

Aspecte ale transcrierii limbii române vorbite în vederea prelucrării computerizate DIANA GHIDO

Institutul de Lingvistică
„Iorgu Iordan – Al. Rosetti”, București

Interesul crescând pentru studiul limbii vorbite se justifică prin numeroasele sale aplicații: reevaluarea gramaticii sau elaborarea unor dicționare de expresii pe baza unor exemple din vorbirea reală, spontană (nu preluate din textele literare sau inventate), predarea limbilor străine din perspectiva variantelor stilistice ale limbii, pentru o mai bună adaptare a discursului la situația de comunicare, studiul diverselor aspecte etnopragmatice etc. În plus, multe studii de limbă vorbită au în vedere astăzi recunoașterea vocală și perfecționarea interacțiunii om – mașină, pornind de la interacțiunea verbală interumană.

1. În ultimele două decenii se constată că numărul de **corpusuri de limbă vorbită** pentru diverse limbi a crescut considerabil. Multe dintre ele sunt specializate, autorii urmărind: particularitățile discursului didactic, învățarea limbii materne, limbajul adolescenților, abordări interculturale ale comunicării etc.

1.1. Stadiul cercetărilor în această direcție diferă foarte mult de la o limbă la alta. Dacă pentru limba engleză se publica în 1980 primul corpus oficial de limbă vorbită, astăzi ea se bucură de numeroase astfel de corpusuri (sau incluzând un subcorpus consistent de acest gen): *London-Lund Corpus*, *The British National Corpus*, *ICE BG Corpus*, *Corpus of London Teenage Language*, *COBUILT Bank of English* ș.a.m.d.

În ceea ce privește limbile romanice, amintim, pentru limba franceză, *GARS-ESA 6060* al CNRS și *Corpus de referință al francezei vorbite*, pentru italiană, *Lessico di frequenza dell'italiano parlato*, *CHILDES ITALIA*, *LIR del MURST* etc., pentru spaniolă, *Corpus Oral de Referencia del Español Contemporaneo*, la care se adaugă o serie de alte corpusuri specializate pe studiul lexicului, al dialogului, al discursului public etc.; pentru limba portugheză, cel mai semnificativ este *Corpus de referință do português contemporaneo*. Există și rețele internaționale pentru schimb de corpusuri, cum ar fi *The Network of European Reference Corpora* sau, dedicat exclusiv limbilor romanice, proiectul C-ORAL-ROM (*Corpora for Spoken Romance Languages*), în care sunt cuprinse limbile italiană, franceză, spaniolă și portugheză.

1.2. În cazul studiilor de **limbă română vorbită**, destul de reduse la număr, adesea autorii au recurs la abordări pragmatice ale unor texte dialectale (cu limitările inerente tipului de interacțiune pe care îl reflectau, și anume ancheta dialectală). Alți autori, cum ar fi Georgeta Ghiga (1999), au realizat studii pe baza unui corpus individual, nepublicat ca atare. Anul 2002 a marcat însă publicarea a

două volume de transcrieri de română vorbită: *Corpus de română vorbită (CORV). Eșantioane* (Dascălu Jinga, 2002) și *Interacțiunea verbală în limba română. Corpus (selectiv). Schiță de tipologie* (Ionescu-Ruxăndoiu, 2002). În cele ce urmează, ne vom referi la cele două lucrări folosind siglele *CORV* și, respectiv, *IVR*.

1.3. Preocupările noastre legate de transcrierea în vederea prelucrării computerizate au apărut în urma participării în 2001 la proiectul *Interacțiunea verbală în limba română. Corpus și tipologie*, coordonat de prof. univ. dr. Liliana Ionescu-Ruxăndoiu. În ianuarie 2002, am avut onoarea de a citi în manuscris textul volumului *Corpus de română vorbită (CORV). Eșantioane*, prilej cu care am ascultat și înregistrările corespunzătoare textelor transcrise. Luând contact cu două sisteme de transcriere pentru româna vorbită, am descoperit o serie de aspecte deosebit de interesante legate de pluralitatea opțiunilor pentru reprezentarea grafică a materialului înregistrat audio, de problemele practice de limitare a interpretării în transcriere, de consistență internă și flexibilitate a sistemului de convenții de notare.

Suntem de părere că întrebuițarea unui sistem de transcriere care să permită o cât mai bună prelucrare a datelor cu ajutorul computerului nu vizează doar niște aplicații pe termen lung. Valorificarea optimă a unei colecții de transcrieri este posibilă deja prin facilitățile de căutare complexe existente în cadrul programului *Microsoft Word* (în versiunea din pachetul *Office 97* sau o versiune ulterioară) cu care este familiarizat orice utilizator de computere. Pentru a putea fișa materialul în funcție de obiectivul cercetării (de la statisticile privind frecvența relativă și/sau absolută a unor unități lexicale și până la selectarea tuturor ocurențelor unui fenomen surprins în transcrieri, a suprapunerilor, spre exemplu) este necesar ca notațiile definite să fie riguroase, clare și aplicate sistematic.

Exigențele cercetărilor similare realizate pentru alte corpusuri de limbă vorbită din lume sunt deosebit de mari. Culegerea corpusului, arhivarea și transcrierea sunt numai primii pași în studiul limbii vorbite. Arhivarea înregistrărilor audio pe suport digital (CD) este obligatorie pentru a trece la o treaptă superioară de prelucrare a materialului, și anume alinierea textului transcris la secvența sonoră corespunzătoare (*text-to-speech alignment*) cu ajutorul unui software conceput în acest scop. În cadrul proiectului C-ORAL-ROM (care se va încheia la sfârșitul anului 2003), pentru corpusurile corespunzătoare limbilor romanice reprezentate în proiect se realizează alinierea transcrierii la sunet (aproximativ 50h de înregistrări pentru fiecare dintre cele patru limbi), cu segmentarea în unități intonaționale (*parsing*) a fiecărui text. Mai mult, textul aliniat se etichetează pe niveluri de analiză lingvistică (*textual tagging*) și studiile de limbă vorbită incluse în proiect pornesc de la aceste date riguros arhivate (Cresti, 2000).

Sinteza și recunoașterea vocală – necesare pentru a trece de la interfața grafică a comunicării dintre om și inteligența artificială la o interacțiune bazată

(și) pe comenzi rostite –, dar și alte aplicații extralingvistice de interes larg (v. Huang *et al.*, 2000), depind în mare măsură de asemenea corpusuri de limbă vorbită și de prelucrarea lor computerizată. Pentru a atinge un asemenea obiectiv este nevoie, pentru fiecare limbă, nu numai de existența unui corpus de limbă vorbită și de transcrierea lui, ci și de definirea parametrilor acustici specifici sunetelor limbii respective.

Ne propunem să analizăm inventarul de fenomene lingvistice sau extralingvistice codificate în prezent în transcrierile de română vorbită, inventarul de semne grafice întrebuințate (luând în considerare normele pe care le impune obiectivul prelucrării computerizate ulterioare) și mijloacele tehnice de identificare, căutare și selectare a simbolurilor grafice cu ajutorul programului *Microsoft Word*. Ne vom opri în prezenta contribuție numai asupra aspectelor privind **u t i l i z a r e a p a r a n t e z e l o r** în transcrierile existente, și, respectând opțiunea autorilor pentru inventarul de fenomene notate, vom lua în discuție o reorganizare a corespondențelor dintre aceste fenomene și simbolurile întrebuințate. Prelucrarea computerizată nu a constituit obiectivul nici unuia dintre cele două volume de transcrieri de română vorbită, însă considerăm că este util ca transcrierile existente să poate fi folosite cât mai curând și în acest scop. Computerul în general, și editoarele de text curente în special, sunt deja instrumente puternice de analiză, care permit fișarea materialului într-un timp mult mai scurt și după parametri exacti. În elaborarea statisticilor de orice tip, dar și pentru verificarea oricăror ipoteze privind corelarea a două sau mai multe fenomene marcate în transcriere, inteligența artificială este de neînlocuit.

2. Cele două corpusuri de română vorbită cuprind, fiecare, zeci de ore de înregistrări audio. Lipsa unor mijloace tehnice corespunzătoare pentru realizarea unor înregistrări audio și video, dar și contextul specific românesc postdecembrist (în care diversitatea interacțiunilor verbale surprinse este uneori limitată din cauza suspiciunii față de înregistrări a multora dintre posibili subiecți) încă își pune amprenta asupra **metodologiei culegerii corpusului**. Ca urmare, sintaxa mixtă – corelarea componentei paraverbale și/sau nonverbale cu informația transmisă verbal – rămâne de cele mai multe ori neconsemnată sistematic.

Metodologia culegerii corpusului și cea a prelucrării lui pot avea, în opinia noastră, influențe antagonice asupra sistemului de transcriere: pe de o parte, cu cât aparatele de înregistrare sunt mai performante (prin aceasta înțelegând și flux de informații complex, audio și video), cu atât procesul transcrierii trebuie să filtreze și să sistematizeze mai multe date. Pe de altă parte, tehnologia prelucrării datelor din transcrieri include, așa cum aminteam mai sus, o serie de programe (*software*) care permit alinierea textului transcris la înregistrarea audio corespunzătoare, sau alinierea textului la imagine (în analiza limbajelor mimico-gestuale), sau alinierea simultană a sunetului, imaginii și transcrierii (v. Linguistic Annotation <http://www ldc.upenn.edu/annotation/>). Această aliniere ar permite o transcriere

simplificată, cum este cazul sistemului folosit de proiectul Lablita (Cresti 2000, 205-225), întrucât cercetările bazate pe corpus se pot face consultând simultan transcrierea și înregistrarea.

În 1991, Orletti / Testa reproșau transcrierilor faptul că urmăresc în cea mai mare parte verbalul (în detrimentul paraverbalului și nonverbalului):

„La ricerca ha, quindi, pur utilizzando come dati di base interazioni reali, concentrato gli interessi sugli aspetti verbali dell'interazione, è stata, diremo, fondamentalmente verbo-centrica, e anche quando si è occupata di strategie comunicative, pur affermando la rilevanza di comportamenti comunicativi non verbali, ha descritto soprattutto comportamenti verbali e, all'interno di questi, comportamenti riconducibili alla produzione di materiale lessicale. Conseguentemente, le trascrizioni sono state ugualmente verbo-centriche, mostrando la tendenza a privilegiare i dati verbali nelle trascrizioni dei dati interazionali, a riportare solo sotto forma di commento i comportamenti non verbali e a trascurare i comportamenti verbali non lessicali come varie forme di *ehm, uhm, ecc*”. (Orletti / Testa, 1991, 252)

Astăzi facem observația că sistemele noastre de transcriere încă trebuie să linearizeze discursul complex, încercând să noteze și celelalte componente ale comunicării. Lipsa accesului la tehnologia de prelucrare este însoțită, deocamdată, în cercetările asupra interacțiunii verbale în româna vorbită de lipsa mijloacelor tehnice adecvate pentru culegerea datelor. Astfel, deși ambele sisteme de transcriere analizate prevăd convenții de notare a elementelor nonverbale, materialul de acest tip rezultat în transcrieri este relativ redus, iar fluxul de informații urmărit consecvent rămâne cel verbal (paraverbalul este adeseori recuperat).

3. Stocarea unor înregistrări de limbă vorbită pe bandă magnetică sau chiar pe suport digital nu este suficientă pentru a putea face studii bazate pe acest material. Caracterul secvențial al comunicării orale nu permite confruntarea datelor și corelarea diverșilor factori care influențează desfășurarea unei interacțiuni verbale. **Necesitatea transcrierii** este evidentă, iar sistemul de convenții definit este responsabil pentru consemnarea consecventă și neambiguă a fenomenelor vizate de obiectivul cercetării. Atributele pe care trebuie să le aibă un sistem de transcriere funcționează de cele mai multe ori antagonic, un exemplu elocvent fiind dezideratul de a nu pierde, în procesul „traducerii” în scris a conținutului înregistrărilor, informații posibil relevante, dar de a evita, în același timp, ca textele transcrise să fie prea încărcate și greu de urmărit. Toate aceste aspecte au condus la proliferarea sistemelor de transcriere, la analize permanente și perfecționări numeroase, problema transcrierii fiind considerată fundamentală în *corpus linguistics*: „Central to the modern study of spoken discourse is the problem of transcription”. (Du Bois *et al.*, 1988, 3)

Definirea sistemului este de cele mai multe ori confruntată cu dificultățile practice ale realizării unui număr cât mai mare de transcrieri efective, pe cât posibil

diversificate, în limita obiectivelor de cercetare propuse. Procesul transcrierii rămâne însă susceptibil de un grad oarecare de subiectivism, fiind orientat către premise teoretice explicite sau implicite: „The process of discourse transcription is never mechanical, but crucially relies on interpretation within a theoretical frame of reference to arrive at functionally significant categories”. (Du Bois, 1991, 72)

3.1. Sistemul de transcriere folosit, fie creat, fie selectat dintre cele existente, depinde în mare măsură și de destinatarul unui astfel de text. Există trei mari tipuri de **destinatari**: specialiști (lingviști), nespecialiști și... inteligența artificială.¹ Am menționat inteligența artificială alături de receptorii umani, întrucât prelucrarea computerizată ridică o serie de probleme care trebuie avute în vedere încă din faza de elaborare a sistemului de transcriere – de pildă, în problema inventarului de semne grafice utilizate, pentru care se recomandă, în general, codul ASCII. (Du Bois, 1991, 87) Sistemul de transcriere către care tindem, prin sugestiile noastre, are în vedere lingviștii și inteligența artificială ca destinatari

3.2. În **elaborarea sistemelor de transcriere** există două aspecte: selectarea inventarului de fenomene lingvistice care vor fi urmărite și codificate în transcriere și stabilirea semnelor grafice prin care vor fi marcate acele fenomene.

3.2.1. Du Bois observă că, în ceea ce privește **inventarul de fenomene**, cele mai multe sisteme de transcriere notează: a) cuvintele rostite, b) identitatea vorbitorului pentru fiecare intervenție (*turn*), c) succesiunea cronologică enunțurilor, c) intervențiile și unitățile intonaționale, d) conturul intonațional, e) emfaza, f) fluctuații ale ritmului vorbirii precum tempo-ul, pauza în vorbire sau lungirea unor sunete, g) zgomote nonverbale, h) particularități deosebite ale vorbirii care definesc o anumită secvență, i) evenimente extralingvistice care sunt relevante pentru interacțiunea verbală și j) comentarii (sau mijloace de evidențiere) privind transcrierea însăși (Du Bois 1991, 76). Opțiunea pentru a marca sau nu un anumit fenomen rămâne însă legată de obiectivele stabilite de fiecare cercetare în parte (Orletti / Testa, 1991, 250).

3.2.2. Pentru cel de-al doilea aspect al creării unui sistem de transcriere, **inventarul de semne** care să codifice fenomenele selectate, Du Bois *et al.* (1988, 81-87) propun cinci principii generale: definirea clară, explicită a categoriilor codificate în sistem, accesibilitatea transcrierii, robustețea sistemului, economia și adaptabilitatea.²

Principiile enunțate de Du Bois (1991) pentru **e l a b o r a r e a** unui sistem de transcriere se regăsesc, în linii mari, în principiile de **s e l e c t a r e** a unui sistem dintre cele definite deja, așa cum apar în Orletti / Testa (1991, 267-271): *compresività vs specializzazione, attendibilita, leggibilita, consistenza interna, flessibilita, trasversalita, riproducibilita.*

Interesant este modul în care anumite principii sunt reformulate de-a lungul timpului, în funcție de obiectivele nou apărute. Spre exemplu, problema accesibilității este discutată de Du Bois din perspectiva scrierii și citirii unui text

transcris, autorul recomandând valorificarea unor sisteme de convenții existente: „drawing on existing traditions for representing speech in writing, whenever viable conditions can be found” (1991, 81). Tot din perspectiva accesibilității, s-a invocat și ușurința cu care semnele alese pot fi introduse pe calculator (*ease of data entry*), sau, chiar ca prim obiectiv, posibilitățile de utilizare a unor baze de date astfel constituite (*usability, not readability*). (O'Connell / Kowal, 1994, 102)

Precizăm că foarte multe dintre principiile enunțate mai sus pentru elaborarea unui sistem pornesc de la premisa că inventarul semnelor grafice folosite în transcriere trebuie să fie inclus în sistemul ASCII, care permite transferul datelor (al textelor transcrise, cu toate notațiile incluse) și prelucrarea computerizată. ASCII și Unicode sunt standarde de reprezentare a informației textuale în computer. Inventarul Unicode este mult mai mare decât al celui alt sistem, însă trebuie reținut că nici unul dintre ele nu codifică (și nu „păstrează” la transferul de date) anumite posibilități de tehnoredactare care constau în schimbarea unor proprietăți ale unor semne grafice, și nu alte semne grafice propriu-zise: „One should avoid using notational resources which are not standardly represented across platforms, such as boldface, italics, underlining, special fonts (especially proportional fonts), margin shifts, a.s.o. as the *sole* marker of crucial contrasts between categories”. (Du Bois, 1991, 89) Așadar, nu va putea fi inclus în prelucrarea computerizată un fenomen căruia îi corespunde o notație numai prin sublinierea caracterelor, îngroșarea sau schimbarea dimensiunii a corpului de literă etc., pentru că aceasta nu se păstrează în trecerea de la o platformă³ la alta.

Notațiile propuse de Du Bois *et al.* se încadrează în inventarul standardului ASCII redus. Deși sugestiile din analiza noastră sunt incluse în inventarul ASCII (cu excepția sistemului IPA), vom avea în vedere standardul Unicode, care îl include pe primul, din două motive principale: a) are un inventar de câteva sute de ori mai mare decât ASCII (permițând mai multă flexibilitate în notații) și b) ambele sisteme de transcriere pentru româna vorbită includ deja semne care fac parte din Unicode și nu fac parte din ASCII (vezi conturul intonațional non-terminal: ascendent, ↑, și, respectiv, descendent, ↓, precum și notația prevăzută în *CORV* pentru observațiile cercetătorului, →, dar și diacriticele românești). Motivul pentru care Du Bois propunea în 1988 (iar apoi în studiul din 1991) folosirea standardului ASCII redus este faptul că Unicode a apărut abia în 1991, fiind ulterior folosit la scară largă.

4. Orletti / Testa (1991) disting două mari **tipuri de sisteme de transcriere**, urmând direcțiile inaugurate de Jefferson (1974) și, respectiv, Gumperz (1982).

4.1. Sistemul de notații propus în 1974 de Sacks, Schegloff și Jefferson și perfecționat ulterior de **Jefferson** (Jefferson, 1978), a avut ca obiectiv analiza conversației. Transcrierea era concepută ca parte integrantă a procesului de analiză și interpretare a datelor și ca o încercare de a reprezenta în scris

interacțiunea verbală. Caracterul secvențial al interacțiunii verbale și ilustrarea lui sunt definitorii pentru sistemul Jefferson. În ultimele decenii acest sistem a cunoscut numeroase îmbunătățiri și adaptări.

Sistemele de transcriere folosite pentru limba română vorbită urmează linia propusă de Jefferson, Dascălu Jinga precizând chiar că sistemul utilizat în *CORV* este „jeffersonian” (*CORV*, 32). Această „filiație” este firească, având în vedere și similitudinea obiectivelor urmărite în analiza interacțiunii verbale. Sistemele de transcriere pentru româna vorbită prezintă o serie de diferențe în raport cu sistemul lui Jefferson (1978) – adaptări, rafinări ale convențiilor – cele mai semnificative fiind, în opinia noastră, cele legate de notarea sistematică a unor elemente de prozodie și raportul textelor transcrise cu ortografia standard.

4.2. Setul de convenții propus de **Gumperz** (1982) are ca principal obiectiv analiza comunicării interculturale. Sistemul ESF, folosit de Orletti / Testa (1991) într-un studiu intercultural (*SSLA – Spontaneous Second Language Acquisition*), urmează acest model.

În analiza transcrierii românei vorbite vom reveni la cele două tipuri de sisteme, propuse de Jefferson și, respectiv, Gumperz, întrucât considerăm utilă rediscutarea unor probleme specifice limbii române actuale folosind mijloace de reprezentare consacrate în sistemele sus-amintite.

5. Pentru a oferi o vedere de ansamblu asupra claselor de fenomene și tipurilor de paranteze pe care le folosește fiecare dintre cele două sisteme de transcriere a românei vorbite, *CORV* și *IVR*, am optat pentru prezentarea lor într-un tabel (v. **Tabelul nr. 1**). În prima coloană sunt trecute diverse tipuri de paranteze, la care am adăugat și barele oblice, folosite într-o manieră asemănătoare parantezelor, ca o structură din două elemente simetrice (identice, de fapt, în cazul barelor) ce izolează o secvență grafică de lungime variabilă: primul element al acestei structuri este bara precedată de blanc și urmată imediat de caractere grafice, iar ultimul element este așezat imediat după caracterele grafice și urmat de blanc sau de unul dintre semnele: ?,., ↓, ↑ sau # (ce marchează conturul intonațional și pauza în rostire).

5.1. În *CORV* se folosesc: paranteze pătrate, [*text*], paranteze rotunde, (*text*), și paranteze unghiulare, <*text*>. Parantezele pătrate sunt întrebuințate pentru: 1) transcrierea fonetică (cuprinzând simboluri din inventarul IPA), 2) marcarea suprapunerilor (trecute între rânduri, fără caractere grafice în intervalul dintre paranteze), 3) componenta paraverbală: [*iși drege vocea*], 4) componenta nonverbală [*gest afirmativ cu capul*] și 5) diverse observații privind înregistrarea și desfășurarea interacțiunii verbale: [*scurtă ștergere involuntară a înregistrării*] (*CORV*, 95) sau [*Oprirea vorbitorului și întreruperea înregistrării, pentru că sună telefonul în încăperea*] (*CORV*, 93). Am precizat care este „conținutul” parantezelor, pentru a evidenția faptul că nu se pot face confuzii între cele trei mari tipuri de utilizări ale parantezelor drepte: cu semne din alfabetul fonetic, (1),

cu blankuri, (2), și cu litere din ortografia curentă, (3)-(5). Cu toate acestea, suntem de părere că este de dorit să se folosească un singur tip de paranteze pentru un tip de informații. Pluralitatea semnificațiilor pe care le are folosirea parantezelor drepte în transcriere rezultă din convergența unor convenții anterioare, preluate din coduri diferite; spre exemplu, sistemul IPA este consacrat, dar și folosirea parantezelor drepte în notarea suprapunerilor este frecvent întâlnită în corpusurile dedicate analizei conversației (v. Jefferson 1978, Du Bois *et al.* 1988 și Du Bois 1991 etc.).

Parantezele unghiulare nu se folosesc decât pentru cuvinte care au fost rostite efectiv în interacțiunea verbală propriu-zisă, fie „marcate” paralingvistice (5), fie secvențe incerte (6) sau indescifrabile (7). Această convenție a fost propusă de Du Bois *et al.* (1988, 20-23) pentru a reliefa elemente paralingvistice, considerate, într-o primă fază, irelevante în sistemele „jeffersoniene”. Flexibilitatea notației derivă din modul descriptiv și virtual nelimitat în care se pot alege mărcile și prefixul care să le codifice; acest fapt se poate observa și din modul în care a fost valorificată în sistemele românești. În *CORV* se notează: ritmul vorbirii, (lent <*L text L*> sau rapid <*R text R*>), înălțimea vocii, (ridicată <*Î text Î*> sau joasă <*J text J*>), intensitatea, (puternică <*F text F*> sau slabă <*P text P*>), șoptitul <*ȘOP text ȘOP*>, imitarea modului de a rosti al altcuiva <*IM text IM*>, râsul concomitent cu rostirea <*@ text @*> sau rostirea marcată <*MARC text MARC*>. În *IVR*, se marchează, în plus, oftatul concomitent cu rostirea <*OF text*>, și secvențele rostite zâmbind <*z text*>; la acestea se adaugă o informație privind caracterul planificat, nespontan al unor comunicări orale, și anume lectura unui text: <*CT text*>.

Considerăm că ar fi utile câteva observații legate de prezentarea mărcilor paralingvistice. Mai întâi, reluăm remarca pe care o fac autorii celor două sisteme, și anume că mărcile paradiscursive folosite în transcriere au un caracter relativ, raportându-se la particularitățile de rostire ale aceluiași vorbitor în cursul aceleiași înregistrări. Altfel, presupunând că s-ar putea face transcrierile numai după măsurători exacte și după un reper oarecare de rostire, textul transcris ar fi nu numai încărcat, ci și ineficient. Reducând la absurd, vocea tuturor participanților de sex feminin ar avea particularitatea „înălțime ridicată”, sau majoritatea subiecților foarte în vârstă ar prezenta o intensitate slabă a vocii, rostire „piano”.

Tabelul nr. 1

Tip de paranteze	<i>CORV</i>		<i>IVR</i>	
	Semnificația	Exemple	Semnificația	Exemple

[]	1) transcriere IPA	„vecinic” [ve'tɨi nik] te te-ncurca↑ (163)	1) [marchează începutul suprapunerii unor intervenții succesive.	A: student la petrol↓ [aici? B: [nu. la bucurești (27)
	2) plasate între rânduri, notează secvențe care se suprapun	GP: Da. De acord [] VJ: Că acolo diferența era enormă (157)	2) întreruperea pasajului transcris	[...]
	3) fenomene paraverbale	VC: [râde] (251)		
	4) fenomene nonverbale	AB: [gest afirmativ cu capul] (269)		
	5) diverse observații privind înregistrarea	[scurtă ștergere involuntară a înregistrării] (95)		
< >	6) mărci paralingvistice	VL: <R președintele României↓ domnul Emil Constanti <Î nescu Î> R> (276)	3) mărci paralingvistice	B: <_ io văd așa↓<@ că toate> problemele sî:nt bu:ne:> (191)
	7) transcriere incertă	SFI: <? Nu prea știu.??> (166)		
	8) secvență indescifrabilă	CJ: Da↓ <xxxxxxxx> (71)		
()	9) scurte explicații necesare înțelegerii textului	MV: ce (zice) V-au venit niște bani din țară↓ (115)	4) transcriere incertă	A: (ca un fel de invitație) pentru oameni d-ăștia (35)
	10) pauze foarte lungi	LDJ: Nu era încălzire? Iarna? GD: (3 sec.) (86)		
	11) transcriere pseudofonetică (cuvinte străine și acronime)	GD: La căminul I.O.V. (iove) (86) Heidelberg (haidălberg) (74)	5) secvență indescifrabilă	A: (xxx) B: nu încă. (38)
	12) întreruperea pasajului transcris	(...) plasat între rânduri (<i>passim</i>)		
	13) notații specializate: heterocorectare (K), autocorectare (AK) și eroare necorectată (sic!).	VJ: (K) Nu↓ Lățești. (56) CJ: în proteș- (AK) în procesul lui Pătrășcanu. (56) IS: Vă vor place (sic!) (270)		

(())			6) comentariile cercetătorului	((<i>între timp sosise în stație un microbuz</i>)) (27)
			7) fenomene paraverbale	((<i>ride</i>)) (31) ((<i>își drege vocea</i>))
			8) fenomene nonverbale	((<i>se uită la ceas</i>)) (27)
/ /			9) transcriere pseudofonetică (pentru cuvintele în limbi străine și abrevieri)	A: <i>am văzut în /vog/</i> (53) B: <i>firma /secea/</i> (91)

O a doua observație se referă la posibilitatea de a nota particularități izolate cu o convenție asemănătoare, fără a risca să îngreuneze asimilarea sistemului de transcriere prin adoptarea unor notații prea numeroase. Am întâlnit un astfel de caz în transcrierile noastre, când unul dintre participanți fredonează câteva cuvinte dintr-o melodie cunoscută, pentru ca imediat după aceea să treacă la adresarea directă față de un alt participant. Efectul acestei treceri rapide a fost acela că ultimele cuvinte din melodia respectivă nu au mai fost fredonate, ci rostite. Exemplul nostru vizează două probleme: caracterul imprevizibil al duratei unei astfel de secvențe și imprecizia notării lui cu un gerunziu de tipul ((*fredonând*)) plasat înaintea textului corespunzător acelei rostiri particulare. Du Bois *et al.* propun, în astfel de cazuri, încadrarea între paranteze unghiulare a secvenței respective și notarea, după transcrierea ei, a „mărcii”, coindexat: < *text I* > <fredonat I>.

În fine, din prezentarea anterioară a mărcilor pentru care a optat fiecare dintre sistemele menționate rezultă și valorificarea diferențiată a opțiunilor de redactare computerizată. După cum aminteam la punctul 3, nici scrierea cu aldine, nici poziția literei față de rând nu constituie informații valide în prelucrarea computerizată, dar, fiind folosite auxiliar, ambele pot fi utile în înlesnirea lecturii. Diferențierea secvențelor grafice corespunzătoare „mărcării” (care pot fi selectate de utilizator sau, dimpotrivă, eliminate, păstrând doar textul „brut” al cuvintelor rostite în dialogul transcris) se face definind acel număr limitat de caractere (<Î, <F, <ȘOP, <@ etc.) care preced textul propriu-zis.

Parantezele rotunde sunt folosite pentru a izola de textul transcris comentariile cercetătorului (8), dar și în transcrierea pseudofonetică, (9). În plus, parantezele rotunde sunt folosite pentru a semnală întreruperea pasajului transcris, cu (...), v. pct. (10), precum și pentru a izola niște notații specializate de tipul (K), (AK), (sic!), v. pct. (11). Ultimul tip de convenție valorifică tradiția notării cu secvența (K) a fenomenului de autocorectare în transcrierile textelor dialectale.

5.2. În *IVR* sunt folosite: parantezele pătrate: *[text]*, parantezele rotunde simple (*text*) și duble (*((text))*) și scrierea între bare oblice */text/*.

Paranteza pătrată „deschisă” *[text]* marchează începutul fiecăreia dintre secvențele rostite simultan de vorbitori diferiți (suprapuneri). Întreruperea intervenției în curs de către un alt participant este considerată un caz particular al suprapunerii și se notează implicit, atunci când semnul [nu este urmat de nici un text, pe rândul următor fiind notată tot cu *[text]* intervenția celui care preia rolul de emițător.

Întreruperea pasajului transcris se notează cu *[...]*.

Parantezele rotunde simple se folosesc în pentru transcrierea secvențelor incerte (*este*) sau indescifrabile (*xxx*) din rostirea unui participant, iar cele duble pentru componenta nonverbală: (*(se ridică brusc de pe scaun)*), pentru fenomene paraverbale: (*(tușește)*) și alte observații necesare înțelegerii textului: (*(între timp sosise în stație un microbuz)*) (*IVR*, 27).

Pe lângă paranteze, sistemul prevede și izolarea transcrierilor pseudofonetice cu ajutorul barelor oblice, ca în */edvărtaizing/* (*IVR*, 37). Utilizarea diferitelor tipuri de paranteze din sistemul *IVR* este foarte asemănătoare cu aceea din sistemul propus de Jefferson în 1978: paranteze rotunde simple pentru transcriere incertă și pentru secvență indescifrabilă (care în Jefferson nu are un șir de *x* între paranteze, ci doar blankuri), paranteze duble pentru componenta nonverbală și cea paraverbală, precum și pentru alte informații care nu reflectă rostirea din dialog, ci comentariile cercetătorului.

6. Analiza noastră are la bază câteva deziderate: a) importanța consistenței interne a unui sistem de transcriere (atât pentru a fi mai ușor de urmărit de către utilizatori, cât și pentru a putea trece la prelucrarea computerizată a datelor), b) valorificarea unor deprinderi de lectură și evitarea folosirii cu alt sens a unor semne grafice frecvent întrebuințate în ortografia curentă, c) definirea unor norme de redactare (succesiunea caracterelor grafice și non-grafice) astfel încât, pentru orice transcriere în parte, fiecare utilizator să își adapteze sistemul de transcriere: se pot elimina anumite paranteze, cum este cazul mărcilor paralingvistice, păstrându-se numai textul cuprins între paranteze sau, mai mult, se pot elimina complet diverse tipuri de paranteze, corespunzând unor tipuri precise de informații cum ar fi elementele nonverbale, spre exemplu. Ultima operațiune este necesară în cazul în care dorim să facem analize statistice, precum debitul verbal al participanților în funcție de situația de comunicare, rol, sex etc. și trebuie eliminate acele cuvinte care apar în transcriere fără să corespundă rostirii din dialogul înregistrat. Ca principiu supraordonat celor sus-menționate, am avut în vedere permanent respectarea fenomenelor pe care autorii au decis să le surprindă în textele transcrise, propunând numai reorganizarea lor în clase care să corespundă sistematic unor tipuri de paranteze.

6.1. Folosirea unor sisteme de transcriere auxiliare a fost considerată necesară, în cazul limbii române vorbite, dar și pentru alte limbi, întrucât complexitatea limbii vorbite a evidențiat, în numeroase situații, insuficiența mijloacelor grafice întrebuintate în ortografia curentă. Atât în *CORV*, cât și în *IVR*, autorii optează, spre exemplu, (și) pentru o transcriere pseudofonetică în cazul abrevierilor. Astfel, o secvență grafică de tipul RTL poate fi rostită ca *er-te-el* sau *er-te-le*. Redarea în scris a cuvintelor străine a fost considerată, la rândul ei, problematică, ortografierea din limba sursă oferind indicii insuficiente asupra pronunțării sale (care adesea variază de la un vorbitor la altul).

CORV folosește două asemenea sisteme auxiliare: IPA și transcrierea pseudofonetică, iar *IVR* numai pe cel din urmă. Cu toate acestea, există anumite situații în care, la rândul lor, sistemele auxiliare se dovedesc insuficiente. Dacă o secvență precum *Harun Tazieff (harun tazief)* (*CORV*, 77) nu pare să ridice probleme, în alte situații aproximarea pronunției cu ajutorul semnelor din ortografia curentă este mai dificilă. În *dantele de Bruges (briuj)*, dincolo de faptul că nu se mai poate distinge pronunțarea ca în limba sursă de orice variantă de adaptare fonetică, există posibilitatea ca unii vorbitori să o rostească bisilabic. Considerăm că asemenea fenomene ar fi interesante din punctul de vedere al preferinței pentru hiat sau diftong în româna actuală, dar și în schițarea unor probleme legate de gradul de instruire a vorbitorilor. În alte cazuri, transcrierea pseudofonetică se face folosind semnele IPA: *Jean Francois Revel (jă frânsoa revel)* (*CORV*, 75), */uipatrõ/* (*IVR*, 89). Uneori se folosesc alte soluții pentru a reda foneme nespecifice limbii române: */edvărtaizing/* (*IVR*, 37) sau */paundț/* (*IVR*, 115), rămânând însă ambiguu dacă vorbitorul le-a rostit ca în limba engleză, în cazurile prezentate, sau nu. Pe de altă parte, transcrierea pseudofonetică nu dă informații asupra accentului și silabației; în */menegimentu/* (*IVR*, 254) putem avea patru sau cinci silabe. Un caz interesant este transcrierea lui O.K., care este și cuvânt străin, și abreviere (**//ochei//*).

IPA este folosit în *CORV*, dar numai în cazuri excepționale, „când interacțiunea verbală vizează însăși pronunțarea sau necesită sugerarea cât mai precisă a acesteia” (*CORV*, 33).

Suntem de părere că ar fi utilă întrebuintarea alfabetului fonetic și în cazurile în care se folosea transcrierea pseudofonetică, pentru a sugera adaptarea fonetică a unor cuvinte noi sau foarte noi (xenisme), putându-se astfel analiza în funcție de diferiți parametri sociolingvistici. Sistemele „jeffersoniene” consideră, în general, că transcrierea fonetică nu este necesară pentru analiza conversației; cele care sunt dedicate studiului achiziționării unei limbi străine (v. Orletti / Testa, 1991) acordă o atenție deosebită redării cât mai fidele a pronunției, urmând linia propusă de Gumperz. În cazul limbii române, există avantajul major al ortografiei sale fonetice (față de limba engleză, de pildă, unde apar o serie de dificultăți în redarea unor fenomene frecvente, precum lungirea unui sunet căruia de fapt nu-i corespunde o

literă anume în transcriere). În contextul socio-istoric actual însă, limba română, scrisă sau vorbită, este „invadată” de o serie de cuvinte de origine străină (în special din limba engleză) și credem că ar fi interesant de notat consecvent pronunția acestor cuvinte la diferiți vorbitori, pentru a surprinde dinamica fenomenului.

De altfel, și restul transcrierii în ambele volume este „pseudofonetic” (sau un sistem fonetic neconvențional, așa cum este numit în Orletti/Testa, 1991, 260), în sensul că nu corespunde ortografiei standard, ci încearcă să redea rostirea: *am crezt că e aceeași atmosferă (IVR, 73)*, sau notarea frecventă a rostirilor de tipul *dă* ('de'), *dân/dîn* ('din') etc.

Un alt aspect care ar putea prezenta interes în studiul dinamicii limbii române actuale este notarea semivocalelor și pseudovocalelor, interesante din punct de vedere morfonologic. Semnalăm că acestea pot fi notate în transcrieri folosind convențiile curente pentru aceste sunete, care se pot „traduce” pentru calculator în secvențe grafice care să permită prelucrarea datelor.

Eliminarea literei x din transcrierea rostirii ar putea aduce, la rândul său, un plus de informație în analiza grupurilor [ks] și [gz], în condițiile în care se constată rostirea unuia în locul celuilalt la diverși vorbitori. În plus, aceasta ar permite evitarea inexactității în marcarea emfazei (se scrie *eXACT*, *EXtraordinar*, dar cele două consoane codificate prin x aparțin unor silabe diferite) și ar permite ca x să apară numai pentru redarea unei secvențe indescifrabile.

6.2. Notarea suprapunerilor cu paranteze coindexate, plasate în text, este propusă în 1988 de Du Bois *et al.* Considerăm că ar fi o îmbunătățire a acestei convenții dacă s-ar folosi acoladele (păstrând parantezele drepte pentru IPA, o convenție cu caracter mai general) și indexarea s-ar face cu un șir de numere crescătoare, constant, până la sfârșitul transcrierii respective. Du Bois *et al.* (1988) propuneau coindexarea numai în cazul unor suprapuneri numeroase într-o anumită porțiune, iar după ce nu ar mai exista ambiguitate în privința secvențelor rostite simultan, să se reia număratoarea de la 1. Avantajul numerotării până la sfârșit este evident în cazul prelucrării computerizate: se pot „extrage” automat toate secvențele cuprinse între paranteze și pot fi analizate precis, în funcție de conturul intonațional, mărcile paralingvistice (în suprapunerile mai lungi este posibil ca cel puțin unul dintre vorbitori ridică vocea), sau relațiile dintre participanți.

Întreruperile se pot nota ca un caz particular, în care primul element este {i} (*i* fiind indicele numeric: 1, 2, 3...*i*,...*n*) și se va nota la sfârșitul rândului corespunzător intervenției întrerupte, iar al doilea este {i}, notat la începutul rândului, după sigla participantului care preia rolul de emițător (**A**; *ieri de CE te-ai supărat și-ai ple- {4} B; {4} *ba n-am plecat supărat*, spre exemplu).*

6.3. Mărcile paradiscursive au fost propuse de Du Bois *et al.* în 1988 (20-23), care oferă și sugestii de notare a lor. Marcarea începutului și sfârșitului unei secvențe rostite cu anumite particularități cu ajutorul parantezelor unghiulare

plasate în text a fost preluată atât în *CORV*, cât și în *IVR*. Așa cum semnalăm, nici convenția grafică a îngroșării literelor, nici scrierea unei secvențe mai sus sau mai jos față de restul caracterelor din rând nu constituie un mijloc suficient de identificare a fenomenului urmărit. În ambele sisteme însă identificarea computerizată se poate face prin respectarea secvenței: paranteză unghiulară urmată de o literă sau un grup de litere dintr-un inventar definit în convenții. Din păcate, opțiuni de transcriere mai economice sau mai simple, precum cele folosite în *IVR* (<SOP, P text> text) care pot fi citite relativ ușor de un receptor uman, prezintă dificultăți majore în prelucrarea cu ajutorul inteligenței artificiale. Combinațiile de mărci (la care se adaugă ordinea permisivă de tipul: <i,j text> text> sau <j,i text> text>, pentru două mărci <i> și <j> care ar caracteriza o anumită secvență) sunt foarte numeroase și nu permit statistici exacte.

Am putut urmări, spre exemplu, în *CORV* numărul de ocurențe al fiecăreia dintre mărcile definite în sistem și am obținut următoarele date: 236 de apariții pentru marca <Î text Î>, 93 pentru <R text R>, 60 pentru <MARC text MARC>, 54 pentru <J text J>, 43 pentru <@ text @>, 35 pentru <F text F>, 30 pentru <P text P>, 7 pentru <L text L>, 5 pentru <CIT text CIT>, 3 pentru <SOP text SOP>.

Menționăm, cu această ocazie, câteva probleme de redactare. Pentru a permite prelucrarea computerizată, este necesar să se noteze simbolul mărcii respective la începutul și sfârșitul secvenței, cu semnul <, și, respectiv, > pentru fiecare marcă în parte. Pentru a păstra unitatea grafică a cuvântului, în cazul în care apar două mărci succesive de tipul: <J Transilvania propriu- J><Î zisă Î> (*CORV*, 89), se impune notarea fără blanc între marca paradiscursivă și textul corespunzător rostirii, la începutul și la sfârșitul marcării. Scopul de a nu îngreuna lectura, urmărit în ambele volume românești, poate fi realizat prin combinarea celor două mijloace grafice folosite: îngroșarea <J text J> și, respectiv, poziția față de rând <j text>: <jtextj>.

Secvența incertă din transcriere considerăm că este preferabil să fie marcată ca în *CORV*, întrucât astfel parantezele unghiulare ar încadra întotdeauna un text corespunzător rostirii. Semnalăm, cu această ocazie, existența unor mijloace moderne de prelucrare a sunetului în format digital, care permit reducerea zgomotului de fond și/sau amplificarea artificială a unei sonore pentru a limita, pe cât posibil, numărul transcrierilor incerte. Aceeași operație poate reprezenta o soluție și pentru unele dintre secvențele indescifrabile. Cu toate acestea, în cazul în care informația nu se poate recupera, merită menționat că se poate nota, de cele mai multe ori, conturul intonațional și pentru aceste secvențe. Sugestia noastră ar fi adoptarea convenției folosite de Du Bois *et al.* (1988) și, ulterior, de Du Bois (1991), potrivit căreia fiecare semn x ar nota o silabă din porțiunea indescifrabilă, iar nu un sunet. Segmentarea în cuvinte este aproape imposibilă în absența semnificatului, dat fiind fluxul continuu al vorbirii. În cazul în care, pentru

înlesnirea lecturii sau când se urmăresc alte obiective în analiza materialului transcris, se dorește eliminarea parantezelor rotunde simple și se păstrează transcrierea incertă și semnalarea cu x a fiecărei silabe indescifrabile, acest lucru este posibil.

6.4. În cadrul reorganizării unor elemente definite și a unor notații pentru acestea, considerăm că ar fi un câștig dacă am exploata obișnuințele de lectură ale utilizatorului, și anume folosirea parantezelor. Folosirea parantezelor rotunde pentru secvențe de text nesigure sau indescifrabile ca în *IVR* prezintă câteva inconveniente, întrucât aceste paranteze reflectă în general în ortografia curentă raportul informație principală – informație secundară. Acele cuvinte care nu au putut fi transcrise cu certitudine nu sunt mai puțin importante pentru construirea enunțului, ci doar accidental au ajuns să fie o informație nesigură. Optăm, în acest caz, pentru notațiile din *CORV*, unde parantezele unghiulare notează numai cuvinte rostite în interacțiunea verbală (deci informație obiectivă, nu metatranscriere), putându-se marca suplimentar orice calitate vocii. Spre exemplu, o transcriere de tipul <*p*<*xxx*> *text**p*>, în care o rostire „piano” împiedică distingerea unei secvențe, este probabilă.

Suntem de părere că ar contribui la o mai bună organizare a transcrierii și la o asimilare mai ușoară a convențiilor de transcriere dacă s-ar nota diferit elementele nonverbale față de cele paraverbale. Pentru cele din urmă propunem parantezele simple (marcarea calității vocii păstrând paranteze unghiulare simple), iar pentru nonverbal parantezele duble. În acest fel, atenția acordată de utilizator informațiilor din interiorul parantezelor poate fi de același tip cu extragerea informației la o lectură obișnuită: textul astfel izolat este parte integrantă din textul per ansamblu, dar de ordin secundar. Nonverbalul și paraverbalul nu sunt notate deocamdată în transcrierile de română vorbită decât cu rol secundar.

Legat de problema utilizării parantezelor în transcriere, propunem ca, în cazul în care se va opta pentru notarea în text a unor fenomene precum trasul aerului în piept sau expirația audibilă, să se folosească convențiile lansate de Du Bois (1991): (*H*) pentru „inspiră adânc”, (*Hx*) pentru „expiră”, întrucât acestea izolează fenomenele vocale nonverbale de transcrierea rostirii propriu-zise (și anume folosind constant același tip de paranteze, cele rotunde simple). Semnificația unor fenomene de acest gen este discutată în cadrul multor sisteme de transcriere: „The reason for distinguishing vocal tract noises made by speech event participants as a special category is that participants often use this channel to give each other subtle cues about aspects of the on-going linguistic interaction, e.g. breathing in to signal the purpose to speak next. Crickets chirping and microphones rustling do not consistently carry such interpersonal meanings for humans.” (Du Bois *et al.*, 1988, 25) În sistemele din *CORV* și *IVR*, fenomenele paraverbale discutate mai sus se notează astfel: *inspiră adânc* între paranteze

pătrate și, respectiv, paranteze rotunde duble (dar astfel vor fi trecute laolaltă cu observații precum defectarea microfonului etc.). Pledăm așadar pentru surprinderea acestor fenomene în transcriere, dar cu ajutorul unor convenții cât mai simple, care să ocupe puțin spațiu grafic și să fie în concordanță cu notațiile pentru fenomene similare.

Propunem, de asemenea, notarea râsului ca în sistemele Du Bois *et al.* (1988) și Du Bois (1991), adică inserarea câte unui semn @ pentru fiecare „silabă” de râs. Acest lucru ne va permite să marcăm durata relativă a secvenței respective (față de notația din *CORV*, unde trecerea între paranteze, în text, a cuvântului *râde* nu oferea informații de acest tip), dar fără a introduce noi „cuvinte grafice” (adică niște unități care nu corespund de fapt cuvintelor din rostirea participanților). Se permite astfel ca în cazul în care un subiect ar rosti efectiv, ironic, *ha-ha*, să nu se confunde cu râsul propriu-zis, mesajul său fiind cu totul diferit. În *IVR* s-a recurs în general la „transcrierea” râsului: *hăhă* (*IVR*, 41), *hîhîhî* (*IVR*, 44) și chiar <@ *hî hî hî*> (*IVR*, 172).

6.5. Notarea paraverbalului cu paranteze duble, ca în *IVR*, ar permite, ca și în cazul utilizării altor paranteze pentru un singur tip de fenomene, fișarea materialului lingvistic pe baza transcrierii în format electronic sau, dimpotrivă, eliminarea sistematică a acestui tip de informații.

Un caz aparte îl reprezintă tăcerea. În prezent, este marcată sub diferite forme, ca pauză lungă (folosind semnul pentru pauză de două sau mai multe ori): ### în *IVR*,... în *CORV*, sau între paranteze simple, precizând durata în secunde: (3 sec.) (*CORV*, 86), ori paranteze duble: ((*tace*)) (*IVR*, 27, 102), ((*pauză*)) (*IVR*, 27). Suntem de părere că ar fi o soluție notarea tăcerii prin repetarea semnului # sau, pentru pauze foarte lungi, împreună cu tipul de paranteze folosit pentru componenta nonverbală: #((5s)). Un element suplimentar ar putea fi precedarea unei paranteze care specifică durata pauzei de semnul stabilit pentru marcarea pauzei în rostire (optăm pentru #, ca în *IVR*, pentru că semnul întrebuințat în *CORV* este, în prelucrarea computerizată, identic cu simbolul pentru contur descendent terminal; diferența dintre. și. este aldin ~ alb, inoperantă pentru inteligența artificială). În acest caz, este important ca între # și ((*Xs*)) să nu fie introdus blankul.

O altă problemă este plasarea notației pentru pauză în interiorul intervenției unui participant sau între intervenții (între rânduri). Uneori distincția între goluri, discontinuități și tăceri semnificative (Ionescu-Ruxăndoiu, 1999, 36) nu este ușor de aplicat (v. Orletti / Testa, 1991, 273). Soluția propusă de Jefferson (1978, xiii) pentru asemenea situații este, în opinia noastră, preferabilă, întrucât limitează interpretările din etapa transcrierii.

6.6. În ceea ce privește comentariile cercetătorului (glosări, observații privind înregistrarea etc.), dar și marcarea întreruperii pasajului transcris, considerăm că soluția folosirii barelor oblice /text/ este preferabilă aceleia de a combina tipuri de paranteze: ([, {[etc. De asemenea, folosirea notațiilor specializate care includ litere *sic!*, *AK*, *K* ar putea fi izolată cu același tip de semne, /text/, permițând o lectură mai ușoară, dar și excluderea lor, în funcție de interesele celui care utilizează transcrierea. Menționăm că există și alte așa-numite „notații specializate”, pentru fenomene precum *false start*, semnul \perp , sau *latching*, notat cu =, dar simbolurile nu sunt caractere alfanumerice (litere sau cifre) și nu a fost necesară izolarea lor în text cu ajutorul parantezelor.

6.7. În cele două sisteme de transcriere pentru româna vorbită notarea numelor proprii în transcrierile de limbă vorbită este abordată diferit. În *CORV* autoarea optează pentru marcarea în text a numelor proprii, folosind convenția din ortografia standard (majuscula). În *IVR* numele proprii nu sunt marcate. Pe de-o parte, transcrierea urmărește redarea rostirii și din acest punct de vedere nu se justifică simboluri suplimentare pentru semnalarea numelor proprii. În plus, convenția din ortografia curentă se suprapune cu notarea emfazei (care se face folosind majusculele), ducând uneori la ambiguitate, în cazul vocalelor inițiale (v. *procesul de integrare-n Uniunea EuroPEAnă*, *CORV*, 228). Pe de altă parte, nemarcarea numelor proprii poate crea dificultăți în înțelegerea textului. Un exemplu ar fi secvența: *o să văd codru* (*IVR*, 177), în care nu este vorba de o excursie în pădure, ci de o persoană („O să văd, Codru[ța]”), fapt care reiese din lectura atentă a textului transcris: *nu știu codru↓ oricum↓ mai mă hotărăsc↓ și: ((bip)) te sun↓ da:?*

Marcarea numelor proprii este, în opinia noastră, importantă, din mai multe motive. Din punct de vedere pragmatic, acestea trimit obligatoriu la cunoștințe comune locutorului și interlocutorului (Bidu-Vrăncianu *et al.*, 2001, 415). Este vorba, în acest caz, de alt act de comunicare decât dialogul transcris; emițătorul este autorul transcrierii, iar receptorul este cel care citește și, eventual, utilizează transcrierile. Așadar este greu de anticipat care dintre informații sunt cunoscute, mai ales atunci când nu este un antroponim, ci un titlu de lucrare, numele unei instituții etc. Nemarcarea numelor proprii în text ar face necesară o listă de note explicative pentru fiecare dintre transcrieri, în timp ce autorii volumelor de acest tip preferă o linearizare a informației din comunicarea orală.

Din punct de vedere gramatical, clasa numelor proprii prezintă o serie de particularități, iar posibilitatea de a le analiza sistematic în limba vorbită este un argument demn de luat în calcul. Propunerea noastră este ca acestea să se marcheze, dar nu cu majusculă, din considerente de consistență internă a sistemului de transcriere, ci cu încadrarea între bare oblice (*backslash*) a numelui:

CE legătură avem noi cu \ușă interzisă\). În plus, nemarcat în transcriere, un nume propriu la singular, precedat de articolul hotărât, ar face dificilă decodarea corectă a enunțului în cazul utilizărilor metaforice ale numelor proprii. „Notorietatea” referentului inițial al numelui propriu metaforizat, condiție a metaforizării (Miron-Fulea, 2002, 346), se poate aplica în cazul participanților la dialogul înregistrat, dar nu în cazul utilizatorilor transcrierii. Autorul înregistrării/transcrierii are, de cele mai multe ori, informații suplimentare în raport cu receptorul textului transcris, întrucât în antologiile de acest tip se publică, în general, numai fragmente din interacțiunea verbală propriu-zisă.

7. În cele ce urmează, vom prezenta succint câteva **funcții de căutare automată** în textul transcrierii. Accesul la text în format electronic ne permite să folosim funcții de căutare prevăzute în editoarele de text. În *Microsoft Word*, spre exemplu, selectând succesiv următoarele opțiuni: *Edit*, (*Find and*) *Replace*, *More*, *Use Wildcards* vom putea defini oricare dintre șirurile de caractere (și, implicit, fenomenele astfel codificate), pentru a le identifica în text, număra sau exclude din transcrieri. După ce selectăm opțiunea *Use Wildcards*, în *Special* putem afla mai multe despre codul folosit de calculator pentru a identifica șirul de caractere dorit. Nu este suficient să copiem exact secvența grafică din text și să o inserăm în *Find*, ci trebuie să respectăm sintaxa impusă de calculator. Astfel, secvența *[a-z]* înseamnă orice literă de la *a* la *z*, *@* - repetarea unității anterioare de oricâte ori, iar prin combinarea lor, *[a-z]@*, vom obține orice cuvânt, de orice dimensiune, dar fără alte semne în interiorul său, cum ar fi : pentru lungirea silabei. Dacă dorim să includem și această variantă în funcția de căutare apelăm la secvența *?@*, unde semnul *?* înseamnă orice caracter (unul și numai unul). În *Special* vom găsi o listă de astfel de corespondențe; semnalăm însă faptul că o serie de simboluri grafice: *<*, *>*, *!*, *@*, *?*, *[*, *]* etc. au alte semnificații în *Use Wildcards*. Pentru a le include totuși în șirurile de caractere pe care dorim să le identificăm în text, trebuie ca în căsuța de la *Find* fiecare semn din *Special* folosit cu altă valoare decât în lista data să fie precedat de ** (*backslash*).

Tabelul nr. 2

Tip de paranteze	Clasă de fenomene	Exemple	Avantaje
1 [] și IPA	rostirea cuvintelor străine	<i>['ædvãtaizɨn]</i>	<input type="checkbox"/> precizia notației <input type="checkbox"/> valorificarea unei convenții anterioare și de largă circulație
Tip de paranteze	Clasă de fenomene	Exemple	Avantaje
2 { }	suprapuneri (eventual și întreruperi)	<i>A; unde# {am fost I} eu vara trecutã. B; {ai fost I}</i>	<input type="checkbox"/> precizia notației <input type="checkbox"/> eliminarea dificultăților tehnice în transferul de date <input type="checkbox"/> prelucrare computerizată

				eficientă
3	<>	secvență grafică corespunzătoare rostirii: (1)mărci paradiscursive, (2) secvență neclară și transcriere incertă și (3) secvență indescifrabilă (fiecare x corespunde unei silabe rostite)	(1) \transil<ɾvaniaɨ>\ (2) <ɾacoloɾ> (3) <xx>	<input type="checkbox"/> păstrarea unității grafice a cuvântului <input type="checkbox"/> precizia notației în cazul marcării a două sau mai multe mărci pentru aceeași secvență rostită <input type="checkbox"/> posibilitatea realizării unor statistici computerizate <input type="checkbox"/> posibilitatea selectării automate a uneia sau mai multor secvențe marcate <input type="checkbox"/> posibilitatea eliminării automate a parantezelor de acest tip, păstrându-se doar textul corespunzător rostirii. <input type="checkbox"/> înlesnirea lecturii
4	()	elemente paraverbale: (1) descrierea în cuvinte a fenomenului și (2) convenții pentru fenomenele mai frecvente: (@@), (H), (Hx) etc.	(1) A; cred că noi (tușește) (2) A; (H) domnule \pleșu\↓	<input type="checkbox"/> posibilitatea eliminării automate a notațiilor respective dacă nu corespund obiectivelor utilizatorului <input type="checkbox"/> posibilitatea realizării unor statistici <input type="checkbox"/> înlesnirea lecturii
5	(())	elemente nonverbale; tăcerea, cu #(durata în secunde))	A; ((se apropie de microfon)) <ɾstimați colegiɾ> #((3s))	<input type="checkbox"/> posibilitatea eliminării automate a notațiilor respective <input type="checkbox"/> posibilitatea realizării unor statistici <input type="checkbox"/> înlesnirea lecturii
6	//	metatranscriere: (1) comentariile cercetătorului, (2) notații specializate: /K/, /AK/, /sic!/, (3) întreruperea pasajului transcris	(1) A; conviețuirea a fost posibilă /sună telefonul/ /.../ A; regele lor \ștefan\ i-a creștinat. /14,5 sec./ (2) A; v-ar place (sic!)	<input type="checkbox"/> posibilitatea eliminării automate a notațiilor respective <input type="checkbox"/> posibilitatea realizării unor statistici <input type="checkbox"/> înlesnirea lecturii
7	\\	marcarea numelor proprii	A; CE legătură avem noi cu ușa interzisă\.	<input type="checkbox"/> notarea unei informații importante fără a periclita consistența internă a sistemului (v. utilizarea majusculor pentru emfază)

Vom oferi o listă de expresii corespunzătoare celor din **Tabelul nr. 2**, astfel încât, inserându-le în *Find what* din *Find and Replace*, să fie identificate corect în textul transcrierii. Completând căsuța corespunzătoare lui *Find* cu șirul de caractere indicat, putem face două operații: numărarea ocurențelor fenomenului respectiv în transcriere (se trece toată expresia de la *Find what* între paranteze

rotunde, iar la *Replace with* se scrie numai `|I`, adică orice expresie rezultată în urma căutării automate va fi înlocuită cu ea însăși) și excluderea unor fenomene care nu prezintă interes pentru o anumită cercetare bazată pe transcriere (se tastează un blank în *Replace with*), cum ar fi elementele nonverbale, spre exemplu.

7.1. Pentru a căuta în text numai cuvintele străine, notate cu IPA, folosim secvența `|/?@|`.

7.2. Suprapunerile și întreruperile notate ca în tabel pot fi căutate cu `|{/?@|}`.

7.3. Pentru a iniția o căutare automată a mărcilor paralingvistice folosim `|<?@|>` sau, pentru fiecare marcă în parte, de exemplu, cu `|<ȘOP?@ȘOP|>`. Secvența *ȘOP* va fi înlocuită, la fiecare căutare, cu prefixul corespunzător tipului de marcă: *Î, J, F, P, R* etc.

Transcrierea incertă este codificată în *Find what* astfel: `|<|??@|?|>`, iar secvențele indescifrabile cu `|<x@|>`.

7.4. Elementele paraverbale notate în transcriere pot fi identificate cu `|/!(/|/(/!(/|@)|`. Pentru acestea am avut în vedere excluderea posibilității ca în urma căutării automate să obținem și parantezele simple „incluse” în notarea celor duble.

7.5. Identificarea elementele nonverbale se poate face cu `|/(/?@|)|`.

7.6. Comentariile cercetătorului pot fi găsite în text cu `|V?@V|`, notațiile specializate cu `|VKV|`, `|AKV|` și, respectiv, `|sic!V|`, iar întreruperea pasajului transcris cu `|V...V|`.

7.7. În cazul în care analizăm numele proprii care apar în transcrieri, scriem în *Find what* secvența `\\?@\\`.

8. Concluzii. Faptul că prelucrarea computerizată poate constitui un instrument de lucru puternic și eficient, inclusiv în domeniul științelor umaniste, este un loc comun astăzi. Fișarea materialului după parametri bine stabiliți (eventual corelați) și realizarea statisticilor pot prelua deja o parte migăloasă și consumatoare de timp din munca specialiștilor. Pentru aceasta este necesar, însă, ca datele introduse în calculator să fie compatibile cu inteligența artificială, neglijarea sau nerespectarea unor reguli minore de redactare putând împiedica o bună „colaborare” om – mașină.

Ar fi util, în opinia noastră, ca efortul cercetătorilor de a surprinde în scris complexitatea comunicării orale, prin intermediul transcrierilor, să fie contrabalansat de o sistematizare automată a datelor din corpus.

În ceea ce privește reorganizarea simbolurilor folosite pentru a codifica diferite fenomene ce apar în interacțiunea verbală, în contribuția de față am propus mai multe clase de elemente pentru care să se folosească diferite tipuri de paranteze: 1) informație „neverbală”, care ține de interacțiunea propriu-zisă: paraverbal (*tușește*), b) nonverbal (*(se ridică de pe scaun)*); 2) informație verbală,

care ține de interacțiunea propriu-zisă: a) transcriere IPA pentru cuvinte străine, b) mărci paraverbale <*ftext_F*>, c) secvențe incerte <*?text?*>, d) secvențe indescifrabile <*xxxx*>; 3) observațiile cercetătorului: a) /*comentariu*/, b) întreruperea secvenței transcrise /.../, c) unele notații specializate: /*K*/, /*AK*/, /*sic!*/; 4) fenomene interacționale: suprapunerile {*text n*} și întreruperile {*n*}.

NOTE:

Linguistic Annotation <http://www ldc.upenn.edu/annotation>

MIRON-FULEA, Mihaela, „Numele proprii metaforice în limba română actuală”, în Gabriela PANĂ DINDELEGAN (coord.), *Aspecte ale dinamicii limbii române actuale*, București, Editura Universității din București, 2002, p. 337-348.

O'CONNELL, Daniel C. și Sabine KOWAL, „Some Current Transcription Systems for Spoken Discourse: A Critical Analysis”, în *Pragmatics*, 1994, 4, p. 81-107.

ORLETTI, Franca și Renata TESTA 1991. „La trascrizione di un corpus di interlingua: aspetti teorici e metodologici” în *Studi italiani di linguistica teorica e applicata*, XX, 1991, 2, p. 243-283.

ASPECTS OF SPOKEN ROMANIAN TRANSCRIPTION. A COMPUTERIZED ANALYSIS PERSPECTIVE

The aim of our study is to approach the process of transcription from the perspective of computerized analysis, which enables researchers to make a virtually infinite number of statistics, to correlate various linguistic elements or just check their hypotheses on the correlation of specific phenomena. Our analysis is focused on the use of brackets, square brackets, braces a.s.o. in the transcription of spoken Romanian, corresponding to the categories of phenomena encoded. We have defined a number of types of information given in a transcription: information corresponding to the actual verbal interaction which is transcribed (verbal, vocal nonverbal sounds or nonverbal elements) and to the transcriber's perspective, respectively. Also, the study provides tools for a computerized analysis, if the conventions used in the transcriptions do not flout internal consistency and they are written correctly (see the misuse of space, the order of symbols, etc.).

¹ Du Bois detaliază primele două categorii: „Who will use the transcriptions? Discourse researchers, of course, in all their variety. But these days their interest in discourse is shared by an everwidening circle. Grammarians and general linguists use transcriptions as sources of linguistic data on a range of topics, and to follow the action in theories grounded in discourse; computational linguists use them to test speech recognition protocols against actual language use; language teachers use them to illustrate realistic uses of spoken language; social scientists use them for understanding the nature of social interaction; curious folks find it intriguing to look closely at how people really talk; and the students of any of these may use transcriptions to learn more about their field of study. And, as we shall see, one of the most important groups of users is the transcribers themselves. A good transcription system should be flexible enough to accommodate the needs of all these kinds of users”. (1991, 74)

² „DEFINE GOOD CATEGORIES: 1. Define transcriptional categories which make the necessary distinctions among discourse phenomena., 2. Define sufficiently explicit categories., 3. Define sufficiently general categories., 4. Contrast data types.
MAKE THE SYSTEM ACCESSIBLE: 5. Use familiar notations., 6. Use motivated notations (iconicity and internal consistency), 7. Use easily learned notations., 8. Segregate unfamiliar notations., 9. Use notations which maximize data access., 10. Maintain consistent appearance across modes of access.

MAKE REPRESENTATIONS ROBUST: 11. Use widely available characters., 12. Avoid invisible contrasts., 13. Avoid fragile contrasts.

MAKE REPRESENTATIONS ECONOMICAL: 14. Avoid verbose notations., 15. Use short notations for high frequency phenomena., 16. Use discriminable notations for word-internal phenomena., 17. Minimize word-internal notations., 18. Use space meaningfully.

MAKE THE SYSTEM ADAPTABLE: 19. Allow for seamless transition between degrees of delicacy., 20. Allow for seamless integration of user-defined transcription categories., 21. Allow for seamless integration of presentation features., 22. Allow for seamless integration of indexing information., 23. Allow for seamless integration of user-defined coding information”. (Du Bois *et al.* 1988, 81-97)

³ ASCII și Unicode, standarde de reprezentare a informației textuale, permit transferul datelor în computer, indiferent de platformă. Prin platformă se înțelege orice combinație posibilă de sisteme de operare (cum ar

fi Windows 98, Windows 2000, Linux, Mac-OS etc.) și tipul de computer (IBM-PC, Macintosh etc.). ASCII are un inventar de 256 (2^8) unități. 128 dintre acestea (ASCII redus) codifică alfabetul englez și un set limitat de semne de punctuație: a) valorile numerice cuprinse în intervalul 0-31 și 127 codifică semne non-grafice (cum ar fi trecerea pe un rând nou, de pildă), b) 32 - pauza dintre cuvinte sau blankul și c) valorile de la 33 la 126 codifică semne grafice: semnele de punctuație, cifrele și literele (minuscule și majuscule). Valorile cuprinse în intervalul 128-255 sunt folosite, pentru fiecare limbă în parte, pentru a codifica semnele grafice specifice. Aceasta înseamnă că atribuirea unui cod numeric (128-255) se face *diferit* pentru celelalte semne care nu sunt incluse în alfabetul englez, iar *ă*-ul românesc nu va fi recunoscut de un editor de text suedez, spre exemplu.

Unicode are un inventar de 65.536 (2^{16}) unități și fiecare simbol are o valoare numerică unică (deci poate fi transferat și recunoscut de la o platformă la alta, dar și de la o limbă la alta). Dat fiind numărul foarte mare de unități, Unicode include literele specifice ortografiei standard a majorității limbilor (în cazul românei, și diacriticele), inclusiv ideograme. Toate simbolurile incluse de *Microsoft Word* (folosind comanda *Insert*, opțiunea *Symbol* și fontul *Times New Roman*) într-un inventar foarte accesibil fac parte din Unicode.

Singurul dezavantaj posibil al standardului Unicode față de ASCII este faptul că ocupă, comparativ, mai mult spațiu de stocare (ceea ce este firesc în raport cu inventarul său), însă nesemnificativ pentru tehnologia actuală.

Bibliografie:

- BIDU-VRÂNCEANU, Angela, Cristina CĂLĂRAȘU, Liliana IONESCU-RUXĂNDIOIU, Mihaela Mancaș, Gabriela PANĂ DINDELEGAN, *Dicționar de științe ale limbii*, București, Nemira, 2001.
- CRESTI, Emanuela, *Corpus di italiano parlato*. Vol. I, II, Firenze, 2000.
- DASCĂLU JINGA, Laurenția, *Corpus de română vorbită (CORV)*. Eșantioane, București, Oscar Print, 2002.
- DU BOIS, John W., Susanne CUMMING, Stephan SCHUETZE COBURN, „Discourse Transcription”, în S. A. Thompson (ed.) *Discourse and Grammar (Santa Barbara Papers in Linguistics, 2)*, p. 1-71, 1988.
- DU BOIS, John. W., „Transcription Design Principles for Spoken Discourse Research”, în *Pragmatics*, 1991, 1, p. 71-106.
- GHIGA, Georgeta, *Elemente fatice ale comunicării în româna vorbită*, București, Editura Alcris, 1999.
- HUANG, Xuedong, Alexandro ACERO și Hsiao-Wuen HON, *Speech Processing*, www.clsp.jhu.edu/courses/zilla, 2000.
- IONESCU-RUXĂNDIOIU, Liliana, *Conversația: structuri și strategii. Sugestii pentru o pragmatică a românei vorbite*, ediția a II-a, București, ALL, 1999.
- IONESCU-RUXĂNDIOIU, Liliana (coord.) *Interacțiunea verbală în limba română actuală. Corpus (selectiv). Schiță de tipologie*, București, Editura Universității din București, 2002.
- JEFFERSON, Gail 1978. „Explanation of transcript notation”, în J. SCHENKEIN (ed.) *Studies in the Organization of Conversational Interaction*, New York /San Francisco /London, 1978, p. XI-XVI.