UNELE ASPECTE DE EXPLOATARE A DOCUMENTELOR WEB ÎN SCOPURILE MORFOLOGIEI COMPUȚATIONALE

Mircea Petic
Institutul de Matematică şi Informatică al AŞM
mirsha@math.md

Abstract. *The paper deals with the computational derivational morphology mechanisms which permit automatic acquisition of the lexical resources for the Romanian language. In this context, the possibilities of derivational sequence construction using Web documents are subject to review. The article ends with several approaches in automatic new generated words validation.*

Cuvinte-cheie: *documente Web, resurse lingvistice computaționale, prefix, sufix*

I. Introducere

Dicționarele moderne se confruntă cu unele deficiențe, devenind astfel obiecte de cercetare pentru lexicografi. Avînd în vedere că dicționarele sînt permanent completate cu intrări noi, grație dezvoltării limbii, sarcina elaborării unui vocabular complet rămîne una practic imposibilă. Completarea automată sau/şi semiautomată a resurselor lingvistice computaționale cu cuvinte generate în baza celor deja existente reprezintă o sursă importantă de îmbogățire a vocabularului prin mijloace exclusiv interne, în cazul acestui articol va fi studiată derivarea lexicală.

Soluționarea problemei de derivare automată necesită efectuarea unor studii preliminare, care ne-ar permite să deducem ulterior anumite legități referitoare la comportamentul afixelor în limba română. Pentru efectuarea acestor studii sînt necesare resurse lingvistice computaționale corespunzătoare. Cea mai reprezentativă resursă lingvistică computațională constă în documentele Web disponibile pe Internet. În această ordine de idei scopul acestui articol constituie studierea unor aspecte de exploatare a documentelor Web în scopurile morfologiei computaționale, şi anume: exploatarea documentelor Web la stabilirea ordinii de derivare a cuvîntului şi validarea derivatelor generate automat.

II. Exploatarea documentelor Web la stabilirea ordinii de derivare a cuvîntului

Dat fiind faptul, că un cuvînt poate fi derivat cu mai multe afixe, este semnificativă ordinea în care un derivat s-a obținut, adică ordinea în care au fost alipite afixele. Procesul de stabilire a ordinii de derivare a cuvîntului constă în identificarea morfemelor constituente, formarea derivatelor posibile din aceste părți constituente, stabilirea posibilelor derivate valide prin verificarea prezenței lor în documentele electronice existente pe Internet cu un motor de căutare (în cazul nostru fiind aplicat cel oferit de Google).

În continuare vor fi examinate mai multe exemple pentru care se va cerceta ordinea de derivare a cuvîntului.

Exemplul 1. Cuvîntul *antimuncitoresc* este un derivat valid pentru limba română. Acesta este format dintr-un prefix (*anti-*), o rădăcină ((a) *munci*) şi două sufixe lexicale simple (*-tor* şi *-esc*). Astfel pornind de la această structură a cuvîntului este posibil de format mai multe derivate, şi anume: muncit*or*, muncitor*esc*, *anti*munci, *anti*muncit*or* şi *anti*muncit*oresc*. Pentru a stabili care din aceste derivate pot fi considerate valide, s-a verificat prezența acestor cuvinte derivate în documentele electronice existente pe Internet cu motorul de căutare Google.com (Tabelul 1.) [1].

Tabelul 1. Date pentru derivatul antimuncitoresc.

Derivatul	Numărul de pagini găsite
muncitor	601000
muncitoresc	65300
antimunci	0
antimuncitor	1
antimuncitoresc	125

Datele din Tabelul 1 sugerează ideea că derivarea, pornind de la rădăcina ***munci*** pînă la obținerea cuvîntului ***anti***munci***toresc***, se efectuează în următoarea succesiune de transformări:

$[\text{munci}]_{\text{verb}} \rightarrow [\text{muncitor}]_{\text{subst, adj}} \rightarrow [\text{muncitoresc}]_{\text{adj}} \rightarrow [\text{antimuncitoresc}]_{\text{adj}}$.

Arborele de derivare pentru cuvîntul ***anti***munci***toresc*** este prezentat în Figura 1.

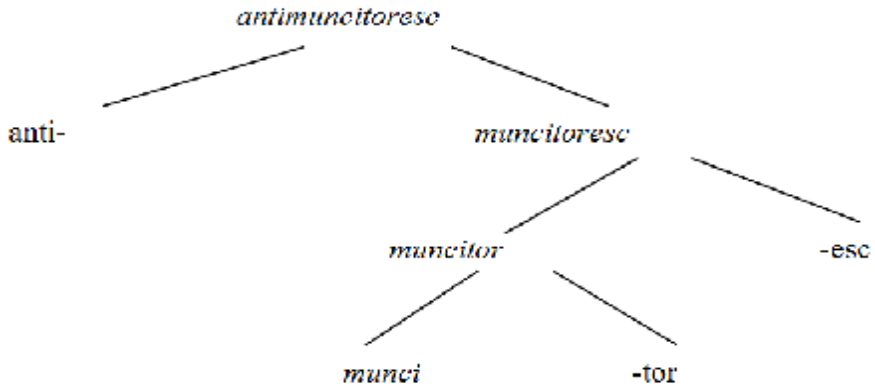

Figura 1 Arborele de derivare pentru cuvîntul antimuncitoresc

Arborele de derivare din Figura 1 este unic şi nu are alternative.

Exemplul 2. În cazul cuvîntului derivat ***descentralizator*** situația este alta. Derivatul este format dintr-un prefix (***des-***), o temă (***central***) şi două sufixe lexicale (***-iza*** şi ***-tor***). Astfel pornind de la această structură a cuvîntului este posibil de format mai multe derivate, şi anume: central***iza***, centraliza***tor***, ***des***central, ***des***central***iza*** şi ***des***central***izator***. Pentru a stabili care din aceste derivate pot fi considerate valide, s-a procedat ca şi în cazul precedent, adică s-a verificat prezența cuvintelor derivate în documentele electronice existente pe Internet (Tabelul 2.).

Tabelul 2 Date pentru derivativul descentralizator

Derivatul	Numărul de pagini găsite
centraliza	21000
centralizator	792000
descentral	0
descentraliza	5520
descentralizator	1260

Din datele de mai sus se vede că de la verbul format central***iza*** este posibil de a obține două derivate: centraliza***tor*** şi ***des***centraliza. De la acestea din urmă, putem forma cu uşurință, cuvintele derivate centralizator şi descentralizator (Figura 2) [1]. S-a constatat că pentru cuvîntul derivat ***des-***

centralizator pot fi doi arbori de derivare. Dincolo de exemplele analizate, mai există un caz ce diferă de aceste două [1].

Exemplul 3. Un caz aparte îl vor constitui derivatele semianalizabile, de felul *redeschidere* şi *reînchidere*. Ambele, la prima vedere, sînt formate din două prefixe (*re-* şi *des-*/*în-*), o rădăcină (*chide*) şi un sufix lexical (*-re*). De fapt, nu este atestată o astfel de rădăcină, de aceea drept rădăcină pentru analiză se vor lua temele deschide/închide. Astfel, pentru a decide structura arborelui de derivare, s-a procedat ca şi în cazul cuvintelor derivate anterior, verificîndu-se prezența cuvintelor derivate în documentele electronice existente pe Internet. Datele statistice sînt prezentate în Tabelul 3.

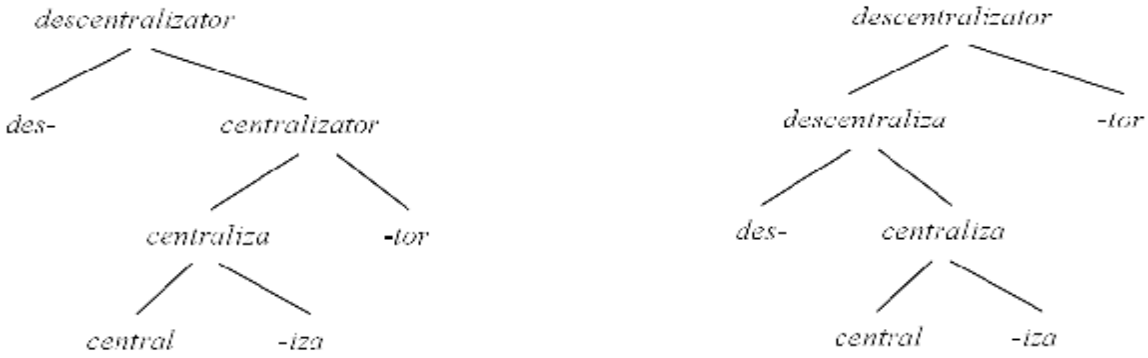

Figura 2 Arborii de derivare pentru derivatul descentralizator

Aşadar, succesiunea de derivare va fi:

$[deschide]_{verb} \rightarrow [deschidere]_{subst} \rightarrow [redeschidere]_{subst}$ sau

$[deschide]_{verb} \rightarrow [redeschide]_{verb} \rightarrow [redeschidere]_{subst}$.

Ambele variante sînt posibile, dar, ținînd cont de datele din Tabelul 3, preferabilă este prima variantă. În cazul cuvîntului *reînchidere*, se obține:

$[închide]_{verb} \rightarrow [închidere]_{subst} \rightarrow [reînchidere]_{subst}$ sau

$[închide]_{verb} \rightarrow [reînchide]_{verb} \rightarrow [reînchidere]_{subst}$.

La fel, prima variantă este mai des folosită.

Tabelul 3 Datele statistice pentru cuvintele redeschidere/reînchidere.

Derivatul	**Numărul de pagini găsite**
redeschidere	13700
redeschide	117000
deschidere	3920000
reînchidere	1150
reînchide	529
închidere	686000

În rezultat s-a constatat că, în caz general, arborele de derivare nu este unic, acest proces fiind posibil de realizat pe căi diferite. Acest rezultat a fost luat în considerare la elaborarea algoritmilor de procesare a dicționarelor în scopul identificării cuvintelor derivate. În acest sens se evidențiază utilitatea verificării prezenței cuvintelor derivate în documentele electronice pe Internet, exercițiu care este util în procesul de validare a cuvintelor derivate generate automat.

III. Validarea derivatelor generate automat

Derivatele noi, care sînt generate prin mijloace automate, ar trebui să fie corecte din punct de vedere morfologic şi semantic. Unul din procedeele de validare a derivatelor constă în examinarea manuală a fiecărui cuvînt generat, în corespundere cu cerinţele regulilor morfologice şi semantice. Garantînd calitatea rezultatului (în cazul cînd procedeul este efectuat de către un specialist în domeniu) ne confruntăm cu dezavantajele specifice unui lucru manual: resurse considerabile de timp, precum şi posibilitatea comiterii unor erori [2].

Ideea de validare automată a fost inspirată de sistemul *GederiF* [3], pentru limba franceză. În cazul acestui sistem validarea automată a cuvintelor derivate generate s-a realizat prin verificare în corpusuri (*Encyclopedia Universalis* şi *La Banque des Mots*), precum şi cu motorul de căutare (*Yahoo.com*).

Aşa cum nu dispunem de multe (şi nu prea reprezentative) corpusuri pentru limba română, deci utilizarea resurselor acumulate în Internet devine destul de importantă. Evident validarea nu se poate reduce doar în verificarea prezenţei cuvintelor derivate în documentele electronice existente pe Internet. Este necesar de precizat faptul, că verificarea în Internet trebuie să se realizeze doar pentru documentele în limba română. În acest caz însă ne confruntăm cu o serie de dificultăţi. Chiar utilizînd opţiunea cu privire la limba setată este posibil să se găsească cuvinte în alte limbi. Este cazul cuvintelor *maciza* (limba spaniolă), *bariza* (limba arabă), *neautomobil* (limba cehă), *nemonolit* (limba croată) care au fost găsite de către motorul de căutare *Google* în cazul căutării cuvintelor respective pentru limba română. În plus, apar deficienţe create de o eventuală segmentare a cuvîntului căutat. Astfel, de exemplu, la încercarea de a valida în acest mod verbul fataliza s-a găsit fraza „...o fată, Liza...", în loc de cristianiza s-a găsit Cristian Iza. Există şi cuvinte, care reprezintă substantive proprii, în particular denumiri de companii, validitatea cărora trezeşte dubii, de exemplu, SRL „Daniza" şi SRL „Cariza" găsite de către motorul de căutare, nu pot confirma validitatea verbelor omonime generate automat.

Dincolo de cele expuse mai sus apare şi dificultatea stabilirii condiţiilor în care un cuvînt este valid. S-ar părea, că numărul de apariţii ale cuvîntului ar fi un criteriu obiectiv. De exemplu, pentru cuvîntul *catiza* s-au găsit mai multe intrări, dar cu greu s-ar găsi argumente pentru a-l valida.

Mai pot apărea cazuri, în care se formează un cuvînt derivat cu alt sens, de exemplu, *negros* format de la adjectivul gros, ar trebui să aibă sensul „subţire". În DEX există un astfel de cuvînt *negros* dar are alt sens:„brun, brunet, negricios" [1].

Chiar dacă se ţine cont de cele enumerate mai sus, mai există un criteriu care nu pare să fie obiectiv, anume numărul de apariţii ale cuvîntului în paginile Web. Apare întrebarea: care ar putea fi acest număr rezonabil de acceptanţă a derivatelor generate? Am putea presupune că acest număr ar trebui să fie de la derivate la derivate diferit.

Astfel apare ideea stabilirii relaţiei între numărul de pagini Web găsite, rădăcina şi afixul de derivare, pe care o vom numi *indice de popularitate*. Acest indice îl putem defini prin următoarea formulă:

$$I(a,r) = \frac{N(a,r)}{N(r)}, \qquad (1)$$

unde r – rădăcina derivatului,

ar – cuvîntul derivat cu afixul a şi rădăcina/tema r,

$N(r)$ – numărul de pagini în limba româna găsite pe Internet pentru cuvîntul r,

$N(a,r)$ – numărul de pagini în limba română găsite pe Internet pentru cuvîntul derivat,

$I(a,r)$ – indicele de frecvenţă a afixului a pentru o rădăcină r concretă.

De exemplu, pentru *r=redeschide, a=re, ar=redeschidere, N(r)=117000, N(a,r)=13700, I(a,r)=N(a,r)/N(r)=13700/117000=0.117094*.

Astfel, calculînd *indicele de popularitate* conform formulei (1), putem observa atît derivate mai frecvente ca rădăcina, cît şi mai puţin frecvente. Totodată, acest indice ar putea servi pentru

validarea cuvintelor cu afixe, punînd condiția ca indicele de popularitate să nu fie mai mic decît indicele minimal pentru derivatele deja validate. De exemplu ținînd cont de datele din Tabelul 4 s-ar lua valoarea 0,00246, drept cel mai mic indice pentru prefixul *ne-*.

In Tabelul 4 este calculat indicele de popularitate pentru prefixul *ne-*. Se observă, că chiar şi pentru 5 derivate, diferența între valorile indicelui diferă mult. Astfel, raportul dintre indicele de popularitate pentru cuvintele neabordabil şi neaccesibil este de circa 211. Acest lucru nu poate permite ca indicele de popularitate să fie considerat în calitate de parametru al unui afix sau derivat.

Tabelul 4. Indicele de popularitate pentru prefixul ne-.

Derivatul	***N(r)***	***N(ar)***	***I(a)***
neabil	*247000*	*85800*	*0,34737*
neaplicabil	*138000*	*7190*	*0,05210*
neabordabil	*7970*	*4140*	*0,51945*
neacceptabil	*356000*	*1930*	*0,00542*
neaccesibil	*671000*	*1650*	*0,00246*

Din alt punct de vedere, în rețeaua Internet pot fi găsite documente neverificate, din acest motiv ele nu sînt complet credibile. Pentru a mări gradul de credibilitate ar trebui să ne asigurăm că ele sînt într-adevăr în limba română, să excludem segmentările, să cunoaştem partea de vorbire a cuvîntului găsit, etc [4].

În această ordine de idei, folosind posibilitățile unui motor de căutare (*Google.com*), a fost elaborată o aplicație Web, scrisă în limbajul PHP, care extrage numărul de pagini în limba româna găsite pe Internet pentru cuvîntul derivat solicitat. Utilitatea aplicației constă în faptul că poate funcționa nu doar cu un cuvînt, dar cu o listă de cuvinte, ceea ce economiseşte timpul în procesul de validare. Prin urmare, în baza numărului de pagini în limba româna găsite pe Internet pentru un cuvînt, este posibil de a divide derivatele generate în trei categorii. Prima categorie conține cuvintele care nu sînt găsite pe Internet. A doua categorie constă din derivatele care apar mai puțin de o limită de frecvență *n*, în cazul nostru am stabilit în mod empiric *n=1000*. Derivatele cu o frecvență mai mare ca *n* sînt înregistrate în al treilea grup. Vom admite, că cuvintele cu frecvența mai mare ca *n*, sînt valide. Cele, care se conțin în al doilea grup, pot fi considerate valide după ce au fost verificate de specialistul lingvist. Derivatele care nu se regăsesc pe Internet nu vor fi considerate valide [5]. Ideea clasificării pretinde a fi o metodă mixtă de validare, deoarece necesită şi verificarea manuală pentru cuvintele din categoria a doua de cuvinte.

Drept exemplu pentru aplicarea tehnicii de validare a fost considerat derivarea cu sufixul *–ime*. În urma aplicării algoritmului de generare automată a adjectivelor cu sufixul *–ime* a fost obținută o listă de 2841 de posibile derivate. În urma verificării frecvenței acestor cuvinte doar 120 au o frecvență nenulă, ceea ce constituie 4,22%. Tabelul 5 ilustrează rezultatele înregistrate în procesul de validare.

Datele prezentate în Tabelul 5 demonstrează faptul că nu există o metodă universală de validare cu ajutorul documentelor Web, dar permite filtrarea unui număr semnificativ de cuvinte derivate, restul necesitînd o verificare manuală. În cazul sufixului *–ime*, au fost înregistrate 38 de cuvinte derivate valide din 84 ceea ce constituie 45%, in cazul celor cu frecvență mai mare de 1000, acuratețea este de 83%. Deşi acuratețea e mai ridicată la creşterea frecvenței de apariției a cuvîntului totuşi asta nu garantează rezultatul perfect, aşa cum documentele Web nu sînt sigure din punct de vedere lingvistic. Totuşi este o posibilitate de a reduce considerabil timpul în procesul de validare a cuvintelor generate automat.

Tabelul 5. Datele statistice referitor la derivatele cu sufixul *–ime* validate.

Valoarea lui n	*Numărul de derivate generate*	*Numărul de derivate valide*	*Acuratețea de validare*
0	*2721*	*0*	*0%*
1 – 100	*52*	*25*	*48%*
101 – 1000	*32*	*13*	*41%*
1001 – 10000	*5*	*1*	*25%*
10001 - ...	*31*	*29*	*93%*
TOTAL 1 - ...	*120*	*68*	*57%*

IV. Concluzii

Resursele Web constituie un suport important pentru cercetare a problemelor din domeniul procesării limbajului natural şi anume la soluționarea problemelor de morfologie derivațională. Dat fiind faptul, că un cuvînt poate fi derivat cu mai multe afixe, este semnificativă ordinea în care un derivat s-a obținut, adică ordinea în care au fost alipite afixele, pentru depistarea unor eventuale modele de derivare, folosite ulterior la generarea automată a cuvintelor. În acest sens se evidențiază utilitatea verificării prezenței cuvintelor derivate în documentele electronice pe Internet, exercițiu care este util în procesul de validare a cuvintelor derivate generate automat.

Procesul de validare a derivatelor generate automat este unul care ridică mai multe întrebări şi dificultăți. Totuşi este imposibil de a neglija aspectul de credibilitate documentelor Web în procesul de validare automată. În acest context validarea cuvintelor folosind corporă existentă pare a fi cea mai bună soluție. Oricum momentan nu există o corporă universală şi reprezentativă care ar permite evitarea folosirii documentelor Web.

V. Referințe

1. Cojocaru S., Boian E., Petic M. Stages in automatic derivational morphology processing. In: Knowledge Engineering, Principles and Techniques, KEPT2009, Selected Papers. Cluj-Napoca: Presa Universitară, 2009, pp. 97-104.

2. Grigoriadou M. The Software Infrastructure for the Development and Validation of the Greek Wordnet. In: Romanian Journal of Information Science and Technology, vol. 7, Numbers 1-2, 2004, pp. 89-105.

3. Fiammetta N., Dal G. GéDériF: Automatic generation and analysis of morphologically constructed lexical resources. Second International Conference on Language Resources and Evaluation (LREC). Athens, Greece, May 31 – June 2, 2000, pp. 1447–1454.

4. Petic M. Automatic derivational morphology contribution to Romanian lexical acquisition. In: Special issue: Natural Language Processing and its Application. Research in Computing Science. Mexico, vol. 46, 2010, pp. 67-78.

5. Petic M. Developing a derivatives generator. In: Computer Science Journal of Moldova, vol. 18, nr. 1(52), 2010, p. 82-96.