Rețele de analiză a relațiilor de intertextualitate

Ioan Cristian GHIBAN
Universitatea "Politehnica" Bucureşti
313 Splaiul Independenței, Bucureşti, 060032, Romania
ioan.ghiban@cs.pub.ro

Ştefan TRĂUŞAN-MATU
Universitatea "Politehnica" Bucureşti
313 Splaiul Independenței, Bucureşti, 060032, Romania
Institutul de Cercetări în Inteligență artificială
Caleqa 13 Septmebrie nr. 13, Bucureşti, România
stefan.trausan@cs.pub.ro

REZUMAT

Pentru a găsi particularități ale polifoniei bahtiniene, ale modului în care se influențează firele de discurs între ele, considerăm utilă abordarea intertextualității din perspectiva teoriei grafurilor şi a statisticii. Articolul de față prezintă modele teoretice, tehnici de prelucrare cu aplicațiile lor şi rezultate care certifică drept importantă abordarea bazată pe rețele a intertextualității. În final, sunt introduse două aplicații originale de analiză supervizată şi nesupervizată a textelor antice de natură filosofică şi religioasă. Ele sunt utile în detecția automată a influențelor intertextuale şi în valorificarea unor resurse textuale din Encliclopedia Catolică prin facilitarea procesului de supervizare al conținutului discursiv.

Cuvinte cheie

Intertextualitate, teoria rețelelor sociale, prelucrare limbaj natural

Clasificare ACM

H5.2. Information interfaces and presentation: Natural language

INTRODUCERE

În lingvistică, de obicei, pentru că abordările structuraliste se concentrează pe semne şi cel mult pe relațiile dintre ele, există un inconvenient cu privire la texte şi interdependențele lor. Textele erau considerate entități discrete şi de sine stătătoare pierzându-se imaginea de ansamblu chiar şi când analizele erau făcute pe corpusuri. Ținând cont de acest lucru, Julia Kristeva [5] a introdus noțiunea de intertextualitate probabil influențată de teoria polifonică a lui Bahtin [1].

Intertextualitatea reprezintă totalitatea modurilor în care un text presupune cunoaşterea altor texte [23]. În general, ea este acceptată cu privire la orice formă de discurs, implicând toate sistemele semiotice de reprezentare a cunoştințelor, inclusiv cel al textului scris. Fiecare autor al unui discurs foloseşte un anumit context de cunoştințe preluate de la alți autori. Rezultatul său va fi un „mozaic" de citate, de secvențe de discurs, articulate de noile cunoştințe ale autorului.

Conceptul filosofic de intertextualitate a influețat studiile făcute în domeniile căutării informațiilor (Information Retrieval – IR) şi, în consecință, Stubbs [22] precizează că textele sunt alcătuite în funcție de ce s-a spus în texte anterioare, iar analizele trebuie să cuprindă şi relațiile de intertextualitate. Odată cu apariția webului, alte tipuri de documente au ieşit la iveală: wiki, weblog, forum, chat, grupuri de ştiri, liste de mail şi altele, referite ca hipertexte.

În scopul extragerii automate a relațiilor de intertextualitate, a legăturilor dintre texte sunt necesare extragerea de cunoştințe, clusterizarea şi sumarizarea textelor.

„Ipoteza legătură-conținut" a lui Menczer [18] dovedeşte statistic că textul dintr-o pagină web este similar semantic conținutului din paginile care se leagă la el, iar această similaritate scade exponențial pe măsură ce accesăm linkurile pornind de la pagina centrală. Acest gen de studiu al intertextualității poate fi aplicat şi asupra comunicării comunităților de practică (e.g. comunități ştiințifice, comunități tehnice) dacă acestea prezintă un aparat bibliografic care inter-referențiază la fel ca linkurile web.

Când vorbim de intertextualitate este foarte important să nu restricționăm diversele legături dintre texte doar la cele binare. Analizele trebuie să ia în calcul întreaga rețea de relații.

Lucrarea de față va prezenta, mai întâi, aspecte teoretice şi practice cu privire la rețelele intertextuale (i.e. rețele de analiză a relațiilor de intertextualitate), iar în final, va descrie două implementări software de analiză a unor astfel de rețele, particularizate pentru texte cu conținut filosofic.

APARAT TEORETIC

Intertextualitatea are la bază legături de coeziune şi putem vorbi în funcție de acestea de intertextualitate referențială şi intertextualitate tipologică. Coeziunile tipologice [15] se bazează pe o tipologie comună (e.g. vocabular, gen, temă, adresă URL) ce caracterizează textele după clase, iar cele referențiale se bazează pe referințe între texte evidențiate în mod explicit (e.g. hiperlinkuri, citări). Dintr-o perspectivă web, nodurile rețelelor intertextuale pot fi şi ele analizate la nivele diferite de conceptualizare [15]: ca texte simple, ca e-texte încadrate într-o rețea ierarhică corespunzând adresei lor URL sau ca hipertexte când sunt luate în calcul hiperlinkurile.

Intertextualitate referențială

Studii din domeniul intertextualității referențiale au descris mai multe subdomenii dependente de metoda folosită în analiza legăturilor între texte. Până acum, referitor la legăturile bibliografice, există patru direcții [15]:

- Infometrica se ocupă cu teorii asupra modului în care informațiile sunt transferate în diverse tipuri de rețele. Această ramură este una generală, următoarele fiind mai specifice.
- Bibliometrica pune accent pe statisticile bibliografice din literatură şi pe citările dintre documente. Ea foloseşte rețele de citări pe baza legăturilor dintre documentele care citează şi documentele citate.
- Scientometrica, este un subdomeniu al bibliometricii centrat pe rețele de documente ştiințifice (e.g. articole, publicații la conferințe, referate, cărți etc.). Aceasta oferă suport pentru găsirea de articole cu o anumită tematică, de media, oameni de ştiință, invenții sau articole cu un impact mare în domeniu.
- Webometrica, aplică teoria bibliometricii asupra hipertextelor web luând drept citări hiperlinkurile dintre documente, aşa numitele *sitări*. Unele diferențe apar de exemplu datorită hiperlinkurilor bidirecționale.

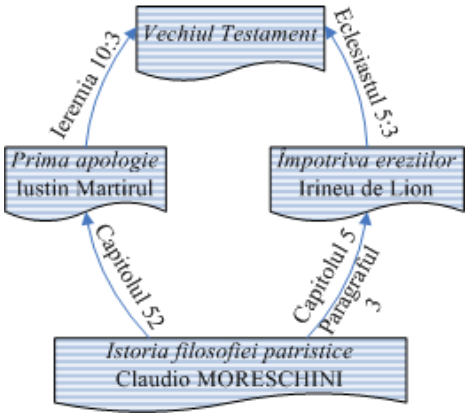

Figura 1. Bibliometrica intertextuală

Dacă două documente au cel puțin un document comun pe care îl citează împreună, scientometrica şi alte domenii asociate folosesc termenul de cuplare bibliografică (de exemplu, în Fig. 1 "Prima apologie" şi „Împotriva ereziilor" sunt cuplate bibliografic prin citarea comună a Vechiului Testament), iar dacă ambele sunt citate în acelaşi document atunci se numesc co-citate (în Fig. 1 "Prima apologie" şi „Împotriva ereziilor" sunt co-citate pentru că „Istoria filosofiei patristice" le citează pe ambele). Aceste două relații determină rețele de cuplări bibliografice (CB) şi respectiv rețele de co-citări (CC). La legăturile rețelelor pot fi atașate ponderi potrivit numărului de citări. În funcție de citările comune (în cazul CB) sau de documentul de citare în comun (în cazul CC) pot fi găsite clustere de apartenență. Ambele tipuri de rețele pot fi combinate într-o singură rețea şi dacă toate documentele în relații de co-citare sunt distincte de documentele în relație cuplaj bibliografic poate fi obținut un graf bipartit (Figura 2).

Se pot face analize ținând cont de ferestre de timp cuprinzând data de publicare a documentelor citate sau ale celor care citează. În acest fel, Redner [19] şi Raan au putut studia relațiile dintre co-citările apărute într-o anumită perioadă şi documentele care citau dintr-o perioadă mai târzie, sau invers, documentele cuplate bibliografic şi cele citate, aparținând unei perioade anterioare.

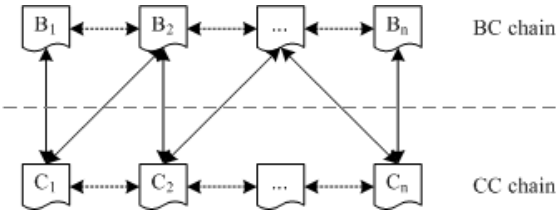

Figura 2. Rețele CB şi CC reduse la un graf bipartit

Intertextualitate tipologică

În intertextualitatea tipologică, sunt utilizate grafuri pentru a descrie relațiile dintre diversele unități lexicale sau conceptuale. Pentru construirea de sisteme lexicale precum grafurile de colocații este folosită procesarea nesupervizată, iar cu ajutorul expertizei unor lexicografi, folosind procesare supervizată, se pot construi ontologii terminologice (e.g. WordNet), tezaure (e.g. Roget) şi alte sisteme bazate pe grafuri care folosesc relații de sens (de exemplu, sinonimie sau polisemie) sau relații conceptuale (hiponimie, meronimie etc.). Prin urmare, există trei tipuri de grafuri lexicale supervizate:

1) Grafuri conceptuale folosind ca muchii relații conceptuale şi ca noduri concepte (e.g. synset-uri în WordNet)

2) Grafuri de tip tezaur având ca muchii relații de sens şi ca noduri cuvinte

3) Grafuri de asociere folosind cuvinte de pornire pentru descrierea nodurilor şi relațiile de asociere cu alte cuvinte pentru muchii. În aceste caz, experții lingviști folosesc unități lexicale ca puncte de plecare şi asociază sens sau cuvinte înrudite ca formă în funcție de anumite experimente cognitiv-lingvistice rezultând grafuri direcționate cu muchii de la cuvintele start către cele asociate.

În mod nesupervizat, se pot construi diferite tipuri de grafuri: grafuri de propoziții, grafuri de co-apariții, grafuri colocaționale şi alte tipuri. Primul referă atât noțiuni lexicografice cât şi propoziționale. Astfel unitățile lexicale se leagă între ele dacă apar în comun în cel puțin o propoziție (co-apariție). Muchia de legătură este direcțională indicând dependențe sintactice: sursa numită modificator (e.g. un substantiv manifestat ca subiect) şi ținta numită centru (e.g. un verb manifestat ca predicat cu modificatorul dat de subiect). În ce priveşte noțiunile propoziționale graful leagă propozițiile dacă acestea împart cel puțin două cuvinte de coeziune. Studii au fost realizate asupra acestui tip de grafuri de către Hoey [11].

Legat de construirea grafurilor colocaționale, în primul rând, utilizând metrici specifice [13], sunt găsite co-aparițiile interesante. Statisticile calculate ajută apoi la

înlănțuirea co-aparițiilor care partajează termeni cu proximitate semantică, iar în următoarea etapă se aplică filtre prag pentru a reduce co-aparițiile la colocații. De asemenea, cu aceste valori, se obține graful final structurat în jurul unui nucleu (*kernel*) şi al mai multor sociolecte marginale sau terminologii cu o tematică specifică. Acestea din urmă vor avea mai probabil unități legate direct cu cele din nucleu decât cu cele din propriul grup. În acest mod, de obicei, nucleul mediază căile dintre unitățile aceluiaşi grup sociolect cu ajutorul termenilor comuni cu rolul de *hub*-uri (i.e. noduri care asigură partea predominantă a coeziunii rețelei).

Utilizată frecvent în această construire de grafuri, Analiza Semantică Latentă (LSA) porneşte de la premisa că unitățile lexicale cu tematică, subiect sau domeniu similar partajează, cu probabilitate mare, termeni comuni sau înrudiți. Este folosit aici modelul spațiului vectorial al lui Salton [16] care reprezintă textele ca vectori de termeni, procesări asupra acestui spațiu ducând la obținerea de clase lexicale sau clustere şi rețele asociative. Ultimele sunt diferite de reprezentările intertextualității sub formă de înlănțuiri [15], obținute din ierarhii ordonate de texte (i.e. ținând cont de anumiți parametri de categorizare a textelor).

Prin hipertexte, în plus față de interpretarea legată de www menționată anterior, se poate de asemenea înțelege un text central şi rețeaua de alte texte înrudite. Potrivit metodei de căutare "lățime m-adâncime n" (*breadth m-depth n*) [16] a lui Salton, generarea de rețele hipertext pentru un text central este realizată pe baza LSA. Pornind de la textul central, se adaugă la rețea repetat, în n paşi, de fiecare dată câte m texte. La un pas oarecare k, aceste m texte sunt selectate în funcție de similaritatea lor cu textele adăugate anterior, la pasul k-1. Astfel, se iau cele mai similare m texte. Ele trebuie să fie şi în conformitate cu textul central, depăşind un anumit prag de similaritate cu acesta.

Din punctul de vedere al coeziunii, există o distincție între acest tip de hipertext şi rețelele asociative. De exemplu, să considerăm trei texte scurte A, B, C (e.g. A={u, v}, B={v, z}, C={z, w}) astfel încât A este înrudit cu B în anumiți termeni comuni pe care-i folosesc (e.g. $A \cap B = \{v\}$), B este înrudit cu C în alți termeni comuni decât cei prin care era B cu A (e.g. $B \cap C = \{z\}$) şi nu există nici un termen comun între A şi C (e.g. $A \cap C = \emptyset$). Rețeau asociativă înlănțuieşte A cu C prin B, dar A şi C nu sunt deloc înrudite rezultând o discontinuitate semantică printr-o modificare de subiect (*topic shifting*). Pentru a găsi costul căii între A şi C, ținând cont de proprietatea de intranzitivitate, nu se pot adăuga pur şi simplu costurile legăturilor de la A la B la C: A şi B sunt similare dintr-un alt motiv decât cel al similarității dintre B şi C. Aşadar, trebuie ținut cont şi de coeziunea dintre A şi C, iar ulterior de întreaga cale cu care C trebuie comparat. Vor exista două tipuri de legături: de coeziune (e.g. de la A la C) şi asociative, rezultate din analiza LSA (e.g. de A la B sau de la B la C). În acest scop, modelul prezentat de hipertext al lui Salton face un pas înainte în rezolvarea problemei. El adaugă un nou nod în rețea nu numai bazându-se pe coeziunea cu ultimele m texte adăugate, ci şi relativ la textul central, ca prim nod al oricărei căi la nodul curent. Se poate merge şi mai departe cu estimarea prin calcule relative la alte noduri din căile posibile dar cu inconvenientul unui consum crescut de resurse. Drept consecință noi estimatori sunt necesari pentru reducerea riscului de ne-coeziune de-a lungul căilor, iar Mehler [16] propune un nou model.

Modele Small World

În concordanță cu legea lui Zipf [30], analizând statistic distribuția cuvintelor în discursuri (i.e. texte coezive şi coerente), frecvența unui cuvânt este invers proporțională rangului său. Aceasta înseamnă că într-un text neartificial dacă probabilitatea celui mai frecvent cuvânt ar fi 0.06, cuvântul cu al doilea rang va fi utilizat în 0.03 dintre cazuri (i.e. 0.06 / 2), al treilea în 0.02 (i.e. 0.06 / 3) ş.a.m.d., această distribuție conturând doar una din condițiile pe care un text natural trebuie să le îndeplinească. De asemenea, alte restricții de naturalețe limitează nu numai textele individuale, ci şi relațiile lor intertextuale.

Topologiile şi caracteristicile statistice ale rețelelor intertextuale sunt dependente de genurile de discurs ale documentelor în studiu (e.g. citările într-o comunitate ştiințifică leagă documentele diferit de modul în care o fac linkurile hipertext în wiki-uri). Începând cu Milgram [17], au fost făcute studii încercând să unifice proprietățile rețelelor intertextuale sub aşa numita proprietate *Small World* (SW) (e.g. aceasta este manifestată şi de rețelele de socializare). Ea oferă metrici care diferențiază grafurile aleatorii de cele intertextuale naturale, neartificiale, ținând cont de existența (a) unui nivel de clusterizare mult mai ridicat şi (b) a unei distanțe mai mici între două noduri alese la întâmplare într-un graf intertextual.

Un model SW a fost propus de Watts şi Strogatz [25] (modelul WS), maximizând coeficientul de clusterizare CWS (a) şi minimizând distanța geodezică medie L (b):

$$\text{(a)} \quad C_{WS}(G) = \frac{1}{n}\sum_{i=1}^{n} C_{v_i}(G) \in [0,1] \tag{1}$$

$$\text{(b)} \quad L(G) = \frac{1}{\binom{|V(G)|}{2}} \sum_{\{v,w\}\in[V]^2} \delta(v,w) \tag{2}$$

$$\text{unde} \quad C_{v_i}(G) = \frac{adj(v_i)}{\binom{d_G(v)}{2}} = \frac{adj(v_i)}{d(v_i)(d(v_i)-1)/2} \in [0,1] \tag{3}$$

este clusterizarea nodului vi, adj(vi) este numărul de muchii între vecinii lui vi, d(vi) este gradul nodului vi şi δ este distanța între v şi w. În consecință, CWS indică probabilitatea ca vecinii unui nod v ales la întâmplare să fie vecini între ei, iar L indică cât de repede se schimbă o variabilă precum tematica sau genul documentului de-a lungul legăturilor (e.g. Wikipedia are un L mic şi tematica se schimbă repede prin accesarea linkurilor).

În generarea automată a unei SW folosind acest model trebuie să se pornească de la un graf regulat, dar această particularitate poate fi considerată un dezavantaj (puține

rețele intertextuale pot fi aproximate cu un graf regulat). Din acest motiv alte modele statistice încearcă să potrivească mai bine modelul natural al rețelelor intertextuale și, de exemplu, Barabási și Albert [2], luând legea lui Zipf ca paradigmatică, propun modelul de atașare preferențială (modelul BA) pentru a descrie dinamica creșterii rețelei intertextuale.

$$P(k) \sim k^{-\gamma} \quad (4)$$

Distribuția expune că un nod selectat la întâmplare va avea un grad k cu o probabilitate $k^{-\gamma}$, unde de cele mai multe ori $\gamma \in [1.5,3.5]$.

ANALIZA DINAMICII LIMBAJULUI

Există multe studii cu privire la aceste domenii de intertextualitate referențială sau tipologică, iar în interesul acestui articol se află exemple de unelte softuri pentru analiza bazată pe rețele a intertextualității.

Sistemul HSCM

Având ca scop găsirea dinamicii utilizării limbajului într-o anumită societate de-a lungul unor perioade de timp ce se pot extinde chiar până la milenii, ca o consecință a cercetărilor din domeniu axate în special pe metode probabilistice [12] (e.g. gramatici probabilistice sau analize statistice ale colocațiilor), dar care desconsideră procesele pe termen lung implicate de dinamica limbajului, o echipă germană (incluzând istorici și ingineri software, A. Mehler fiind unul dintre ei) a implementat în 2007 un sistem numit *Historical Semantics Corpus Management System* (HSCM). El ia ca intrare corpusuri ordonate cronologic (studiile au fost făcute pe "Patrologia Latina") și găsește modificările semantice ale cuvintelor (în mod statistic cu privire la co-apariții) în texte diacronice, dând indicii asupra schimbărilor sociale care ar fi putut avea loc între anii în care textele au fost scrise.

HSCM combină un sistem de interogare a unui corpus cu un lematizator de Latină orientat pe liste și generează rapoarte cu variate posibilități de personalizare asupra folosirii cuvintelor în texte (e.g. diagrame referitoare la matricea termen-document, diverse tipuri de filtre sau de afișare a ieșirilor). Sistemul găsește co-aparițiile determinate de cuvântul specificat sau de fraza de interogare și oferă prin diagrame de frecvență indicații de stabilizare, marginalizare și transformare a modelelor lingvistice utile pentru domeniul istoric (Figura 3).

Sistemul de interogarea corpusurilor necesită corpusuri din *Patrologia Latina Database* (PLD) folosind anumite credențiale (accesul la această bază nu este gratuit pentru toate resursele) și primește texte de intrare într-un format PLD SGML-DTD. După aceasta se realizează o conversie în format PLD XML-DTD având ca rezultat un fișier PLD XML. Apoi se face o conversie la standardul *Text Encoding Initiative* (TEI P5) (i.e. definește diverse taguri pentru descrierea componentelor textuale precum cuvinte, propoziții, paragrafe ș.a.m.d.) astfel încât textele PL devin conforme acestui standard. În final se segmentează structura textelor astfel procesate. Acest ultim pas ține cont de structura logică a documentului (LDS), găsește părțile de vorbire (PoS), lemele și realizează o recunoaștere a entităților numite (Figura 4).

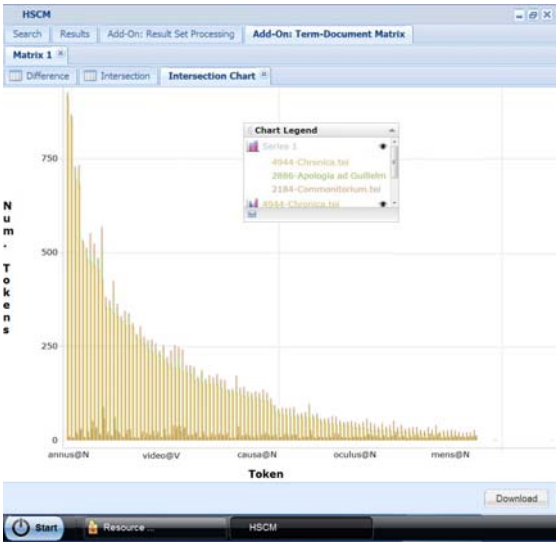

Figura 3. Frecvența cuvintelor în trei texte analizate cu HSCM [27]

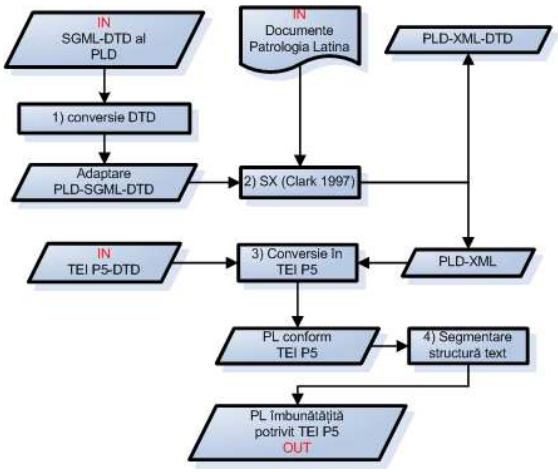

Figura 4. Sistemul de interogare de corpus HSCM [14]

Procesare bazată pe nori semantico-sociali

Studii ulterioare asupra dinamicii limbajului au fost făcute cu ajutorul procesării bazate pe nori prin intermediul câtorva sute de noduri de procesare într-o structură eterogenă (nodurile nefiind toate de același tip) disponibilă în Frankfurt. Proiectul este numit "*Social-Semantic Cloud Computing*" și este interesat de originea, dezvoltării și proliferării artefactelor semantice [21] în cadrul rețelelor sociale descentralizate (e.g. Wikipedia). El dă totodată indicii referitoare la proprietățile SW sau de dezvoltare ale unor astfel de rețele pe perioade lungi de timp prin simularea ciclului lor de viață.

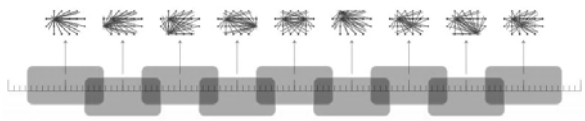

Figura 5. Dinamica rețelei sociale [28]

Până acum, rețelele intertextuale au fost examinate pe fragmente, dar folosind acest sistem ele pot fi investigate holistic, iar referitor la dinamica rețelei, imaginea nu mai este limitată doar la cadrele anumitor momente de timp.

Rețele lingvistice

Un alt sistem, numit "*Linguistic Networks*" [29] furnizează unelte web pentru generarea de rețele lexicale şi de propoziții asupra corpusurilor. Nodurile şi legăturile dintre ele sunt construite prin intermediul unor abordări sintactice (i.e. părți de vorbire), semantice (i.e. înțeles) sau pragmatice [14] (i.e. înțeles în context).

Sistemul este proiectat să funționeze pe corpusuri distincte (e.g. "*Patrologia Latina*") cu limba lor specifică şi oferă diverse tipuri de rețele (e.g. de cuvinte, leme, propoziții sau texte). El furnizează alte detalii precum lema cuvântului în dicționar, în ce formă (cea dinainte de stemming) sau cu ce PoS a fost cuvântul întâlnit, cele mai frecvente co-apariții, vecini stânga sau dreapta (i.e. cuvintele poziționate înainte sau după cuvântul curent în aceeaşi colocație cu el) filtrate sau nu după rangul frecvenței. De asemenea, afişarea rețelei poate fi personalizată prin etichete ale ponderilor pe muchii, culori distincte ale nodurilor în funcție de PoS sau diverse tipuri de structurare ale grafurilor (e.g. circular, organic, ierarhic). Sunt disponibile şi alte grafice cu vecinii: nori de cuvinte şi distribuția vecinilor (rang vs. semnificație).

CÂTEVA REZULTATE

Rezultate interesante au fost obținute prin cercetările făcute de Glänzel şi Czerwon [9] cu privire la intertextualitatea referențială, în special asupra rețelelor CB. Între documentele analizate câteva erau documente nucleu, dispunând de un număr de legături mai mare decât media. Cercetările au identificat în acest set de documente publicații neobservate pentru o perioadă lungă de timp, iar apoi, brusc, citate intens, aşa numitele "sleeping beauties". Studiind rețelele CB aceştia au găsit o modalitate de a descrie teme de cercetare într-un mediu ştiințific folosind clustere de documente. Nu după mult timp, odată ce metoda LSA a devenit mai răspândită, documentele înrudite tematic erau clusterizate printr-o astfel de metodă. Alte studii au fost demarate de către Small [20] în privința rețelelor CC şi a modalității în care discipline distincte se inter-referențiază prin citări inter-disciplinare.

Cercetări în domeniul intertextualității tipologice au fost făcute pe *British National Corpus* prin intermediul grafurilor de co-apariții şi colocații de către Ferrer I Cancho şi Solé [6] urmate de Widdows şi Dorow [26]. Prima echipă a construit grafurilor de co-apariții legând cuvinte ce apar în aceeaşi propoziție şi în cadrul unei ferestre de trei tokeni. A doua echipă a aplicat în plus restricții dependențelor sintactice (i.e. referitoare la PoS). Ca rezultate, ei au observat că ambele tipuri de grafuri (co-apariții şi colocații) se potrivesc proprietății de SW în termenii propuşi de modelele WS şi BA. De exemplu, ei au găsit o distribuție de tip putere cu exponentul $\gamma \in [1.5,3.5]$ pentru primele 5,000 de noduri după gradul de conectivitate, respectând aşadar modelul BA. Pentru grafurile colocaționale, o distribuție de probabilitate logaritmică a fost raportată asupra unor articole de ştiri germane de către Bordag et al. [4].

Studii au fost efectuate şi asupra grafurilor de propoziții. Ferrer I Cancho et al. [7] au folosit propoziții cehe, germane şi române cu adnotări care specificau interdependențele lor. Aceştia au raportat că şi grafurile propoziționale sunt conforme modelului WS şi BA. Mai mult, a fost observată o particularitate numită combinație dezasortativă (*disassortative mixing*): cuvintele cu un grad mare de conectivitate (e.g. substantive uzuale, cuvinte funcționale etc.) au tendința să fie legate cu cele slab conectate.

Referindu-se la conectivitate între hipertextele de tip wiki, Bianconi & Barabasi [3] au găsit conformitate cu ceea ce se numeşte efectul Matthew: două hipertexte au probabilitate mai mare să se conecteze dacă cel puțin unul din ele este mai evident (i.e. are o publicitate sau conectivitate mai mare) şi dacă lemele lor prezintă coeziune. Evidența unui articol a_i se numeşte *fitness* şi este descrisă ca $\eta(a_i) = \eta_i$. Probabilitatea P_i ca un articol nou a_j să se lege la un articol a_i de grad k_i şi fitness η_i este:

$$P_i = \frac{\sigma(a_i, a_j)}{\sum_n \sigma(a_i, a_n)} \frac{\eta_i k_i}{\sum_n \eta_n k_n} \quad (5)$$

unde σ reprezintă coerența între lemele celor două articole.

Pe lângă hipertexte wiki pe web, a fost considerată şi blogosfera (întreaga rețea a blogurilor). Pentru a construi un corpus cuprinzând în jur de 100.000 de bloguri, Glance et al. [8] au folosit un program de recoltare de URL-uri, un crawler blog care indexa bloguri, un aliniator de timp pentru variațiile fusului orar la scrierea pe bloguri şi software de text mining care întoarce orientarea tematică. Folosind un astfel de corpus, Herring et al. [10] a ajuns la concluzia că şi rețelele blog au proprietăți ale SW – de exemplu ataşamentul preferențial datorat tendinței utilizatorilor de a-şi lega blogul la altele deja puternic conectate. Accentul în studiul lui Herring a fost pe diadele blog (i.e. perechi de bloguri mutual conectate), pe "interacțiunea textuală" dintre ele prin schimbul verbal similar unei "conversații" între bloggeri. Concluzia sa a fost că acest tip de legături mutuale este rar, întâlnit în special în clustere mici de bloguri puternic conectate. Mai mult, Herring caracterizează legăturile în blogosferă ca fiind mai degrabă rare şi de aici dificultatea găsirii pentru bloguri a unei distribuții de probabilitate de tip putere. Astfel, Tricas [24] a propus o lege de distribuție folosind exponentul $\gamma \approx 0.58$ care se potriveşte parțial.

În legătură cu discontinuitatea semantică şi cu modelul lui Mehler de a rezolva această problemă, au fost făcute teste [16] pe un corpus de aproximativ 500 de ştiri în germană. Modelul a dat rezultate mai bune decât metoda clasică a

arborilor de acoperire minimă (MST). De exemplu, pornind de la un articol despre un meci de fotbal care a fost difuzat la TV, articolele cele mai similare sunt adăugate gradual potrivit celor două metode: MST şi cea a lui Mehler rezultând doi arbori de coeziune.

Folosind MST primele rezultate sunt similare arborelui de coeziune al lui Mehler, dar în final se ajunge la subiecte diferite de cele din articolul rădăcină şi nu este găsit nici un articol referitor la difuzarea TV. Arborele de coeziune al lui Mehler reuşeşte să clusterizeze textele pe două ramuri: prima legată doar de fotbal şi a doua legată şi de difuzarea TV.

Tabelul 1. Rezultatele clusterizării folosind MST şi modelul Mehler. Subiecte din articole: Fotbal, Alte subiecte, difuzare TV

MST			Modelul lui Mehler
Subiect			
Nun braucht der FC Bayern ei…	F	F	Nun braucht der FC Baye…
└Klinsmann erlöst die Bayern …	F	F	└Klinsmann erlöst die Ba…
└Tapferer Optmismus nach …	F	F	└Tapferer Optmismus n…
├Münchner Geschenke zu …	F	F	├Münchner Geschenk…
└Schaumschläger im Super…	F	F	└Schaumschläger im …
├Ansturm auf die Klagem…	F	F	├Ansturm auf die Kla…
└Erst Badener Lied, dann ...	F	F	├Erst Badener Lied, d...
└Keine Fortschritte im Fl…	F	F	│ └1:1 bei Feyenoord ...
└1860 und die Not ...	F	F	│ └Mehmet Scholl (F...
├Wirklich kein Anlaß …	F	F	│ └Kleine Fortschritt...
│ └Im Landeamflug a…	A	F	│ └Wirklich kein A..
│ └Massenkarambol…	A	F	│ └1860 und die ...
├Borussia Dortmund	F	F	│ └Borussia Dor...
└Meister über Meiers…	F	F,TV	└Taktik des Tauschens
└Peter Neururer Iös…	F	F,TV	└Premiere will [...] C...

Proiectul "*Linguistic Networks*", după cum am prezentat anterior, crează reţele lexicale şi propoziţionale pe baza corpusurilor. Rezultatele sunt folositoare pentru determinarea unor artefacte semantice noi care înlocuiesc în timp pe cele vechi, obţinând astfel dinamica lexicală a conceptelor reprezentate de artefacte. În reţelele generate, artefactele vor apărea drept noduri vecine, iar cele noi vor deveni mai importante odată cu trecerea timpului (Figura 6).

În acelaşi scop, rezultate utile pot fi calculate cu sistemul HSCM. Cu ajutorul său, istoricii pot compara lucrări scrise în secole diferite pentru a observa cum folosirea artefactelor lexicale evoluează în configuraţii diverse.

De exemplu, există o ipoteză larg acceptată [12] că termenul latin "ordo" (i.e. rang, clasă, ordin [31]) ar fi fost foarte important în perioada medievală pentru a exprima concepte legate de lume şi structura sa socială. De asemenea, există lucrări, precum "De Civitate Dei" a lui Augustin, care au influenţat puternic operele de după ele cu privire la conceptualizarea lui "ordo". HSCM poate da indicii despre folosirea lui "ordo" în opera de secol V a lui Augustin şi în alte opere religioase mai târzii, precum cele scrise în secolul XII (Figura 7).

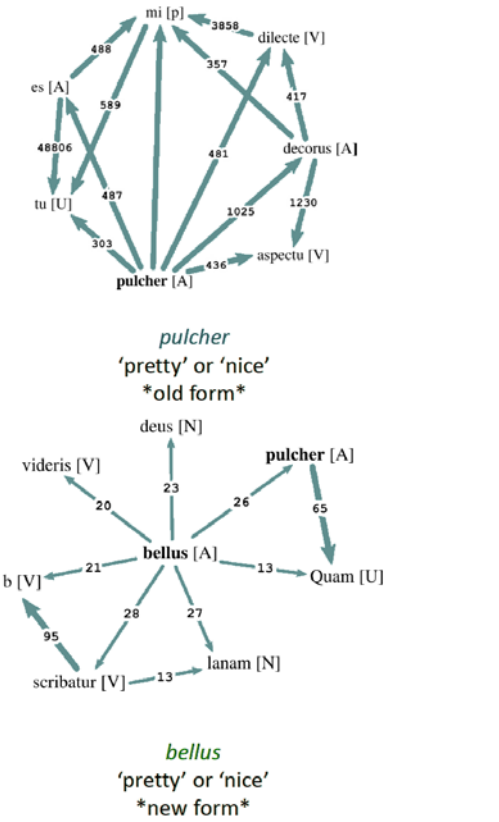

Figura 6. Exemplu de dinamica a limbajului – adjectivul "bellus" înlocuieşte treptat "pulcher" pentru a se exprima "frumos" [14]

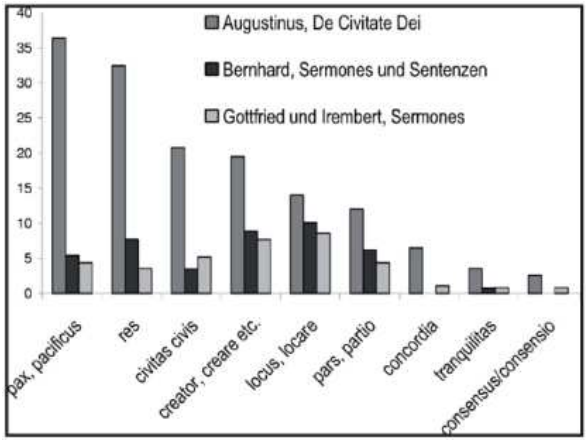

Figura 7. Analiza folosirii lui ordo în trei corpusuri (unul de secol V şi două de secol XII) prin intermediul colocaţiilor [12]

Pentru figura de mai sus, HSCM a găsit în opera lui Augustin colocaţiile de două cuvinte conţinând "ordo". Apoi, ia din aceste colocaţii celelalte cuvinte ca ţintă (e.g. "pax", "res") şi găseşte toate colocaţiile lor. După o

ordonare în funcție de procentul colocațiilor conținând “ordo” în toate colocațiile cuvintelor țintă se fac calcule şi asupra celorlalte două scrieri mai târzii. Se observă că “ordo” nu mai apare aşa de des alături de “pax, pacificus” sau „res”, iar istoricii pot trage concluzii referitoare la aceste detalii.

ANALIZA INTERTEXTUALĂ A TEXTELOR ANTICE

Pentru a studia intertextualitatea am implementat o aplicațe folosind o tehnică nesupervizată (LSA) şi o alta pentru a ajuta experții umani în procesul de analiză supervizată.

Datorită disponibilității online, în vederea analizei nesupervizate, am ales şi noi, ca şi cei de la HSCM cu Patrologia Latina, corpusuri reprezentând scrieri antice de natură filosofică, anume Platon, Aristotel, Epicur şi alți filosofi ai antichității precum şi scrieri de factură religioasă, limitându-ne la patristica creştină (începând cu secolul II până în secolul VIII). Alegerea făcută a avut ca motiv şi prezența unei critici îndelungate cu privire la influențele filosofiilor antice asupra scrierilor patristice. Astfel am putea dispune de informații supervizate referitoare la relațiile de intertextualitate în cadrul corpusului de analizat.

S-au accesat 17 situri care publică gratis texte de acest gen (e.g. *gutenberg.org*, *tertullian.org*, *newAdvent* etc.), am descărcat aproximativ 1800 de fişiere corespunzând diverselor secțiuni în cadrul scrierilor şi s-a alcătuit un corpus de texte (cărți, epistole, discursuri, imnuri) pentru care critica a manifiestat interes.

În construirea corpusului, deoarece formatele textelor difereau, pentru a ajunge la un format comun, cât şi pentru a păstra structura textelor pe secțiuni (aceasta fiind utilă pentru că indică secvențe de text cu acelaşi subiect), a fost necesară implementarea unei aplicații de conversie (în Python). Din cele 17 situri, pentru a ne permite o lărgire ulterioară a corpusului local, am ales 4 pentru care aplicația permite conversia directă a fişierelor html, recunoscând după taguri structura textului. Pentru fisierele din celelalte situri s-a folosit formatul txt, structura scrierilor fiind păstrată prin recunoaşterea titlurilor cu ajutorul expresiilor regulate. În final au rezultat 280 de cărți în format TEI P5.

Procesare de text bazată pe LSA

Am implementat în Python o aplicație de consolă care găseşte cele mai frecvente cuvinte, cele mai semnificative din punct de vedere semantic şi de asemenea similaritățile dintre documente folosind asupra lor un spațiu vectorial ca cel al lui Salton. Implementarea procesează fişiere conținute în unul sau mai multe directoare. Utilizatorul poate grupa fişiere în aceste directoare în funcție de o anumită proprietate (e.g. acelaşi autor). Fiecare fişier conține paragrafe şi se poate considera ca document întregul conținut al fişierului sau un singur paragraf (aceasta este utilă pentru matricea de frecvență a termenilor).

Procesarea documentului (scris în engleză) este făcută cu ajutorul mediului NLTK. Se porneşte cu recunoaşterea de cuvinte folosind expresii regulate urmată de filtrarea cuvintelor stop, tokenizare cu corpusul Brown (secțiunea religie), stemming folosind stemmer-ul Porter şi WordNet. În tokenizarea cuvintelor au fost folosite trei tagere: trigram (găseşte PoS-ul pe baza cuvântului curent şi a precedentelor două), bigram şi unigram. Dacă nu se găseşte nici un rezultat cuvântul este etichetat drept “Substantiv”. Aplicația va reține doar cuvintele de interes: substantive, adverbe, verbe şi adjective.

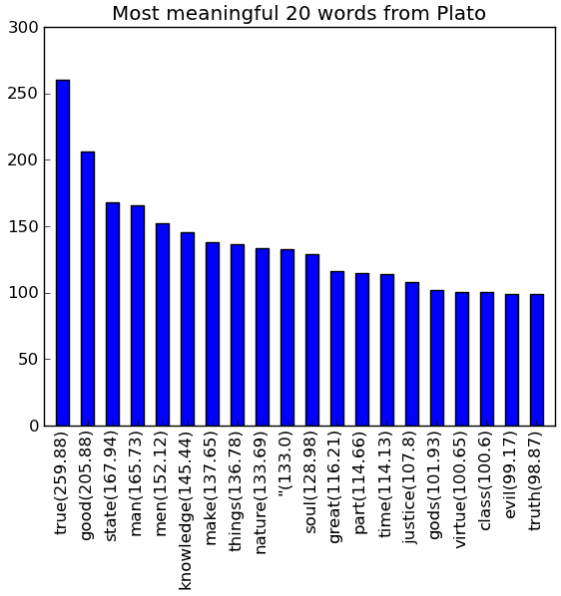

Figura 8. Cele mai semnificative 20 de cuvinte de la Platon

Ca filtrare de cuvinte, după ce matricea de frecvență a termenilor a fost calculată, aplicația va reține doar cuvintele care au cel puțin 5 apariții în oricare din documente. Apoi, este folosită metoda TF-IDF iar în final descompunerea în valori singulare folosind funcția svd disponibilă în librăria Numpy.

Pentru a câştiga timp la procesarea textelor de către utilizator, aplicația salvează rezultatele în patru paşi: documentele parsate, spațiul lexical – după calcularea matricei de frecvență a termenilor, spațiul semantic – după descompunerea în valori singulare şi ultimul pas dat de salvarea statisticilor finale sub forma dorită de utilizator. Astfel procesările ulterioare vor putea începe de la un pas deja efectuat.

Rezultatele finale apar ordonate după scor sub forma unor liste de perechi {element statistic, scor} (i.e. elementul statistic poate fi un cuvânt sau un text). De exemplu, pentru primele 2 texte similare cu cele ale lui Platon scrise de alți autori avem rezultatul:

```
[Plato|TheApology, Justin|TheSecondApology-
(0.6475); Plato|TheRepublic.7,
Irenaeus|AgainstHeresies.6-(0.6095)]
```

Adică Apologia lui Platon cu “A doua Apologie” a lui Iustin au un scor de 0.6475 fiind cele mai similare două lucrări, iar capitolul 7 din Republica lui Platon cu capitolul 6 din „Împotriva ereziilor” a lui Irineu este pe locul secund. Rezultatele sunt disponibile şi grafic.

Limbajul de implementare este Python 2.7 şi a fost folosit NLTK 2.0b9 pentru WordNet, corpusul Brown (religie), stemmer-ul Porter, tokenizare şi etichetare.

Testarea a fost făcută pe operele Sf. Iustin Martirul şi Filosoful (103-165), ale Sf. Irineu de Lyon (a doua jumătate a secolului al II-lea), Origen (185-254) şi lucrările lui Platon. Comparând documentele Sf. Iustin cu operele lui Irineu şi Platon, în primele 10 cele mai similare lucrări au fost opt scrise de Sf. Irineu şi două de Platon: "Cratylus", similară cu "A doua apologie" a lui Iustin şi "Lysis" similară cu "Discurs către greci" a lui Iustin.

În prima şi a doua apologie Sf. Iustin a scris păgânilor, iar în a doua (fiind adresată grecilor) foloseşte mulţi termeni uzuali mediului filosofic elen: "adevăr", „viciu", „virtute", „bine", „suflet", „filosof". De asemenea, Iustin încearcă în "Discurs către Greci" să convingă că religia creştină este pentru întreaga umanitate şi foloseşte din nou termeni greceşti specifici.

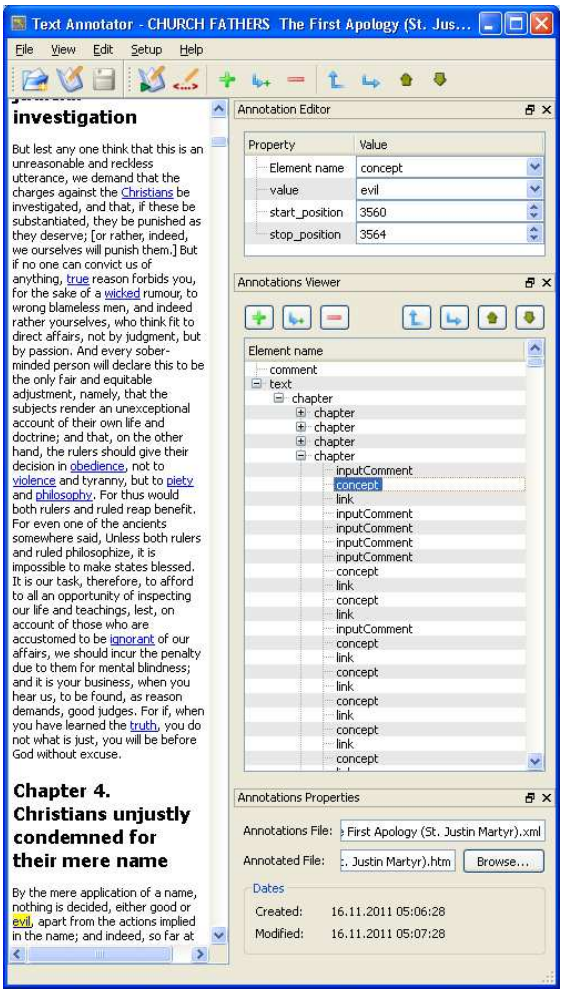

Figura 9. Rezultate în "Text Annotator" după procesarea "Prima apologie" scrisă de Sf. Iustin Martirul şi Filosoful

Text Annotator

Multe scrieri patristice pot fi găsite la adresa *http://www.newadvent.org/fathers/*. Resursele postate pe acest sit sunt hipertexte care includ diverse tipuri de comentarii (utilizand taguri specifice) asupra textului original. Aceste comentarii etichetează cuvinte-cheie sau sunt ancore către detalii şi se pot dovedi utile în găsirea conceptelor din text.

Astfel, era necesară o aplicaţie care să extragă comentariile din situl menţionat anterior. "Text Annotator" (Figura 9) a fost implementat pentru a ajuta utilizatorii în adnotarea pe html sau fişiere txt folosind un set dat de comentarii. El este configurat de asemenea, în funcţie de tagurile html găsite, să adnoteze automat documentele descărcate de la situl specificat (i.e. documente cu formatul *www.newadvent.org*).

Textele de intrare sunt de obicei structurate în cărţi, capitole, paragrafe şi alte elemente de interes pentru utilizator şi, folosind acest format de imbricare, aplicaţia construieşte un arbore de adnotări. După aceasta, se poate insera, muta sau şterge o adnotare şi strânge substructurile pentru a gestiona uşor arborele.

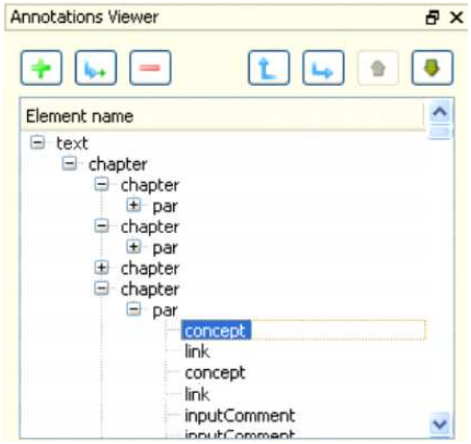

Figura 10. Arborele de adnotări rezultat

Fiecare adnotare poate fi editată manual prin schimbarea valorii atributelor.

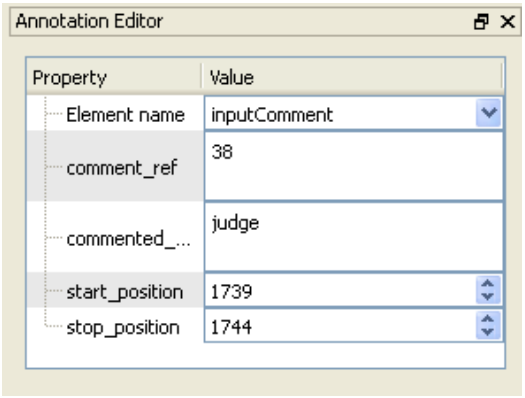

Figura 11. Detalii pentru adnotarea selectată

Totodată atributele însele pot fi modificate (în termeni de nume şi tip) prin modificări în fişierul de configurare Elements.xml. De aici se pot adăuga noi tipuri de adnotări cu atributele şi valorile lor posibile.

Aplicația este implementată în Python 2.7 şi câteva din bibliotecile folosite sunt: html5lib, libxml2, libxslt, PyQt4.

Direcții de cercetare viitoare

Îmbunătățiri ale aplicației de procesare bazate pe LSA pot fi făcute pentru a găsi citate. În scrierile antice, pasajele preluate erau de multe ori diferite de textul inițial la nivelul mai multor termeni. Aceste diferențe apar de exemplu datorită traducerilor făcute de-a lungul timpului (de aici şi celebra expresie „traduttore, traditore"). O căutare automată de astfel de pasaje, care nu ține cont de semantica textului şi caută doar reproduceri fidele ale unor înşiruiri de cuvinte, va rata găsirea multor citări.

Mai pot fi făcute îmbunătățiri în algoritmul de găsire a afilierii la un curent filosofic pentru un text dat. Antrenând aplicația pe intrări care se cunosc că aparțin diferitelor curente, ar fi utilă găsirea de expresii specifice unui anumit curent. De exemplu, să se găsească expresiile uzuale stoicilor dar nu şi epicureicilor (e.g. "rațiuni seminale"). Aceste rezultate ajută la catalogarea mai precisă a unui text dat spre analiză.

În privința "Text Annotator" este utilă adăugarea unui modul de procesare a limbajului natural care să ofere informații sintactice cu privire la conceptele găsite. De exemplu, la stadiul actual, aplicația diferențiază conceptul "*sacrificed*" de „*sacrificing*", dacă cele două cuvinte ar apărea drept concepte în hipertextul de intrare. Prin găsirea rădăcinilor cuvintelor s-ar putea realiza o statistică a conceptelor găsite şi s-ar putea compara cu rezultatele din aplicația LSA.

REFERINȚE

1. Bahtin, M.M., *Problemele poeticii lui Dostoievski*. Ed. Univers, 1970
2. Barabási, A.-L., Albert, R., and Jeong, H. , *Scale-free characteristics of random networks: The topology of the World Wide Web*. Physica A, 281:69–77. 1999
3. Bianconi, Ginestra; Barbasi, Alberto – Laszlo, "*Bose-Einstein condensation in Complex Networks*", In: Physical Review Letters, 86(24); 5632-5635
4. Bordag, S., Heyer, G., and Quasthoff, U. (2003). *Small worlds of concepts and other principles of semantic search*. In Unger, H. and Böhme, T., editors, Innovative Internet Computing Systems Second International Workshop (IICS '03), Berlin. Springer
5. Daniel Chandler, *Semiotics for Beginners*, http://www.aber.ac.uk/media/Documents/S4B/sem09.html, 06.12.2011
6. Ferrer i Cancho, R. and Solé, R. V. (2001). *The small-world of human language. Proceedings of the Royal Society of London*. Series B, Biological Sciences, 268(1482):2261–2265
7. Ferrer i Cancho, R., Solé, R. V., and Köhler, R. (2004). *Patterns in syntactic dependency-networks*. Physical Review, E(69):051915.
8. Glance, N., Hurst, M., and Tomokiyo, T. (2004). BlogPulse: *Automated trend discovery for weblogs*. In WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics
9. Glänzel, W. and Czerwon, H.-J. (1996). *A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level*. Scientometrics, 37(2):195–221
10. Herring, S. C., Kouper, I., Paolillo, J. C., Scheidt, L. A., Tyworth, M., Welsch, P., Wright, E., and Yu, N. (2005). *Conversations in the blogosphere: An analysis "from the bottom up"*. In Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)
11. Hoey, M., *Patterns of Lexis in Text*. Oxford University Press, Oxford
12. Bernhard Jussen et al., *A Corpus Management System for Historical Semantics*, Sprache und Datenverarbeitung. International Journal for Language Data Processing 31(1-2), 2007
13. Manning, C. D. and Schütze, H., Foundations of *Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts
14. M. Alexander Mehler et al., *Time Series of Linguistic Networks by Example of the Patrologia Latina*
15. M. Alexander Mehler, *Large Text Networks as an Object of Corpus Linguistic Studies*
16. M. Alexander Mehler, *Text minning with the help of cohesion trees*, in Wolfgang Gaul et al., Classification, automation, and new media: proceedings of the 24th annual, 2002, Springer
17. Milgram Apud M. Alexander Mehler, *A network perspective on intertextuality*, in Peter Grzybek et al., Exact methods in the study of language and text, 2007, Mouton de Gruyter series
18. Menczer F., *Lexical and semantic clustering by web links*. Journal of the American Society for Information Science and Technology, 55(14):1261–1269, 2004
19. Redner, S. (1998). *How popular is your paper? An empirical study of the citation distribution.* European Physical Journal, B(4):131–134
20. Small, H. (1999). A passage through science: Crossing disciplinary boundaries. Library Trends, 48(1):72–108.
21. Gerry Stahl, Computer Support for building collaborative knowledge, MIT Press, 2006
22. Stubbs M., *Words and Phrases. Corpus Studies of Lexical Semantics*., Blackwell, Oxford, 2001
23. Sanna-Kaisa Tanskanen, *Collaborating towards Coherence*, John Benjamins Publishing Company, Amsterdam, 2006
24. Tricas, F., Ruiz, V., and Merelo, J. J. (2004). *Do we live in a small world? Measuring the Spanish-speaking blogosphere*. In Blogtalk 2.0, June 5-6, Wien
25. Watts, D. J. and Strogatz, S. H., *Collective dynamics of 'small-world' networks*. Nature, 393:440–442, 1998
26. Widdows, D. and Dorow, B. (2002). *A graph model for unsupervised lexical acquisition*. In 19th International Conference on Computational Linguistics, August 24 – September 1, 2002, Taipeh, Taiwan
27. Prezentare a HSCM. http://www.hucompute.org/ressourcen/historical-semantics

28. Procesarea bazată pe nori. http://www.hucompute.org/ressourcen/cloud-computing
29. Site-ul proiect "*Linguistic Networks*". http://www.linguistic-networks.net/
30. Detalii cu privire la legea lui Zipf. http://linkage.rockefeller.edu/wli/zipf/
31. Dictionar englez-latin. http://www.freedict.com/onldict/lat.html