Instrumente de Procesare a Limbajului Natural Pentru Studierea Limbilor Străine

Bobicev V., Carcea L.
Catedra Informatica Aplicată
Universitatea Tehnică a Moldovei
Cişinău, Moldova
victoria.bobicev@gmail.com, carcea@mail.utm.md

***Abstract* — This paper presents an online concordancer that was created at the Applied Informatics department for teaching and learning of French language. It also contains examples of possible activities for teachers and students that can be used in the educational process.**

Besides tools specifically developed for learning and based on classical theories of language teaching, information technology offer other fundamentally new resources and tools that can be used in the educational process.

Corpora are the most important resources that are not yet fully utilized in language teaching.

The most difficult problem is the effective use of corpora. The contribution of TAL (Traitement Automatique des Langues / Natural Language Processing) is necessary for the development of tools and methods which simplify use of corpora. The concordancer is one of such tools. It can effectively process large amounts of text and reveal the necessary elements: keywords, phrases or grammatical constructions.

***Termeni cheie* — Concordansier, predarea limbii, studierea limbii, corpusuri de texte, procesarea limbajului natural.**

I. Introducere

Tehnologiile informaţionale oferă noi posibilităţi, care nu sunt pedeplin aplicate în domeniul educaţional. Vom menţiona diferite aspecte de utilizare oferite de învăţământul la distanţă şi tehnologiile informaţiei şi a comunicaţiei:

- posibilitatea de a obţine în orice loc sau orice moment cunoştinţe actuale şi moderne;
- studii în regim autonom cu materiale în format electronic la un calculator personal, sau un dispozitiv mobil;
- posibilitatea de a dezvolta resurse Web educaţionale;
- crearea unei comunităţi de utilizatori distribuiţi în reţele sociale, ceea ce conduce la activităţi educaţionale globale;
- posibilitatea de a interacţiona la distanţă, de a obţine ajutor sau a fi evaluat;

Furnizorul principal de informaţii şi sfaturi sub toate aspectele de utilizare a tehnologiei de învăţare a limbilor este EUROCALL [2]. Această asociaţie profesională a fost oficializată în 1993. Trei direcţii de dezvoltare au fost identificate pentru EUROCALL:

- utilizarea calculatoarelor în sala de studii;
- pregătirea profesorilor;
- dezvoltarea şi evaluarea produselor soft;

II. Procesarea Automată a Limbajului Natural

Procesarea automată a limbajului natural (PALN) este o disciplină plasată la frontierele lingvisticii, informaticii şi a inteligenţei artificiale. Ea ţine de aplicaţiile produselor soft şi tehnicilor informatice în toate aspectele limbajului natural (uman).

Să nu se confunde cu lingvistica informatică, care se bazează pe o modelare simbolică sau matematică a limbajului natural.

Primele cercetări au demaral la mijlocul secolului XX în scopul traducerii automate a textului. În continuare cercetătorii din inteligenţa artificială presupuneau că prelucrarea limbajului natural prin intermediul calculatoarelor este o problemă de scurt timp. Lipsa de rezultate a curmat acest punct de vedere şi utilizarea algoritmilor euristici. Au fost necesare cercetări în domeniul fomalizării gramaticelor limbajeor umane. Lucrările lui Chomsky [1] au condus la ipoteza de similitudine între limbajele naturale şi cele formale informatice. Dar şi gramaticele formale au întâmpinat numeroase obstacole în cazul limajului uman. A urmat încercarea de a aplica modele bazate pe cunoştinţe, care nu au permis obţinerea rezultatelor rapide şi universale. În consecinţă cercetătorii au acceptat dificultatea problemelor procesării limbajului natural şi au devenit mai modeşti.

Totuşi, creşterea vertiginoasă a circulaţiei documentelor, analiza, prelucrarea şi traducerea lor forţează căutarea soluţiilor viabile şi imediate a problemei automatizării acestor procese.

Apariţia Internetului şi utilizarea lui masivă provoacă urgent elaborarea instrumentelor ce ar putea prelucra informaţia de care au nevoie cercetătorii.

În aceste condiţii a apărut pe scenă şi încearcă să rezolve problemele existente, fără a intra în complexitatea funcţionării limbajului natural procesarea automată alimbajului natural (PALN). În general se aplică metode statistice, care conduc la rezultate satisfăcătoare. De altfel aceste metode pot fi uşor realizate în formă de produs soft. Condiţia principală a succesului este cantitatea majoră (aproape gigantică) necesară pentru crearea modelului statistic adecuat. Tocmai acest lucru a şi devenit disponibil la momentul actual.

III. Corpusuri de Texte

Corpusurile sunt în acelaşi timp materie primă pentru studiul limbilor şi a textelor cât şi mijloace pentru a testa modelele propuse. Un corpus este reprezentat printru-n număr

mare de materiale lingvistice: cărţi, reviste, rapoarte la lurări de laborator ale studenţilor, înregistrări audio şi video,discuţii, etc. Au apîrut mai multe publicaţii ce ţin de lingvistica corpusului înainte de era calculatorului, dar este problematic să găseşti dovezi pornind de la corpus după ce a fost imprimat pe hârtie. Chomski a criticat corpusul ca mod de studiere a limbii. Mai mult ca atât această critică nu era unicul reproş. Argumentul principal era caracterul infinit al limbii umane.În toate lucrările de la debutul lingvisticii corpusului au fost susţinute două ipoteze fundamentale, pentru moment imperfecte: frazele unui limbaj natural sunt limitate şi pot fi colectate şi numărate.

Însă numărul de fraze al unul limbaj natural este nu numai mare, el tinde spre infinit. Chomski a estimat că singura modalitate de elaborare a unei gramatici a unei limbi este posibilă prin descrierea regullilor sale şi nu prin enumărarea frazelor sale.

Alţi cercetători [8] afirmă corpusul este o metodă mult mai puternică din punct de vedere ştiinţific, fiindcă el este deschis unei verificări obiective a rezultatelor. Francis şi Kucera [6] au început să lucreze asupra celebrului corpus Brown, care a durat circa două decenii. Pas cu pas corpusul a devenit lizibil şi pentru calculator, ceea ce a simplificat considerabil manipulare şi utilizarea sa. Lingvistica corpusului modern defineşte corpusul ca un ansamblu de texte care au patru caracteristici:

- el trebuie sa fie lizibil pentru calculator;
- el trebuie să fie reprezentativ pentru domeniul elaborat;
- el trebuie să fie finit;
- toate textele trebuie să fie codificate şi adnotate în acelaş mod;

Majoritatea corpusurilor îndeplinesc aceste cerinţe, dar unele sunt în creştere continue. De exemplu, corpusurile ce conţin colecţii de ediţii periodice electronice cum ar fi reviste şi ziare. Unele corpusuri sunt create în scopuri diverse. Corpusurile reprezentative au tendinţa de a reprezenta ansamblul limbii cu toată diversivitatea sa particulară. Există corpusuri naţionale pentru unele limbi, cum ar fi British National Corpus[1] or Russian National Corpus[2].

Corpusurile speciale reprezintă fenomenele specifice ale unei limbi. De exemplu corpusul operelor lui Alexandre Dumas, sau corpusul corespondenţei de serviciu.

Diverse metode de creare a corpusurilor au fost dezvoltate, dar mai este mult de lucru pentru a obţine un corpus calitativ. Cele mai preţioase se consideră corpusurile marcate sau adnotate, dar marcajul calitativ necesită un lucru manual de profesionişti calificaţi. Adnotarea corpusului înseamnă adăugarea informaţiilor lingvistice interpretative acestui corpus. De exemplu, o adnotare de tip comun este adăugarea etichetelor ce indică clasa morfologică a cuvintelor, altfel spus marcarea categoriilor gramaticale şi poate fi utilă pentru căutarea tuturor formelor unui cuvânt.

Alte tipuri de adnotare sunt: adnotarea lexicală, adnotarea fonetică, adnotarea semantică, adnotarea sentimentelor, etc.

Corpusurile sunt utilizate în lexicografie fiindcă exemplele pot fi uşor extrase şi organizate pentru analiza ulterioară cu un produs soft. Ele sunt instrumente bune pentru cercetarea sintactică din motivul cuantificării reprezentând toată varietetea unei limbi. Avantajul principal este faptul că datele păstrate sunt în mare parte naturaliste. În acest mod corpusul constituie una din sursele fiabile de origine naturală ce poate fi examinată.

Totuşi corpusurile de texte sunt resurse foarte importante care nu sunt pe deplin valorizate şi nu sunt din plin utilizate în studierea şi predarea limbilor.

IV. CONCORDANSIERE

Utilizarea eficace a concordansierelor este dificilă motivul fiind volumul lor foarte mare. Pe de altă parte filologii nu sunt obişnuiţi cu aceste surse de informaţii. Utilizarea eficace a corpusurilor în lingvistică, filologie, lexicografie în studierea şi predarea limbilor este posibilă cu instrumente şi metode specifice. Cocordansierul reprezintă unul din aceste instrumente. El permite scoaterea din joc a unor cantităţi mari de informaţie şi vizulizarea elementelor necesare: cuvinte-cheie, fraze, sau construcţii gramaticale.

Termenul 'concordanţă' provine din engleză şi semnifică o listă alfabetică de cuvinte dintr-o carte sau un set de cărţi în care fiecare cuvânt poate fi găsit şi deseori explicată utilizarea sa. Elaborarea cocordansierelor în mod manual fără calculator a fost o problemă enormă şi dificilă. Astăzi un calculator poate produce un concordansier în câteva minute.

Una din primele referinţe la utilizarea concordansierelor electronice a fost [4], dar şi alte exemple cu utilizare practică în sala de studii sunt descrise de [7]. În 1991 Tim Johns a ameliorat utilizarea concordansierului în clasa de studii cu conceptul "studierea pilotată de date" (Data Driven Learning - DDL) [5]. DDL încurageză elevii să lucreze în baza propriilor reguli vis-a-vis de sensul şi utilizarea cuvintelor cu ajutorul unui concordansier pentru a găsi exemple intr-un corpus de texte autentice. De asemenea este posibil ca profesorul poate să gasească colocaţii tipice şi să genereze exerciţii bazate pe exemplele găsite.

În DDN procesul de studiere nu se bazează numai pe iniţiativa profesorului de a veni cu subiecte, materiale metodice, însuşirea explicită a regulelor, dar şi pe descoperirea propriilor reguli, principii şi modele de utilizare a limbii străine.

Principalul avantaj al concordansierului este rezumat prin fraza frecvent utilizată "cunoaşteţi cuvântul după vecinii săi" sau "spunemi cine sunt vecinii ca să-ţi spun cine eşti". Cu alte cuvinte vom găsi nu numai sensul unui cuvânt dar şi utilizarea sa cu alte cuvinte în textul dat.

În acest mod o concordanţă este un cuvânt numit cuvânt-cheie scos dintr-o lucrare în limbaj autentic (corpus) reprezentat cu părţi contextuale în care el se întâlneşte.

[1] http://www.natcorp.ox.ac.uk/

[2] http://www.ruscorpora.ru/en/

Cuvântul este recunoscut ca cuvânt-cheie în concordanţă de conext sau KWIC-concordanţă.

V. FUNCŢII TIPICE ALE UNUI CONCORDANSIER

Un concordnasier are ca regulă mai multe funcţii (instrumente) pentru predarea şi studierea unei limbi străine:

- Lista de cuvinte. Generează lista cuvintelor dîn textul dat şi le ordonează la afişare. Permite vizualizarea celor mai frecvente cuvinte în text şi celor care se întâlnesc o singură dată.
- Lista cuvintelor-cheie. Acest instrument selectează cuvintele frecvente în textul dat cu algoritmi sofisticaţi. Pentru început sunt eliminate cele mai frecvente cuvinte, care deobicei sunt prepoziţii, articoşe, pronume, etc. Apoi se elimină cele mai rar întâlnite, care nu sunt specifice textului dat. În final se compară frecvenţa cuvintelor rămase cu frecvenţa lor în baza corpusului. Cuvintele care au frecvenţa mai mare în text de cât în corpusul de referinţa sunt considerate cuvinte-cheie.
- Clustere. Acestă funcţie selectează clustere în baza unei condiţii de căutare. De obicei se selectează expresii curente în textul dat.
- Concordanţa cu un cuvânt şi cu un frqgment. Acest instrument va afişa rezultatele căutării în format KWIC (cuvinte-cheie în context). Concordanţele cu un fragment pot fi utilizate pentru căutarea cuvintelor care au aceiaşi rădăcină.
- Co-occurence cu un cuvânt şi cu un fragment.

Toate aceste instrumente sunt bazate pe statistică şi produc un rezultat bun în cazul unui volum mare de texte. Există totuşi şi o problemă cu concordansierul on-line: Volumul textelor transmis pin Internet. Traficul insuficient poate provoca probleme pentru texte mari.

Concordancierul poate fi utilizat pentru predarea şi studierea (învăţarea) limbilor străine în diferite maniere:

- Profesorul poate utiliza concordansierul pentru a căuta exemple autentice, colocaţii tipice etc;
- Profesorul poate genera exerciţii bazate pe exemple din diferite corpusuri;
- Eevii pot lucra cu reguli gramaticale de căutare a cuvintelor-cheie şi a caracteristicelor lexicale elaborate de ei însăşi. În funcţie de nivelul lor ei pot să formuleze întrebări ce ţin de reguli în baza observaţiilor modelelor limbii studiate.
- Elevii pot fi mult mai activi în învăţarea vocabularului: ei pot fi invitaţi să descopere semnificaţii noi, să obsreve colocaţii obişnuite sau relaţiile cuvintelor cu sintaxa, ba chiar să critice propunerile vocabularului;
- Elevii pot fi invitaţi să se pronunţe asupra utilizării limbii în general în baza propriilor analize a unui corpus de texte.

Există mai multe activităţi posibile în clasă, dar care trebuie pregătite de profesor cu concordansierul şi textele respective.

Site-ul Gobuild Web sugerează generarea unui cuvânt-cheie apoi eliminarea lui. Elevii trebuie să gasească acest cuvânt, ceea ce îi va încuraja să determine cuvântul din context şi nu din dicţionar. Se generează o listă KWIC pentru un cuvânt particular şi se va întreba care este prepoziţia care precede sau cre urmează după acest cuvânt-cheie.

Iată câteva exemple de activităţi posibile cu concordansierul în clasă:

Activitatea 1: Determinaţi cuvântul misterios.

Scopul activităţii: Familiarizarea studenţilor cu concordanţa KWIC şi cu importanţa conextelor din stânga din dreapta când se operează cu cuvinte cheie.

Foaie de lucru. Citiţi grila din tabelul 1 unde cuvântul fără sens "gloup" a fost înscris în locul unui cuvânt corect. Sarcina este să decideţi care este cuvântul corect.

Activitatea 2: Donc on peut dire que…

Scopul activităţii: conştientizarea elevilor cu deosebirile stilistice ale cuvântului "Donc" în limba franceză în texte scise formal faţă de cele neformale.

Foaie de lucru. În franceză puteţi plasa un cuvânt ca "Donc" la începutul unei fraze (de exemplu "Donc on peut dire que..."). Dar puteţi să-l întâlniţi des între verbul principal şi cel care urmează după verbul principal (de exemplu " On peut donc dire que ..."). Este unul mai bun ca altul? Unde trebuie inserat "Donc"? Pentru a cunoaşte acest lucru analizaţi listele din tabelul 2.

Activitatea 3: Le roman s'agit d'un amour malheureux.

Scopul activităţii: a implica elevii în activitatea în care ei deduc regula utilizării expresiei «s'agit» care nu este utilizată nici o dată cu alt subiect decât cu «il» impersonal.

Foaie de lucru "Le roman (ou "le poème" ou "la pièce") s'agit d'un amour Malheureux ... " Profesorii de franceză vor corecta întotdeauna primele două cuvinte în astfel de fraze în eseurile studenţilor. Pentru a explica reacţia cadrelor didactice analizaţi textele din tabelul 3 şi decideţi cum trebuie utilizată fraza «s'agit».

VI. CONCLUZII

Procesul de predare şi învăţare a limbilor străine asistate de calculator ne sugerează concluzia că calculatorul poate servi ca o varietate de utilizări pentru predarea limbilor. Calculatorul poate fi un tutore care propune exerciţii ce ţin de limba studiată; un stimulant pentru discuţii şi interacţiune; un instrument pentru a scrie şi pentru cercetare. Cu apariţia Internetului calculatorul poate deveni şi un mijloc de comunicare mondială şi o sursă de materiale autentice nelimitate.

Totuşi după cum a fost subliniat in [3], utilizarea calculatorului nu este o metodă. Mai degrabă putem vorbi de un mediu în care o varietate de metode, de abordări de filosofii pedagogice pot fi lansate. Eficacitatea predării asistată de calculator nu poate să rezide în mediul său, dar numai în modul în care aduce profit.

TABEL I. EXEMPLE PENTRU EXERCIȚIU 1.

1. pport critique sur certaines utilisations abusives de la	**gloup**	est devenu un geste banal plus qu'une décision.
2. que pour beaucoup d'entre nous le fait d'allumer une	**gloup**	.
3. laquelle on est pris pour gens qui "s'abrutissent à la	**gloup**	", dans une proportion croissante depuis 1896
4. Tous les grands moments de	**gloup**	superposent un message recherché et un messa
5. sieurs postes et l'augmentation du temps de diffusion (	**gloup**	du matin et de la nuit).

TABEL II. EXEMPLE PENTRU EXERCIȚIU 2.

Il devient	donc	difficile de proposer un plan de gestion de ces troupeaux.
Elle permet	donc	la circulation des avions de grandes dimensions en cas de besoin.
Ce droit exclusif n'atténue	donc	pas du tout les droits de ces derniers puisque ils ont accès à toutes les
Les autochtones ont	donc	priorité quant à la récolte.
Il serait	donc	intéressant de comparer des données plus récentes afin de
Ces chiffres révèlent	donc	une tendance à la baisse entre 1976 et 1980 mais sûrement aussi
Les données furent	donc	suggéré de répartir les territoires de chasse selon des zones "

.Cher Mr Je viens	donc	vous écrire pour vous demander si je peux avoir de l'octroi pour
suite pourvu que je sois certain d'avoir ma prime	Donc	je me fie entièrement à vous pour régler cette affaire-là et je
mais cette année la Compagnie Fraser l'achète.	Donc	espérant recevoir une bonne réponse de vous le plus tôt
et ils demandent déjà un bon prix	donc	s'il vous plaît enseignez-moi les moyens à prendre pour le
Bien cordialement. Soyez	donc	assez bon de me dire s'il y a encore de bons lots à prendre
Je connais la terre étant fils de cultivateur.	Donc	Monsieur le Curé je sais que si vous le voulez je pourrais aller
bien si vous pouviez venir inspecter ce chemin	donc	je veux pas vous ennuyer avec cela

TABEL III. EXEMPLE PENTRU EXERCIȚIU 3.

1. e de s'adapter au monde contemporain. Il	s'agit	de savoir si l'on table, oui ou non
2. es, de Beurs ni de Blacks (hélas). Il ne	s'agit	pas d'une bande dessinée mais
3. is gaulois. On en use à présent quand il	s'agit	d'évoquer les solutions apportées
4. oins comme ami, simplement parce qu'il	s'agit	de quelqu'un de différent
5. une femme. Et réciproquement. Quand il	s'agit	de cette différence-là il y a
6. olution partiellement dans le prototype. Il	s'agit	de définir les autorisations
7. uation de l'élève faite par le système. Il	s'agit	donc de recueillir ces informa

Una din principalele caracteristici ale învăţării asistate de calculator este individulizarea procesului de învăţare. Această funcţie este utilizată în abordările mai recente care favorizează abordarea centrată pe profesor. Ultima este caracterizată prin utilizarea programelor de concordanţă în clasa orelor de studiere a limbilor – abordare numită data Driven learning.

BIBLIOGRAFIE

[1] CHOMSKY, N., La Linguistique cartésienne suivi de La Nature formelle du langage, Éditions du Seuil, 1969.

[2] DAVIES, G., The History of EUROCALL: an article produced to celebrate the dawning of the new millennium, 2000 - http://www.camsoftpartners.co.uk/EuroHist.htm

[3] GARRETT, N., Technology in the service of language learning: Trends and issues. Modern Language Journal, 75(1), 74-101, 1991.

[4] HIGGINS, J. & JOHNS, T. Computers in language learning, London: Collins, 1984.

[5] Johns, T. & P. King (Eds.), (1991), Classroom Concordancing. English Language. Research Journal, 4. KETTEMANN, B. On the use of concordancing in ELT. TELL & CALL, 1995.

[6] Henry Kučera; W. Nelson Francis Computational Analysis of Present-Day American English International Journal of American Linguistics, Vol. 35, No. 1 (Jan., 1969), pp. 71-75.

[7] LAMY, M-N. & KLARSKOV MORTENSEN, H. J., Using concordance programs in the Modern Foreign Languages classroom. Module 2.4 in Davies G. (ed.) Information and Communications Technology for Language Teachers (ICT4LT), Slough, Thames Valley University, 2011 [Online]: http://www.ict4lt.org/en/en_mod2-4.htm

[8] LEECH, G., Corpora and theories of linguistic performance. In Svartvik J. (ed.) Directions in corpus linguistics, Berlin: Mouton de Gruyter, 1992.