mihai.dascalu@cs.pub.ro, larise.stavarache@ro.ibm.com, stefan.trausan@cs.pub.ro

Philippe Dessus, Maryse Bianco

Laboratorul de Ştiințe ale Educației, Universitatea Grenoble Alpes

38040 Grenoble CEDEX 9 France

philippe.dessus@upmf-grenoble.fr, maryse.bianco@upmf-grenoble.fr

REZUMAT

Eforturile de cercetare în ceea ce priveşte analiza complexităţii textuale automate se concentrează în principal asupra limbii engleze, în prezent existând doar câteva adaptări pentru alte limbi. Pornind de la un fundament solid în ceea ce priveşte analiza discursului şi analiza complexităţii textuale bazată pe un model de evaluare validat pentru limba engleză, în lucrarea de faţă introducem un model de analiză textuală pentru limba franceză antrenat pe 200 de documente extrase din manuale şcolare franţuzeşti, pre-clasificate în cinci clase de complexitate textuală. Factorii de analiză acoperă o multitudine de dimensiuni şi sunt distribuiţi în următoarele categorii principale: de suprafaţă, factori specifici analizei sintactice, morfologice, semantice şi analiza discursului, metrici care ulterior sunt combinate prin intermediul maşinilor cu vector suport (Support Vector Machines – SVM) pentru îmbunătăţirea preciziei. La nivel global, adiţional faţă de factorii pur cantitativi de suprafaţă, anumite părţi de vorbire şi diverse metrici de coeziune s-au dovedit a fi predictori de încredere raportat la nivelul de dificultate al materialelor şcolare, creând astfel o bază consistentă pentru a construi modele de încredere de evaluare automată a complexităţii textuale franceze.

Cuvinte cheie

coeziunea textelor, evaluarea complexităţii textuale, lizibilitatea, metode kernel, maşini cu vector suport (Support Vector Machines)

Clasificare ACM

I.2.7 Natural Language Processing.

INTRODUCERE

Notarea automată a eseurilor, identificarea de modele şi şabloane textuale sau evaluarea dificultăţii textelor reprezintă subiecte de interes în prezent în domeniul analizei textuale, a căror aplicabilitate se adresează atât mediilor academice, cât şi celor comerciale, din prisma alinierii optime la nivelul de înţelegere al cursanţilor. Pe de-o parte, grupurile de cercetare se concentrează pe analiza şi inter-conectarea volumelor mari de texte extrase din web-ul social într-o continuă extindere, în timp ce piaţa comercială este axată pe dezvoltarea de instrumente de analiză care ar putea îmbunătăţi predicţia nivelului de înţelegere şi ar putea induce un control mai bun al fluxului de informaţii de la sursă către publicul ţintă.

Un dezavantaj major al eforturilor de analiză automată a complexităţii textuale provine din lipsa de conţinut suficient în cazul unor limbi precum limba franceză, în comparaţie cu limbile pentru care există astfel de resurse (spre ex., engleza). Prin urmare, analiza multi-lingvală a complexităţii textuale este încă un domeniu puţin explorat care ar putea reliefa rezultate surprinzătoare raportat la următoarele dimensiuni ale evaluării: complexitatea vocabularului, morfologia şi structura discursului. Indiferent de nivelul de analiză adresat, scopul final presupune îmbunătăţirea nivelului de înţelegere al cursanţilor.

Limba engleză are un vocabular dinamic, în continuă expansiune, iar dificultatea acesteia poate fi evaluată drept medie din punctul de vedere al timpului de învăţare pentru vorbitorii non-nativi. Pornind de la factori specifici de analiză automată, în această lucrare vom prezenta un model de evaluare a complexităţii textuale aferent limbii franceze dintr-o perspectivă nouă. Scopul nostru este de a confirma premisele analizei lingvistice pe limba engleză studiată în extenso [1, 2, 3] care demonstrează viabilitatea predicţiei automate a complexităţii unui text, dar şi de a sublinia particularităţile lingvistice şi factorii relevanţi în ceea ce priveşte evaluarea înţelegerii şi a trăsăturilor textuale specifice limbii franceze.

Bazat pe scenariul clasic de învăţare în care elevilor le sunt prezentate materiale specifice selectate de către un profesor, scopul studiului presupune evaluarea dificultăţii textelor în vederea alinierii optime cu nivelul de înţelegere al elevilor, asigurând astfel compatibilitatea dintre materialele prezentate şi cursanţi.

Totodată, evaluarea nivelului de înţelegere al elevilor poate fi realizată folosind diverse strategii precum: colectarea de feedback în timp real, teste şi examinări, evaluarea interactivităţii sau teme obligatorii. Indiferent de modalitatea de scalare a rezultatelor obţinute folosind aceste strategii în raport cu practicile actuale din sistemul de învăţământ, este evident că un factor central, cu implicaţii majore vizavi de înţelegerea cursanţilor, este reprezentat de *subiectivitatea* tutorelui. Astfel, profesorul poate fi indus în eroare atât de rezultate intermediare, cât şi de numeroşi factori externi, precum regulamentele şcolare. Un alt aspect problematic care accentuează subiectivismul rezidă în *timpul de evaluare*; astfel, evaluarea unui număr mare de teste într-un interval de timp foarte scurt este predispusă la numeroase erori umane, cu atât mai mult când analiza este realizată de un singur evaluator.

Scopul acestei lucrări este de a sprijini profesorii în evaluarea materialelor de curs, a testelor şi a materialelor opționale disponibile în limba franceză într-un mod cât mai precis şi totodată echitabil față de elevi, într-o manieră reproductibilă, aplicabilă asupra unui volum mare de texte într-un interval scurt de timp şi a cărei trasabilitate poate fi controlată.

Din punctul de vedere al structurii, următoarea secțiune prezintă o vedere de ansamblu asupra metodelor şi sistemelor de evaluare a complexității textuale, clasificările lor, precum şi o comparație cu alte sisteme automate similare de evaluare a complexității textuale. A treia secțiune este centrată pe descrierea modelului de analiză automată a complexității textuale pentru limba franceză, în timp ce a patra secțiune se concentrează pe validarea modelului, reliefarea factorilor de impact major şi interpretarea rezultatelor. Ulterior, ultima secțiune este axată pe conturarea concluziilor şi pe definirea direcțiilor ulterioare de cercetare.

O VIZIUNE DE ANSAMBLU ASUPRA EVALUĂRII COMPLEXITĂȚII TEXTUALE

Măsurarea automată a complexității textuale a reprezentat şi reprezintă o provocare a ultimilor ani pentru profesori şi cercetători deopotrivă. Secolul al XX-lea este caracterizat de un progres puternic la nivel tehnologic care s-a repercutat în toate domeniile şi în care sistemul educațional a trebuit să îşi adapteze rapid metodologiile pentru a fi la curent cu toate schimbările. Astfel, metodologiile de predare, notarea lucrărilor scrise sau a temelor de clasă trebuie să fie adaptate la un număr din ce în ce mai mare de elevi care vor să învețe si să aibă acces rapid la informații. Elevii trăiesc într-un sistem dinamic înconjurat de activități sociale şi de jocuri cognitive, dar totodată receptivitatea lor se confruntă cu un factor de retenție din ce în ce mai scăzut şi disociat. Astfel, deşi pe termen scurt memoria lor funcționează foarte bine pe baza unor asocieri rapide, pe termen lung interconectarea informațiilor se pierde, principalul factor fiind volumul mare de informații care îi înconjoară [4].

Diametral opus, profesorii trebuie să îşi adapteze stilul de predare pentru alinierea materialelor la nivelul de înțelegere şi lectură al cititorilor, cu accent pe retenție şi pe realizarea unei reprezentări coerente a unui text care să faciliteze înțelegerea [5, 6]. Astfel, în vederea realizării acestui obiectiv, tutorii trebuie mai întâi să îşi facă o imagine de ansamblu a grupului de lucru, să evalueze complexitate textuală a materialelor utilizate pentru predare şi să îşi pună frecvent şi proactiv probleme de măsurare a percepției individuale în ceea ce priveşte uşurința de lectură şi de înțelegere a elevilor.

Din intersecția celor două perspective rezultă una din problemele majore ale sistemului de învățământ actual şi anume alinierea materialelor prezentate la nivelul grupului țintă datorită percepției diferite a indivizilor în raport cu materialul prezentat. Percepția poate fi alterată din varii cauze precum: cunoştințele anterioare în domeniul respectiv, familiarizarea cu limba, motivația personală a individului sau interesul pentru subiectul prezentat. Cu alte cuvinte, factorii principali care influențează lizibilitatea şi facilitează înțelegerea unui text pot fi sumarizați astfel: nivelul de educație al auditoriului, capacitățile cognitive şi experiențele sale anterioare. Pornind de la factorii menționați anterior, nivelul de complexitate măsurat trebuie adaptat folosind un model cognitiv al cititorului care să permită maximizarea percepției şi realizarea de asocieri coerente ale informației prezentate.

În concordanță cu alinierile anterioare, instrumentele software de analiză care se concentrează pe măsurarea automată a complexității textuale şi pe notare, trebuie să fie adaptate, în sensul că, pentru un public-țintă, nivelurile estimate de complexitate textuală obținute prin aplicarea unor metrici predefinite trebuie să fie relevante şi adecvate la situația analizată. Suplimentar, există o multitudine de factori care ar trebui luați în considerare: viteza de asimilare a cursanților (uneori corelată cu vârsta acestora), fondul cultural al subiecților, interesul personal, motivația, accesul la informație, dar şi capacitatea tutorelui de distribuție a informației corespunzător la gradul de înțelegere a studenților.

Complexitatea unui text are un impact mare asupra înțelegerii şi a capacității de retenție a cursanților pe termen lung. Adițional, realizarea de conexiuni coerente la nivelul informațiilor prezentate este un factor determinant raportat la zona proximei dezvoltări (ZPD) [7] a fiecărui elev. ZPD este strâns legată de motivația personală a elevului, întrucât lipsa constantă de înțelegere poate cauza deficiențe mari în capacitatea acestuia de asimilare şi înțelegere, pe când o lipsă de provocare poate duce în timp la frustrare personală sau la respingerea colectivului. În acest context, tutorii/profesorii au de înfruntat în mod constant o sarcină dificilă în privința evaluării complexității textuale şi în prezentarea materialelor adecvate pentru cursanți.

În plus, timpul joacă un rol foarte important, iar profesorii ar trebui să pună accent pe interacțiune şi pe transmiterea eficientă a cunoştințelor, recurgând chiar la schimbări frecvente de materiale ca urmare a rezultatelor slabe obținute în clasă. Pentru a putea facilita acest lucru într-un timp relativ scurt, utilizarea de instrumente software dedicate devine o necesitate vizavi de oferirea de suport prin reliefarea trăsăturilor textuale şi sublinierea anumitor caracteristici în cadrul materialelor de învățare.

Inițiativa de standardizare aparținând Common Core State [8] afirmă că analiza complexității textuale joacă un rol determinant în evaluarea fiecărui student în vederea pregătirii/admiterii la liceu. Astfel, următoarele trei dimensiuni concentrează viziunea asupra complexității textuale: analiza cantitativă, analiza calitativă şi orientarea cititorului şi a activității educaționale. Prima dimensiune cuprinde factorii cantitativi (spre exemplu, frecvența cuvintelor sau lungimea propozițiilor) şi reprezintă cea mai directă şi palpabilă metodă de calcul. Factorii calitativi sunt centrați pe nivelurile de înțelegere, structură, convenționalism de limbaj, claritate şi complexitate a termenilor. În cele din urmă, orientarea poate fi considerată cea mai dificilă dimensiune întrucât aceasta adresează cunoştințele elevilor, motivația şi interesele lor.

Pornind de la ideea că o analiză viabilă nu poate fi realizată pe baza doar a unei singure dimensiuni sau factor

de evaluare, setul de metrici luate în considerare şi agregate automat reprezintă un element cheie pentru evaluarea corectă şi precisă a complexităţii textuale. Marea majoritate a sistemelor existente se bazează exclusiv pe factori cantitativi simpli. Drept referinţă, există o mare varietate de soluţii utilizate în diverse programe de studiu în limba engleză [9] – spre exemplu, *Lexile* (MetaMetrics) [10, 11], *ATOS* (Renaissance Learning) [12], Degrees of Reading Power: *DRP Analyzer* (Questar Assessment, Inc.) [13, 14], *REAP* (Universitatea Carnegie Mellon) [15, 16], *SourceRater* (Educational Testing Service) [17], the *Pearson Reading Maturity Metric* (Pearson Knowledge Technologies) [18], *Coh-Metrix* (Universitatea din Memphis) [19, 20]. În schimb, există un singur sistem care adresează evaluarea limbii franceze - *Dmesure* (Universitatea Catolică Louvain-La-Neuve) [21, 22].

Dmesure integrează factori de complexitate textuală la nivel lexical şi sintactic preponderent aplicaţi asupra unor texte destinate vorbitorilor non-nativi de franceză. Metricile anterioare sunt agregate prin diverşi clasificatori (spre exemplu, arbori de decizie, regresie logistică multinomială, optimizări de tip bagging & boosting sau metode kernel: maşini cu vector suport – SVM) [23] în vederea generării automate de exerciţii de limbă adecvate nivelului cititorului. Textele au fost manual clasificate în conformitate cu scara CEFR (Common European Framework of Reference for Languages – Cadrul European Comun de Referinţă pentru Limbi Străine), iar validările au fost efectuate din prisma predicţiei clasificatorilor raportat la clasificarea manuală a textelor.

Extragerea unor metrici valide pentru analiza automată a complexităţii textuale nu trebuie privită ca un înlocuitor al profesorilor, ci drept suport care să îi sprijine în adaptarea facilă la schimbările sistemului educaţional şi care să le permită reacţia agilă şi integrată la dezvoltarea societăţii. Avantajele majore ale metodelor automate includ reducerea timpului de aşteptare a rezultatelor pentru teste şi examene, evaluarea la o scară mai largă, posibilitatea de a efectua evaluări mai frecvent, fără a pierde accentul pe interacţiune şi interactivitate şi fără a diminua calitatea evaluării globale.

Această lucrare îşi propune să găsească factori şi şabloane care oferă măsurători valide ale complexităţii textuale pentru texte în limba franceză, accentuând totodată impactul factorilor de profunzime care adresează morfologia, semantica şi analiza discursului.

MODELUL PROPUS DE ANALIZĂ A COMPLEXITĂŢII TEXTUALE PENTRU LIMBA FRANCEZĂ

Analiza automată a complexităţii textuale este evaluată relativ la măsurătorile obţinute şi la metricile existente pentru limba engleză [1, 24, 25], dar poate fi fără sens în absenţa unui model antrenat şi validat pe un corpus specific. Astfel, pornind de la experimentele anterioare propunem un model multi-dimensional de analiză a complexităţii textuale adaptat limbii franceze, integrând atât metrici clasice de suprafaţă, cât şi factori derivaţi din tehnicile de analiză şi notare automată, morfologie şi sintaxă, dar şi semantică şi analiza discursului [2, 25]. În final, subseturi de factori sunt agregate automat prin intermediul SVM-urilor [26] care s-au dovedit a fi cea mai eficientă metodă de clasificare automată [27].

În primul rând, analiza de suprafaţă este centrată pe analiza elementelor individuale precum cuvinte, fraze şi paragrafe utilizând exclusiv statistici simple. Modelul de analiză textuală la acest nivel se bazează pe studiile lui Page [28] privind notarea automată a eseurilor şi includ următoarele categorii inspiratele din clasele lui Slotnick [29]: fluenţă, formule standard de calcul a lizibilităţii, structură, dicţie, precum şi entropia aplicată la nivel de cuvinte şi caractere.

O subcategorie specifică de factori este centrată pe analiza complexităţii cuvintelor luând în calcul diferite metrici precum: numărul de silabe, distanţa dintre forma inflecţională, lemă şi rădăcină (stem). Specificitatea unui concept este evidenţiată prin proporţionalitatea inversă a numărului său de apariţii în documentele analizate din corpusul de antrenare (în cazul nostru, articole din ziarul “Le Monde” cuprinzând aproximativ 24 de milioane de cuvinte), dar şi de distanţa în ierarhia de concepte sau polisemia cuvântului desprinse din ontologia lexicalizată WOLF [30].

Ulterior, în cadrul analizei sintactice şi morfologice se regăsesc statistici cu privire la diverse părţi de vorbire (prepoziţiile, pronumele, substantivele şi adverbele sunt cele mai relevante), la arborele de analiză sintactică (spre exemplu, adâncimea maximă şi dimensiunea sa), precum şi o adaptare a metricilor CAF (Complexitate, Acurateţe, Fluenţă) [31] pentru franceză. Informaţiile la nivel de cuvânt sunt centrate pe forma pronumelui (singular, plural, prima, a doua sau a treia persoană) şi sunt identificate prin utilizarea de cuvinte cheie predefinite.

În final, cea mai interesantă categorie de factori este centrată în jurul analizei semantice şi a discursului, în care graful de coeziune şi legăturile coezive aferente [2, 32], lanţurile lexicale [33], elementele de conectică şi marcajele predefinite sunt componentele centrale în reprezentarea şi analiza discursului. Din punct de vedere computaţional, coeziunea este determinată de mix-ul dintre măsurătorile efectuate constând în măsurarea similarităţii semantice în ontologii lexicalizate [34], similaritatea de tip cosinus aplicată pe vectorii din spaţiul semantic al analizei semantice latentă (LSA – Latent Semantic Analysis) [35], precum şi inversul disimilarităţii Jensen-Shannon calculată pe baza distribuţiilor de probabilităţi din alocarea Dirichlet latentă (LDA – Latent Dirichlet Allocation) [36].

VALIDAREA MODELULUI

Modelul descris în secţiunea anterioară cuprinde în total 54 de factori individuali grupaţi pe categorii şi a fost antrenat pe 200 de documente extrase din manuale şcolare primare franţuzeşti clasificate în cinci clase de complexitate. Cele cinci clase utilizate sunt mapate direct pe cele cinci clase primare din intervalul 6-11 ani, astfel: CP – “Cours préparatoire“ (clasa întâi), CE1, CE2 – “Cours élementaire” (clasa a II-a şi clasa a III-a), CM1 şi CM2 – “Cours moyen” (clasele a IV-a şi clasa a V-a) din sistemul de învăţământ francez.

Tabelul 1. Dimensiunile complexității textuale: concordanță exactă şi adiacentă (EA/AA).

Dimensiunea	Categorie de factori de complexitate textuală	*EA* medie	*AA* medie
Analiza de suprafaţă	Fluenţă	.67	.93
	Complexitatea structurii lexicale	.73	.80
	Dicţie	.33	.73
	Entropie la nivel de caracter şi cuvinte	.53	.87
	Complexitatea cuvintelor	.60	.87
Morfologie şi sintaxă	Metrica CAF (Complexitate, Acurateţe, Fluenţă) balansată	.67	.93
	Părţi de vorbire specifice	.60	.93
	Nivelul de informare al fiecărui cuvânt	.60	.80
	Complexitatea arborelui de parsare	.73	.93
Semantică şi analiza discursului	Dificultatea discursului reliefată în graful de coeziune	.53	.80
	Conectori ai discursului	.40	.67
	Lanţuri lexicale	.60	.80

Fiecare document a fost validat manual, clasificat şi notat de experţi în lingvistică, iar corpusul selectat a fost suficient de cuprinzător pentru antrenarea SVM-ului. În final s-a aplicat validarea în cruce folosind 3-fold cross-validation [37] pentru determinarea preciziei şi concordanţei de tip exact (*EA* – Exact Agreement) sau adiacent (*AA* – Adjacent Agreement) [21], în raport cu procentul prin care SVM-ul a obţinut predicţii identice sau similare cu clasele de complexitate manual predefinite. Ca specificitate a analizei am optat pentru utilizarea unui kernel radial (gaussian) de grad 3 şi pentru integrarea unei metode de căutare de tip grid pentru sporirea eficienţei SVM-ului prin selectarea parametrilor optimi aferent kernel-ului (*gamma* şi *C*).

Tabelul 1 prezintă rezultatele sintetice pentru fiecare categorie de factori de complexitate textuală, în timp tabelul 2 prezintă factorii individuali ordonați descrescător după valoarea concordanţei exacte (*EA*).

Integrarea factorilor anteriori a demonstrat că măsurătorile efectuate cu ajutorul SVM-ului au fost precise şi că modelul prezentat este viabil (*EA* = .733 şi *AA* = .933 prin utilizarea tuturor factorilor de complexitate textuală implementați). În comparație cu rezultatele obținute pe modelul de analiză a limbii engleze (*EA* = .779 şi *AA* = .997) [3], putem concluziona că rezultatele obținute în evaluarea curentă au fost foarte bune şi promițătoare pentru un prim experiment. De menționat este că rezultatele obținute pentru vocabularul englez au avut la bază 1000 de documente pre-clasificate pe baza scorului DRP [13] şi structurate în şase clase de complexitate, în timp ce modelul francez a beneficiat doar de adnotarea manuală a 200 de documente în cinci clase de complexitate.

Tabelul 2. Factorii cu impactul cel mai mare în antrenarea modelului de complexitate textuală franceză (valori EA/AA maximale).

Factorul de complexitate textuală	*EA* medie	*AA* medie
Coeziunea medie dintre fraze şi paragraful corespunzător	.80	.93
Numărul normalizat de virgule	.73	1
Metrica CAF balansată globală	.67	.93
Lungimea medie a cuvintelor	.60	1
Sofisticarea sintactică din cadrul metricii CAF	.60	.87
Numărul normalizat de fraze	.53	.93
Lungimea medie a frazelor	.53	.93
Numărul mediu de fraze per paragraf	.53	.60
Dimensiunea medie a paragrafelor exprimată în numărul de caractere	.53	.73
Sofisticarea lexicală din cadrul metricii CAF	.53	.87
Numărul de prepoziţii	.53	.80
Scorul mediu al fiecărui paragraf obţinut prin intermediul grafului de coeziune	.53	.67
Numărul de relaţii cauzale	.53	.67
Numărul de cuvinte din document	.47	.80
Numărul de paragrafe	.47	.67
Numărul mediu de cuvinte per frază	.47	.80
Diversitatea lexicală din cadrul metricii CAF	.47	.80
Distanţa medie dintre lemă şi stem pentru fiecare cuvânt utilizat	.47	.80
Lungimea medie a lanţurilor lexicale	.47	.80
Coeziunea medie dintre paragrafe şi document	.47	.73
Coeziunea medie între frazele din acelaşi paragraf	.47	.73
Numărul de pronume persoana a treia plural	.47	.73

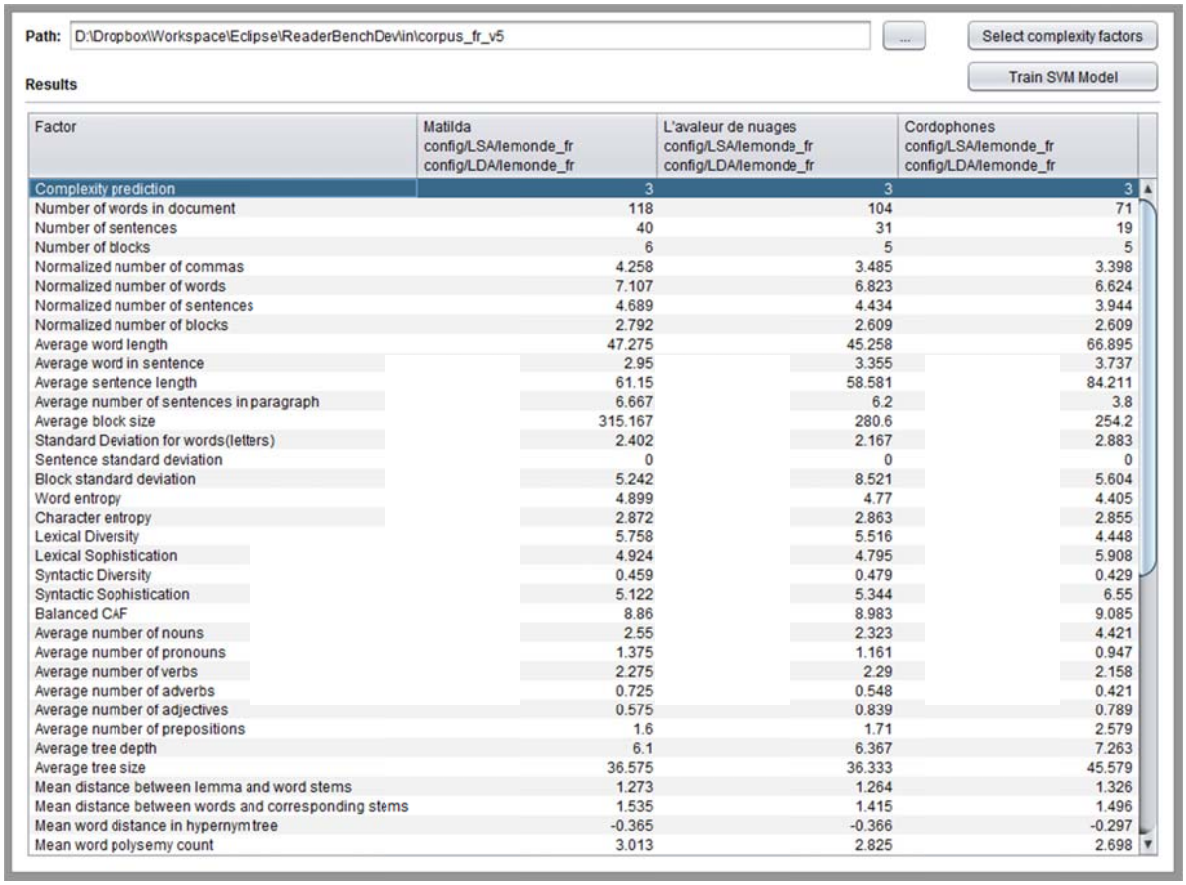
Suplimentar, modelul antrenat de complexitate textuală poate fi aplicat cu uşurință asupra unor noi texte în vederea evaluării dificultății acestora. Astfel, sistemul pune la dispoziție interfața prezentată în Figura 1, care oferă tutorelui o imagine comparativă la nivelul tuturor factorilor de complexitate textuală utilizați în realizarea predicției automate.

CONCLUZII ŞI DIRECȚII VIITOARE DE CERCETARE

Pornind de la rezultatele individuale prezentate în Tabelul 2 şi în corelație cu viziunea de ansamblu creionată în Tabelul 1 se constată că, suplimentar față de factorii pur cantitativi de suprafață, diverşi factori care adresează morfologia, coeziunea şi analiza discursului, pot fi considerați predictori de încredere raportat la nivelul de dificultate al materialelor şcolare. Astfel, modelul propus reprezintă o bază consistentă pentru construirea de modele de încredere de evaluare automată a complexității textuale franceze.

Direcțiile de cercetare imediat următoare vizează integrarea de factori suplimentari şi migrarea anumitor factori disponibili în prezent doar pentru limba engleză, precum şi extinderea corpusului de documente cu

materiale de dificultate elevată din ciclurile gimnaziale şi liceale.



Factor	Matilda config/LSA/lemonde_fr config/LDA/lemonde_fr	L'avaleur de nuages config/LSA/lemonde_fr config/LDA/lemonde_fr	Cordophones config/LSA/lemonde_fr config/LDA/lemonde_fr
Complexity prediction	3	3	3
Number of words in document	118	104	71
Number of sentences	40	31	19
Number of blocks	6	5	5
Normalized number of commas	4.258	3.485	3.398
Normalized number of words	7.107	6.823	6.624
Normalized number of sentences	4.689	4.434	3.944
Normalized number of blocks	2.792	2.609	2.609
Average word length	47.275	45.258	66.895
Average word in sentence	2.95	3.355	3.737
Average sentence length	61.15	58.581	84.211
Average number of sentences in paragraph	6.667	6.2	3.8
Average block size	315.167	280.6	254.2
Standard Deviation for words(letters)	2.402	2.167	2.883
Sentence standard deviation	0	0	0
Block standard deviation	5.242	8.521	5.604
Word entropy	4.899	4.77	4.405
Character entropy	2.872	2.863	2.855
Lexical Diversity	5.758	5.516	4.448
Lexical Sophistication	4.924	4.795	5.908
Syntactic Diversity	0.459	0.479	0.429
Syntactic Sophistication	5.122	5.344	6.55
Balanced CAF	8.86	8.983	9.085
Average number of nouns	2.55	2.323	4.421
Average number of pronouns	1.375	1.161	0.947
Average number of verbs	2.275	2.29	2.158
Average number of adverbs	0.725	0.548	0.421
Average number of adjectives	0.575	0.839	0.789
Average number of prepositions	1.6	1.71	2.579
Average tree depth	6.1	6.367	7.263
Average tree size	36.575	36.333	45.579
Mean distance between lemma and word stems	1.273	1.264	1.326
Mean distance between words and corresponding stems	1.535	1.415	1.496
Mean word distance in hypernym tree	-0.365	-0.366	-0.297
Mean word polysemy count	3.013	2.825	2.698

Figura 1. Vizualizarea comparativă a trei texte în limba franceză raportat la factorii de complexitate textuală selectaţi de către utilizator

Mulţumiri

Cercetările prezentate au fost susţinute de finanţarea Agence Nationale de la Recherche (DEVCOMP) şi de către proiectul FP7-REGPOT-2010-1 264207 ERRIC-Empowering Romanian Research on Intelligent Information Technologies.

REFERINŢE

1. Dascalu, M., Trausan-Matu, S., Dessus, P., Bianco, M., şi Nardy, A., 2013. ReaderBench, o platformă integrată pentru analiza complexităţi textuale şi a strategiilor de lectură. In Conf. Nat. de Interactiune Om-Calculator (RoCHI 2013), T. Stefanut and C. Rusu Eds. MatrixRom, Bucuresti, Romania, 39–46.
2. Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., şi Nardy, A., 2014. Mining texts, learners productions and strategies with ReaderBench. In Educational Data Mining: Applications and Trends, A. Peña-Ayala Ed. Springer, Switzerland, 335–377.
3. Dascalu, M., Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence. Springer, 2014.
4. Bereiter, C., Education and mind in the knowledge age. Lawrence Erlbaum Associates, Mahwah, NJ, 2002.
5. Blanc, N. şi Brouillet, P., Mémoire et compréhension: Lire pour comprendre. Editions InPress, Paris, France, 2003.
6. Graesser, A.C., Singer, M., şi Trabasso, T. (1994) Constructing inferences during narrative text comprehension. Psychological Review 101 (3), 371–395.
7. Vygotsky, L.S., Mind in society. Harvard University Press, Cambridge, MA, 1978.
8. National Governors Association Center for Best Practices, C.o.C.S.S.O., 2010. Common Core State Standards National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C.
9. Nelson, J., Perfetti, C., Liben, D., şi Liben, M., 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. Council of Chief State School Officers.
10. Stenner, A.J., 1996. Measuring reading comprehension with the Lexile Framework. MetaMetrics, Inc.
11. Stenner, A.J., Koons, H.H., şi Swartz, C.W., 2009. Closing the text complexity gap: Reconceptualizing the text complexity continuum. MetaMetrics, Inc.
12. Borman, G.D. şi Dowling, N.M., 2004. Testing the Reading Renaissance Program Theory: A Multilevel Analysis of Student and Classroom Effects on

Reading Achievement. University of Wisconsin-Madison.
13. Koslin, B.L., Zeno, S.M., Koslin, S., Wainer, H., şi Ivens, S.H., The DRP: An effectiveness measure in reading. College Entrance Examination Board, New York, NY, 1987.
14. Zeno, S.M., Ivens, S.H., Millard, R.T., şi Duvvuri, R., The educator's word frequency guide. Touchstone Applied Science Associates, Inc., Brewster, NY, 1995.
15. Heilman, M., Collins-Thompson, K., Callan, J., şi Eskenazi, M., 2006. Classroom success of an intelligent tutoring system for lexical practice and reading comprehension. In 9th Int. Conf. on Spoken Language Processing ISCA, Pittsburgh, PA, 4.
16. Dela Rosa, K. şi Eskenazi, M., 2011. Self-Assessment of Motivation: Explicit and Implicit Indicators in L2 Vocabulary Learning. In 15th Int. Conf. on Artificial Intelligence in Education (AIED2011), G. Biswas, S. Bull, J. Kay and A. Mitrovic Eds. Springer, Auckland, New Zealand, 296–303.
17. Sheehan, K.M., Kostin, I., Futagi, Y., şi Flor, M., 2010. Generating automated text complexity classifications that are aligned with published text complexity standards. Educational Testing Service.
18. Landauer, T.K., Kireyev, K., şi Panaccione, C., (2011) Word maturity: A new metric for word knowledge. Scientific Studies of Reading 15, (1), 92–108.
10. Graesser, A.C., McNamara, D.S., Louwerse, M.M., şi Cai, Z., (2004) Coh-Metrix: Analysis of text on cohesion and language. Behavioral Research Methods, Instruments, and Computers 36, (2), 193–202.
11. McNamara, D.S., Louwerse, M.M., McCarthy, P.M., şi Graesser, A.C., (2010) Coh-Metrix: Capturing linguistic features of cohesion. Discourse Processes 47, (4), 292–330.
12. François, T. şi Miltsakaki, E., 2012. Do NLP and machine learning improve traditional readability formulas? In First Workshop on Predicting and improving text readability for target reader populations (PITR2012) ACL, Montreal, Canada, 49–57.
13. François, T., 2012. Les apports du traitement automatique du langage à la lisibilité du français langue étrangère. In Centre de Traitement Automatique du Langage Université Catholique de Louvain, Faculté de Philosophie, Arts et Lettres, Louvain-la-Neuve, Belgium.
14. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., şi Steinberg, D., (2008) Top 10 algorithms in data mining. Knowledge and Information Systems 14, (1), 1–37.
15. Dascalu, M., Trausan-Matu, S., şi Dessus, P., 2012. Towards an integrated approach for evaluating textual complexity for learning purposes. In 11th Int. Conf. in Advances in Web-Based Learning (ICWL 2012), E. Popescu, R. Klamma, H. Leung and M. Specht Eds. Springer, Sinaia, Romania, 268–278.
16. Dascalu, M., Dessus, P., Trausan-Matu, S., Bianco, M., şi Nardy, A., 2013. ReaderBench, an environment for analyzing text complexity and reading strategies. In 16th Int. Conf. on Artificial Intelligence in Education (AIED 2013), H.C. Lane, K. Yacef, J. Mostow and P. Pavlik Eds. Springer, Memphis, USA, 379–388.
17. Cortes, C. şi Vapnik, V.N., (1995) Support-Vector Networks. Machine Learning 20, (3), 273–297.
18. Hsu, C.-W. şi Lin, C.-J., (2002) A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks 13, (2), 415–425.
19. Page, E., (1968) Analyzing student essays by computer. International Review of Education 14, (2), 210–225.
20. Slotnick, H., (1972) Toward a theory of computer essay grading. Journal of Educational Measurement 9, (4), 253–263.
21. Sagot, B., 2008. WordNet Libre du Francais (WOLF) INRIA, Paris.
22. Schulze, M., 2010. Measuring textual complexity in student writing. In American Association of Applied Linguistics (AAAL 2010) Waterloo Centre for German Studies, Atlanta, GA, 590–619.
23. Trausan-Matu, S., Dascalu, M., şi Dessus, P., 2012. Textual complexity and discourse structure in Computer-Supported Collaborative Learning. In 11th Int. Conf. on Intelligent Tutoring Systems (ITS 2012), S.A. Cerri, W.J. Clancey, G. Papadourakis and K. Panourgia Eds. Springer, Chania, Grece, 352–357.
24. Galley, M. şi McKeown, K., 2003. Improving word sense disambiguation in lexical chaining. In 18th International Joint Conference on Artificial Intelligence (IJCAI'03), G. Gottlob and T. Walsh Eds. Morgan Kaufmann Publishers, Inc., Acapulco, Mexico, 1486–1488.
25. Budanitsky, A. şi Hirst, G., (2006) Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics 32, (1), 13–47.
26. Landauer, T.K. şi Dumais, S.T., (1997) A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. Psychological Review 104, (2), 211–240.
27. Blei, D.M., Ng, A.Y., şi Jordan, M.I., (2003) Latent Dirichlet Allocation. Journal of Machine Learning Research 3, (4-5), 993–1022.
28. Geisser, S., Predictive inference: an introduction. Chapman and Hall, New York, NY, 1993