Revista Română de Interacţiune Om-Calculator 4(2) 2011, 109-130

Identificarea automată a afixelor româneşti. Studiu de caz: identificarea sufixelor

Verginica Barbu Mititelu

Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română – ICIA Bucureşti
Calea 13 Septembrie, nr. 13, 050711, Bucureşti
E-mail: vergi@racai.ro

Rezumat. Din perspectiva identificării automate a cuvintelor derivate şi a bazelor lor de derivare în wordnetul românesc, în vederea îmbogăţirii acestuia cu relaţii derivaţionale şi etichete semantice asociate acestora, prezentăm în acest articol rezultatele unui studiu de caz al cărui scop a fost identificarea automată a sufixelor cu care se formează cuvinte în limba română. O prezentare succintă a fenomenului derivării aşa cum se manifestă el în română deschide acest articol, anticipând şi provocările pe care le întâmpină studiul nostru. Pentru identificarea automată a sufixelor în cuvinte am folosit arborii de sufixe generalizaţi sub forma cărora am reprezentat lemele dintr-un lexicon electronic românesc. Am impus o serie de filtre, stabilite prin metoda observaţiei, cu care am îmbunătăţit rezultatele. Pentru evaluarea sufixelor identificate automat am folosit liste de referinţă de sufixe, grupate în funcţie de partea de vorbire a cuvântului obţinut prin derivare. Evaluarea rezultatelor am făcut-o în trei etape, prin comparare doar cu lista de sufixe, apoi cu lista de sufixe îmbogăţită cu sufixoide şi, în final, cu lista unificată a sufixelor şi sufixoidelor pentru părţile de vorbire care prezintă omonimie în română. Am prezentat felul în care precizia şi acurateţea algoritmului variază în funcţie de limita impusă asupra productivităţii sufixelor. Studiul de faţă pune în evidenţă ansamblul de cunoştinţe necesare pentru recunoaşterea structurii morfematice a cuvintelor, dificultăţile întâmpinate în acest proces, dar şi importanţa acestui studiu pentru cercetările lingvistice, pentru domeniul Inteligenţei Artificiale, în sarcini precum regăsirea informaţiei, rezumarea textelor, răspunsul automat la întrebări şi toate celelalte care se bazează pe prelucrarea limbajului natural.

Cuvinte cheie: sufix, derivare, identificarea radicalului cuvintelor, analiză morfematică, arbore de sufixe generalizat.

1. Introducere

Vocabularul (i.e. totalitatea cuvintelor) unei limbi se modifică de-a lungul timpului, iar explicaţia rezidă în schimbările care se petrec în viaţa oamenilor, a căror comunicare ţine pasul cu aceste modificări.

Compartimentul limbii cel mai susceptibil de a suferi variaţii în timp este vocabularul. Cuvintele existente se pot perima, îşi pot schimba sau îmbogăţi sensurile, iar alte cuvinte pot apărea. În lingvistică se vorbeşte despre două modalităţi de îmbogăţire a vocabularului: o modalitate internă (reprezentată de derivare, compunere şi schimbarea categoriei gramaticale) şi alta externă (împrumutul); alături de ele mai există calcul lingvistic, un procedeu mixt. Acestea sunt principalele mijloace de îmbogăţire a vocabularului, alături de ele existând şi altele, considerate secundare şi pe care le lăsăm deoparte aici, întrucât obiectul preocupărilor noastre îl formează derivarea.

Aceasta se defineşte ca procedeul de formare a unui cuvânt nou având ca bază un alt cuvânt existent în limbă (Marouzeau, 1933: 63 ş.u.). Are două subtipuri: derivarea progresivă sau propriu-zisă şi derivarea regresivă. Cea dintâi constă din adăugarea unor afixe (sufixe sau prefixe) la un cuvânt bază. În funcţie de poziţia din cuvântul bază în care se ataşează afixul, vorbim despre prefixare (afixul, i.e. prefixul, se ataşează înaintea bazei, de ex. *străbun*<prefixul *stră-*+*bun*) sau sufixare (afixul, i.e. sufixul, se ataşează la sfârşitul bazei, de ex. *scriitor*<*scrie*+sufixul *-tor*). Derivarea regresivă constă în suprimarea unor afixe de la un cuvânt bază (de ex., *deranj*<*deranja*).

Derivarea este studiată de lingviştii interesaţi de evoluţia vocabularului, de identificarea afixelor unei limbi şi de caracterizarea acestora sub diverse aspecte (etimologie, sens, variaţii fonetice, productivitate etc.).

Derivarea prezintă interes şi pentru inginerii din domeniul prelucrării limbajului natural. Sarcini precum regăsirea informaţiei (engl. *information retrieval*), răspunsul automat la întrebări (engl. *question answering*) şi altele folosesc informaţii despre cuvintele care aparţin aceleiaşi familii lexicale (cuvinte derivate de la aceeaşi bază cu afixe diferite). Asemenea informaţii se obţin prin reducerea cuvintelor derivate la baza lor, prin înlăturarea tuturor afixelor conţinute; dacă în urma acestei înlăturări se obţine acelaşi şir de litere, atunci cuvintele de la care s-a pornit au mari şanse de a aparţine aceleiaşi familii lexicale[1] şi sunt tratate împreună. În unele limbi, printre ele şi româna, în acest şir de litere sunt permise alternanţe vocalice şi/sau

[1] Cele două cuvinte nu aparţin aceleiaşi familii lexicale dacă bazele lor nu reprezintă acelaşi cuvânt, ci sunt omonime.

consonantice. De exemplu, cuvântul *băieţel* este derivat de la baza *băiat* cu sufixul *-el*; rădăcina prezintă atât o alternanţă vocalică (ia:ie), cât şi una consonantică (t:ţ).

În acest articol ne-am propus să investigăm în ce măsură este posibilă identificarea automată a afixelor limbii române pornind de la un inventar de cuvinte la formele lor de bază şi supunându-le unui proces de analiză şi filtrare a rezultatelor.

Teoretic, putem folosi acelaşi algoritm (cu adaptări impuse doar de poziţia din cuvânt asupra căreia ne concentrăm, i.e. început sau sfârşit) pentru identificarea prefixelor şi sufixelor. Practic, însă, am constatat că rata de eroare este mult mai mare în cazul în care încercăm să identificăm prefixele. Acest lucru ne-a obligat să ne rezumăm în acest articol la un studiu de caz, şi anume posibilitatea identificării automate a sufixelor în cuvinte.

Pentru început, se impun câteva precizări terminologice. O distincţie importantă este aceea între rădăcină (engl. *root*) şi radical (engl. *stem*). Rădăcina unui cuvânt este segmentul comun tuturor cuvintelor care formează o familie lexicală. Este un morfem indivizibil, minimal, purtător de sens lexical. De exemplu, în cazul familiei de cuvinte *pădure*, *pădurice*, *pădurar*, *păduros*, *împăduri*, *împădurire*, *împădurit*, *reîmpăduri*, *reîmpădurire*, *reîmpădurit* rădăcina este *pădur*, comună tuturor membrilor familiei. Radicalul se defineşte ca acea parte a cuvântului în structura căreia se regăseşte rădăcina plus/minus toate afixele derivative. Astfel, în cazul membrilor familiei lexicale de mai sus, avem următorii radicali: *pădur* (identic cu rădăcina, deci) în cazul lui *pădure*, *păduric* (rădăcină plus sufix) în cazul lui *pădurice*, *pădurar* (rădăcină plus sufix) în cazul lui *pădurar*, *păduros* (rădăcină plus sufix) în cazul lui *păduros*, *împădur* (prefix plus rădăcină) în cazul lui *împăduri*, *împădurir* (prefix plus rădăcină plus sufix plus sufix) în cazul lui *împădurire* şi aşa mai departe. În concluzie, în cazul cuvintelor nederivate, rădăcina şi radicalul sunt identice, iar în cazul cuvintelor derivate, radicalul conţine rădăcina şi toate afixele derivative.

În Prelucrarea Limbajului Natural (PLN) se face distincţia între lematizare (engl. *lemmatization*) (i.e. identificarea formei de bază, de dicţionar a unui cuvânt flexionat) şi identificarea rădăcinii cuvintelor (engl. *stemming*). De exemplu, de la şirul de caractere *sfinţeniei* obţinem prin lematizare *sfinţenie*, iar prin steming rezultatul trebuie să fie *sfânt*, deoarece acestuia i s-a ataşat sufixul *-enie* pentru a forma *sfinţenie*. În cazul

cuvintelor nederivate, precum *sfântului*, ambele procese dau acelaşi rezultat: atât prin lematizare, cât şi prin stemming obținem *sfânt*.

2. Sistemul derivativ românesc

În limba română principalul mijloc intern de formare a cuvintelor noi este derivarea, iar dintre subtipurile sale cel mai frecvent întâlnit este sufixarea, după cum atestă lucrări ale specialiştilor din momente diverse din evoluția limbii (Puşcariu 1940, Hristea 1978, Avram 1989, Stoichițoiu-Ichim 2007). Prefixarea nu a fost foarte productivă în trecut, însă în cercetările care se concentrează asupra ultimelor decenii autorii constată o creştere a productivității acestui fenomen lingvistic, explicabil, printre altele (după părerea noastră), prin dorința sau nevoia vorbitorilor de concizie în comunicare.

Prefixele se ataşează la începutul unui cuvânt direct sau prin modificarea corpului lor fonetic, prin fenomene de asimilare (de exemplu, *-în* devine *-îm* în contextul unei consoane labiale) sau disimilare (de exemplu, *des-* devine *de-* în *desăra*). Ele nu modifică structura fonetică a bazei.

Sufixele se ataşează la sfârşitul cuvântului bază direct sau în urma unor fenomene fonetice petrecute în corpul bazei: eliminări de sunete (de ex. *bucur<u>a</u>>bucuros*), alternanțe vocalice sau/şi consonantice (de ex. *sf<u>â</u>n<u>t</u>>sf<u>i</u>n<u>ț</u>enie*). În general, rădăcina unui derivat este o formă de dicționar (sau lemă, după engl. *lemma*). Există însă şi cazuri când aceasta este o formă flexionară (de ex. forma de plural a unor substantive: *noduri>noduros*, participiul verbelor: *plâns>plânset*).

Sufixarea poate schimba clasa morfologică a bazei. Astfel, cu ajutorul sufixelor putem forma substantive de la alte substantive (*prunc>pruncie*), de la verbe (*vorbi>vorbire*), de la adjective (*tineri>tineret*) sau de la adverbe (*împrejur>împrejurime*); verbe putem obține prin sufixare de la alte verbe (*linge>linguşi*), de la substantive (*muşama>muşamaliza*), de la adjective (*orb>orbeca*) sau de la adverbe (*înluntru>înluntroşa*); adjective obținem de la alte adjective (*nițel>nițeluş*), de la substantive (*copil>copilăresc*), de la verbe (*întuneca>întunecos*), de la adverbe (*anevoie>anevoios*) sau de la numerale (*patru>pătrar*); adverbe obținem de la alte adverbe (*încet>încetinel*), de la substantive (*cruce>cruciş*), de la verbe (*fura>furiş*) sau de la adjective (*chior>chiorâş*).

Prefixarea singură nu schimbă, de obicei, clasa morfologică a bazei.

În studiul nostru am avut în vedere numai clasele deschise de cuvinte, adică substantivele, adjectivele, verbele şi adverbele, conştienţi de limitarea analizei. Dată fiind însă productivitatea slabă a derivării implicând restul părţilor de vorbire, considerăm că rezultatele nu sunt decât în mică măsură afectate.

Alături de prefixe şi sufixe, în lingvistică se discută şi despre prefixoide şi sufixoide: elemente formative care dau impresia unor prefixe, respectiv sufixe. Sunt foarte slab productive în limba română, dar se pot recunoaşte în structura unor cuvinte create în alte limbi sau după modelul cuvintelor din alte limbi. Sfera lor de întrebuinţare este limitată, în general, la domeniul ştiinţific.

Din punct de vedere funcţional, ca şi afixele, afixoidele se pot ataşa la începutul sau la sfârşitul unui cuvânt. Din punct de vedere lexico-semantic, rezultă un nou cuvânt, la al cărui sens contribuie sensul tuturor elementelor participante la formarea sa.

Spre deosebire de afixe care se ataşează numai la cuvinte, afixoidele se pot combina între ele, rezultând astfel cuvinte formate exclusiv din prefixoide şi sufixoide: *kinetoscop*. În plus, sufixoidele pot deveni prefixoide şi invers, schimbându-şi locul în structura cuvântului rezultat: *rusofil* versus *filorus*.

Din punctul de vedere al productivităţii, afixoidele sunt mult mai slab productive decât afixele, care participă la formarea a numeroase cuvinte în română.

În general, afixele nu au utilizare restrânsă la anumite domenii de activitate sau stiluri funcţionale. Excepţie fac trei sufixe din domeniul medical: *-ită*, *-om*, *-oză*; creaţiile spontane de felul lui *chiulangită* demonstrează statutul lor de sufixe recunoscute de vorbitori, care le transferă în limba comună, chiar în aspectul ei neîngrijit. În schimb, afixoidele sunt utilizate preponderent în limbajele ştiinţifice.

Date fiind aceste deosebiri între afixe şi afixoide, lingviştii români tratează cuvintele formate cu afixe ca fiind derivate, iar pe cele cu afixoide ca fiind la limita între compunere şi derivare, mulţi înclinând spre includerea lor în cel dintâi fenomen.

3. Resurse şi instrumente folosite

Studiul efectuat se bazează pe puține resurse lingvistice pe care le-am avut la dispoziție, în comparație cu cantitatea mare şi diversă a informațiilor pe care le foloseşte lingvistul pentru a analiza morfematic un cuvânt. Considerăm că cea mai importantă cunoştință este de tip semantic: cuvintele derivate şi baza lor au un conținut semantic parțial comun, conținut cu care participă baza de derivare. Chiar şi un elev din şcoala primară va analiza cuvântul *bălos* ca derivat de la *bale* cu sufixul *-os*, nu de la *bal* cu sufixul *-os*, chiar dacă sufixul *-os* formează frecvent adjective de la baze substantivale, iar alternanța fonetică a:ă este frecventă. Studiul nostru nu a avut la dispoziție informații despre semantica cuvintelor implicate, fapt ce şi-a lăsat amprenta în mod vizibil asupra rezultatelor.

3.1 Resurse lingvistice

Baza studiului nostru o reprezintă un lexicon ce conține formele flexionare ale cuvintelor din limba română, însoțite de forma lor de dicționar (lema) şi de descrierea morfosintactică. Lexiconul a fost creat în cadrul ICIA în proiectul MULTEXT-EAST şi conține 1.222.832 de intrări (la data scrierii acestui articol), reflectând toate părțile de vorbire. O intrare constă într-o formă ocurență (flexionată în cazul cuvintelor flexionare), lema acesteia (cuvântul-titlu) şi descrierea morfosintactică (Tufiş et al., 1997). Iată un fragment:

Tabelul 1. Fragment din lexiconul cu cuvinte-titlu, forme flexionare şi etichete morfosintactie

Formă flexionată	**Cuvânt-titlu**	**Etichetă morfosintactică**
circuitului	circuit	Ncmsoy
circul	circ	Ncmsry
circul	circula	Vmip1s
circulă	circula	Vmip3
circulă	circula	Vmis3s

Prima coloană conține forme ocurență. În cea de-a doua se află lemele sau cuvintele-titlu corespunzătoare formelor din prima coloană, iar apoi urmează descrierea morfosintactică a acestora. Astfel, *circuitului* are lema *circuit* şi din descrierea morfosintactică aflăm că este substantiv (N) comun

(c) de genul masculin[2] (m), la singular (s), are formă de caz oblic (o) şi este articulat hotărât (y).

Nu am folosit acest lexicon în întregime, ci ne-am rezumat doar la formele de bază ale cuvintelor aparţinând claselor deschise (substantive, adjective, verbe, adverbe). Am ignorat, deci, formele flexionare, cu riscul asumat de a nu regăsi fapte de tipul *colţuri>colţuros*, adică derivate de la forme flexionare; considerăm însă că sufixul *-os*, cu acelaşi sens, se combină şi cu alte cuvinte, la forma lor de bază, deci va putea fi regăsit (şi chiar a fost regăsit) în acelea: de exemplu: *băţ>băţos*. De asemenea, nu vom putea şti în câte derivate apare fiecare sufix. Dar, după cum se va vedea mai jos, în prezentarea rezultatelor, nici acest lucru nu afectează studiul de caz, el interferând cu alte aspecte.

Pe lexicon am rulat un algoritm de identificare a sufixelor. Rezultatele au fost comparate, în vederea evaluării, cu o listă de sufixe creată manual, pe baza unor lucrări de referinţă despre derivare în limba română (*Formarea cuvintelor în limba română* 1970, 1978, 1989, Pascu 1916, Coteanu 2007, Philippide 2011, Tudose 1978, Iordan 1939, *Studii şi materiale despre formarea cuvintelor în limba română* 1959, 1967, 1969). Listele acestea create manual le numim liste de referinţă (engl. *gold standard*) şi pornim de la asumpţiunea că ele au caracter exhaustiv. În tabelul de mai jos ilustrăm numărul de sufixe conţinute de fiecare listă.

Tabelul 2. Numărul de sufixe conţinute de listele standard.

Tipul de sufixe	**Numărul de sufixe din lista de referinţă**
substantivale	260
adjectivale	104
verbale	104
adverbiale	14
TOTAL	482

[2] Substantivul *circuit* este de genul neutru. Conform specifica□iilor de adnotare, substantivele neutre la singular sunt marcate ca având genul masculin, iar la plural ca având genul feminin. Distinc□ia între substantivele masculine, feminine □i neutre se face astfel: substantivele marcate cu m □i la singular □i la plural sunt masculine, cele adnotate cu f □i la singular □i la plural sunt feminine, iar cele adnotate cu m la singular □i cu f la plural sunt neutre.

3.2. Descrierea algoritmului

Pentru identificarea automată a sufixelor în cuvintele din lexicon am decis să lucrăm cu arbori de sufixe (engl. *suffix trees*) (Ukkonen, 1995). Aceştia prezintă două avantaje majore: rapiditate şi utilizare eficientă a memoriei.

Un arbore de sufixe este o reprezentare arborescentă a unui string (în cazul nostru, un cuvânt) şi a tuturor sufixelor sale. Sufixele sunt considerate a fi ceea ce rămâne dintr-un cuvânt după eliminarea succesivă a câte unei litere de la începutul acestuia. Astfel, pentru cuvântul *banană* avem reprezentarea arborescentă din Figura 1. Sufixele acestui cuvântul sunt: *banană*, *anană*, *nană*, *ană*, *nă*, *ă*.

Menţionăm că figurile 1 şi 2 au fost create cu ajutorul Generalized Suffix Tree Java Applet disponibil la adresa http://illya-keeplearning.blogspot.com/2009/06/
generalized-suffix-trees-java-applet.html (consultată în august 2011).

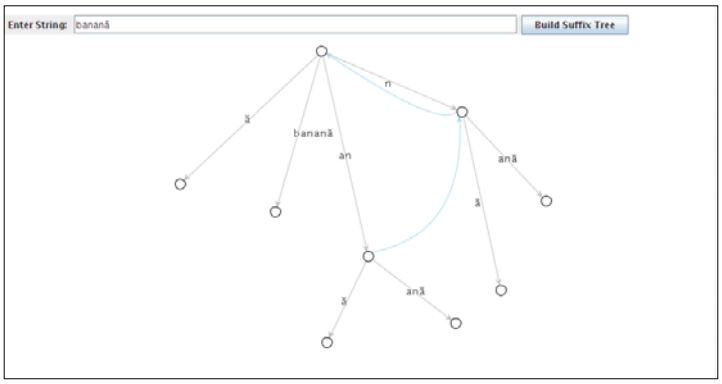

Figura 1. Arborele de sufixe pentru cuvântul banană.

Ipoteza studiului nostru este că sufixele sunt productive şi că există, deci, în lexicon o mulţime de cuvinte formate cu acelaşi sufix. De aceea, mult mai potrivit pentru studiul nostru este un tip special de arbori de sufixe, şi anume arborii de sufixe generalizaţi (engl. *generalized suffix trees*) (Gusfield, 1999). Aceştia se construiesc pentru o mulţime de cuvinte şi reprezintă toate sufixele acestora. Iată cum arată arborele de sufixe generalizat pentru cuvintele *cădere* şi *citire*. De obicei, pentru a construi un asemenea arbore, se adaugă un simbol ca marcator de sfârşit de cuvânt. Cuvintele din mulţimea pentru care se creează arborele sunt împărţite în sufixe, dar pentru

toate aceste cuvinte se construieşte un singur arbore. Sufixele comune (aici - *re*) sunt recunoscute ca fiind acele sufixe care diferă doar prin simbolul terminal (aici $ şi #).

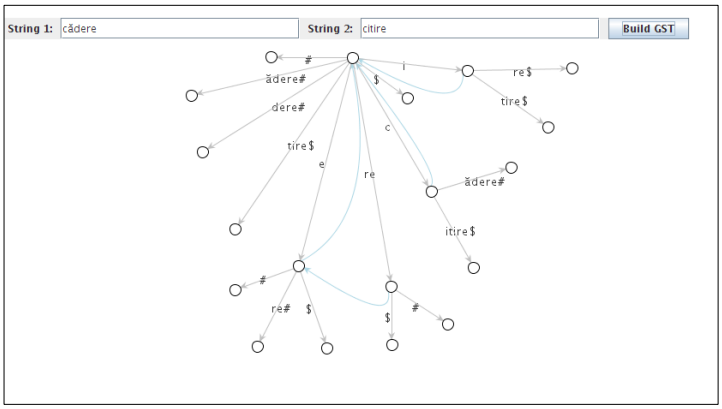

Figura 2. Arborele de sufixe generalizat pentru cuvintele cădere şi citire.

Primul pas în identificarea sufixelor îl constituie crearea unui index Lucene pentru lexiconul cu care lucrăm. Acest index conține documente. Fiecare document are trei câmpuri: forma de bază sau lema cuvântului, rădăcina şi eticheta morfosintactică. Am impus trei restricții asupra cuvintelor din documentele din index: o restricție asupra părților de vorbire (am lucrat numai cu clasele de cuvinte deschise, adică cu substantive, verbe, adjective şi adverbe), o restricție morfologică (am selectat numai lemele acestor cuvinte, adică: forma de nominativ singular nearticulat a substantivelor, forma de infinitiv a verbelor, de masculin singular nominativ a adjectivelor; adverbele au o singură formă în română, nu fac parte dintre părțile de vorbire flexibile) şi o restricție de formă (adică am exclus: cuvintele scrise cu cratimă, de ex. *floarea-soarelui*, cu liniuță de subliniere – engl. *underscore* – de ex. *praf_de_puşcă*, termeni consideraţi compuşi prin alăturare, cu puncte, de ex. *pp.* cu lema *pagină*, care a fost adnotat ca substantiv, în contradicție cu toate celelalte prescurtări care sunt marcate cu o etichetă specială pentru prescurtări, şi anume Y, sau cuvintele cu apostrof, de ex. *dom'* pentru *domn*).

Am creat acest index pentru a face mai rapidă căutarea şi pentru a putea modifica tipul căutării, atunci când este nevoie.

În Tabelul 2 prezentăm numărul de leme cu care am lucrat, grupate pe părți de vorbire. Chiar dacă nu este un fenomen extrem de frecvent, omonimia (i.e. relația între cuvinte cu aceeaşi formă, dar sens diferit) există în română şi se stabileşte între diverse părți de vorbire: substantive şi adjective (de exemplu *nebun*), adverbe şi adjective (de pildă *liniştit*), adjectiv, adverb şi substantiv (precum *frumos*) etc. În consecință, aceeaşi lemă poate avea mai multe adnotări morfosintactice. Pe ultima linie a acestui tabel am notat numărul de leme unice cu care am lucrat.

Tabelul 3. Numărul de leme utilizate, distribuite pe părți de vorbire.

Partea de vorbire	**Numărul de leme**
substantiv	67255
verb	8473
adjectiv	35285
adverb	1255
TOTAL	112.268
LEME UNICE	103.856

Indexul Lucene este folosit pentru a alege cu uşurință categoria gramaticală cu care să lucrăm şi pentru a valida radicalii obținuți după îndepărtarea afixelor identificate.

Radicalul cuvintelor derivate cu sufixe se obține după următoarea procedură (inspirată din algoritmul Snowball (http://snowball.tartarus.org/algorithms/romanian/stemmer.html) de identificare a rădăcinilor din cuvinte româneşti): (i) atunci când *i* şi *u* apar în poziție intervocalică sunt protejate prin marcarea lor cu literă majusculă, pentru a nu fi eliminate; (ii) cel mai lung dintre şirurile *a*, *e*, *i*, *ie*, *ă* este înlăturat; (iii) *i* şi *u* sunt de-protejate prin transformarea lor înapoi în litere mici.

Apoi extragem din indexul Lucene lista tuturor cuvintelor aparținând unei părți de vorbire. Eliminăm primele două litere din acestea (deci nu vom trata cuvintele alcătuite doar din două litere; ele nu sunt susceptibile de a fi derivate), iar formele astfel scurtate le introducem într-un arbore de sufixe generalizat. Menționăm că am folosit implementarea *open source* a algoritmului de arbori de sufixe generalizați de la carrot[2] (disponibilă pe situl project.carrot2.org).

Apoi am căutat în arbore ramuri comune pentru mai multe cuvinte. Acestea sunt sufixele potențiale, pe care le-am introdus într-o tabelă (engl. *map*), împreună cu cuvintele în care apar, i.e. cuvintele probabil derivate cu

aceste sufixe. Am inspectat manual această tabelă şi am constatat că este absolut necesară aplicarea unor filtre în procesul de identificare a sufixelor.

În primul rând, am impus o restricție asupra lungimii sufixelor, bazată pe observațiile noastre asupra listei standard de sufixe verbale: ele nu trebuie să fie mai lungi de cinci caractere pentru verbe, adjective şi adverbe şi şase caractere pentru substantive.

În al doilea rând, validăm radicalii cuvintelor: verificăm dacă şirul de caractere rămas după îndepărtarea potențialului sufix este un radical sau o lemă în indexul Lucene. Păstrăm doar acele sufixe care apar cu cel puțin un radical valid sau cu o lemă validă.

În al treilea rând, atunci când sunt posibile mai multe sufixe pentru acelaşi cuvânt (de ex. *spălătoreasă* are două sufixe: *-tor* şi *-easă*) păstrăm ca valid sufixul cel mai scurt. Procedăm astfel considerând că în cazul unui cuvânt cu două sufixe, derivarea s-a produs succesiv, în două etape. Suntem conştienți că astfel nu mai identificăm eventualele sufixe compuse din limba română.

În al patrulea rând, păstrăm doar sufixele care apar în cel puțin zece cuvinte, nu neapărat valide toate. Motivația este că scopul nostru este să găsim sufixe productive pe teren românesc. În finalul acestui articol prezentăm şi variații ale preciziei şi acurateței algoritmului în funcție de modificarea limitei de productivitate a sufixelor: am rulat programul impunând, pe rând, productivitate 20, 30 şi 40 asupra potențialelor sufixe.

În al cincilea rând, am observat că sufixele de patru sau cinci caractere care apar în mai puțin de şapte radicali valizi nu sunt cu adevărat sufixe şi le-am eliminat.

Pentru fiecare sufix, am calculat numărul total de cuvinte în care apare şi numărul de cuvinte (cu radical sau lemă) valide în care apare. Apoi am calculat un scor pentru fiecare sufix, împărțind numărul de ocurențe valide la numărul total de ocurențe. Acest scor reprezintă probabilitatea ca respectivul şir de litere aflat la finalul unui cuvânt să reprezinte un sufix.

4. Rezultate şi interpretări

Prezentăm mai jos câteva exemple din fişierul de sufixe substantivale:

oar (19 / 35 = 0.54): [balans+oar, ben+oar, bud+oar, cont+oar, cul+oar, eboş+oar, ferm+oar, fum+oar, hematoz+oar, hidroz+oar, lav+oar, muz+oar, patin+oar, pis+oar, polis+oar, poş+oar, raz+oar, remont+oar, urin+oar]

uleț (79 / 147 = 0.54): [bob+uleț, ciob+uleț, cioc+uleț, colț+uleț, corb+uleț, corn+uleț, cort+uleț, coș+uleț, crâng+uleț, cub+uleț, cuib+uleț, drug+uleț, dâmb+uleț, foc+uleț, frig+uleț, ghiob+uleț, golf+uleț, grup+uleț, hud+uleț, ied+uleț, joc+uleț, lemn+uleț, moș+uleț, muc+uleț, murg+uleț, mânz+uleț, nor+uleț, om+uleț, orz+uleț, partid+uleț, piept+uleț, plug+uleț, pod+uleț, pom+uleț, porc+uleț, prost+uleț, prunc+uleț, prânz+uleț, puf+uleț, punct+uleț, pâlc+uleț, roib+uleț, rug+uleț, runc+uleț, răc+uleț, schit+uleț, scoc+uleț, sfert+uleț, sfânt+uleț, smoc+uleț, snop+uleț, solz+uleț, somn+uleț, soț+uleț, spin+uleț, stog+uleț, strop+uleț, struț+uleț, strâmb+uleț, sturz+uleț, stâlp+uleț, stârc+uleț, sur+uleț, sânt+uleț, sân+uleț, tiv+uleț, toc+uleț, tren+uleț, trunchi+uleț, tub+uleț, turc+uleț, turn+uleț, urs+uleț, vin+uleț, șip+uleț, șir+uleț, șoim+uleț, șorț+uleț, ștuc+uleț]

easă (136 / 173 = 0.79): [arhondăr+easă, arăm+easă, aurăr+easă, baron+easă, blănăr+easă, boier+easă, bor+easă, brutăr+easă, brânzăr+easă, bucătăr+easă, bumbăcăr+easă, băbăr+easă, băcăn+easă, cenuşăr+easă, ciorăpăr+easă, ciubotăr+easă, cizmăr+easă, clucer+easă, cofetăr+easă, cojocăr+easă, colonel+easă, covrigăr+easă, croitor+easă, ctitor+easă, cucul+easă, culbec+easă, cusător+easă, cârciumăr+easă, cârnățăr+easă, călcător+easă, cămătăr+easă, căpitan+easă, cărturăr+easă, doctor+easă, doftor+easă, dubălăr+easă, fermecător+easă, florăr+easă, franzelăr+easă, făinăr+easă, gal+easă, giuvaierger+easă, gornic+easă, grădinăr+easă, găinăr+easă, hătmăn+easă, irod+easă, isprăvnic+easă, jimblăr+easă, jitar+easă, jud+easă, judec+easă, judecător+easă, județ+easă, jupân+easă, legător+easă, lenjer+easă, leurd+easă, lăptăr+easă, maior+easă, mezelăr+easă, meşter+easă, mir+easă, morăr+easă, moşier+easă, măcelăr+easă, neder+easă, negustor+easă, neguțător+easă, notăr+easă, năsturăr+easă, ofițer+easă, olăr+easă, ospătăr+easă, pair+easă, pantofăr+easă, papucăr+easă, pescăr+easă, pețitor+easă, pităr+easă, plugăr+easă, plăpumăr+easă, poftitor+easă, polcovnic+easă, pop+easă, popicăr+easă, portăr+easă, porumbăr+easă, postelnic+easă, postăvăr+easă, potcovăr+easă, potropop+easă, preot+easă, primăr+easă, profesor+easă, protopop+easă, puric+easă, pâinăr+easă, pârcălăb+easă, păcurăr+easă, păhărnic+easă, pălărier+easă, rotăr+easă, rudăr+easă, serdăr+easă, servitor+easă, sfârnăr+easă, sluger+easă, spițer+easă, spoitor+easă, spălător+easă, spătăr+easă, stolnic+easă, strungăr+easă, stăpân+easă, sânger+easă, săpunăr+easă, sărăr+easă, tinichiger+easă, tutunger+easă, tâmplăr+easă, tăbăcăr+easă, vier+easă, vod+easă, voievod+easă,

voivod+easă, vornic+easă, vrăjitor+easă, văcăr+easă, zidăr+easă, zlătăr+easă, învăţător+easă, şoric+easă, şuncăr+easă, şurubăr+easă, şătrăr+easă]

Primul este un exemplu de terminaţie care nu este sufix în limba română: *-oar*. El este un fals sufix, identificat de algoritm întrucât şirurile de caractere rămase prin îndepărtarea lui *-oar* din cuvintele enumerate sunt leme sau radicali valizi în limba română. Cu toate acestea, nu există nicio legătură derivativă între aceşti radicali şi cuvintele de la care s-a pornit, deşi una semantică poate fi găsită în anumite cazuri: între *balans* şi *balansoar*, între *fum* şi *fumoar* etc. Vorbitorii de limbă română nu au decupat din aceste împrumuturi terminaţia *-oar* pe care să o folosească apoi la crearea altor cuvinte, adică să o trateze ca pe un sufix.

În schimb, *-uleţ* şi *-easă* sunt sufixe în limba română, chiar dacă nu în toate cuvintele validate de algoritm şi enumerate mai sus.

După aplicarea filtrelor descrise mai sus am obţinut următoarele rezultate:

Tabelul 4. Rezultatele după rularea algoritmului şi aplicarea filtrelor.

Partea de vorbire	**Total sufixe identificate**	**Sufixe corecte**	**Precizie (% corecte din total)**	**Acurateţe (% corect identificate din listele standard)**
substantive	888	161	18,13	61,92
verbe	209	81	38,76	77,88
adjective	320	54	16,88	51,92
adverbe	46	3	6,52	21,43

Printre rezultate găsim şi false pozitive. Unele terminaţii ale cuvintelor sunt identificate ca sufixe greşite. De exemplu, *metresă* este identificat ca derivat de la *metru* cu sufixul *-esă*. Explicaţia rezidă în precaritatea diversităţii resurselor lingvistice folosite: singura resursă pe care se bazează este lexiconul de forme. Nu foloseşte nicio resursă fonetică sau semantică. Un verb ca *noroi* este analizat ca fiind format din *nor* cu sufixul *-oi*. Cuvintele derivate şi bazele lor au în comun un anumit conţinut semantic. Or, cum algoritmul nostru nu foloseşte nicio informaţie de tip semantic, astfel de analize sunt inevitabile.

Parcurgând manual listele de sufixe identificate, am constatat următoarele:

- Efectuând şi o validare manuală a rezultatelor, am identificat şi sufixe inexistente în listele de referinţă. De exemplu, *-ăt* în *strigăt*

(<*striga*+*-ăt*) sau *-atec*, variantă a sufixului *-atic* : *singuratec*. Mult mai interesant este cazul sufixului *-iadă* (nediscutat în lucrările de specialitate), care a devenit productiv în română mai târziu, creând cuvinte precum *universiadă*, *mineriadă*. Algoritmul nostru identifică aceste sufixe, dar nu le regăseşte în lista de referinţă, ca atare nu le validează.

- Lipsa unei resurse fonetice care să îmbunătăţească rezultatele duce la analize greşite: *barochist* este analizat în *baroc* + *-hist*. În limba română grupurile de litere *ch* şi *gh* nu reprezintă niciodată două sunete. Algoritmul poate fi îmbunătăţit adăugând acest filtru: ch sau gh nu se separă astfel încât c sau g să aparţină rădăcinii, iar h sufixului.
- Unele cuvinte prezintă omonimie: aceeaşi formă are valori aparţinând mai multor clase morfologice. Neincluderea unui sufix decât în lista standard a unei părţi de vorbire, ignorând astfel conversiunea, scade precizia şi acurateţea. Soluţia noastră a fost să efectuăm o evaluare şi prin unificarea listelor de referinţă ale părţilor de vorbire ce prezintă omonimii între ele. Rezultatele se găsesc mai jos, în Tabelul 5.
- Printre sufixele identificate, mai ales pentru substantive şi adjective, se află şi multe sufixoide, pe care nu le incluseserăm în lista standard. Ca atare, ele nu au fost validate de algoritm. De exemplu: *-fag* din *fotofag*, *-cid* din *zoocid*, *-urie* din *alcoolurie* şi altele. Pentru unele, analiza este greşită: în *aglutinogen* este identificat sufixul *-ogen* (corect ar fi *-gen*), în *anglofob* sufixul *-ofob* (corect este *-fob*); lingviştii trec vocala de legătură (în exemplele de mai sus *o*) la partea iniţială a cuvântului, nu la sufixoid. În fond, atunci când această vocală nu face parte din radicalii cuvintelor, ea poate fi decupată alături de elementul din faţă sau de cel următor.

Iniţial, lista noastră de referinţă a inclus numai sufixe. Întrucât unei maşini îi este greu să distingă între sufix şi sufixoid (mai ales că unele au statut controversat, unii lingvişti considerându-le sufixe, alţii sufixoide), am decis să includem şi sufixoidele alături de sufixe, pentru a verifica în ce măsură se modifică rezultatele.

În consecinţă, listele standard ale sufixelor se modifică după cum urmează: cea a substantivelor a fost îmbogăţită cu 151 de sufixoide,

ajungând astfel la 411 elemente, iar cea a adjectivelor cu 73 de sufixoide, numărând în final 177 de elemente. Celelalte liste (ale sufixelor verbale şi adverbiale) au rămas nemodificate. Comparând rezultatele algoritmului cu aceste noi liste standard constatăm următoarele rezultate în cazul substantivelor şi adjectivelor (pentru celelalte părți de vorbire nu înregistrăm modificări, în mod evident):

Tabelul 5. Evaluarea prin luarea în considerare şi a sufixoidelor.

Partea de vorbire	**Total sufixe identificate**	**Sufixe corecte**	**Precizie (% corecte din total)**	**Acurateţe (% corect identificate din listele standard)**
substantive	888	188	21,17	45,74
adjective	320	77	24,06	43,50

Comparând datele din Tabelul 4 cu cele din Tabelul 5 reiese că precizia a crescut uşor (peste 3% pentru substantive şi cu aproape 8% pentru adjective), în detrimentul acurateței (care a scăzut cu aproape 17% pentru substantive şi cu peste 8% pentru adjective).

Pentru a acoperi şi cazurile de omonimie între părți de vorbire diferite, am comparat sufixele identificate de algoritm cu listele de referință unificate astfel: substantive cu adjective şi adjective cu adverbe. Adică, lista sufixelor substantivale identificate de algoritm am comparat-o cu lista unificată a sufixelor (şi sufixoidelor) substantivale şi adjectivale, lista celor adjectivale am comparat-o tot cu lista unificată a sufixelor (şi sufixoidelor) substantivale şi adjectivale, lista sufixelor adverbiale am comparat-o cu lista unificată a sufixelor (şi sufixoidelor) adjectivale şi adverbiale. Am obținut datele din tabelul următor:

Tabelul 6. Evaluarea prin luarea în considerare şi a omonimiilor

Parte de vorbire	**Total sufixe (şi sufixoide) unice în lista standard unificată**	**Total sufixe găsite**	**Bune găsite**	**Precizie (% Bune găsite din Total găsite)**	**Acurateţe (% Bune găsite din lista standard)**
substantive	513 (subst.+adj.)	888	214	24,10	41,72
adjective	513 (subst.+adj.)	320	89	27,81	17,35
adverbe	183 (adj.+adv.)	46	16	34,78	8,74

şi în felul acesta observăm, comparând datele din Tabelul 6 cu cele din Tabelele 4 şi 5, că precizia creşte, pentru substantive, cu aproape 3% faţă de datele din Tabelul 5, deci faţă de cazul cu luarea în considerare şi a sufixoidelor, respectiv cu 7% faţă de cazul iniţial; pentru adjective, precizia creştecu aproape 4% faţă de cazul din Tablul 5, deci cu aproape 11% faţă de cazul iniţial; pentru adverbe precizia creşte de peste 5 ori faţă de cazul iniţial; pentru toate părţile de vorbire acurateţea scade: cu 4%, respectiv cu 20% pentru substantive faţă de cazul din Tabelul 5, respectiv din Tabelul 4; pentru adjective ea scade dramatic, de la 45% în Tabelul 5, respectiv de la 51% în Tabelul 4, la 17% în ultimul caz; şi la adverbe scade de 3 ori faţă de cazul iniţial.

În graficele din figurile 3, 4 şi 5 prezentăm variaţia preciziei şi acurateţei în funcţie de limita impusă asupra productivităţii sufixelor: toate datele din tabelele de mai sus sunt redate pentru productivitate 10. Adică permitem un număr minim de 10 cuvinte în care potenţialul sufix să apară, pentru a fi luat în considerare în procesul de validare. În graficele de mai jos am analizat variaţiile celor două mărimi atunci când am impus o productivitate diferită asupra prefixelor, adică atunci când ele apar în 10, 20, 30, respectiv 40 de cuvinte. După cum se observă, cu cât creştem limita impusă productivităţii sufixelor, cu atât precizia este mai mare, iar acurateţea mai mică. Modul în care ele variază diferă de la o parte de vorbire la alta, după cum se poate observa după formele curbelor în cele trei figuri. În cazul adverbelor, oricum numărul de sufixe găsite este foarte mic; dacă modificăm limita impusă productivităţii sufixelor, găsim tot 3 sufixe corecte pentru productivitate 20, iar pentru productivitate 30 găsim doar 2, în vreme ce pentru productivitate 40 găsim doar un sufix. Pentru adverbe nu am mai redat un grafic.

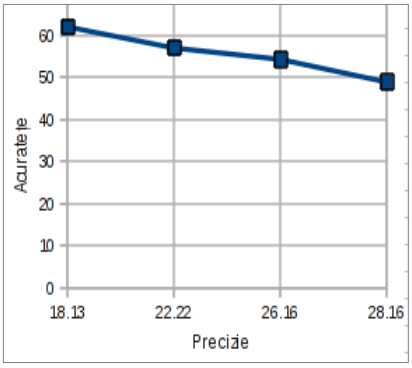

Figura 3. Evoluția preciziei şi acurateței în cazul substantivelor.

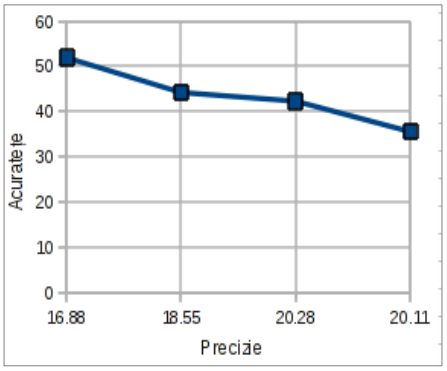

Figura 4. Evoluția preciziei şi acurateței în cazul adjectivelor.

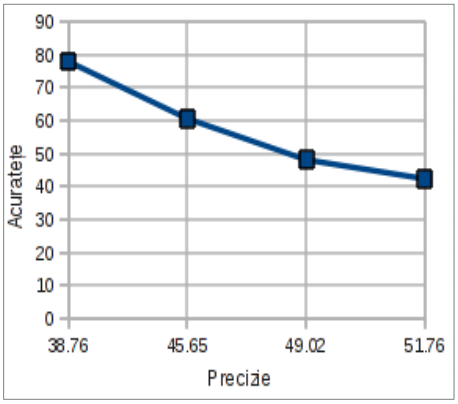

Figura 5. Evoluția preciziei şi acurateței în cazul verbelor.

5. Constatări finale

Analiza morfematică a cuvintelor prezintă interes atât pentru lingvişti (datori, printre multe altele, să dea explicaţii despre modul de apariţie sau formare a cuvintelor), precum şi pentru inginerii care, propunându-şi să rezolve variate sarcini ce implică prelucrarea textelor, conştientizează valoarea legăturii lexico-semantice dintre cuvintele din aceeaşi familie lexicală şi se angajează la îmbunătăţirea rezultatelor activităţii lor prin valorificarea informaţiilor oferite de aceasta.

Limba română este, în continuare, deficitară în ceea ce priveşte resursele lingvistice şi instrumentele de prelucrare a ei: corpusurile existente nu au dimensiuni foarte mari, nu acoperă domenii şi stiluri funcţionale variate, nu au adnotări pe mai multe niveluri (morfologic, sintactic, semantic, pragmatic etc.). În diverse aplicaţii se simte nevoia unui analizor sintactic, fie el şi de suprafaţă.

Având ca scop final al unui proiect mai amplu dezvoltarea unei reţele semantico-derivaţionale pentru limba română, ce va putea fi folosită şi ca dicţionar de cuvinte derivate, ne-am propus să întocmim, pentru început, o listă cu afixele româneşti. O provocare în realizarea acestui obiectiv a constituit-o încercarea de identificare automată a acestora într-un dicţionar de leme. În acest articol am prezentat modul în care am efectuat acest experiment şi rezultatele obţinute.

Bazându-ne exclusiv pe informaţii lexicale (am avut la dispoziţie un lexicon de cuvinte româneşti din care am ales să lucrăm numai cu formele de bază ale substantivelor, adjectivelor, verbelor şi adverbelor), ignorând orice fel de informaţie de tip fonetic (alternanţe), grafic (convenţii de notare a unor sunete) şi semantice (relaţii semantice între cuvinte), şi folosind metoda arborilor de sufixe generalizaţi am extras terminaţii ale cuvintelor, adică potenţiale sufixe. Pentru evaluarea metodei, prin calcularea preciziei şi acurateţei, am comparat aceste terminaţii cu listele de referinţă ale sufixelor româneşti, pe care le-am extras din lucrările dedicate derivării, mai exact sufixării, în limba română. Aceste liste de referinţă sunt disponibile la adresa www.racai.ro/~vergi în categoria Research.

Analizând rezultatele obţinute constatăm câteva lucruri esenţiale: în primul rând, se observă imediat care sufixe sunt productive şi care nu, precum şi cât de productive sunt.

În al doilea rând, scorurile calculate de noi (prin împărţirea numărului total de cuvinte în care apare un sufix la numărul de cuvinte valide)

reprezintă probabilitatea ca terminaţia respectivă să fie sufix într-un cuvânt oarecare. Această informaţie şi cea de dinainte sunt extrem de importante în diversele sarcini din Inteligenţa Artificială, în care identificarea derivatelor şi reducerea lor la rădăcină au un rol important (de exemplu, rezumarea automată, răspunsul automat la întrebări etc.)

În al treilea rând, algoritmul a identificat şi sufixe despre care lingviştii încă nu au scris (vezi *-iadă*). Astfel, am îmbogăţit listele de referinţă ale sufixelor româneşti, este adevărat, cu un număr mic de elemente (*-ăt*, *-atec*, *-iadă*), însă ele sunt valoroase.

În al patrulea rând, se observă consecinţele de neevitat ale nefolosirii informaţiilor semantice: este identificat sufixul *-ier* din mai multe cuvinte (*afişier*, *textier*), iar la validare se acceptă, în mod eronat, *vistier* ca derivat de la *vist* (joc de cărţi).

În al cincilea rând, se observă tratarea identică a derivării progresive, a celei regresive, a substituţiei de sufixe şi a împrumuturilor: ceea ce contează este posibilitatea decupării sufixului dintr-un şir de caractere, nu şi proveniența sa. Astfel, *problematiza*, considerat de lingvişti derivat prin substituţia sufixului *-ic* din *problematic* cu *-iza*, este analizat, în absenţa unui radical valid *problemat*, ca derivat de la *problemă* cu sufixul *-atiza*.

Aşa cum am menţionat pe parcursul articolului, metoda de lucru mai poate fi îmbunătăţită. În plus, o nouă provocare poate fi identificarea prefixelor cuvintelor, după un algoritm asemănător, aducând ajustările necesare.

Din cunoştinţele noastre, aceasta este prima raportare a unui experiment de identificare a sufixelor româneşti în cuvinte derivate. Alte experimente au vizat recunoaşterea cuvintelor româneşti derivate, cunoscut fiind un inventar de afixe (Petic, 2011).

□i pentru alte limbi, s-a manifestat interesul pentru procesul de reducere a derivatelor la baza lor, util pentru îmbunătăţirea rezultatelor în domeniul regăsirii informaţiei (Porter, 1980). Totuşi, lipsa resurselor lingvistice pentru diverse limbi a impulsionat cercetările legate de găsirea unei metode de identificare a morfemelor în cuvinte (Goldsmith, 2001, Majgaonker şi Siddiqui, 2010, Freitag, 2005 ş.a.). Metodele variază de la învăţarea nesupervizată (Goldsmith, 2001), la folosirea grupurilor de cuvinte (Freitag, 2005), a regulilor de segmentare (Porter, 1980), a împărţirii cuvintelor în n-grame dintre care să se recunoască sufixele (Majgaonker şi Siddiqui, 2010).

În contextul cercetărilor despre wordnet, îmbogățirea acestuia cu relații derivaționale a urmărit fie identificarea perechilor bază-derivat dat fiind un inventar de afixe (în general sufixe), pentru care au fost generate apoi etichetele semantice (Fellbaum et al., 2007, Bilgin et al., 2004), fie generarea derivatelor de la baze existente în wordnet cu ajutorul unor sufixe productive (Pala şi Hlavackova, 2007, Kahusk et al., 2010).

Interesul pentru identificarea sufixelor unui cuvânt se explică, din perspectivă lingvistică, prin nevoia de a crea, într-un mod rapid şi ieftin, resurse lingvistice pentru un număr cât mai mare şi cât mai divers de limbi naturale. Din perspectiva interacțiunii om-calculator, această etapă de prelucrare a limbii are utilizări în mai multe subdomenii: generarea limbii naturale, răspunsul automat la întrebări, sumarizarea, regăsirea informației şi altele.

Mulțumiri

Această lucrare a fost realizată în cadrul proiectului "Valorificarea identităților culturale în procesele globale", cofinanțat de Uniunea Europeană şi Guvernul României din Fondul Social European prin Programul Operațional Sectorial Dezvoltarea Resurselor Umane 2007-2013, contractul de finanțare nr. POSDRU/89/1.5/S/59758. Mulțumim pentru observațiile formulate pe marginea manuscrisului dlui academician Grigore Brâncuş şi recenzorilor anonimi ai RRIOC, precum şi lui Emanuele Pianta pentru discuțiile fructuoase purtate pe marginea acestui experiment.

Referințe

Avram, M., *Introducere în studiul sufixelor*, în Graur, Al., Avram, M. (coord.), Formarea cuvintelor în limba română, Bucureşti, Editura Academiei, vol. III 1989.

Bilgin, O., Cetinoglu, O., Oflazer, K., *Morphosemantic Relations in and across Wordnets: A Study Based on Turkish*, în P. Sojka, K. Pala, P. Smrz, C. Fellbaum, P. Vossen (eds.), Proceedings of GWC, 2004.

Coteanu, I., *Formarea cuvintelor în limba română: derivarea, compunerea, conversiunea*, ed. Narcisa Forăscu, Angela Bidu-Vrănceanu, Bucureşti, Editura Universității din Bucureşti, 2007.

Fellbaum, C., Osherson, A., Clark, P. E., *Putting Semantics into Wordnet's "Morphosemantic" Links*, în Proceedings of the 3rd Language and Technology Conference, Poznan, 2007.

Freitag, D., *Morphology induction from term clusters*, în Proceedings of the ninth conference on computational natural language learning, p. 128-135, 2005.

Goldsmith, J., *Unsupervised Learning of the Morphology of a Natural Language*, în Computational Linguistics, vol. 27, nr. 2, p. 153-198, 2001.

Graur, Al. şi Avram, M. (coord.), *Formarea cuvintelor în limba română,* Bucureşti, Editura Academiei, vol. I 1970, vol. II 1978, vol. III 1989.

Gusfield, D., *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, Cambridge University Press, SUA, 1999.

Hristea, Th., *Romanian Vocabulary and Etymology în Current Trends in Romanian Linguistics*, RRL, no. 1-4, 1978.

Iordan, I., *Sufixe româneşti de origine recentă (neologisme)*, în Buletinul Institutului de filologie română „Alexandru Philippide", vol VI, Iaşi, p. 1-59, 1939.

Kahusk, N., Kerner, K., Vider, K., *Enriching Estonian WordNet with Derivations and Semantic Relations*, în Proceedings of the 2010 Conference on Human Language Technologies – The Baltic Perspective, 2010.

Majgaonker, M. M., and Siddiqui, T. J., *Discovering Suffixes: A Case Study for Marathi Language*, în International Journal on Computer Science and Engineering, vol. 2, nr. 8, p. 2716-2720, 2010.

Marouzeau, J., *Lexique de la terminologie linguistique*. Librairie orientaliste Paul Geuthner, Paris, 1933.

Pala, K., Hlavackova, D., *Derivational Relations in Czech Wordnet*, în Proceedings of the Workshop on Balto-Slavonic, 2007.

Pascu, G., *Sufixele româneşti*, Bucureşti, Ed. Academiei Romane, 1916.

Petic, M., *Automatizarea procesului de creare a resurselor lingvistice computaţionale*, teză de doctorat, Chişinău, Academia de □tiinţe a Moldovei, Institutul de Matematică şi Informatică, 2011.

Philippide, Al., *Istoria limbii române*, Iaşi, Polirom, 2011.

Porter, M., *An Algorithm for Suffix Stripping*, în Program, 14, nr. 3, p. 130-137, 1980.

Puşcariu, S., *Limba română*. Vol. I Privire generală, Fundaţia pentru Literatură şi Artă "Regela Carol II", Bucureşti, 1940, vol. II. Rostirea, Editura Academiei RSR, Bucureşti, 1959.

Stoichiţoiu-Ichim, A., *Vocabularul limbii române actual. Dinamică, influenţe, creativitate*, Bucureşti, Editura BIC ALL, 2007.

Studii şi materiale privitoare la formarea cuvintelor în limba română, vol I, Editura Academiei, vol I 1959, vol IV 1967, vol V 1969.

Tudose, C., *Derivarea cu sufixe în româna populară*, Bucureşti, Editura Universităţii din Bucureşti, 1978

Tufiş, D., Barbu, A.-M., Pătraşcu, V., Rotariu, G., Popescu, C., *Corpora and Corpus-Based Morpho-Lexical Processing*, în D. Tufiş, P. Andersen (eds.), Recent Advances in Romanian Language Technology, Bucureşti, Editura Academiei Române, p. 35-56,

1997.
Ukkonen, E., *On-line construction of suffix trees*, în Algorithmica, 14(3), p. 249-260, 1995.