

Corpusul electronic

un instrument fiabil pentru analiza lingvistică a terminologiei

Coancă Mariana
Universitatea Româno-Americană
București

Rezumat

Constituirea corpusului electronic depinde în mare parte de tipul de cercetare pe care dorim să-l efectuăm, iar rezultatele cercetării depind de constituirea corpusului. Folosirea unui corpus este indispensabilă pentru analiza lingvistică a terminologiei, deoarece corpusul ne poate oferi, prin intermediul textelor electronice, informații cu privire la detectarea, sursa și sensul termenilor din domeniul investigat.

Cuvinte-cheie

terminologie, texte electronice, corpus, analiză lingvistică

1. Terminologia în contextul actual

„Importanța socio-culturală, economică, pedagogică a **terminologiilor** este în creștere, ceea ce explică **abordarea interdisciplinară** și **interesul mai larg pentru** diverse domenii, dincolo de o anumită specialitate științifică. Se vorbește în acest sens de *laicizarea progresivă a cunoașterii*, tocmai din această perspectivă interdisciplinară” (Angela Bidu-Vrănceanu, 2007: 71).

Curentul normalizator, care se ocupă cu standardizarea termenilor, consideră **terminologia** o disciplină preocupată de *vehicularea cunoștințelor științifice și tehnice*, conform *triumfului terminologic* (ibidem: 57), care îi impune prin asocierea denumirilor cu definițiile dintr-un anumit domeniu. Stabilirea unor termeni specifici la acest nivel, este baza descrierilor lingvistice din orientarea descriptiv-lingvistică în **terminologie**.

În studii recente Angela Bidu-Vrănceanu consideră **terminologia** „disciplina care se ocupă de *comunicarea specializată*, realizată *fără echivoc* într-un anumit *domeniu științific, tehnic, profesional*.” Autoarea numește **terminologia specialiștilor** **terminologie internă (T1)**. De la această valoare, se ajunge ușor la desemnarea concretă a *ansamblului de unități* (lingvistice sau de alt fel), utilizate efectiv în comunicarea specializată dintr-un anumit domeniu (ibidem: 18-19). Astfel, de la o **terminologie a specialistului, internă**, interpretată strict în raport cu problemele comunicării specializate s-a ajuns, în ultimele decenii, la o

terminologie externă, de interes mai larg și *pentru nespecialiști* (ibidem: 19). „În **T1 (terminologia internă)** rolul lingvisticii e secundar, limitat la atribuirea de nume, la impunerea formelor lor (cu explicarea originii sau formării lor), problemele conceptual semantice căzând în sarcina specialiștilor. Deși condiționată de **T1 (terminologia internă)**, **T2 (terminologia externă)** se preocupă mai ales de problemele decodării și utilizării sensului specializat la mai multe nivele ale comunicării” (ibidem: 19).

Terminologia externă (T2), în sens larg, se constituie în lucrări din ultimul deceniu și își propune descrierea termenilor ca fapte naturale (nu convenționale ca în **T1**), similar cu cuvintele lexicului comun” (ibidem: 20). **Terminologia externă** are în vedere două nivele ale cercetării. Aceeași autoare (ibidem: 20-23) menționează că *primul nivel* urmărește stabilirea *specificului lingvistic* al termenilor din diferite domenii, descriși după o grilă lingvistică unitară, paradigmatică și sintagmatică. *Al doilea nivel*, are în vedere relațiile dintre LC și LS, prin tendința actuală de extindere a celui dintâi, dincolo de comunicarea specializată. Nivelele de analiză țin seama de criterii pragmatice: interesează nu numai *emittătorul* (mai mult sau mai puțin specialist), ci și *destinatarul* (care este tot specializat în textele specializate, sau un public eterogen, în cel de-al doilea caz). Analizând o gamă variată a condițiilor de uzaj, **terminologia externă** intercon condiționează nu numai cu **semantica**, **lexicologia**, ci și cu **pragmatica**” (Angela Bidu-Vrănceanu, 2007: 23). Astfel, teoria terminologică a evoluat de la o concepție esențialmente sistematică și normativă, care privilegia conceptele, spre o viziune mai realistă și mai complexă *a terminologiei, ca instrument de comunicare lingvistică* de diferite tipuri. Conceptul nu este o realitate fixă, independentă de discurs. Prin urmare, o *perspectivă lingvistică și comunicațională* este foarte importantă în descrierea termenilor din diverse domenii.

2. Corpus. Tipuri de corpus

Dacă activitatea terminologică se ocupă cu sistematizarea termenilor unui anumit domeniu, având ca obiectiv stabilirea relațiilor conceptuale și analizarea frecvenței, atunci instrumentul esențial pentru cercetarea și analiza terminologiei este **corpusul**, care este considerat un instrument foarte util pentru activitatea terminologică și traductivă (Aston, 2000: 21-29). Un corpus este definit și limitat de criteriul de timp și spațiu. Prin urmare, există o relație reciprocă de dependență între corpus și scopul cercetării: și unul și celălalt se condiționează reciproc. Noțiunea de **corpus** este utilizată de câțiva ani și este departe de a fi unanimă, fiind îndelung dezbătută. În lucrarea *Corpus, Concordance, Collocation*, Sinclair

definește *corpusul* „o colecție de texte, menită să caracterizeze starea și varietatea unui limbaj “ (1991: 171-172). Cu toate acestea, corpusul trebuie să aibă patru trăsături: să reprezinte un ansamblu **important** de texte **autentice**, compilate în format **electronic**, după **criterii precise** (Bowker și Pearson, 2002: 9).

În plus, *corpusul de texte specializate* prezintă avantajul de a furniza termenii și de a oferi explicații ale acestora (Kocourek, 1991: 23). Acest aspect este de o importanță majoră pentru disciplinele tehnice, în cadrul cărora *termenii* desemnează concepte precise pe care trebuie să le identificăm, pentru a recunoaște *termenii* și sinonimele, împrumuturile, abrevierile specifice unui domeniu specializat, tehnic. Constituirea corpusului presupune și selectarea textelor electronice. Colectarea textelor poate avea obiective variate, precum vulgarizări în revistele care îi informează permanent pe specialiștii din domeniul investigat, elaborarea dicționarelor sau descrierea lexicului respectiv și, ulterior, crearea unui instrument fiabil pentru traducerea termenilor.

Conform (Bowker și Pearson, 2002: 11-13) se pot distinge următoarele tipuri de corpus:

- *Corpus de referință* (general reference corpus) și *corpus specializat* (special purpose corpus). Un corpus de referință prezintă un limbaj comun și permite observații de tip general. Acest tip de corpus conține o mixtură de texte (jurnale, reportaje, dezbateri, etc.). Corpusul specializat prezintă aspectul specializat al vocabularului dintr-un anumit domeniu, pe baza unui anumit tip de texte, și nu în ultimul rând, pe baza limbajului membrilor unui grup social (de exemplu, tinerii). Prin utilizarea simultană a celor două tipuri de corpus (corpusul de referință și corpusul specializat), se pot identifica diferențele dintre limbajul specializat și cel comun (general).
- *Corpus scris* și *corpus oral - audio*. Un corpus oral este format din înregistrări de discursuri orale (conversații, emisiuni, cursuri etc.), iar un corpus scris conține texte scrise. Unele corpusuri, precum British National Corpus, sunt disponibile atât în varianta scrisă, cât și în cea audio.
- *Corpus monolingv* (general și specializat) și *corpus multilingv* (general și specializat). Spre deosebire de un corpus monolingv, un corpus multilingv conține texte în cel puțin două limbi străine. Corpusul multilingv poate fi împărțit în corpus paralel (uni-direcțional și bi-direcțional) și corpus comparabil. Un corpus paralel este format din texte în limba sursă și traducerea acestora în limba țintă. Un corpus

comparabil nu conține traduceri, ci texte scrise în limba sursă care au același subiect, aceeași funcție de comunicare (manual de instrucțiuni, raport tehnic, etc.) și au o dată de publicație relativ recentă.

- *Corpus sincron* și *corpus diacronic*. Un corpus sincron reprezintă imaginea uzului unui limbaj pe o perioadă limitată de timp. Un corpus diacronic analizează evoluția unui limbaj pe o perioadă lungă de timp.
- *Corpus deschis* și *corpus închis*. Unui corpus deschis i se adaugă în mod constant texte. Acest tip de corpus este utilizat, de obicei, în lexicografie. Un corpus închis rămâne neschimbat.
- *Corpus de învățare a unei limbi străine*, pentru cei care studiază o limba străină (learner corpus). Acest tip de corpus conține texte scrise de cei care învață o limba străină (non-native speakers). Prezintă interes faptul că se pot efectua comparații cu textele scrise de vorbitorii nativi.
- *Corpus ad hoc*.
- *Corpus echilibrat*. Un corpus echilibrat este format din diverse tipuri de texte (tehnice, științifice, texte vulgarizate, texte științifice – vulgarizate), pentru a înlesni căutarea termenilor principali, a termenilor vedetă. În cadrul unui corpus echilibrat găsim mai multe tipologii textuale, iar structurile perifrastice sunt diferite. Un exemplu de corpus echilibrat poate fi corpusul dedicat domeniului informaticii.

3. Procesul de documentare și utilizare a corpusului

Procesul de documentare presupune folosirea resurselor oferite de Internet. Pentru a completa textele electronice, se pot căuta pe Internet glosare specializate care ne-ar putea ajuta să înțelegem termenii tehnici întâlniți, ca să putem găsi un echivalent pentru termenul întâlnit. Un alt motiv care explică de ce se apelează atât de des la instrumentele oferite de Internet, este reprezentat de o *interdisciplinaritate accentuată a terminologiilor*, care este caracterizată de o combinație variată de termeni și de concepte care provin din discipline, domenii diferite. Astfel Internetul devine o sursă importantă, fiabilă de documentare, pentru realizarea cercetării. În plus, evoluția continuă ce caracterizează limbajele specializate reprezintă un motiv în plus, care ne determină să apelăm la resursele internetului, ca sursă fundamentală de documentare și luăm în considerare faptul că printre caracteristicile esențiale ale acestui mijloc de diseminare a informației, apare și actualizarea constantă a materialelor puse la dispoziție.

Dacă activitatea terminologică consideră textele ca fiind mediul natural al *termenilor*, lexicografia este esențialmente normativă și nu ține cont de folosirea reală a termenilor, pentru că această folosire nu este tot timpul în conformitate cu normele. Exemplele de dicționare de limbă sunt, de altminteri, foarte des inspirate din texte literare și nu din limba vorbită. Astfel, limbajul dicționarilor nu era limbajul așa cum trebuie vorbit, ci așa cum considera lexicograful că trebuie vorbit (Cabré, 1999: 37-38). Se poate spune că această modalitate era exact inversul activității de cercetare pe baza corpusului: mai degrabă se elabora o normă, o descriere a limbajului, apoi se aplica la limbajul „real” (care intră sau nu în acest cadru prestabilit), decât să se plece de la limbaj pentru a găsi un model. Exemplele din dicționare erau, cel mai adesea, enunțuri menite să justifice un model, o teorie și nu o manifestare autentică a limbajului.

Utilizarea *corpusului* și a *corpusurilor electronice* în special, permite stocarea multor informații și reprezintă un procedeu mai recent în redactarea dicționarilor. *Corpusul* poate furniza *contexte* care ilustrează *sensul și folosirea termenilor*. De asemenea, contextul ne permite, de asemenea, să constatăm schimbările de sens și să urmărim evoluția limbajului comun sau specializat (utilizarea cuvintelor noi, dispariția cuvintelor sau sensurilor etc.). Dacă este actualizat în mod regulat, *corpusul* constituie un foarte bun „observator” al limbajului specializat, interdisciplinar. Documentarea este o activitate de sine stătătoare în cadrul activității terminologice. Unii autori consideră că „unitățile terminologice fac parte din *limbajele de specialitate* și apar natural în *textele de specialitate*” (Cabré 1999: 80). Ni se pare de necontestat faptul că, fiind vorba de limbaje de specialitate, nu se repune în discuție „legitimitatea” textului, existența lui în raport cu dicționarul. Dacă ne întrebăm unde se găsesc termenii specializați, răspunsul logic pare să fie: în textele specializate, în discursurile specialiștilor. Dacă începuturile terminologiei au fost strict onomasiologice, dezvoltarea *lingvisticii de corpus* i-a determinat pe terminologi să se orienteze spre semasiologie.

4. Condiții de realizare a corpusului

Conform lui Sinclair (1991: 171) un *corpus* elaborat în mod adecvat oferă informații privind uzul autentic al *termenilor* dintr-un anumit domeniu, ajutând la rezolvarea problemelor lexicale, frazeologic, sintactice pe care le presupune și traducerea specializată. În procesul de elaborare a corpusului, trebuie să îndeplinim acele cerințe pe care lingvistica modernă le consideră indispensabile pentru realizarea corpusurilor: *reprezentativitate, mărime finită, standard de referință*. În ceea ce privește prima cerință, corpusul ar trebui să garanteze o reprezentare

adecvată a varietății lingvistice în chestiune și să reflecte uzul real al *limbajului comun sau specializat*. Pentru a garanta acest aspect, punctul de plecare trebuie să fie un nucleu constituit dintr-o serie de documente, care tratează subiectele respective din domeniul de interes sau printr-o serie de varietăți lingvistice.

O a doua caracteristică a *corpusurilor* este, așa cum s-a afirmat, numărul finit de texte pe care acestea îl conțin. Această limitare va trebui făcută printr-o selecționare atentă a textelor semnificative din punct de vedere cantitativ, menită să satisfacă cerința de *reprezentativitate* menționată mai sus. Mai adăugăm și faptul că a doua caracteristică a *corpusurilor* nu este deloc statică și că cerințele de înțelegere a terminologiei în chestiune, precum și necesitatea de a individualiza echivalențe au dus la îmbogățirea nucleului inițial cu texte care sunt de mare interes, conform celor afirmate de unii autori: „un corpus (...) trebuie să fie cât mai amplu” (Sinclair 1991: 18). Acesta vorbește și despre *trimming* și de “adding extra texts where there are gaps – adăugarea de noi texte când acestea nu sunt suficiente” pentru realizarea obiectivelor cercetării (ibidem: 20). În ceea ce privește dimensiunile, același autor sugerează că aceste corpusuri trebuie fie alcătuite din texte întregi, nu din simple fragmente. Această soluție este impusă de faptul că unele caracteristici lingvistice, care trebuie observate în texte, nu sunt distribuite uniform în interiorul unui document.

A treia caracteristică a *corpusurilor* este reprezentată de formatul lor *electronic*, care permite o serie de cercetări calitative și cantitative, mult mai greu de realizat cu instrumentele tradiționale, oferind astfel informații mult mai rapide, în ceea ce privește uzul termenilor din domeniul investigat. Corpusul poate fi de folos studiilor despre varietatea lingvistică, putând fi utilizat apoi ca bază pentru cercetări viitoare. În afară de proprietățile deja menționate, în elaborarea *corpusurilor* se pot lua în considerare și alte aspecte care influențează calitatea textelor și care pot satisface cerința de reprezentativitate a varietăților lingvistice de interes și aspectele menționate mai sus. Alegerea textelor se face ținând cont de provenienței acestora, acordând prioritate textelor electronice ce aparțin unor instituții și autori de articole despre domeniul investigat (limbaj comun sau limbaj specializat). O altă cerință în alegerea textelor electronice este cerința cronologică, mai exact se selectează texte electronice recente și actualizate care reflectă într-o manieră adecvată schimbările lingvistice rapide ce caracterizează domeniul investigat. Un alt criteriu fundamental este reprezentat de atenția acordată *destinatarilor* textelor electronice din *corpus*. În cazul unei terminologii specializate, se acordă o mare importanță textelor care sunt destinate atât specialiștilor, cât și nespecialiștilor.

5. Concluzii

Analiza lingvistică se bazează pe principiile de alcătuire a *corpusului*. Prin urmare, există o relație de dependență între *corpus* și scopul cercetării. *Corpusul electronic* este un instrument fiabil, întrucât este verificat prin analiza lingvistică *la nivelul exprimării sintagmatice* (analiza sintagmelor care se întâlnesc frecvent în diferite tipuri de texte electronice și analiza sintagmelor susceptibile de a avea statut de termen) și *la nivelul exprimării paradigmaticice* – relațiile semantice care se stabilesc între *termeni*, făcându-se aprecieri cantitative, pentru a evidenția particularitățile unei terminologii.

Referințe bibliografice

1. Aston, G. "I corpora come risorsa per la traduzione e per l'apprendimento", in S. Bernardini e F. Zanettin (a cura di) *I corpora nella didattica della traduzione*, Bologna: CLUEB, 21-29, 2000.
2. Bidu-Vrânceanu, A. *Lexicul specializat în mișcare. De la dicționare la texte*, Editura Universității din București, 2007.
3. Bowker, L., Pearson J. *Working with Specialized Language: A Practical Guide to Using Corpora*, Routledge, New York, 2002.
4. Cabré, M. T. *Terminology: Theory, methods and applications*. Amsterdam, Philadelphia: John Benjamins Publishing Company, 1999.
5. Kocourek, R. *La langue française de la technique et de la science. Vers une linguistique de la langue savante*. 2^{ème} Édition. Wiesbaden: Oscar Brandstetter Verlag, 1991.
6. Sinclair, J. M. *Corpus, Concordance and Collocation*. Oxford: Oxford University Press, 1991.