Revista Română de Interacțiune Om-Calculator 6 (4) 2013, 339-356

Analiza complexității manualelor şcolare din sistemul de învățământ francez primar

Lucia Larise Stavarache[1], Mihai Dascalu[2], Stefan Trausan-Matu[1,2], Philippe Dessus[3]

[1]Facultatea de Automatică şi Calculatoare, Universitatea Politehnica din Bucureşti
Splaiul Independenței 313, Bucureşti 060042, Bucureşti
E-mail: larise.stavarache@ro.ibm.com, mihai.dascalu@cs.pub.ro, stefan.trausan@cs.pub.ro

[2]Institutul de Cercetări în Inteligență Artificială
Calea 13 Septembrie nr. 13, Bucureşti, România

[3]Laboratoire de Sciences de l'Education, Université Pierre-Mendès France
F-38040 Grenoble CEDEX 9, Franța
E-mail: philippe.dessus@upmf-grenoble.fr

Rezumat. Una din sarcinile dificile din prelucrarea limbajului natural ține de evaluarea complexității sau a dificultății înțelegerii a textelor. Astfel, de un interes aparte sunt instrumentele de analiză automată a textelor care permit realizarea unei predicții a complexității textelor pornind de la factori lexicali, sintactici, morfologici sau chiar semantici în funcție de specificitatea abordării. Totodată, elementele de complexitate sunt dependente de context şi de domeniul de aplicare. Astfel, o analiză pur tehnică care se bazează doar pe metrici fără să ia în considerare constrângerile de psihologie, modele umane, vârstă şi motivația este insuficientă pentru a realiza o predicție adecvată. În plus, alte aspecte ale analizei complexității sunt corelate cu etapele de achiziție (împreună cu acuratețea şi fluența), corelate cu adaptarea mesajului comunicat audienței din prisma corectitudinii, coerenței şi adaptabilității la nivelul acesteia. Adițional, metricile de complexitate textuală reprezintă indicatori importanți de înțelegere şi coerență pentru texte comune întâlnite uzual în Internet, lucrări publicate şi cărți.

Cuvinte cheie: analiza complexității textuale, metrici de complexitate, grupuri şi clase de concepte.

1. Introducere

Fiecare act de comunicare trebuie să fie adaptat audienței, iar mesajul trebuie să fie transmis corect şi coerent indiferent de complexitatea sa. Doar în acest mod mesajul va rezona cu auditoriul şi îşi va atinge scopul. Analiza

complexității textelor presupune o arie vastă care înglobează numeroase domenii precum psihologia cognitivistă, tehnologia informației din prisma metodelor automate de evaluare, precum şi tehnicile de comunicare formală şi informală aplicabile în vederea facilitării înțelegerii conținutului transmis.

Având ca bază exemplele anterioare, putem afirma că analiza complexității textuale poate presupune inclusiv evaluarea nonsens-ului dacă ne raportăm la perioada dadaistă în care poezia reprezenta doar o înşiruire de cuvinte extrase aleator, pornind invers de la cuvânt către mesaj (Ball, 1996). Chiar şi în acest context putem afirma că există grade de complexitate, deoarece cuvintele în sine au propriul lor grad de complexitate.

Scopul principal al unei comunicări scrise sau orale îl reprezintă transmiterea mesajului indiferent de situație, locație sau resurse. Lucrarea îşi propune analiza complexității textelor având ca referință următoarele criterii: calitate, cantitate, analiza dependenței de context şi scop. Factorii utilizați pentru analiză sunt în general împărțiți în metrici standard, respectiv metrici complexe (Graesser et al., 2004). *Metricile standard* sunt cele care se bazează pe valori numerice, care pot fi obținute relativ simplu, fără să se țină cont de context şi pe baza unor valori fixe care fundamentează ecartul de comparare. *Metricile complexe* nu sunt atât de precise întrucât acestea nu se raportează la cifre fixe, ci la indicatori, oferind astfel o abordare mai detaliată asupra analizei.

Astfel, în vederea evaluării gradului de înțelegere al unui text generic extras din manuale şcolare, tema lucrării nefiind cunoscută, lucrarea de față îşi propune sa fundamenteze o abordare complexă pentru calcularea complexității care să ofere o vedere mai clară asupra subiectului, raportându-se la context, locație şi potențiala temă. Abordarea propusă pleacă de la definiția de bază a discursului conform căreia un cuvânt este o combinație de consoane şi vocale care alăturate creează mesaje cu diferite înțelesuri.

Articolul continuă prin prezentarea unor abordări similare, analiza informației extrasă din corpusul de manuale şcolare şi prin detalierea grupurilor extrase din Manulex (Lété et al., 2004) utilizate în evaluarea complexității textuale. În final sunt prezentate rezultate şi studii de caz, urmate de concluzii şi direcții ulterioare de cercetare.

2. Abordări similare de evaluare a complexității textuale

Secțiunea de față își propune prezentarea a trei abordări diferite de evaluare a complexității textuale, fiecare axată pe o altă fațetă a problematicii. Totodată, abordarea noastră este axată pe determinarea diferitelor semnificații pentru aceeaşi expresie, în contexte diferite. Pentru a marca diferențele drastice de percepție care pot apărea apelăm la un exemplu clasic din programare care presupune folosirea sintagmei "Hello World!". Aceasta este în general utilizată pentru implementarea celui mai simplu program într-un limbaj de programare, aplicație care presupune tipărirea la consolă a mesajului anterior. Dacă am privi lucrurile din punctul de vedere al unei persoane care nu a avut nici o tangență cu lumea calculatoarelor, semnificația cuvintelor ar fi complet diferită, iar un răspuns complet diferit ar putea fi: „Hello to you too!". Diametral opus celor două exemple prezentate anterior este răspunsul unui analizor de text care oferă informații conform unor metrici prestabilite şi care nu ține cont de context: „ afirmație exclamativă, camel-case, afirmație exclamativă, nu există verbe, 1 substantiv, 1 adverb".

2.1 Evaluarea automată a eseurilor (AES – Automated Essay Scoring)

Conceptul din spatele evaluării automate a eseurilor (Chodorow & Burstein, 2004) se bazează pe funcții aplicate pe un corpus de documente precum:

- numărul de cuvinte;
- numărul de propoziții subordonate;
- numărul de propoziții;
- numărul de fraze;
- media cuvintelor scrise cu literă mare/cuvinte scrise cu literă mică.

Folosind aceşti indicatori se construieşte un model matematic care este ulterior aplicat şi eseurilor noi care intră în analiză. Această metodă devine tot mai utilizată în universități dar şi în alte instituții precum examenul TOEFL. Evaluarea automată nu este lipsită de marjă de eroare, deoarece alegerea unei metode de rezolvare diferită față de cea standard a unui exercițiu cu răspuns scris poate genera scoruri greşite.

De asemenea, dacă este impusă o limită minimă pentru lungimea unui eseu există posibilitatea ca un eseu mai scurt să primească o notă mai mică decât un eseu mai lung deşi conţinutul primului eseu este mai valoros. Eroarea este inevitabilă în aceste cazuri deoarece lungimea minimă a unui eseu este un indicator setat iniţial în analizor. Dacă în schimb tezele ar fi fost analizate de o comisie de profesori astfel de cazuri nu ar fi trebuit să apară.

Unul din cele mai cunoscute sisteme de evaluare automată care au la bază AES este „E-rater" (Attali & Burstein, 2004). Acesta este un analizor automat folosit cu succes in cadrul GMAT – "Graduate Management Admission Test". Analizatorul are la bază principiile de analiză a limbajelor naturale dar ia în considerare pentru calcularea scorului şi următorii factori: analiza conţinutului din punct de vedere lexical, numărul de greşeli gramaticale, numărul de greşeli mecanice (greşeli datorate supraîncărcări sistemului), numărul de comentarii. E-rater are la bază un dicţionar de fraze, cuvinte stop (un set predefinit de cuvinte care este scos înainte de a începe procesarea şi analiza corpusului) şi un set de euristici pentru identificarea relaţiilor retorice bazate pe sintaxă şi distribuţie în paragraf cât şi pe reguli de analiză a structurii frazelor.

Rezultatele studiului realizat de Powers et al. (2001) au arătat că în anumite condiţii eseurile mai slabe pot obţine note bune. De asemenea gradul de interes al elevilor scădea la aflarea faptului că eseurile lor sunt corectate automat. Pentru a preveni aceste marje de eroare s-a decis folosirea E-rater împreună cu notarea clasica, mai ales pentru examenele de importanţă majoră.

2.2 Extragerea relaţiilor semantice

Adiţional sistemelor automate de notare există şi analizoare semantice (Jurafsky & Martin, 2009, Manning & Schütze, 1999). Ideea din spatele conceptului se bazează pe relaţiile semantice între substantive şi adjective cât şi pe verbele asociate lor. Setul de reguli folosit pentru determinarea complexităţii caută şabloane („pattern matching") de secvenţe de cuvinte şi dependenţe gramaticale la nivel de propoziţie (Gervasi & Ambriola, 2002). Algoritmul din cadrul acestei abordări se bazează pe constrângerile gramaticale combinate cu regula drumului cel mai scurt obţinut folosind un graf orientat, după cum se poate vedea în figura 1.

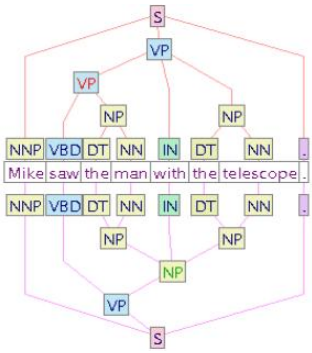

Figura 1. Exemplu de graf de parsare http://www.cs.jhu.edu/~jason/465/

2.3 Coh-Metrix

Coeziunea lingvistică (Halliday & Hasan, 1976, Tangkiengsirisin) este axată pe elementele constitutive ale discursului care îi conferă coerență din punct de vedere lexical, sintactic şi semantic. Metricile sistemului Coh-Metrix (Graesser, McNamara, Louwerse & Cai, 2004, McNamara et al., 2010) analizează textele din diverse perspective precum topică, gramatică, structură, poziție sau vocabular. Scorul final este calculat în funcție de scorul fiecărui factor de coeziune. Procentul fiecărui factor de coeziune din cadrul scorului final este asignat în funcție de scopul şi contextul textului. Metricile sistemului Coh-Metrix sunt împărțite in două: referențiale şi explicite (Graesser et al., 2011). De exemplu, fragmentele de text care nu au multe conexiuni cu celelalte părți ale textului adaugă un procent de complexitate scorului general, deoarece corelațiile lor raportate la context nu pot fi extrase.

Pe lângă regulile prezentate mai sus, metricile sistemului iau în calcul şi metricile clasice de măsurare a complexității precum numărul de: cuvinte, substantive, adjective, lungimea frazei, lungimea textului. Printre formulele consacrate de evaluare a complexității textuale se numără:

1. Flesch Reading Ease (Flesch, 1948):
   206.835-(1.015×ASL)-(86.4×ASW)
   - ASL = media lungimii propoziției (numărul de cuvinte/numărul de propoziții);
   - *ASW* = media numărului de silabe per cuvânt.

2. Flesch-Kincaid Grade Level (Kincaid et al., 1975):
$(0.39 \times ASL) + (11.8 \times ASW) - 15.59$, unde ASL şi ASW au semnificaţiile anterioare.

2.4 Analiză comparativă

Din punct de vedere al diferenţelor între articolul de faţă versus tehnicile şi metodologiile deja existente, următorii factori cheie ies in evidenţă:

- Scopul lucrării este de a implementa un criteriu cantitativ de clasificare şi ordonare a lucrărilor, nu de a evalua o formula standard a complexităţii, în opoziţie cu lucrările prezentate mai sus care îşi propun să realizeze o predicţie cât mai precisă a complexităţii;
- Lucrarea de faţă se axează pe analiza complexităţii manualelor şcolare franceze, domeniu neacoperit de către analizoarele prezentate anterior, atât din punctul de vedere al metricilor pentru limba franceză, cât şi a specificităţii domeniului;
- Contextul este un factor cheie în jurul căreia abordarea propusă gravitează, iar fiecare factor rezultat din metricile implementate este analizat în funcţie de context, ceea ce în abordările anterioare nu reprezintă un punct de interes predominant;
- Scopul articolului nu presupune notarea manualelor şcolare franceze, ci de a identifica elementele care pot afecta negativ înţelegerea mesajului şi asimilarea conţinutului precum: prezenţa unui capitol în manual de complexitate mult mai mare decât complexitatea manualului, care poate duce la asimilarea incorectă a informaţiei, pierderea interesului auditoriului sau confundarea informaţiilor.

3. Corpusul aferent limbii franceze

Pentru a defini un context coerent al abordării considerăm oportună definirea unui set de prezumţii şi constrângeri. Astfel, raportat la specificitatea analizei, se presupune că toată informaţia primită este corectă din punct de vedere gramatical, cât şi din punctul de vedere al sensului conţinutului. Totodată, euristicele folosite în determinarea complexităţii exclud textele incomplete, caracterele speciale şi imaginile. Totuşi,

evaluarea în sine a complexității textuale ia în considerare toate cele trei categorii menționate anterior bazându-se pe metrici adaptate la nivelul contextului analizei.

Pentru a putea fi analizate, textele sunt împărțite în fragmente. Un fragment conține unul sau mai multe paragrafe. Fiecare fragment tratat individual are mesaj coerent. Un fragment poate să reprezinte: un capitol, o definiție, o poveste, o bucată relevantă de informație care conține un mesaj şi poate fi clar identificată. Peste 40 de manuale au fost luate în considerare şi analizate pentru a determina factori care influențează cel mai mult complexitatea în cadrul textelor şcolare din sistemul francez de învățământ.

Textele şcolare sunt grupate în funcție de vârstă şi clasă conform tabelului 1. În total au rezultat peste 1.000 de fragmente din procesarea manualelor şcolare, cu o medie de 200 de fragmente per manual. Media numărului de cuvinte rezultat ~100.000.

Tabelul 1. Indici de măsurare a frecvenței

Ciclul primar		Ciclul gimnazial	
Nivel şcolar	Vârstă	Nivel şcolar	Vârstă
Clasa I	3-6 ani	Clasa V	11-12 ani
Clasa II	6-8 ani	Clasa VI	12-13 ani
Clasa III	8-10 ani	Clasa VII	13-14 ani
Clasa IV	10-11 ani	Clasa VIII	14-15 ani

4. Grupurile Manulex de concepte, factori de analiză a complexității textuale şi euristici de optimizare

4.1 Grupurile Manulex

Manulex (Lété, Sprenger-Charolles & Colé, 2004) este o bază de date online care conține manualele şcolare primare franceze cu aproximativ 2 milioane de cuvinte împărțite în grupurile de complexitate $G_1 \rightarrow G_5$. Pe lângă grupurile de complexitate Manulex mai oferă şi informațiile suplimentare prezentate în cadrul tabelului 2.

Baza de date Manulex conține 48.886 intrări nelematizate şi 23.812 intrări lematizate. Acest conținut online a fost ales pentru această lucrare datorită flexibilității şi preciziei datelor. Analiza complexității textelor din manualele de limba franceză este calculată şi în funcție de distribuția

cuvintelor în grupuri de complexitate. Distribuția cuvintelor în grupuri conform Manulex este prezentată în tabelul 3.

Tabelul 2. Indici de măsurare a frecvenței

Indice	Descriere
F	frecvența generală
D	indicele de dispersie
U	frecvența estimată per 1 milion de cuvinte
SFI	indicele standard de frecvență

Astfel, fiecare text conține mai multe fragmente care la rândul lor conțin mai multe paragrafe. Un paragraf poate conține una sau mai multe fraze iar componența unei fraze este formată din două sau mai multe propoziții. Cuvântul este unitatea indivizibilă în cadrul acestei analize. Scopul acestui criteriu de măsurare este determinarea procentului de complexitate aparținând fiecărui grup.

Tabelul 3. Grupurile Manulex - 1.909.918 cuvinte

Grup	Componență
G_1	172.248 – cuvinte elementare
G_2	351.024 – cuvinte de dificultate medie
G_3-G_5	1.386.546 – cuvinte de dificultate ridicată

La finalul analizei putem să avem de exemplu pentru un text de complexitate redusă precum un manual din ciclul primar, 60% din cuvinte aparținând grupului G_1, 30% din cuvinte aparținând grupului G_2 şi 10% din cuvinte aparținând grupului G_3. Având o distribuție macro a complexității textului analizat putem determina uşor dacă un manual este corespunzător categoriei de vârstă în care este încadrat cel puțin ca o concluzie primară a analizei.

4.2 Distribuția paragrafelor în grupuri de complexitate

După cum s-a precizat în paragraful anterior, analiza complexității globale folosind doar euristici simple nu este foarte precisă. Un exemplu elocvent ar fi capitolele din manuale care conțin informație care depăşeşte nivelul de înțelegere al clasei de studiu căreia i se adresează. Astfel de capitole sau pasaje de informație al căror nivel de complexitate este mai ridicat decât nivelul de înțelegere al studenților pot cauza o discontinuitate a informației şi o diminuare a interesului față de domeniu. Pentru a determina pasajele cu

grad de dificultate peste media conținutului se va analiza distribuția în grupuri de complexitate per paragraf. Prin urmare, un paragraf al cărui grup de complexitate primar este G_3 va fi greu de asimilat într-un text cu grup primar de complexitate G_1.

Totodată analiza la nivel de paragraf oferă o viziune mai detaliată asupra structurii textului, putând determina semantica, dinamica şi structura frazelor pentru diferite paragrafe aparținând unor grupuri de complexitate diferite. Trecerea de la analiza complexității la nivel de text către analiza complexității la nivel de paragraf reprezintă un prim pas între trecerea de la analiza macro către analiza micro.

4.3 Variația grupurilor de complexitate induse de o variabilă auxiliară

Introducerea unei variabile auxiliare reprezintă o formă de a determina elementele care influențează major complexitatea textuală. Astfel putem considera variabila auxiliară ca fiind un delimitator, respectiv complexitatea textului este calculată după ce s-au eliminat cuvintele de lungime mai mică sau mai mare ca variabilă auxiliară.

Un alt mod de a folosi o variabilă auxiliară este de a considera că variabila reprezintă o unitate gramaticală respectiv: verb, substantiv, adjectiv, pronume ş.a.m.d. Eliminând pe rând unitățile gramaticale putem determina unitatea care influențează cel mai mult complexitatea.

Un alt aspect care poate fi determinat folosind cele două modalități prezentate este determinarea dinamicii textului. Astfel putem calcula dinamica unui text combinând următoarele metrici: numărul de verbe din text, distribuția textului în grupuri de complexitate eliminând verbele şi distribuția textului în grupuri de complexitate eliminând cuvintele cu o lungime mai mică de un număr predefinit de caractere. Eliminarea verbelor are ca scop încadrarea textului în cele trei categorii: monolog, dialog, „dialog monologat” şi determinarea unor metrici de diferențiere între cele 3 categorii utilizând frecvența cuvintelor şi impactul complexității verbelor.

4.4 Eliminarea minimului şi a maximului

Doi alți indicatori de complexitate reprezintă minimul şi maximul cuvintelor din punct de vedere al lungimii. Eliminarea minimului reprezintă scoaterea

cuvintelor de legătură dintr-un text. Un astfel de procedeu afectează în proporție foarte mică complexitatea generală deoarece cuvintele cheie care formează contextul rămân. Eliminarea maximului are rezultate mult mai spectaculoase dacă textul nu are cuvinte maximale izolate. Un cuvânt maximal izolat reprezintă un cuvânt foarte lung cu o frecvență generală foarte mică care nu influențează complexitatea generală. Dacă în schimb exista un procent de cel puțin 3-5% cuvinte maximale, influența asupra complexității este majoră. Un alt efect major asupra textului în cazul în care maximul nu este izolat reprezintă pierderea contextului odată ce este eliminat.

4.5 Indicii de frecvență F, D, U şi SFI

Variația celor 4 indici U, F, D şi SFI introduce metrici şi indicatori noi pentru analiza complexității din punct de vedere cantitativ. Astfel un factor F cu valoarea de 90 are o frecvență de apariție de 1 la 10 cuvinte. Dacă combinăm cei 4 indicatori cu regulile de determinare a complexității grupurilor poate fi făcută o a doua observație. Dacă cuvântul al cărui factor F de valoare 90 aparține grupului G_1, complexitate redusă putem afirma că complexitatea sa este redusă. Dacă în schimb cuvântul aparține grupurilor G_3-G_5 de exemplu „mitocondrii" atunci frecvența sporită nu reduce complexitatea textului, chiar o măreşte în majoritatea cazurilor în care audiența nu are cunoştințe despre context şi domeniul de referință al textului.

- ***F*** este singurul indice care nu depinde de ceilalți 3, reprezentând frecvența generală a unui cuvânt. *F* este folosit pentru determinarea raportului între frecvența cuvântului şi distribuția sa în grupuri. Astfel în propoziția "Cuvintele sunt elementele componente ale unei propoziții. Propozițiile sunt elementele componente ale unei fraze." F are următoarele valori:
  - Cuvintele, fraze, propoziții, propozițiile => F=1
  - Elementele, componente, ale, unei, sunt => F=2
- ***D*** este dependent de F şi reprezintă dispersia:

$$D = \log(\sum p_i) - \left[\left(\sum p_i \times \log(p_i)\right) / \sum p_j\right] / \log(n)$$;

  - *n* = numărul de manuale din corpusul Manulex respectiv 54;

- i = numărul manualului din grup;
- p este raportat la i şi reprezintă frecvența cuvântului în carte.

După cum putem observa în formulă, D ia valori în intervalul [0; 1]. Valoarea minimă este obținută când cuvântul aparține unui singur grup, iar cea maximă apare atunci când cuvântul se regăseşte în toate grupurile.

- ***U*** reprezintă frecvența raportată la 1 milion de cuvinte – convergența U este considerată optimă pentru 1 milion de cuvinte:

$$U = (1.000.000/n)\times\left[FD-(1-D)\times f_{\min}\right];$$

- N = numărul de cuvinte;
- FD = frecvența şi dispersia;
- $f_{\min}$ = suma produselor f_i şi s_i împărțită la N.
- f_i =frecvența cuvântului în carte
- s_i= numărul de cuvinte din carte când D=1, U este calculat din exact 1 milion e cuvinte;

- ***SFI*** reprezintă indicele standard de frecvență şi depinde de ceilalți 3 factori.

$$SFI = 10\times\left[\log_{10}(U)+4\right].$$

Un cuvânt cu SFI = 90 are o frecvență de 1/10 cuvinte. O frecvență mare nu reprezintă şi o complexitate redusă întotdeauna, deoarece cuvântul în cauză poate sa fie foarte specific unui anumit domeniu iar înțelegerea poate să fie dificilă pentru 80% din auditoriu.

4.5 Imagini

Precum a fost specificat în prealabil, textele incomplete, caracterele speciale şi imaginile au fost eliminate la analiza inițială a calculului de complexitate textuală. Analiza aportului imaginilor în calculul complexității a fost introdusă ulterior pentru a putea obține rezultate amănunțite în ce priveşte articolele de complexitate similară, adresate aceleiaşi clase primare.

Imaginile ajută la înțelegerea mai rapidă a textelor, deoarece conținutul este asociat cu elemente vizuale. În cazul elevilor din clasele primare imaginile reprezintă un element foarte important de asimilare şi de reținere a

mesajului. Simultan, în cazul elevilor din ciclurile superioare scade nevoia de a asocia elementele în raport cu conținutul textului. Drept comparație între cele doua cicluri putem afirma următoarele:

- în cazul elevilor din ciclurile primare plictiseala se poate instala în momentul neînțelegeri fragmentelor întrucât bagajul de cunoştințe este în formare, iar nevoia de asociere cu elemente cunoscute este mai pregnantă (Macri, 2013);
- în cazul elevilor din ciclurile superioare interesul se poate diminua mult în cazul în care sunt multe imagini, deoarece se creează senzația de uşurință, iar concentrarea scade;

Aportul adus de existența imaginilor sau de lipsa lor din informația analizată se calculează conform următoarelor criterii:

- Aportul numărului de imagini este direct proporțional cu lungimea şi grupul de complexitate al fragmentului. Proporția a fost calculată per fragment în parte (definiția fragmentului poate fi regăsită mai sus);
  - f_i – numărul imaginilor în fragmentul *i*;
  - Lf_i – numărul de cuvinte din fragmentul *i*;
  - *Restricție*: se consideră imagine asociată unui fragment, imaginea situată la începutul fragmentului sau în interiorul acestuia. De asemenea, pentru a fi considerat relevant, un fragment trebuie să conțină minim două propoziții;
  - *Restricție*: fragmentele trebuie să aparțină aceluiaşi grup de complexitate. Astfel, dacă avem un manual cu 12 fragmente distribuite după cum urmează: 6 - G_1, 4 - G_2, 2 – G_3 vor rezulta distribuții pe fiecare grup în parte;
- Euristica folosită: $Img(G) = 1 - \sum \frac{ft}{\sum ft}$, $(0,1]$
  - G – reprezintă grupul de complexitate (G_1, G_2, G_3 – G_5)

Coeficienții de complexitate obținute sunt rafinate ulterior pentru a putea observa diferențele între două manuale aparținând aceleiaşi clase primare care au fost încadrate inițial în aceeaşi categorie de complexitate.

Astfel, drept exemplu prezentăm următorul scenariu în cadrul căruia considerăm două manuale adresate claselor primare CM2 cu următoarea distribuție de complexitate: manualul 1 (G_1 – 60%, G_2 – 28%, G_3 – 12%), respectiv manualul 2 (G_1 – 60%, G_2 – 28%, G_3 – 12%). Având distribuțiile de mai sus este dificil să observăm o diferență clară de complexitate. În acest

caz, analiza modificării coeficientului de complexitate în funcție de numărul de imagini din fiecare manual se dovedeşte utilă: manualul 1 ($Img(G1) = 0,98$, $Img(G2) = 0,96$, $Img(G3) = 0,93$), respectiv manualul 2 ($Img(G1) = 1,00$, $Img(G2) = 1,00$, $Img(G3) = 0,94$).

În această situația, noua distribuție de complexitate devine: manualul 1 (G_1 – 60%*0,98 = 57,60%, G_2 – 28%*0,96 = 26,88%, G_3 – 12%*0,94 =11,24%), manualul 2 (G_1 – 60%*1,00 = 60,00%, G_2 – 28%*1,00 = 28%, G_3 – 12%*0,93 = 11,16%). Astfel, rezultatele anterioare sunt corelate cu observația că în al doilea manual nu regăsim imagini pentru fragmentele din grupurile de complexitate G_1 şi G_3.

5. Rezultate

Validarea factorilor anteriori s-a realizat pe baza unui sub-corpus de 10 manuale aparținând claselor primare CM2 – echivalentul clasei a patra din sistemul de învățământ românesc, vârsta între 10-11 ani. O să numim grupul CM2 în analiza grafică de mai jos, cumulul de 10 manuale şcolare primare care fac parte din studiul de caz.

Tabelul 4. Distribuția generală CM2 în diverse cazuri

Grup	Distribuție generală CM2	Distribuție generală CM2 cu variabila auxiliară > 2	Distribuție generală CM2 cu variabila auxiliară > 3	Distribuție generală CM2 cu variabila auxiliară > 4
G_1	63.00%	55.00%	70.00%	65.00%
G_2	20.00%	25.00%	18.00%	20.00%
G_3-G_5	17.00%	20.00%	12.00%	15.00%

Mediile prezentate în tabelul 4 au fost calculate în urma împărțirii cuvintelor în grupuri de complexitate şi făcând media indicatorului per grup. Întrucât fiecare indicator este raportat la împărțirea grupurilor în cuvinte şi complexitatea lor aferentă, procentele în sine nu sunt imbricate şi este normal ca suma lor să depăşească 100%.

Conform cu tabelul se pot observa diferite oscilații ale grupurilor G_1, G_2 şi G_3. Excluderea cuvintelor de legătura şi a pronumelor (variabila auxiliară >2), măreşte gradul de complexitate al textului, din mai multe cauze:

- Eliminarea cuvintelor de legătură afectează gradul de înțelegere al contextului;
- Totodată 90% din cuvintele <2 caractere sunt încadrate în grupul G_1 conform Manulex;
- Proporția de cuvinte complexe rămâne constantă;
- Sunt folosite astfel de eliminări când se doreşte identificarea cuvintelor cheie şi a elementelor care generează complexitate.

Complexitatea grupului G_1 revine în aceleaşi marje în momentul în care variabila auxiliară creşte deoarece se vor elimina şi cuvinte din grupurile G_2 şi G_3.

Observăm că în urma experimentului eliminările cu variabila auxiliară cea mai mare respectiv 4 converge către distribuția inițială. Astfel, pentru a concluziona în urma acestor rafinări , complexitatea unui text este dispersată conform cu elementele ce o compun. Totodată, eliminarea elementelor din dispersie ajută la localizarea elementelor care influențează major, dar fără a ține seama de conținut şi de auditoriul țintă analiza bazată exclusiv pe distribuția în grupuri de complexitate nu este suficientă. În continuare prezentăm analiza conținutului CM2, dar în funcție de indicii D şi SFI.

Indicatorii de dispersie D şi de indice standard de frecvență SFI prezentați în tabelul 5 sunt folosiți pentru a determina frecvența şi aria de răspândire a cuvintelor în corpus (vezi figura 2). Folosind cei doi indicatori putem să identificăm gradul de complexitate al unui cuvânt raportat atât la corpus, cât şi la gradul de izolare. Astfel, un cuvânt care are indicele D < 30% este prezent foarte rar în corpus, crescând complexitatea generală. De asemenea, o frecvență redusă a cuvântului creşte gradul de complexitate, deoarece sunt mai puține contexte în care acesta poate fi întâlnit şi totodată care pot fi asociate cuvântului.

Tabelul 5. Media Indicatorilor D şi SFI pentru grupul CM2

Grup	Procent Indicator D	Procent Indicator SFI
G_1	72%	67%
G_2	65%	53%
G_3-G_5	54%	29%

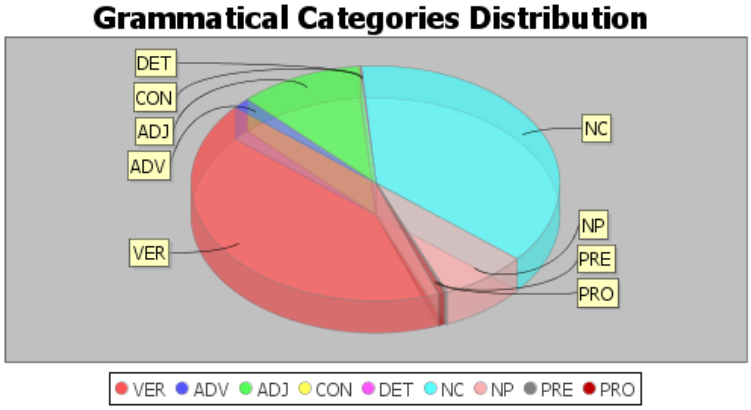

Figura 2. Exemplu de distribuția CM2 pe diverse părți de vorbire

6. Concluzii

În concluzie, lucrarea de față a avut drept scop urmărirea evoluției complexității utilizând atât factori generali, cât şi metrici adaptate manualelor sistemului şcolar francez. Rezultatele acestei lucrări pot fi măsurate folosind indicatori de complexitate al căror unic scop este de a adapta mesajele scrise şi orale la nivelul de înțelegere al audienței pentru a asigura o asimilare cât mai clară şi de durată a cunoştințelor. Totodată, drept extinderi ulterioare vizăm integrarea metricilor definite anterior în modelul multi-dimensional de analiză a complexității textuale propus de Dascalu et al. (2012) şi prezentat în detaliu în Dascalu et al. (2013).

Studiul de caz prezentat analizează evoluția complexității grupurilor în urma unor acțiuni de rafinare macro. Prin rafinarea macro se urmăreşte analiza complexității din punct de vedere metric, iar informațiile obținute trebuie să treacă prin toate procesele de prelucrare descrise mai sus pentru a putea extrage concluzii rafinate şi a putea face trecerea către analiza micro – analiza contextuală. Astfel o primă informație obținută în urma analizei macro reprezintă majoritatea grupului G_1 în corpusul CM2.

Majoritatea grupului G_1 este reflectată şi prin ceilalți indicatori studiați F,D,U şi SFI. Pornind de la aceste informații putem analiza în detaliu semnificația fiecărui indicator şi modul în care aceştia afectează complexitatea în caz real. Astfel indicii D şi SFI indică o frecvență mare a

cuvintelor în corpus, corelând cu apartenența cuvintelor la grupul G_1 putem trage concluzia că textul este în marea lui majoritate de complexitate redusă şi astfel adaptat clasei primare aferentă, respectiv clasa a patra din ciclul primar. Totuşi există elemente care indică şi prezența unor fragmente de complexitate ridicată, dar corelând din nou cu indicii D şi SFI care indică o frecvență mare şi ținând cont de contextul general al grupului G_1, putem afirma că nivelul de complexitate textuală converge către minimul categoriei G_3-G_5, respectiv G_3.

Analizând dinamica textuală a manualelor şcolare analizate şi folosind metodele descrise anterior, abordarea propusă a condus la următoarele concluzii:

- Grupul de complexitate mediu al textului scade dacă verbele sunt eliminate complet;
- Scăderea este treptată dacă verbele sunt eliminate folosind o variabilă auxiliară – lungime, unitate gramaticală;
- Numărul de verbe din text influențează în mod direct complexitatea; astfel, un dialog are complexitate mai mare decât un monolog, o narațiune sau o definiție de concept.

Un alt factor care ajută la accentuarea mesajului şi a evidențierii contextului este reprezentat de eliminarea cuvintelor puțin relevante de „stopwords". Metoda în sine nu reprezintă un factor determinant în analiza complexității, dar corelată cu alte tipuri de metrici ajută la identificarea componentelor cheie ale textului. Disjuncția şi prelucrarea grupurilor Manulex a fost un prim pas de clasificare a textelor şi totodată de asigurare a unicității cuvintelor. Utilizând grupuri disjuncte s-a putut face o analiză granulară a textelor, rezultatul fiind apoi prelucrat şi în funcție de cei factori F, U, D şi SFI cât şi a celorlalte metrici prezentate mai sus.

În altă ordine de idei, euristicele clasice ignoră imaginile dintr-un text, calculând un scor fără să țină cont dacă un text conține doar text sau fiecare paragraf are o imagine asociată. Imaginile au un rol foarte important în asimilarea şi asocierea noțiunilor mai ales pentru ciclurile şcolare primare. Un şcolar în clasa întâi va asimila mult mai uşor conceptul de primată dacă are asociată o imagine relevantă. Memoria vizuală şi corelațiile între informații reprezintă o modalitate de a face informația mai uşor de reținut pentru audiență. De asemenea timpul în care este reținută informația este mai scurt, iar perioada de timp în care noțiunea este reținută se măreşte.

Toate observațiile anterioare sunt luate în considerare drept potențiale direcții ulterioare de cercetare.

Referințe

Attali, Y. and Burstein, J., 2004. Automated essay scoring with e-rater V.2.0. In Annual Meeting of the International Association for Educational Assessment Association for Educational Assessment, Philadelphia, PA, 23.

Ball, H., Flight Out Of Time. University of California Press, Berkeley and Los Angeles, USA, 1996.

Chodorow, M. and Burstein, J., 2004. Beyond essay length: Evaluating e-rater's performance on TOEFL essays. Educational Testing Service.

Dascalu, M., Dessus, P., Trausan-Matu, S., Bianco, M., and 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.

Dascalu, M., Trausan-Matu, S., and 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.

Flesch, R., (1948) A new readability yardstick. Journal of Applied Psychology 32, (3), 221–233.

Gervasi, V. and Ambriola, V., 2002. Quantitative assessment of textual complexity. In Complexity in language and text, M.L. Barbaresi Ed. Plus, Pisa, Italy, 197–228.

Graesser, A.C., McNamara, D.S., and Kulikowich, J., (2011) Coh-Metrix: Providing multilevel analyses of text characteristics. Educational Researcher 40, (5), 223–234.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., and Cai, Z., (2004) Coh-Metrix: Analysis of text on cohesion and language. Behavioral Research Methods, Instruments, and Computers 36, (2), 193–202.

Halliday, M.A.K. and Hasan, R., Cohesion In English. Longman, London, 1976.

Jurafsky, D. and Martin, J.H., An introduction to natural language processing. Computational linguistics, and speech recognition. Pearson Prentice Hall, London, 2009.

Kincaid, J.P., Fishburne, R.P., Rogers, R.L., and Chissom, B.S., Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Chief of Naval Technical Training, Naval Air Station Memphis, 1975.

Lété, B., Sprenger-Charolles, L., and Colé, P., (2004) Manulex: A grade-level lexical database from French elementary-school readers. Behavior Research Methods, Instruments, & Computers 36, 156–166.

Macri, C.V., 2013. Dezvoltarea competențelor de citit-scris în ciclul primar prin utilizarea

strategiilor semi-globale în cadrul proiectelor tematice University Babeş-Bolyai, Cluj-Napoca, Romania.

Manning, C.D. and Schütze, H., Foundations of statistical Natural Language Processing. MIT Press, Cambridge, MA, 1999.

McNamara, D.S., Louwerse, M.M., McCarthy, P.M., and Graesser, A.C., (2010) Coh-Metrix: Capturing linguistic features of cohesion. Discourse Processes 47, (4), 292–330.

Powers, D.E., Burstein, J., Chodorow, M., Fowles, M.E., and Kukich, K., 2001. Stumping e-rater®: Challenging the validity of automated essay scoring. Educational Testing Service.

Tangkiengsirisin, S., 2012. Cohesion and Coherence In Text Language Institute, Thammasat University.

.