Instrumente informatice implicate în cercetarea terminologică

Cristina Varga
(Universitatea Babes-Bolyai, Cluj napoca)

Introducere

Există o mare varietate de instrumente informatice ce intervin în cercetarea lingvistică în general şi în cercetarea terminologică în special. Acest din urmă domeniu este unul dintre domeniile predilecte ale lingvisticii aplicate, printre altele, şi pentru faptul că rezultatele cercetării terminologice se materializează în instrumente de lucru pe teren lingvistic cu o largă utilizare[1]. Elementele fundamentale care stau la baza cercetării în domeniul terminologiei sunt: *crearea, manipularea, explorarea şi gestionarea* de corpusuri de texte.

Analiza şi explorarea unui corpus de texte se face, prin forța lucrurilor, utilizând instrumente electronice, deoarece acestea oferă avantajele vitezei de lucru, exactității şi eficienței, rezultatele procesului de cercetare putând fi utilizate în varii domenii precum: ***terminologia*** (detectarea şi extracția termenilor, selectarea unui context valid pentru un termen); ***predarea limbilor străine***, ***traducere şi interpretare, lexicologie / lexicografie*** (generală sau specializată), ***lingvistică generală, lingvistică contrastivă, redactare de texte, analiza discursului***, ***în procesul didactic*** (în cadrul cursurilor de formare de traducători, interpreți şi terminologi). În toate aceste domenii cercetarea ştiințifică apelează la instrumente electronice capabile să recunoască, să extragă, să compare segmente lingvistice, care apoi sunt interpretate de către specialistul în domeniu în funcție de finalitatea cercetării. Alegerea instrumentelor de lucru, din multitudinea de alternative existente, trebuie să aibă în vedere deci finalitatea cercetării şi gradul de eficacitate pe care îl are acesta într-un context clar determinat.

Acest articol îşi propune să prezinte câteva instrumente electronice complexe, capabile să gestioneze şi să manipuleze corpusuri. Este vorba despre instrumente destinate activității de cercetare, cu distribuție gratuită în Internet, suficient de complexe ca să permită dezvoltarea unui proiect de investigație lingvistică sau terminologică. Sunt create în mediul de cercetare academic şi fac obiectul de studiu al unor cursuri de specialitate în diverse universități europene. Cunoaşterea lor şi pe teren românesc, cunoaşterea facilităților pe care le oferă precum şi a limitelor lor putând constitui atât o ilustrare a direcțiilor de evoluție a instrumentelor electronice în cadrul cercetării lingvistice cât şi un model ilustrativ pentru cei care, implicați fiind în proiecte de cercetare lingvistică pe teren românesc, ar dori să dezvolte instrumente de explorare şi exploatare de corpusuri de texte dedicate limbii române.

Instrumente electronice de investigație lingvistică

Ne propunem ca în acest articol să prezentăm trei instrumente electronice cu aplicații în domeniul cercetării terminologice, programe a căror utilitate, mai ales în ceea ce priveşte cercetarea terminologică, nu a fost încă suficient pusă în evidență pe teren românesc.

---

[1] De cele mai multe ori, analiza şi exploatarea de corpusuri de texte au ca rezultat elaborarea a diverse tipuri de materiale lingvistice (*dicționare generale, dicționare-tezaur, glosare specializate***,** etc.).

Acestea sunt: ***SCP (Simple Concordance Program), Lexico3 şi Corpografo***. Sunt programe cu distribuție gratuită în scopul cercetării, două dintre acestea se instalează local pe stații de lucru, în timp ce al treilea, ***Corpografo*** este un program cu acces on-line[2]. Toate sunt caracterizate printr-o interfață grafică accesibilă, uşor de înțeles şi de utilizat. De asemenea, prezintă un grad de complexitate[3] corespunzător necesităților din domeniul cercetării academice. Fiind de producție diferită[4], fiecare dintre acestea prezintă elemente specifice ce ilustrează liniile de dezvoltare ale cercetării lingvistice în centrul academic de proveniență. Analiza lor contrastivă ne-a permis să facem o ierarhizare a acestor programe în funcție de gradul de complexitate şi de facilitățile pe care le oferă în domeniul cercetării lingvistice, ierarhizare care va face ca prezentarea lor să debuteze cu cel mai simplu dintre ele şi să continue treptat cu cele care au un grad mai mare de complexitate.

SCP (Simple Concordance Program)

Program de analiză de texte şi recuperare de informație dintr-un corpus textual, cu distribuție gratuită, creat de Alan Reed şi ajuns actualmente la versiunea 4.09. Se poate obține de la URL***: http://www.textworld.com/***.

Simple Concordance Program (SCP) este cel mai „simplu" dintre cele trei instrumente electornice care fac obiectul acestui articol. Înainte de a pune în evidență utilitatea sa în contextul cercetării lingvistice, este de remarcat faptul că ***SCP*** s-a dovedit a fi un instrument foarte bun în procesul didactic. Foarte uşor de utilizat şi cuprinzând principalele programe şi funcții de analiză şi explorare de corpus, în unele instituții academice a fost inclus în programul de studii pentru discipline ca: ***Informatica pentru traducători*** sau ***Lingvistica de corpus***[5]. Acest program s-a transformat într-un instrument de lucru excelent pentru studenții[6] care se inițiază în cercetarea terminologică.

Funcțiile sale de bază se referă la recuperarea şi extragerea de informații (*ocurențe, cuvinte cheie*, *termeni*, etc.) dintr-un corpus de texte propus de utilizator şi care trebuie prezentat în format .TXT. Prezentarea unităților lexicale care fac obiectul cercetării se poate face în context (la nivel de frază – KWIC sau la nivel de linie - LINE). Sunt permise de asemenea afişarea listei unităților lexicale care compun textul în diverse moduri: *ordine alfabetică*, în funcție de *frecvența cuvintelor* sau în *ordinea apariției lor în text*. Programul beneficiază, de asemenea de funcții ce permit crearea unui *profil de frecvență* al unei unități lexicale sau prezentarea *proprietăților statistice* ale corpusului analizat.

---

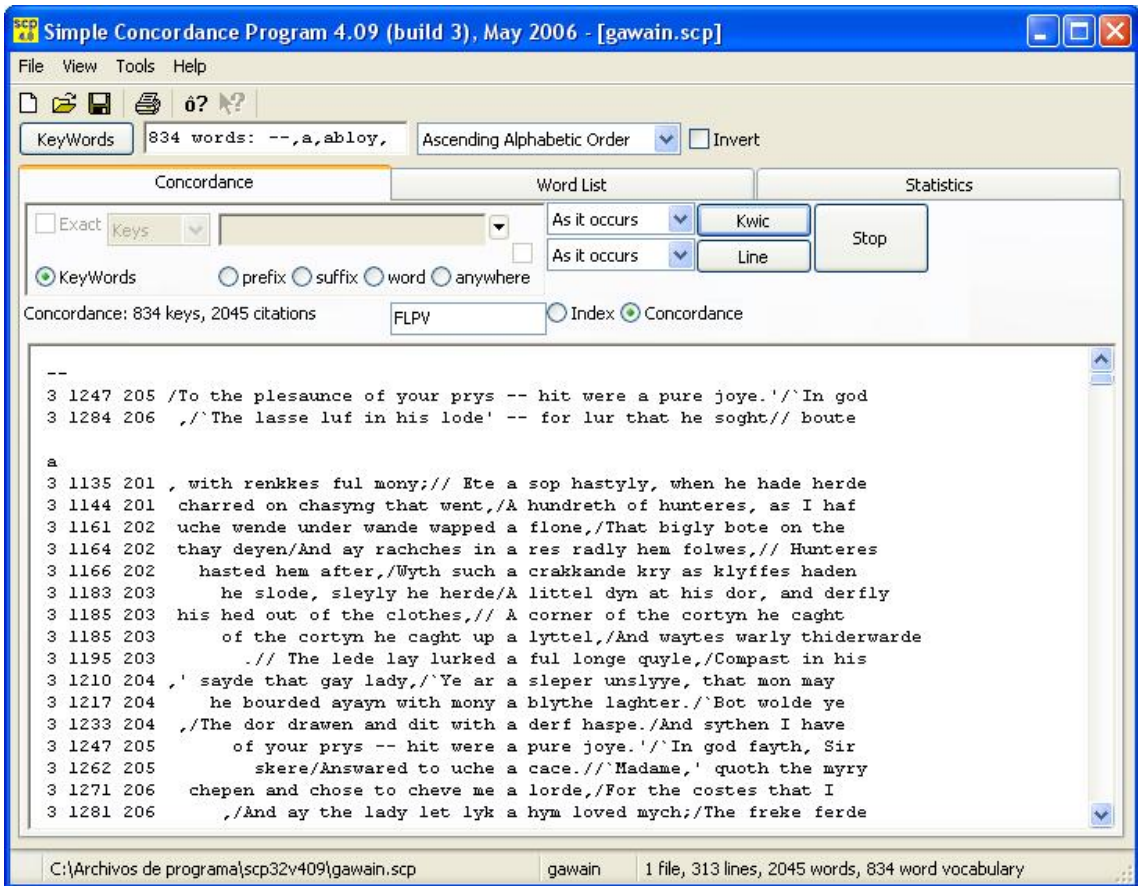
[2] Pentru a putea lucra cu ***Corpografo*** este necesară înscrierea în lista de utilizatori ai acestui program şi obținerea unei chei de acces.

[3] În cazul fiecărui instrument vorbim despre un complex de aplicații unificate într-un pachet de programe, ceea ce individualizează profilul fiecărui instrument.

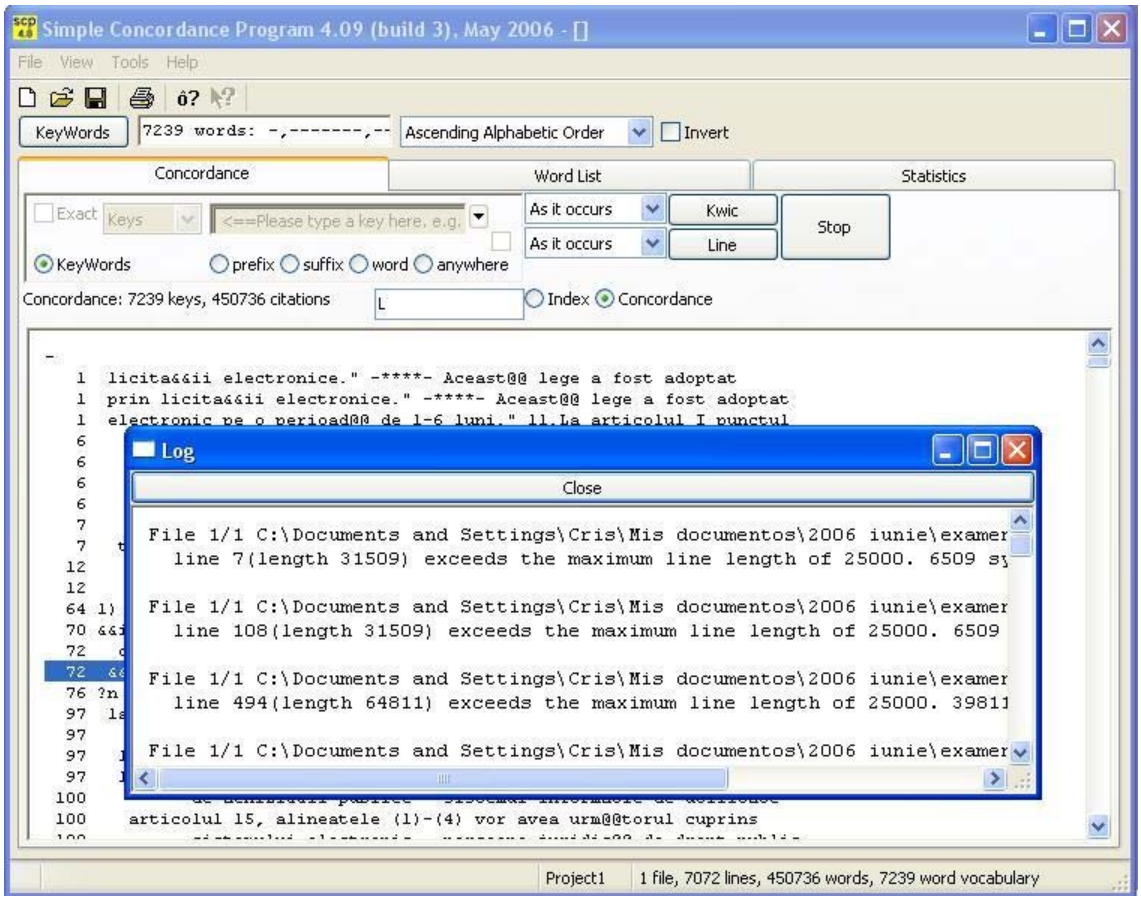
[4] Fiecare dintre aceste programe reprezintă orientările cercetărilor lingvistice din altă țară. Astfel, ***SCP*** este produs de un grup de cercetători din SUA, ***Corpografo*** reprezintă un proiect în desfăşurare a unui grup de cercetare din Portugalia, iar ***Lexico3*** este dezvoltat de către cercetători de la Université Paris 3 – Sorbonne Nouvelle, Franța.

[5] Este cazul ***Universității Pompeu Fabra***, Barcelona, unde acest program se studiază în cadrul a două specializări.

[6] De remarcat faptul că este utilzat şi în centre de cercetare terminologică de prestigiu cum ar fi de exemplu IULA (***Institut Universitari de Lingüística Aplicada***), Barcelona.



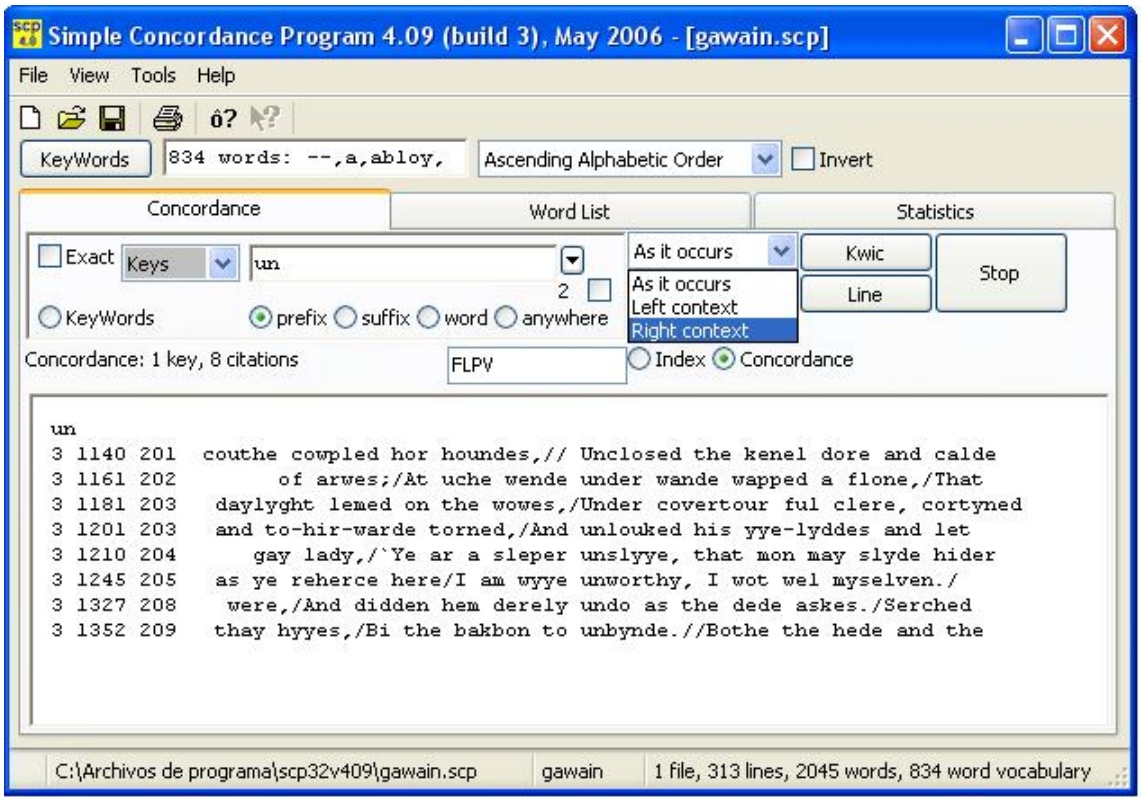
Fiecare sesiune de lucru debutează cu crearea unui *proiect* ***SCP*** sau cu modificarea unuia existent, *proiect* în care se poate încărca un corpus de texte. Există situații în care pot apare mesaje de eroare, este cazul în care lungimea unei linii a corpusului de texte depăşeşte limita de 25.000 de caractere. Semnele care depăşesc această limită vor fi ignorate. Programul semnalează utilizatorului care sunt liniile în cauză şi, de asemenea, se menționează numărul de caractere ignorate în fiecare dintre cazuri, astfel acesta va putea aprecia cât din informația pe care o analizează este ignorată.

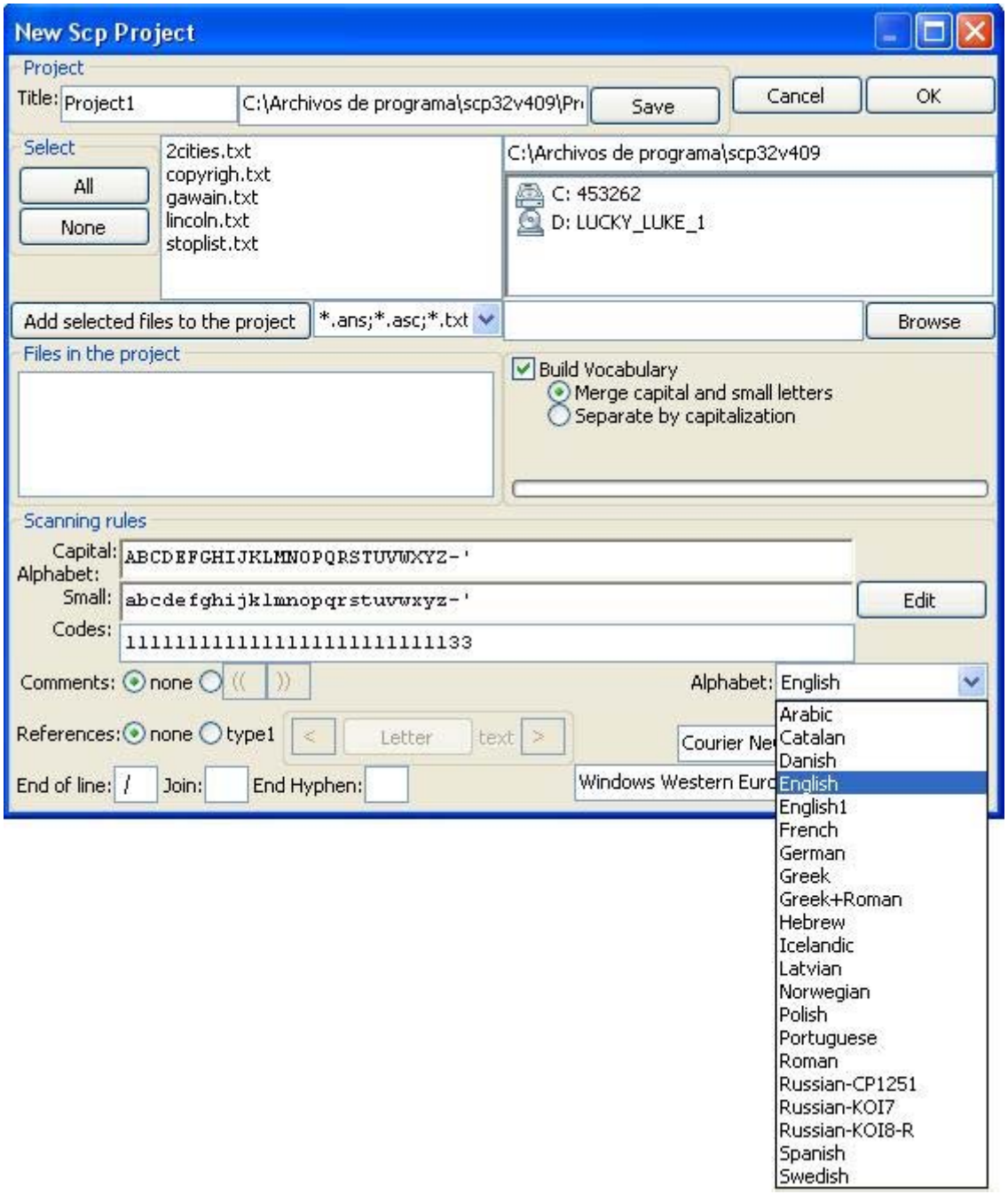


Extragerea informației din corpus se poate face utilizând o listă de *cuvinte cheie*, sau căutând ocurențe în funcție de *prefix, sufix, întregul cuvânt* sau *o parte oarecare a acestuia*. Toate instanțele elementului căutat vor fi afişate în context, incluzând, la cererea utilizatorului *descriptori statistici* Descriptorii statistici se întâlnesc nu doar la nivel de text ci şi la nivelul întregului proiect. Aceştia descriu proiectul din punctul de vedere al unităților lexicale ce compun textul (*frecvența unităților lexicale, număr acestora în text, vocabular cumulativ, statistici de vocabular/ocurențe*), sau la nivel global cu referire la întregul proiect, dar şi la nivelul caracterelor ce compun textul (*frecvență majuscule, minuscule şi simboluri*).

Foarte interesantă din acest punct de vedere este posibilitatea de a explora corpusul pe baza unor *liste de cuvinte* ce pot fi personalizate de către utilizator. Aceste liste de cuvinte sunt create pornind de la inventarul corpusului cu care se lucrează în acel moment. Opțiunea ***keywords*** permite atât selecția manuală a cuvintelor cheie cât şi utilizarea unor formalime care automatizează acest proces. Aceste liste se pot crea pe baza frecvenței unităților lexicale (se poate stabili între anumite limite prin utilizarea operatorilor <, > sau =), în funcție de structura cuvântului (prefixe, sufixe, sau indiferent de poziția unui element în cuvântul căutat) sau după criteriul de lungime a cuvântului (stabilit între anumite limite prin utilizarea operatorilor <, > sau =). Programul permite atât manipularea listelor de cuvinte precum şi importarea sau exportarea acestora în format TXT.

Prezentarea elementelor lexicale recuperate din text se poate face sub forma de ***index*** sau de ***concordanță de tip LINE sau KWIC***, pentru această ultimă opțiune putându-se selecta posibilitatea prezentării elementelor în funcție de contextul de stânga sau de dreapta al cuvântului.


Simple Concordance Program 4.09 (build 3), May 2006 - [gawain.scp]
File View Tools Help
KeyWords
834 words: --,a,abloy,
Ascending Alphabetic Order
Invert
Concordance
Word List
Statistics
Exact
Keys
un
2
As it occurs
Left context
Right context
Kwic
Line
Stop
KeyWords
prefix
suffix
word
anywhere
Concordance: 1 key, 8 citations
FLPV
Index
Concordance
un
3 1140 201 couthe cowpled hor houndes,// Unclosed the kenel dore and calde
3 1161 202 of arwes;/At uche wende under wande wapped a flone,/That
3 1181 203 daylyght lemed on the wowes,/Under covertour ful clere, cortyned
3 1201 203 and to-hir-warde torned,/And unlouked his yye-lyddes and let
3 1210 204 gay lady,/`Ye ar a sleper unslyye, that mon may slyde hider
3 1245 205 as ye reherce here/I am wyye unworthy, I wot wel myselven./
3 1327 208 were,/And didden hem derely undo as the dede askes./Serched
3 1352 209 thay hyyes,/Bi the bakbon to unbynde.//Bothe the hede and the
C:\Archivos de programa\scp32v409\gawain.scp
gawain
1 file, 313 lines, 2045 words, 834 word vocabulary



Limitele programului sunt vizibile în momentul în care limba textelor dintr-un anumit corpus presupune existența în texte a unor caractere speciale, cum este cazul limbii române. ***SCP*** permite utilizarea unui set restrâns de caractere (ANSI / ASCII) deşi numărul limbilor cu care se poate lucra este destul de mare. Din păcate această listă nu cuprinde şi limba română iar fonturile nu sunt adaptate pentru limbă noastră. Acest fapt are ca rezultat dificultăți în explorarea unui corpus în limba română, totuşi, acest lucru nu este imposibil.

Corpografo

Situându-se la un nivel superior în ceea ce priveşte complexitatea şi funcțiile pe care le oferă, ***Corpografo*** este prezentat ca instrument de analiză şi exploatare de corpus textual pentru elaborare de instrumente lingvistice în format digital. Scopul cu care a fost construit este crearea de surse lingvistice complexe cum ar fi: *dicționare generale*, *dicționare-tezaur*, *glosare specializate*,

corectoare ortografice, *aplicații de recuperare a informațiilor* sau de *traducere automată*.

Proiect portughez creat de o echipă formată din: Luís Sarmento, Ana Sofia Pinto, Luís Miguel Cabral, Débora Oliveira, Belinda Maia, Diana Santos[7]. Este vorba despre un proiect lingvistic cu acces gratuit on-line, însă necesitând obținerea unei chei pentru fiecare utilizator şi având de întâmpinat toate dificultățile accesului on-line la o bază de date aflată la mare distanță, ***Corpografo*** poate fi accesat la URL: ***www.linguateca.pt/corpografo***. Cu o interfață uşor de utilizat, care însă prezintă dificultatea unui meniu de comenzi în portugheză, este un proiect în curs de dezvoltare ce doreşte să integreze într-un mediu electronic unic diverse instrumente de analiză textuală ce sunt considerate în mod obişnuit ca entități individuale. Este destinat cercetătorilor, studenților, traducătorilor, etc.

Foarte bine organizat din punctul de vedere al sarcinilor pe care le poate îndeplini, ***Corpografo*** permite *crearea* şi *gestionarea* de corpusuri personalizate, *analiza*, *căutarea* şi *extracția terminologică* pe baza acestor corpusuri. Interfața acestuia prezintă patru meniuri de lucru reprezentate prin categoriile: ***Gestor, Pesquisa, Centro de Conhecimento şi Centro de Comunicaçao***. Fiecare dintre aceste categorii corespunde unei etape de lucru în cercetarea terminologică a unui corpus. Le vom prezenta pe fiecare pe scurt în cele ce urmează.

Gestor reprezintă un spațiu virtual de stocare a fişierelor care contribuie la formarea unui corpus ce stă la baza cercetării terminologice. Crearea corpusului corespunde unei prime etape a cercetării, de modul în care se realizează această etapă de investigare depinzând calitatea şi eficiența interogărilor ulterioare a corpusului creat. Astfel, o mare atenție trebuie acordată, în utilizarea ***Corpografo***, unei etape premergătoare de „pregătire" a textelor înainte de explorarea corpusului.

De asemenea, o mare importanță este acordată organizării informației în cadrul corpusului. Pentru o mai bună gestionare şi un control efectiv al informației conținute într-un corpus, există posibilitatea de a clasifica fişierele din interiorul acestuia în funcție de *tematica* pe care o tratează textul, iar în cadrul acestei categorii există descriptori ca: *mediul, domeniul şi subdomeniul de specializare*. Se poate menționa *sursa textului* (prin menționarea organizației, instituției sau a editurii) precum şi *autorul*, în cazul în care aceste informații sunt repetitive, există, posibilitatea de a relaționa între ele textele care au aceeaşi sursă sau aparțin aceluiaşi autor. Alte informații care însoțesc textul sunt: *numele fişierului, titlul documentului, limba textului, reguli de acces public, data de introducere şi de modificare*, precum şi o *descriere* a textului.

---

[7] Pentru mai multe detalii asupra rolului fiecărui membru al echipei în cadrul proiectului, vezi ***www.linguateca.pt/corpografo***

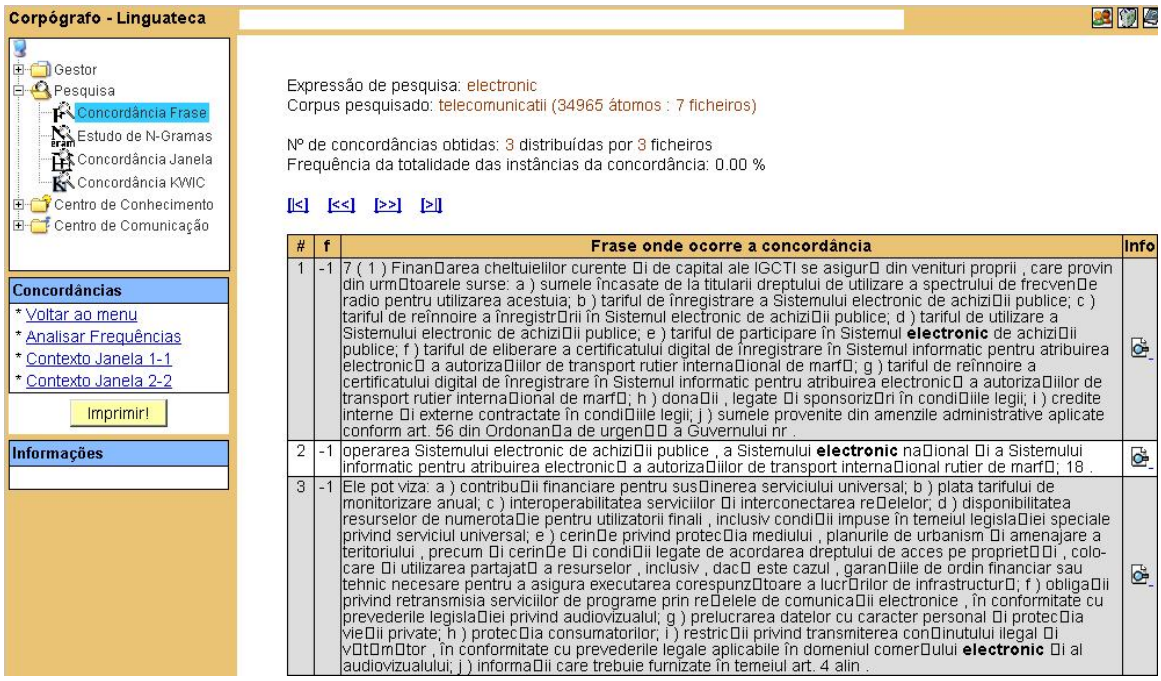
Pentru realizarea operațiunii de „pregătire” a textului, este necesară accesarea opțiunii de *editare a textelor* din meniul ***Texto.*** Se consideră necesară eliminarea tuturor elementelor irelevante din text, cum ar fi: *caractere speciale, referințe bibliografice, erori de ortografie, titluri, subtitluri, note bibliografice* existente în text, precum şi *formulele* ce pot apărea. O opțiune importantă pentru o primă luare de contact cu textul este ***Ver dicionário,*** ceea ce permite obținerea unei liste generale a *atomilor* unui text. O altă opțiune utilă este cea de comparare a două texte.

Odată încheiată etapa de *introducere*, *descriere* şi *pregătire* a materialului lingvistic se trece la crearea corpusului, după care se poate începe explorarea acestuia.

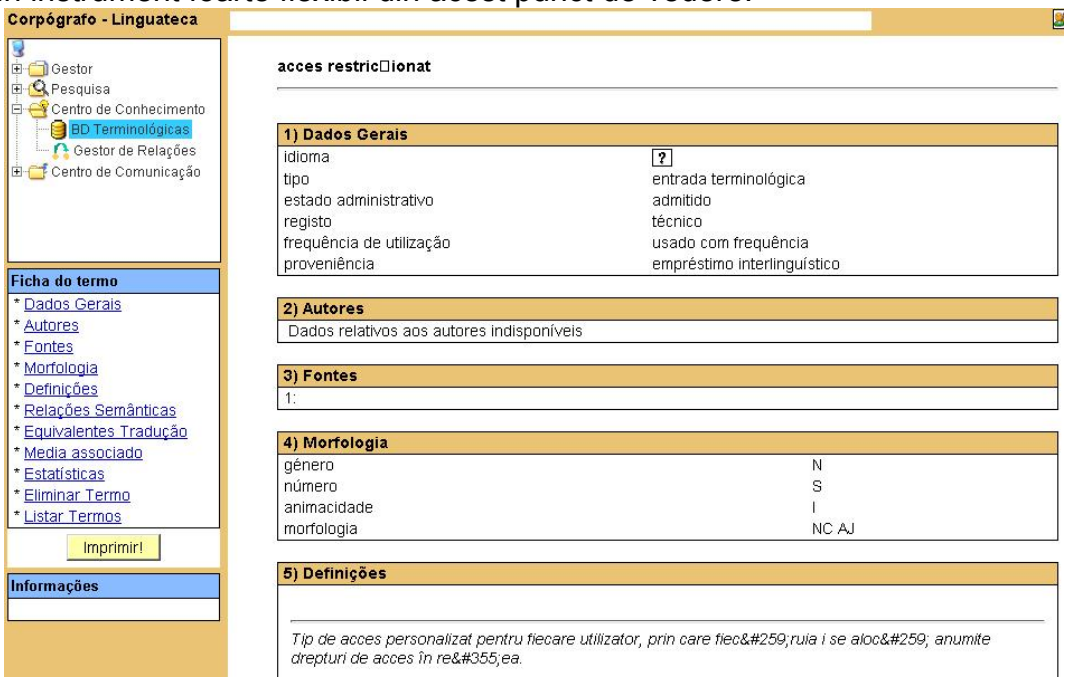
Funcțiile destinate explorării corpusului sunt grupate în meniul ***Pesquisa*** (căutare). ***Pesquisa*** corespunde etapei de cercetare lingvistică propriu-zisă, iar instrumentele pe care le pune la dispoziție permit *căutarea, studierea şi extragerea* de informații dintr-un corpus. O opțiune importantă se consideră a fi căutarea de secvențe de cuvine consecutive de diverse lungimi ce se pot stabili de către utilizator (***N-gramas***) şi care apar frecvent într-un text. Aceasta permite observarea unor structuri şi combinații sistematice de cuvinte, foarte utilă în identificarea termenilor de specialitate.

Foarte importantă la nivelul explorării textului este opțiunea de realizare a diverse tipuri de concordanțe: ***concordanța la nivel de frază***[8], ***concordanță Janela***[9] şi ***concordanță KWIC***[10]. În afară de aceste posibilități de explorare de corpus, ***Corpografo*** prezintă avantajul de a putea efectua căutări şi în funcție de expresii regulare, care, de asemenea, pot constitui baza de pornire în crearea de concordanțe.

[8] Tip de concordanță în care contextul se limitează la fraza în care apare cuvântul căutat.
[9] Tip de concordanță în care care se poate defini numărul de cuvinte care să constituie contextul de stânga şi dreapta al termenului căutat.
[10] Tip de concordanță în care se permite definirea contextului în funcție de numărul de cuvinte sau de caractere.

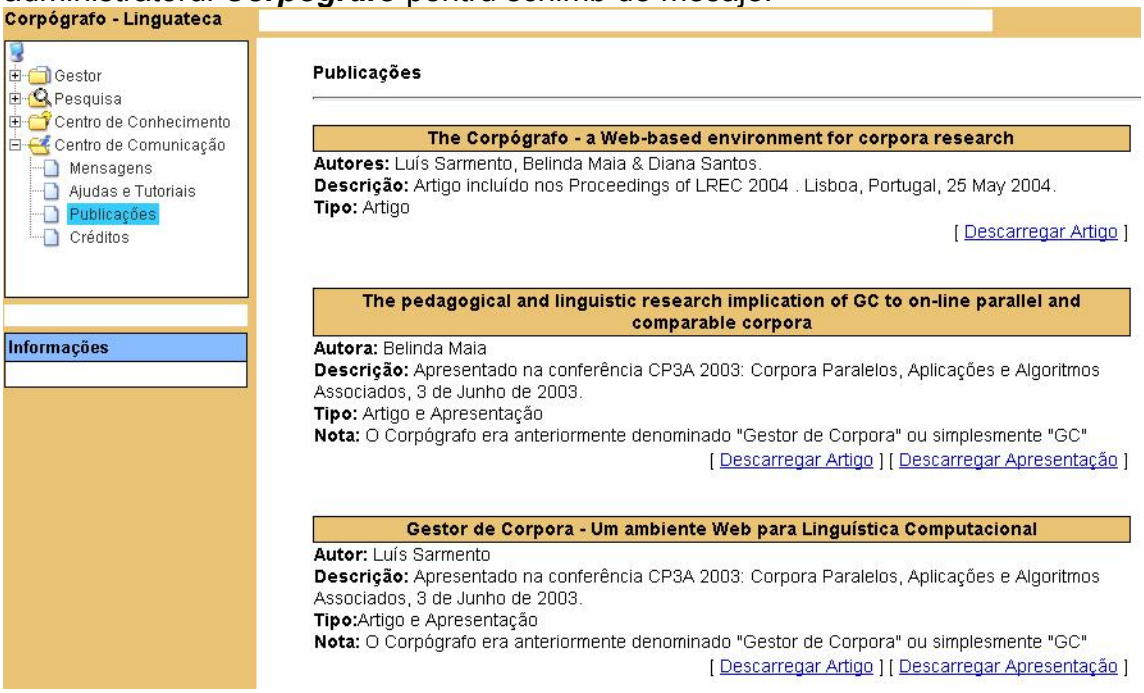


Centro de Conhecimento reprezintă spațiul în care se pot sistematiza şi organiza informațiile obținute în urma explorării unui corpus. Este vorba despre informații de tip *lexical*, *morfologic*, *sintactic* şi *semantic*, ce pot permite crearea de materiale lingvistice cum ar fi: *liste de cuvinte, glosare, reguli de căutare, tipare, relații semantice*, etc. Cu ajutorul ***Corpografo*** se poate crea o reprezentare formală a conceptelor şi informațiilor lingvistice asupra unui domeniu specializat. Acest aspect are aplicații importante în cercetarea terminologică, formalizarea cunoştințelor dintr-un domeniu specializat fiind un element fundamental. La nivel tehnic această secțiune a ***Corpografo*** permite gestionarea de baze de date terminologice. Se insistă asupra faptului că ***Corpografo*** nu doar permite gestionarea unor simple liste de cuvinte ci utilizatorul are posibilitatea de a stabili *relații* între termeni, ceea ce îi permite crearea de *rețele conceptuale multidimensionale, **Corpografo*** dovedindu-se a fi un instrument foarte flexibil din acest punct de vedere.



Centro de Conhecimento permite de asemenea gestionarea şi editarea bazelor de date terminologice, căutarea şi stabilirea de realții între termeni. În ceea ce priveşte căutarea termenilor într-un corpus, utilizatorul are posibilitatea de a stabili lungimea unei secvențe textuale, trecerea de la *forma flexionată* a acesteia la *forma normalizată*, de asemenea se pot omite termenii care deja sunt introduşi în baza de date. După introducerea tuturor termenilor selectați de către utilizator în baza de date, aceştia vor fi descrişi, în baza de date existând descriptori pentru *limbă, descriere morfologică, autor, referință bibliografică,* de asemenea este posibilă *căutarea unei definiții a unității lexicale* în cauză în corpusul de texte ce se explorează, căutarea de eventuale *relații semantice* între termenii bazei de date, *căutarea de echivalente de traducere*, *asocierea de elemente multimedia* pentru termenul respectiv, *consultarea de statistici* cu referire la un anumit termen existent în corpus.

Ultimul meniu, ***Centro de Comunicaçao***, se referă la partea de documentație asupra ***Corpografo*** şi la posibilitatea de a contacta administratorul ***Corpografo*** pentru schimb de mesaje.

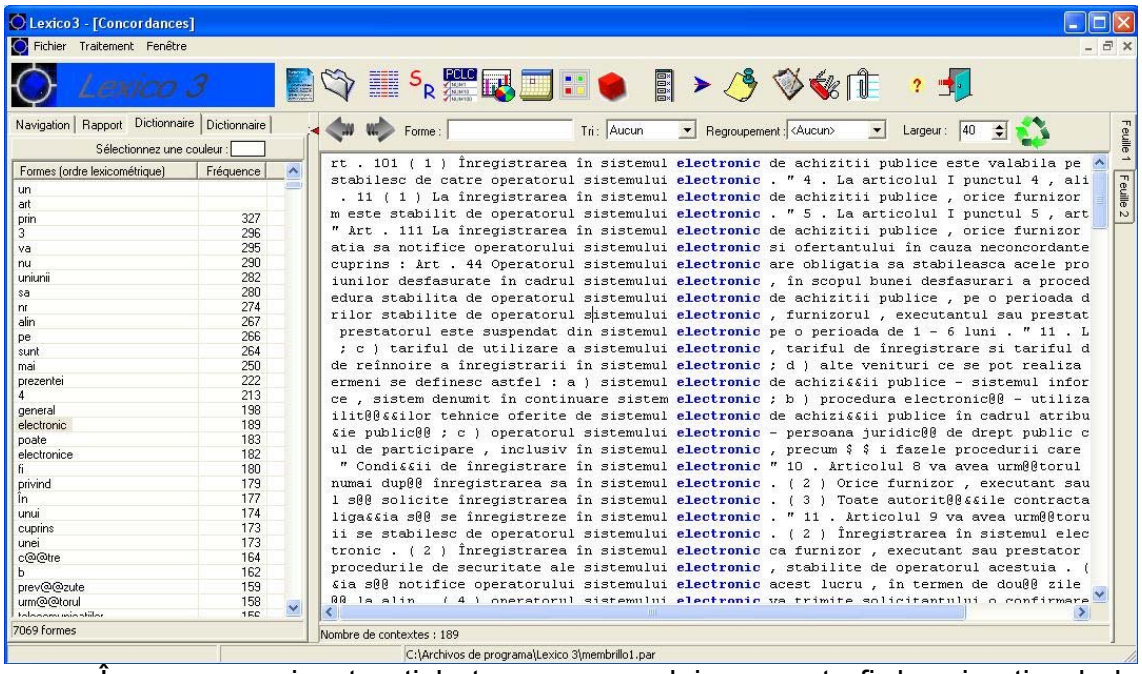


Din punctul de vedere al limbii române, ***Corpografo*** prezintă limite, deoarece limba română nu face parte dintre limbile propuse pentru studiu şi nu are implementate caracterele speciale din limba română. Totuşi utilizarea textelor în limba română este posibilă, dezavantajul fiind afişarea caracterelor speciale sub formă de coduri .HTML. În cazul în care ar exista un interes pentru acest instrument, contactarea membrilor proiectului şi propunerea implementării caracterelor speciale pentru limba română şi a limbii române ca limbă de lucru în acest program, ar fi cea mai simplă soluție. Avantajul cert pe care îl oferă față de celelalte instrumente electronice de explorare de corpus ce fac obiectul acestui articol, este faptul că este foarte flexibil în ceea ce priveşte formatul fişierelor care pot să compună corpusul. ***Corpografo*** acceptând simultan fişiere .DOC, .RTF, .PDF, .TXT, .PS, .HTML ceea ce reprezintă de departe cea mai largă gamă de formate de documente text acceptate până în prezent de un asemenea instrument electronic de acest tip.

Lexico3

Ultimul dintre instrumentele electronice care marchează evoluția actuală în lingvistica aplicată, şi pe care ne-am propus să îl prezentăm este ***Lexico3***, un complex de programe de statistică textuală, după cum îl prezintă autorii săi. Elaborat de către o echipă a Universității Paris 3– Sorbonne Nouvelle, din care fac parte Cédric Lamalle, William Martinez, Serge Fleury şi André Salem. Este un instrument complex, cu distribuție gratuită în scopul cercetării şi testării, care poate fi descărcat de la URL ***http://lexico3.no-ip.org/***.

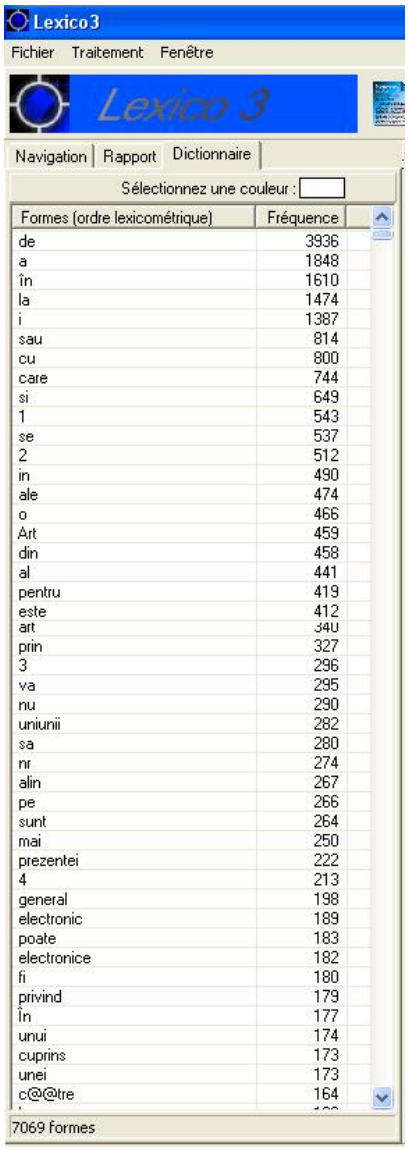
Cu o interfață transparentă şi uşor de manevrat, modul de lucru cu ***Lexico3*** se reduce la introducerea unui corpus de texte în format .TXT într-o bază, după care se poate trece direct, fără alte etape intermediare, la explorarea textului prin utilizarea a diverse instrumente de investigare, analiză şi statistică pe care ***Lexico3*** le pune la dispoziția utilizatorului. Avantajele certe pe care le permite programul în această primă etapă de încărcare a corpusului este faptul că permite o etichetare a corpusului, precum şi faptul că utilizatorului i se cere confirmarea utilizării unui inventar de delimitatori textuali cum ar fi: ***.,:;!?/_-\"'()[]{}§$*** şi care nu vor fi analizați ca şi componente ale textului ci vor avea un rol important în fragmentarea acestuia. Listă care poate fi modificată de către acesta, în funcție de tipul de text pe care doreşte să îl supună analizei statistice.



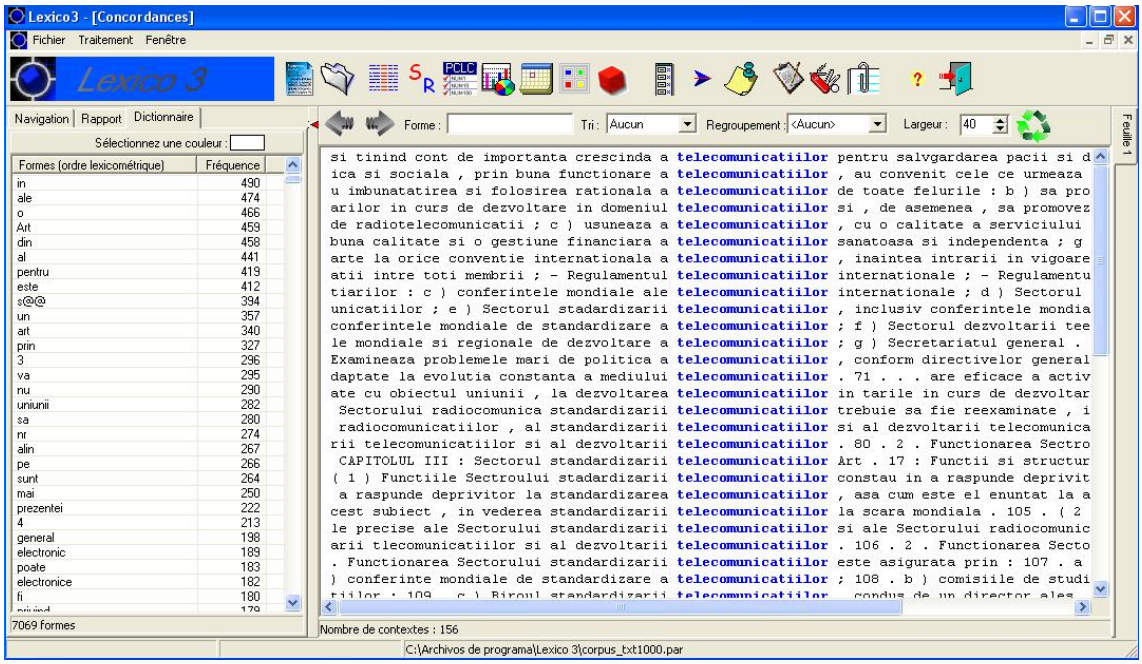
În ceea ce priveşte etichetarea corpusului, ea poate fi de orice tip, de la cea mai simplă la cea mai complexă, în funcție de ceea ce doreşte utilizatorul, etichetele sunt similare cu cele XML, pot fi definite cu cea mai mare libertate, urmând ca apoi să i se menționeze programului care sunt etichetele importante şi ce rol au în text. De exemplu, corpusul de texte demonstrativ al ***Lexico3*** se referă la presa din timpul Revoluției Franceze şi prezintă următoarea etichetare: ***<mois=01><quinzaine=11> <semaine=111> <Sda=1793> <numero=260> <edito=0> <Epg=1><Sat=0>***, destul de transparentă în prima parte, mai puțin transparentă în ceea ce priveşte ultimele patru elemente de etichetare. Utilitatea acestei etichetări se remarcă în momentul în care pentru o ilustrare grafică a distribuției unei ocurențe într-un text se cere delimitarea textului. Ca

delimitatori se pot folosi atât semnele de punctuație cât şi etichetele care au rolul de descriptori într-un text.

Prima operațiune care se poate efectua după introducerea corpusului în bază este consultarea dicționarului, lucru ce permite identificarea unităților lexicale care constituie textul şi frecvența acestora în text.

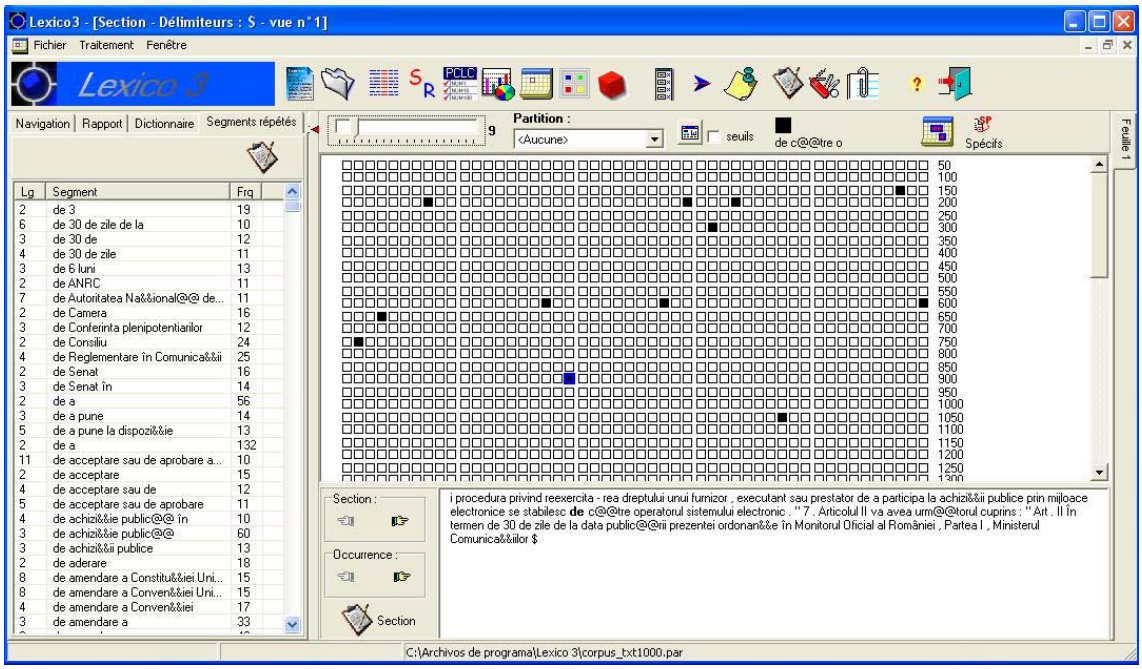


Pasul următor este crearea unei concordanțe. Spre deosebire de instrumentele prezentare anterior, interfața grafică a Lexico3 permite deplasarea elementelor lexicale cu ajutorul mouse-ului, deoarece are implementată funcția ***drag and drop***. Concordanța pe care o poate realiza este una simplă, singurul element deosebit fiind faptul că se poate stabili de către utilizator dimensiunea contextului iar ordonarea ocurențelor se poate face în ordinea apariției în text sau în ordine alfabetică în funcție de contextul de stânga sau de dreapta.



Există de asemenea o opțiune ce permite utilizatorului să afişeze toate formele unei unități lexicale prezente în text, acest lucru putând permite identificarea de câmpuri lexicale, familii de cuvinte sau forme în flexiune, împreună cu frecvența apariției lor în text.

Dacă este vorba despre un corpus etichetat, se poate obține o reprezentare grafică a segmentelor care compun textul, aşa cum au fost ele marcate de către utilizator. Astfel, se poate vedea, de exemplu, distribuția anumitor elemente lexicale, lucru destul de important în cazul analizei unui discurs sau a comparării a două texte. De asemenea pentru a putea studia distribuția şi uzul unei unități/segment lexical repetitiv, în partea inferioară a ecranului, se poate afişa contextul de utilizare.



O altă opțiune foarte utilă se referă la crearea unei liste a segmentelor care se repetă într-un text. Sunt posibile, de asemenea, analiza elementelor specifice unui fragment al corpusului sau o analiză factorială a corespondențelor în text.

Rezultatele analizei se pot păstra într-un raport care poate fi citit cu ***Netscape*** sau ***Internet Explorer***.

Limitele ***Lexico3*** se referă la aceeaşi problemă a acceptării limbii române ca limbă de lucru. Dar putând procesa fişiere în format .TXT, ***Lexico3*** prezintă avantajul de a putea lucra cu fişiere text ce conțin caractere UTF-8. Afişarea lor pe ecran lasă mult de dorit însă informațiile sunt accesibile utilizatorului. O altă limită a programului, care însă poate fi datorată faptului că pe Internet este doar o versiune demo, este limitarea corpusului de texte la 200 de pagini, ceea ce reprezintă un eşantion suficient pentru a studia posibilitățile de lucru pe care le oferă programul ***Lexico3*** dar care în contextul real al cercetării lingvistice şi al analizei de corpus este insuficient.

Concluzii

Prezentarea acestor instrumente electronice de explorare şi exploatare a unui corpus de texte poate forma cercetătorului o idee generală asupra tendințelor actuale de evoluție a instrumentelor electronice cu aplicație în domeniul lingvistic. Se remarcă astfel o trecere de la instrumentele simple care rezolvau o problemă la nivel punctual[11] la crearea de instrumente complexe, cu aplicație multidisciplinară, care unifică mai multe programe într-o suită ce prezintă avantajul fiabilității şi flexibilității. De asemenea, la nivel lingvistic se constată tendința de a elabora instrumente independente de limbă[12] astfel acelaşi instrument putând fi folosit pentru analiza oricărui text. Aşa cum se vede din prezentarea de față, această tendință are anumite limite iar limba română prezintă o problematică complexă în contextul lingvisticii aplicate. Există astfel diverse probleme la nivelul utilizării acestor instrumente de analiză şi recuperare a informației pe teren lingvistic românesc. Totuşi, este de remarcat faptul că nu este imposibil de a utiliza aceste instrumente cu rezultate bune, în momentul în care se conştientizează care sunt aceste probleme.

De asemenea, se constată o tendință de a oferi acces direct utilizatorului unui instrument electronic de analiză textuală, dar un acces controlat. Astfel, ***Corpografo*** nu este doar un instrument pus la dispoziția comunității ştiințifice pentru analiză şi cercetare lingvistică. Prin permiterea unui acces on-line la ***Corpografo*** se pot obține informații asupra interesului comunității ştiințifice asupra acestui instrument de lucru, asupra opțiunilor celor mai des utilizate şi a scopului în care acest instrument este utilizat, a limbilor de lucru, a domeniilor de interes pentru analiză lingvistică, astfel acest instrument se converteşte într-un instrument de măsură a necesităților existente în domeniul cercetării terminologice în special. De asemenea posibilitatea pe care acesta o oferă de a intra în contact direct cu cercetătorii care au proiectat instrumentul, asigură de asemenea feed-back-ul necesar pentru a dezvolta şi îmbunătăți acest instrument electronic.

Lipsa unui asemenea instrument de analiză textuală şi de explorare de corpus, dedicat limbii române sau a soluțiilor de incorporare a acesteia în cadrul altor instrumente de lucru deja existente, ne poate da o idee despre tendințele actuale în cercetarea lingvistică la nivel mondial şi, de asemenea, limitele şi carențele cercetării în acest domeniu la nivel local pentru limba română. Lipsa unor astfel de instrumente stă, în mod cert, la baza lipsei din peisajul lingvistic românesc a unor surse lingvistice de o calitate comparabilă cu cea a celor ce descriu alte limbi, materiale cum ar fi: *dicționare generale, dicționare specializate, dicționare-tezaur, glosare de termeni, corectoare automate, etc.* Considerăm că o îmbunătățire a acestui aspect al cercetării lingvistice şi o extindere a utilizării instrumentelor de analiză şi recuperare a informației bazată pe corpus poate avea ca rezultat o mai bună gestionare a informației lingvistice şi, în consecință o mai bună calitate în ceea ce priveşte elaborarea de resurse şi instrumente lingvistice ce descriu limba română.

---

[11] De exemplu nu se poate afirma că Corpografo sau Lexico3 sunt concordancier-e. Sunt şi concordanciere-e dar mai au şi alte funcționalități care le fac să iasă din această clasificare foarte îngustă.

12 Există instrumente lingvistice cu mai mare tradiție care sunt dedicate doar unei limbi (ex: ***The British National Corpus*** -http://www.natcorp.ox.ac.uk/) sau unui text (ex: concordanța textelor biblice sau cea a Constituției europene).

Bibliografie

1. ***Simple Concordance Program***, http://www.textworld.com/
2. ***Corpografo,*** www.linguateca.pt/corpografo
3. ***Lexico3,*** http://lexico3.no-ip.org/
4. ***The British National Corpus,*** http://www.natcorp.ox.ac.uk
5. ***Concordance biblique***, http://www.lueur.org/bible/bible_rechercher.php, pagina web a Bisericii baptiste protestante din Angers.
6. André Salem, ***Approches quantitatives des corpus textuels***, conferință IULA, Barcelona, 2006.
7. Luigi Sansonetti, ***Exploration textuelle d'interactions verbales entre un adulte et un enfant avec Lexico3***,
8. Andrea Kuncova, Aude Mansondieu, ***Outils de statistique textuelle. Manuel d'utilisation abrégé (Dix premiers pas avec*** *Lexico3*), SYLED-CLA2T, Université de la Sorbonne Nouvelle – Paris 3
9. Belinda Maia, Luís Sarmento, ***Gestor de Corpora – Um ambiente Web integrado para. Linguística baseada em Corpora***, www.linguateca.pt/corpografo
10. Belinda Maia, Luís Sarmento, Diana Santos, ***The Corpógrafo – a Web-based environment for corpora research,*** www.linguateca.pt/corpografo
11. Luís Sarmento (2004), ***Relatório Técnico sobre o Corpógrafo,*** http://poloclup.linguateca.pt/docs/cg/.