# Prefaţă

Născută în a doua jumătate a secolului trecut, concomitent în Europa şi în SUA, lingvistica matematică are acum o întreagă istorie iar dezvoltarea ei este în plin progres. Şcoala românească a avut încă de la început o prezenţă solidă în domeniu (realizând chiar lucrări de pionierat), aşa cum se poate vedea dintr-o lucrare de sinteză apărută în anul 1978 [1]: peste 500 de titluri de peste 120 de autori.

Lucrarea de faţă se referă la unele aspecte cantitative şi formale ale limbajului natural, cu atenţie specială acordată limbii române. Sunt câteva decenii de când nu am mai avut, la Facultatea de Matematică a Universităţii din Bucureşti, o teză de doctorat cu această tematică, în ciuda faptului că s-au prezentat numeroase teze referitoare la modelarea matematică a limbajelor, domeniu în care România se manifestă puternic pe plan mondial.

Sunt două părţi, bine distincte, chiar dacă în strânsă legătură în această lucrare: prima, caracterizată prin obiectul ei, silaba, ocupă capitolele al doilea (Aspecte ale silabei) şi al treilea (Un model contextual-Marcus al silabei); a doua, caracterizată prin atenţia îndreptată spre anumite probleme generale de similaritate, agregare şi clasificare, dar care şi ele sunt confruntate în final cu situaţii punctuale din studiul limbilor naturale, ocupă capitolele al patrulea (Despre similaritate), al cincilea (Metode multicriteriale de decizie) şi al şaselea (Aplicaţii şi rezultate experimentale). O bogată bibliografie de peste 120 de titluri (cele mai multe din ultimii 10 ani) folosită cu maturitate de către autor, încheie lucrarea.

Silabei i s-a refuzat statutul de unitate structurală a limbii, fapt care a contrastat cu acceptarea acestui statut pentru fonem şi morfem. În consecinţă, nici modelarea matematică a silabelor nu a ajuns la complexitatea modelelor propuse pentru fonem şi morfem. În deficit la înţelegerea calitativă, logică, a problemei silabei, cercetarea s-a orientat cu precădere spre aspectele cantitative, statistice. De exemplu, V. V. Ivanov atrăgea atenţia asupra poziţiei mediane pe care silaba o are, între fonem şi morfem: dacă numărul fonemelor este, în orice limbă, cuprins între $10^1$ şi $10^2$ iar cel al morfemelor între $10^3$ şi $10^4$, numărul silabelor se află în intervalul dintre $10^2$ şi $10^3$. Există deci

[1] Marcus, S. Mathematical and computational linguistics and poetics. *Revue Roumain de Linguistique*, XXIII, 559-588, 1978

o armonie combinatorială care urmează principiul minimului efort, aflat la baza multor legităţi cantitative care guvernează limbajul uman. Dar detaliile acestui fenomen se lasă cu greu identificate iar cercetarea din ultimele decenii încearcă să sporească lumina în aceste direcţii.

Tocmai de aceea, Liviu P. Dinu întreprinde, în capitolul al doilea, un tur de orizont al ipotezelor şi rezultatelor propuse de diverşi autori. Nu cunoaştem să existe în literatura recentă o altă privire, la fel de bogată, asupra acestei chestiuni, dar, desigur, o atare sinteză rămâne totdeauna deficitară, atâta vreme cât nu se poate raporta la situaţia din fiecare limbă naturală. În ultima parte a capitolului al doilea, autorul propune un algoritm propriu de segmentare silabică a cuvintelor din limba română, după care este analizată una dintre puţinele încercări, numai parţial reuşite, de formalizare a silabei, încercare aparţinând lui Theo Vennemman (1978). Este astfel pregătit terenul pentru ceea ce se va întreprinde în capitolul al treilea, în care autorul prezintă propria sa încercare de modelare matematică a silabei lexicale (grafice).

Ceea ce autorul întreprinde în capitolul al treilea constituie o iniţiativă dublă: mai întâi, propunerea, în studiul silabei, a folosirii unui instrument pe care nimeni nu s-a gandit până acum să-l asocieze cu o problematică de acest fel; în al doilea rând, perfecţionarea instrumentului respectiv, cu consecinţe posibile în domeniul teoriei gramaticilor formale. Gramaticile contextuale la care recurge Liviu P. Dinu sunt de natură predominant calitativă şi utilizarea lor în studiul silabei confirmă faptul că problemele cantitative sunt organic legate de cele calitative. Cantitatea şi structura se află într-o dialectică foarte fină, ele interacţionează. Ideea ingenioasă a autorului pleacă de la analogia pe care o observă între tăietura silabică şi procesul de generare în gramaticile contextuale; pe această bază, Liviu P. Dinu introduce o variantă nouă de gramatici contextuale totale şi pe care le consideră în două variante. Cu ajutorul lor se definesc gramaticile silabice, care se plasează sub aspect generativ între gramaticile contextuale interne cu selecţie şi cele contextuale totale. Itinerarul propus de Liviu P. Dinu face jonţiunea cu aşa numitele automate "*go-through*" introduse în 2001 de Radu Gramatovici (în cu totul altă ordine de idei), identifică astfel o clasă de automate eficiente pentru despărţirea în silabe şi, pe parcurs, implică, în mod surprinzător, structuri de tipul şirurilor lui Fibonacci. Aceste structuri sunt introduse pentru a investiga cuvintele regulate din limba română, cuvinte care sunt construite pe un alfabet de două elemente, V (vo-

cală) şi C (consoană); ele amintesc de problema pe care şi-o punea A. Markov, în urmă cu aproximativ o sută de ani, când, căutând să modeleze alternanţa consoanelor şi vocalelor în "*Evgeni Onegin*" de Puşkin, a introdus noţiunea pe care mai târziu matematicienii au numit-o "*lanţ Markov*". Şi sub acest aspect, demersul lui Liviu P. Dinu este foarte atractiv. Dar miza sa este, în această privinţă, mai ambiţioasă, prin apropierea pe care o face de teoria lui Levelt şi Indefrey, din 2001, privind producerea vorbirii, teorie în care silabificarea constituie o etapă esenţială. Plecând de la ipoteza unei funţionări mai degrabă paralele decât secvenţiale a creierului în formarea frazelor, este propusă şi o abordare paralelă a segmentării silabice prin introducerea gramaticilor de inserţie cu derivare paralelă. Ştiintele cognitive sunt aici direct vizate.

În capitolul al patrulea se pleacă de la tendinţa limbajului natural de a plasa informaţia cea mai semnificativă în prima parte a mesajului şi se propune, într-o primă fază, o măsură de similaritate a unor clasamente diferite ale aceloraşi obiecte. Se consideră şi cazul în care obiectele considerate în cele două clasamente sunt numai parţial aceleaşi. Sunt valorificate idei dintre cele mai variate, de la clasica distanţă a lui Hamming din teoria codurilor până la măsurile de similaritate folosite recent în studiul genomului uman. Autorul introduce ceea ce numeşte *distanţa rang*, permiţând compararea unor clasamente de obiecte şi lungimi diferite. Extensiunea acesteia la cuvinte permite definirea unei măsuri de similaritate a arborilor. Sunt identificate viitoare extinderi ale acestei măsuri, dintre care remarcăm extinderile posibile la calculul cu cuvinte al lui Zadeh, rezultatele putând fi astfel adaptate la operarea cu percepţii.

Agregarea si clasificarea clasamentelor fac obiectul capitolului al cincilea. În raport cu distincţia care se operează în literatură între metodele (de clasificare şi agregare) supervizate (care necesită o mulţime de antrenament) şi cele nesupervizate, autorul prezintă două metode supervizate de clasificare şi propune propria metodă nesupervizată de clasificare. Toate aceste metode se bazează pe interferenţa deciziilor mai multor clasificatori, autorul demonstrând astfel că este conectat la tendinţa ultimilor 10-15 ani din literatura ştiinţifică de specialitate. Problema determinării autorului unui text controversat se situează printre aplicaţiile posibile ale metodelor de clasificare multicriterială. Metodele de agregare şi clasificare introduse de autor se prevalează de distanţa rang introdusă în capitolul anterior şi sunt comparate cu alte metode. Mai este

propus un algoritm de agregare polinomial şi unul de clasificare, corespunzător metodei autorului. Urmează o analiză interesantă a modului în care metoda de agregare satisface una sau alta din condiţiile de raţionalitate propuse în această privinţă. Reperele sunt constituite aici de faimoasa teoremă de imposibilitate a agregării a lui Arrow şi de cele 17 condiţii de raţionalitate formulate de Gheorghe Păun. Sunt formulate, în încheiere, câteva probleme deschise.

Din ultimul capitol, al şaselea, dedicat unor aplicaţii, se detaşează ca interes studiul comparativ al limbilor romanice, din punctul de vedere al celor mai frecvente silabe în fiecare dintre ele. Cuvintele din vocabularele reprezentative ale acestor limbi au fost descompuse in silabe, acestea din urmă fiind ordonate dupa frecvenţă, după care ierarhiile obţinute au fost studiate din punctul de vedere al distanţei rang dintre ele. O serie de tabele şi grafice prezintă rezultatele obţinute. Se constată că dintre distanţele rang ale limbii române faţă de celelalte limbi romanice cea mai mică este distanţa faţă de limba portugheză. Pe de altă parte, se constată că, din punctul de vedere al structurii silabice, distanţa limbii române faţă de una sau alta dintre limbile romanice occidentale este mai mare decât distanţa dintre două limbi romanice occidentale, oricare ar fi acestea. Aceste rezultate concordă cu aşteptările care provin din alte considerente, de ordin istoric şi structural, dar aduc o serie de precizări de fineţe.

Acesta este, în esenţă, conţinutul lucrării propuse de dl. Liviu P. Dinu; lucrarea demonstrează capacitatea autorului de a întreprinde o cercetare conformă cu cerinţele actuale ale ştiintei, deci bazată pe o informaţie bogată şi la zi şi răspunzând unor probleme semnificative, în imediata continuare a cercetărilor anterioare ale altor autori sau ale sale proprii. Dupa cum s-a putut vedea, autorul are nu numai cunoştinţe ci şi aptitudinea de a lua iniţiative fericite, de a imagina procedee ingenioase.

Sperăm ca această carte să stimuleze interesul studenţilor şi al comunităţii ştiinţifice cu preocupări în domeniul lingvisticii matematice şi computaţionale.

Bucureşti, mai 2004 Acad. prof. univ. dr. docent Solomon Marcus