Aspecte ale transcrierii limbii române vorbite în vederea prelucrării computerizate

DIANA GHIDO

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

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 recunoaşterea vocală şi perfecționarea interacțiunii om – maşină, 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* ş.a.m.d.

În ceea ce priveşte limbile romanice, amintim, pentru limba franceză, *GARS-ESA 6060* al CNRS şi *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 referentia do português contemporaneo*. Există şi 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 şi 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, şi 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). Eşantioane* (Dascălu Jinga, 2002) şi *Interacţiunea verbală în limba română. Corpus (selectiv). Schiţă de tipologie* (Ionescu-Ruxăndoiu, 2002). În cele ce urmează, ne vom referi la cele două lucrări folosind siglele *CORV* şi, respectiv, *IVR*.

1.3. Preocupările noastre legate de **transcrierea în vederea prelucrării computerizate** au apărut în urma participării în 2001 la proiectul *Interacţiunea verbală în limba română. Corpus şi 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). Eşantioane*, prilej cu care am ascultat şi 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 opţiunilor pentru reprezentarea grafică a materialului înregistrat audio, de problemele practice de limitare a interpretării în transcriere, de consistenţă internă şi flexibilitate a sistemului de convenţii de notare.

Suntem de părere că întrebuinţarea unui sistem de transcriere care să permită o cât mai bună prelucrare a datelor cu ajutorul computerului nu vizează doar nişte aplicaţii pe termen lung. Valorificarea optimă a unei colecţii de transcrieri este posibilă deja prin facilităţile 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 fişa materialul în funcţie de obiectivul cercetării (de la statisticile privind frecvenţa relativă şi/sau absolută a unor unităţi lexicale şi până la selectarea tuturor ocurenţelor unui fenomen surprins în transcrieri, a suprapunerilor, spre exemplu) este necesar ca notaţiile definite să fie riguroase, clare şi aplicate sistematic.

Exigenţele cercetărilor similare realizate pentru alte corpusuri de limbă vorbită din lume sunt deosebit de mari. Culegerea corpusului, arhivarea şi transcrierea sunt numai primii paşi în studiul limbii vorbite. Arhivarea înregistrărilor audio pe suport digital (CD) este obligatorie pentru a trece la o treaptă superioară de prelucrare a materialului, şi anume alinierea textului transcris la secvenţa 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 sfârşitul 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 unităţi intonaţionale (*parsing*) a fiecărui text. Mai mult, textul aliniat se etichetează pe niveluri de analiză lingvistică (*textual tagging*) şi studiile de limbă vorbită incluse în proiect pornesc de la aceste date riguros arhivate (Cresti, 2000).

Sinteza şi recunoaşterea vocală – necesare pentru a trece de la interfaţa grafică a comunicării dintre om şi inteligenţa artificială la o interacţiune bazată

(şi) pe comenzi rostite –, dar şi alte aplicaţii extralingvistice de interes larg (v. Huang *et al.*, 2000), depind în mare măsură de asemenea corpusuri de limbă vorbită şi de prelucrarea lor computerizată. Pentru a atinge un asemenea obiectiv este nevoie, pentru fiecare limbă, nu numai de existenţa unui corpus de limbă vorbită şi de transcrierea lui, ci şi 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 întrebuinţate (luând în considerare normele pe care le impune obiectivul prelucrării computerizate ulterioare) şi mijloacele tehnice de identificare, căutare şi selectare a simbolurilor grafice cu ajutorul programului *Microsoft Word*. Ne vom opri în prezenta contribuţie numai asupra aspectelor privind utilizarea parantezelor în transcrierile existente, şi, respectând opţiunea autorilor pentru inventarul de fenomene notate, vom lua în discuţie o reorganizare a corespondenţelor dintre aceste fenomene şi simbolurile întrebuinţate. 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 şi în acest scop. Computerul în general, şi editoarele de text curente în special, sunt deja instrumente puternice de analiză, care permit fişarea materialului într-un timp mult mai scurt şi după parametri exacţi. În elaborarea statisticilor de orice tip, dar şi pentru verificarea oricăror ipoteze privind corelarea a două sau mai multe fenomene marcate în transcriere, inteligenţa 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 şi video, dar şi contextul specific românesc postdecembrist (în care diversitatea interacţiunilor verbale surprinse este uneori limitată din cauza suspiciunii faţă de înregistrări a multora dintre posibilii subiecţi) încă îşi pune amprenta asupra **metodologiei culegerii corpusului**. Ca urmare, sintaxa mixtă – corelarea componentei paraverbale şi/sau nonverbale cu informaţia transmisă verbal – rămâne de cele mai multe ori neconsemnată sistematic.

Metodologia culegerii corpusului şi cea a prelucrării lui pot avea, în opinia noastră, influenţe antagonice asupra sistemului de transcriere: pe de o parte, cu cât aparatele de înregistrare sunt mai performante (prin aceasta înţelegând şi flux de informaţii complex, audio şi video), cu atât procesul transcrierii trebuie să filtreze şi să sistematizeze mai multe date. Pe de altă parte, tehnologia prelucrării datelor din transcrieri include, aşa 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 şi 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 şi înregistrarea.

În 1991, Orletti / Testa reproşau transcrierilor faptul că urmăresc în cea mai mare parte verbalul (în detrimentul paraverbalului şi 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 transcrizioni sono state ugualmente verbo-centriche, mostrando la tendenza a privilegiare i dati verbali nelle transcrizioni 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 observaţia că sistemele noastre de transcriere încă trebuie să linearizeze discursul complex, încercând să noteze şi celelalte componente ale comunicării. Lipsa accesului la tehnologia de prelucrare este însoţită, deocamdată, în cercetările asupra interacţiunii verbale în româna vorbită de lipsa mijloacelor tehnice adecvate pentru culegerea datelor. Astfel, deşi ambele sisteme de transcriere analizate prevăd convenţii de notare a elementelor nonverbale, materialul de acest tip rezultat în transcrieri este relativ redus, iar fluxul de informaţii 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 secvenţial al comunicării orale nu permite confruntarea datelor şi corelarea diveri;ilor factori care influenţează desfăşurarea unei interacţiuni verbale. **Necesitatea transcrierii** este evidentă, iar sistemul de convenţii definit este responsabil pentru consemnarea consecventă şi neambiguă a fenomenelor vizate de obiectivul cercetării. Atributele pe care trebuie să le aibă un sistem de transcriere funcţionează de cele mai multe ori antagonic, un exemplu elocvent fiind dezideratul de a nu pierde, în procesul „traducerii" în scris a conţinutului înregistrărilor, informaţii posibil relevante, dar de a evita, în acelaşi timp, ca textele transcrise să fie prea încărcate şi greu de urmărit. Toate aceste aspecte au condus la proliferarea sistemelor de transcriere, la analize permanente şi perfecţionări 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 dificultăţile 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 şi de destinatarul unui astfel de text. Există trei mari tipuri de **destinatari**: specialişti (lingvişti), nespecialişti şi... 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 lingviştii şi inteligența artificială ca destinatari

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

3.2.1. Du Bois observă că, în ceea ce priveşte **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 şi 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 şi j) comentarii (sau mijloace de evidențiere) privind transcrierea însăşi (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 şi 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, aşa 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 şi citirii unui text

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

Precizăm că foarte multe dintre principiile enunţate 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 notaţiile incluse) şi prelucrarea computerizată. ASCII şi Unicode sunt standarde de reprezentare a informaţiei textuale în computer. Inventarul Unicode este mult mai mare decât al celuilalt sistem, însă trebuie reţinut că nici unul dintre ele nu codifică (şi nu „păstrează" la transferul de date) anumite posibilităţi de tehnoredactare care constau în schimbarea unor proprietăţi ale unor semne grafice, şi 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) Aşadar, nu va putea fi inclus în prelucrarea computerizată un fenomen căruia îi corespunde o notaţie numai prin sublinierea caracterelor, îngroşarea sau schimbarea dimensiunii a corpului de literă etc., pentru că aceasta nu se păstrează în trecerea de la o platformă[3] la alta.

Notaţiile propuse de Du Bois *et al.* se încadrează în inventarul standardului ASCII redus. Deşi sugestiile din analiza noastră sunt incluse în inventarul ASCII (cu excepţia 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 (permiţând mai multă flexibilitate în notaţii) şi b) ambele sisteme de transcriere pentru româna vorbită includ deja semne care fac parte din Unicode şi nu fac parte din ASCII (vezi conturul intonaţional non-terminal: ascendent, ↑, şi, respectiv, descendent, ↓, precum şi notaţia prevăzută în *CORV* pentru observaţiile cercetătorului, →, dar şi diacriticele româneşti). 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 direcţiile inaugurate de Jefferson (1974) şi, respectiv, Gumperz (1982).

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

interacţiunea verbală. Caracterul secvenţial al interacţiunii verbale şi ilustrarea lui sunt definitorii pentru sistemul Jefferson. În ultimele decenii acest sistem a cunoscut numeroase îmbunătăţiri şi 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 „filiaţie" este firească, având în vedere şi similitudinea obiectivelor urmărite în analiza interacţiunii verbale. Sistemele de transcriere pentru româna vorbită prezintă o serie de diferenţe în raport cu sistemul lui Jefferson (1978) – adaptări, rafinări ale convenţiilor – cele mai semnificative fiind, în opinia noastră, cele legate de notarea sistematică a unor elemente de prozodie şi raportul textelor transcrise cu ortografia standard.

4.2. Setul de convenţii 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 şi, 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 şi tipurilor de paranteze pe care le foloseşte fiecare dintre cele două sisteme de transcriere a românei vorbite, *CORV* şi *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 şi barele oblice, folosite într-o manieră asemănătoare parantezelor, ca o structură din două elemente simetrice (identice, de fapt, în cazul barelor) ce izolează o secvenţă grafică de lungime variabilă: primul element al acestei structuri este bara precedată de blanc şi urmată imediat de caractere grafice, iar ultimul element este aşezat imediat după caracterele grafice şi urmat de blanc sau de unul dintre semnele: ?,., ↓, ↑ sau # (ce marchează conturul intonaţional şi pauza în rostire).

5.1. În *CORV* se folosesc: paranteze pătrate, *[text]*, paranteze rotunde, *(text)*, şi paranteze unghiulare, <*text*>. Parantezele pătrate sunt întrebuinţate 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: *[îşi drege vocea]*, 4) componenta nonverbală *[gest afirmativ cu capul]* şi 5) diverse observaţii privind înregistrarea şi desfăşurarea interacţiunii verbale: *[scurtă ştergere involuntară a înregistrării]* (*CORV*, 95) sau *[Oprirea vorbitorului şi întreruperea înregistrării, pentru că sună telefonul în încăpere]* (*CORV*, 93). Am precizat care este „conţinutul" parantezelor, pentru a evidenţia faptul că nu se pot face confuzii între cele trei mari tipuri de utilizări ale parantezelor drepte: cu semne din alfabetul fonetic, (1),

cu blancuri, (2), şi 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 informaţii. Pluralitatea semnificaţiilor pe care le are folosirea parantezelor drepte în transcriere rezultă din convergenţa unor convenţii anterioare, preluate din coduri diferite; spre exemplu, sistemul IPA este consacrat, dar şi folosirea parantezelor drepte în notarea suprapunerilor este frecvent întâlnită în corpusurile dedicate analizei conversaţiei (v. Jefferson 1978, Du Bois *et al.* 1988 şi Du Bois 1991 etc.).

Parantezele unghiulare nu se folosesc decât pentru cuvinte care au fost rostite efectiv în interacţiunea verbală propriu-zisă, fie „marcate" paralingvistic (5), fie secvenţe incerte (6) sau indescifrabile (7). Această convenţie 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 notaţiei derivă din modul descriptiv şi virtual nelimitat în care se pot alege mărcile şi prefixul care să le codifice; acest fapt se poate observa şi din modul în care a fost valorificată în sistemele româneşti. În *CORV* se notează: ritmul vorbirii, (lent <***L*** *text* ***L***> sau rapid <***R*** *text* ***R***>), înălţimea vocii, (ridicată <***Î*** *text* ***Î***> sau joasă <***J*** *text* ***J***>), intensitatea, (puternică <***F*** *text* ***F***> sau slabă <***P*** *text* ***P***>), şoptitul <***ŞOP*** *text* ***ŞOP***>, 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*>, şi secvenţele rostite zâmbind <$_{Z}$ *text*>; la acestea se adaugă o informaţie privind caracterul planificat, nespontan al unor comunicări orale, şi anume lectura unui text: <$_{CIT}$ *text*>.

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

Tabelul nr. 1

Tip de paranteze	*CORV*		*IVR*	
	Semnificaţia	Exemple	Semnificaţia	Exemple

[ ]	1) transcriere IPA	*„vecinic" [ve't͡ʃi nik] te te-ncurca↑ (163)*	1) [marchează începutul suprapunerii unor intervenţii succesive.	***A:** student la petrol↓ [aici?* ***B:** [nu. la bucureşti* (27)
	2) plasate între rânduri, notează secvenţe care se suprapun	***GP:** Da. De acord* [ ] ***VJ:** Că acolo diferenţa era enormă* (157)	2) întreruperea pasajului transcris	*[...]*
	3) fenomene paraverbale	***VC:** [râde]* (251)		
	4) fenomene nonverbale	***AB:** [gest afirmativ cu capul]* (269)		
	5) diverse observaţii privind înregistrarea	*[scurtă ştergere involuntară a înregistrării]* (95)		
< >	6) mărci paralingvistice	***VL:** <**R** preşedintele României↓ domnul Emil Constanti <**Î** nescu **Î**> **R**>* (276)	3) mărci paralingvistice	***B:** <z io văd aşa↓<@ că toate> problemele sî:nt bu:ne:>* (191)
	7) transcriere incertă	***SF1:** <**?** Nu prea ştiu.**?**>* (166)		
	8) secvenţă indescifrabilă	***CJ:** Da↓ <xxxxxxxxxx>* (71)		
( )	9) scurte explicaţii necesare înţelegerii textului	***MV:** ce (zice) V-au venit nişte bani din ţară↓* (115)	4) transcriere incertă	***A:** (ca un fel de invitaţie) pentru oameni d-ăştia* (35)
	10) pauze foarte lungi	***LDJ:** Nu era încălzire? Iarna?* ***GD:** (3 sec.)* (86)		
	11) transcriere pseudofonetică (cuvinte străine şi acronime)	***GD:** La căminul I.O.V. (iove)* (86) *Heidelberg (haidălberg)* (74)	5) secvenţă indescifrabilă	***A:** (xxx)* ***B:** nu încă.* (38)
	12) întreruperea pasajului transcris	*(...)* plasat între rânduri (*passim*)		
	13) notaţii specializate: heterocorectare (K), autocorectare (AK) şi eroare necorectată (sic!).	***VJ:** (K) Nu↓ Lăteşti.* (56) ***CJ:** în proteş- (AK) în procesul lui Pătrăşcanu.* (56) ***IS:** Vă vor place (sic!)* (270)		

(( ))			6) comentariile cercetătorului	*((între timp sosise în staţie un microbuz))* (27)
			7) fenomene paraverbale	*((rîde))* (31) *((îşi drege vocea))*
			8) fenomene nonverbale	*((se uită la ceas))* *(27)*
/ /			9) transcriere pseudofonetică (pentru cuvintele în limbi străine şi abrevieri)	***A:*** *am văzut în /vog/* (53) ***B:*** *firma /secea/* (91)

O a doua observaţie se referă la posibilitatea de a nota particularităţi izolate cu o convenţie asemănătoare, fără a risca să îngreuneze asimilarea sistemului de transcriere prin adoptarea unor notaţii prea numeroase. Am întâlnit un astfel de caz în transcrierile noastre, când unul dintre participanţi fredonează câteva cuvinte dintr-o melodie cunoscută, pentru ca imediat după aceea să treacă la adresarea directă faţă 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 secvenţe şi 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 secvenţei respective şi notarea, după transcrierea ei, a „mărcii", coindexat: < *text 1*> <fredonat *1*>.

În fine, din prezentarea anterioară a mărcilor pentru care a optat fiecare dintre sistemele menţionate rezultă şi valorificarea diferenţiată a opţiunilor de redactare computerizată. După cum aminteam la punctul **3**, nici scrierea cu aldine, nici poziţia literei faţă de rând nu constituie informaţii valide în prelucrarea computerizată, dar, fiind folosite auxiliar, ambele pot fi utile în înlesnirea lecturii. Diferenţierea secvenţelor grafice corespunzătoare „marcării" (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, <ŞOP, <@ etc.) care preced textul propriu-zis.

Parantezele rotunde sunt folosite pentru a izola de textul transcris comentariile cercetătorului (8), dar şi în transcrierea pseudofonetică, (9). În plus, parantezele rotunde sunt folosite pentru a semnala întreruperea pasajului transcris, cu (...), v. pct. (10), precum şi pentru a izola nişte notaţii specializate de tipul (K), (AK), (sic!), v. pct. (11). Ultimul tip de convenţie valorifică tradiţia notării cu secvenţa (K) a fenomenului de autocorectare în transcrierile textelor dialectale.

5.2. În *IVR* sunt folosite: parantezele pătrate: *[text]*, parantezele rotunde simple *(text)* şi duble *((text))* şi 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 şi 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: *((tuşeşte))* şi alte observații necesare înțelegerii textului: *((între timp sosise în stație un microbuz))* (*IVR*, 27).

Pe lângă paranteze, sistemul prevede şi 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 şi pentru secvență indescifrabilă (care în Jefferson nu are un şir de x între paranteze, ci doar blancuri), paranteze duble pentru componenta nonverbală şi cea paraverbală, precum şi 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 uşor de urmărit de către utilizatori, cât şi pentru a putea trece la prelucrarea computerizată a datelor), b) valorificarea unor deprinderi de lectură şi evitarea folosirii cu alt sens a unor semne grafice frecvent întrebuințate în ortografia curentă, c) definirea unor norme de redactare (succesiunea caracterelor grafice şi non-grafice) astfel încât, pentru orice transcriere în parte, fiecare utilizator să îşi 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. şi 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 şi pentru alte limbi, întrucât complexitatea limbii vorbite a evidenţiat, în numeroase situaţii, insuficienţa mijloacelor grafice întrebuinţate în ortografia curentă. Atât în *CORV*, cât şi în *IVR*, autorii optează, spre exemplu, (şi) pentru o transcriere pseudofonetică în cazul abrevierilor. Astfel, o secvenţă 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 pronunţării sale (care adesea variază de la un vorbitor la altul).

CORV foloseşte două asemenea sisteme auxiliare: IPA şi transcrierea pseudofonetică, iar *IVR* numai pe cel din urmă. Cu toate acestea, există anumite situaţii în care, la rândul lor, sistemele auxiliare se dovedesc insuficiente. Dacă o secvenţă precum *Harun Tazieff (harun tazief)* (*CORV*, 77) nu pare să ridice probleme, în alte situaţii aproximarea pronunţiei cu ajutorul semnelor din ortografia curentă este mai dificilă. În *dantele de Bruges (briuj)*, dincolo de faptul că nu se mai poate distinge pronunţarea 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 preferinţei pentru hiat sau diftong în româna actuală, dar şi în schiţarea unor probleme legate de gradul de instruire a vorbitorilor. În alte cazuri, transcrierea pseudofonetică se face folosind semnele IPA: *Jean Francois Revel (jã frãnsoa revel)* (*CORV*, 75), */uipatrõ/* (*IVR*, 89). Uneori se folosesc alte soluţii pentru a reda foneme nespecifice limbii române: */edvărtaizing/* (*IVR*, 37) sau */paundţ/* (*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 informaţii asupra accentului şi silabaţiei; în */menegimentu/* (*IVR*, 254) putem avea patru sau cinci silabe. Un caz interesant este transcrierea lui O.K., care este şi cuvânt străin, şi abreviere (**//ochei//*).

IPA este folosit în *CORV*, dar numai în cazuri excepţionale, „când interacţiunea verbală vizează însăşi pronunţarea sau necesită sugerarea cât mai precisă a acesteia” (*CORV*, 33).

Suntem de părere că ar fi utilă întrebuinţarea alfabetului fonetic şi în cazurile în care se folosea transcrierea pseudofonetică, pentru a sugera adaptarea fonetică a unor cuvinte noi sau foarte noi (xenismele), putându-se astfel analiza în funcţie de diferiţi parametri sociolingvistici. Sistemele „jeffersoniene” consideră, în general, că transcrierea fonetică nu este necesară pentru analiza conversaţiei; cele care sunt dedicate studiului achiziţionării unei limbi străine (v. Orletti / Testa, 1991) acordă o atenţie deosebită redării cât mai fidele a pronunţiei, urmând linia propusă de Gumperz. În cazul limbii române, există avantajul major al ortografiei sale fonetice (faţă de limba engleză, de pildă, unde apar o serie de dificultăţi 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) şi credem că ar fi interesant de notat consecvent pronunţia acestor cuvinte la diferiţi vorbitori, pentru a surprinde dinamica fenomenului.

De altfel, şi restul transcrierii în ambele volume este „pseudofonetic” (sau un sistem fonetic neconvenţional, aşa cum este numit în Orletti/Testa, 1991, 260), în sensul că nu corespunde ortografiei standard, ci încearcă să redea rostirea: *am crezt că e aceiaşi 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 şi pseudovocalelor, interesante din punct de vedere morfonologic. Semnalăm că acestea pot fi notate în transcrieri folosind convenţiile curente pentru aceste sunete, care se pot „traduce” pentru calculator în secvenţe grafice care să permită prelucrarea datelor.

Eliminarea literei x din transcrierea rostirii ar putea aduce, la rândul său, un plus de informaţie în analiza grupurilor [ks] şi [gz], în condiţiile în care se constată rostirea unuia în locul celuilalt la diverşi vorbitori. În plus, aceasta ar permite evitarea inexactităţii în marcarea emfazei (se scrie *eXACT, EXtraordinar*, dar cele două consoane codificate prin x aparţin unor silabe diferite) şi ar permite ca x să apară numai pentru redarea unei secvenţe 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 îmbunătăţire a acestei convenţii dacă s-ar folosi acoladele (păstrând parantezele drepte pentru IPA, o convenţie cu caracter mai general) şi indexarea s-ar face cu un şir de numere crescătoare, constant, până la sfârşitul transcrierii respective. Du Bois *et al*. (1988) propuneau coindexarea numai în cazul unor suprapuneri numeroase într-o anumită porţiune, iar după ce nu ar mai exista ambiguitate în privinţa secvenţelor rostite simultan, să se reia numărătoarea de la 1. Avantajul numerotării până la sfârşit este evident în cazul prelucrării computerizate: se pot „extrage” automat toate secvenţele cuprinse între paranteze şi pot fi analizate precis, în funcţie de conturul intonaţional, mărcile paralingvistice (în suprapunerile mai lungi este posibil ca cel puţin unul dintre vorbitori ridică vocea), sau relaţiile dintre participanţi.

Întreruperile se pot nota ca un caz particular, în care primul element este {i} (*i* fiind indicele numeric: *1, 2, 3...i,...n*) şi se va nota la sfârşitul rândului corespunzător intervenţiei întrerupte, iar al doilea este {i}, notat la începutul rândului, după sigla participantului care preia rolul de emiţător (***A;*** *ieri de CE te-ai supărat şi-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 şi sugestii de notare a lor. Marcarea începutului şi sfârşitului unei secvenţe rostite cu anumite particularităţi cu ajutorul parantezelor unghiulare

plasate în text a fost preluată atât în *CORV*, cât şi în *IVR*. Aşa cum semnalam, nici convenţia grafică a îngroşării literelor, nici scrierea unei secvenţe mai sus sau mai jos faţă 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 secvenţei: paranteză unghiulară urmată de o literă sau un grup de litere dintr-un inventar definit în convenţii. Din păcate, opţiuni de transcriere mai economice sau mai simple, precum cele folosite în *IVR* (<ŞOP, P text> text>) care pot fi citite relativ uşor de un receptor uman, prezintă dificultăţi majore în prelucrarea cu ajutorul inteligenţei artificiale. Combinaţiile de mărci (la care se adaugă ordinea permisivă de tipul: *<i,j text> text>* sau *<j,i text> text>,* pentru două mărci <i> şi <j> care ar caracteriza o anumită secvenţă) sunt foarte numeroase şi nu permit statistici exacte.

Am putut urmări, spre exemplu, în *CORV* numărul de ocurenţe al fiecăreia dintre mărcile definite în sistem şi am obţinut următoarele date: 236 de apariţii 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 <**ŞOP** text **ŞOP**>.

Menţionăm, cu această ocazie, câteva probleme de redactare. Pentru a permite prelucrarea computerizată, este necesar să se noteze simbolul mărcii respective la începutul şi sfârşitul secvenţei, cu semnul <, şi, 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 şi textul corespunzător rostirii, la începutul şi la sfârşitul marcării. Scopul de a nu îngreuna lectura, urmărit în ambele volume româneşti, poate fi realizat prin combinarea celor două mijloace grafice folosite: îngroşarea *<**J** text **J**>* şi, respectiv, poziţia faţă de rând *<ȷ text>*: *<ȷtextȷ>*.

Secvenţa 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, existenţa unor mijloace moderne de prelucrare a sunetului în format digital, care permit reducerea zgomotului de fond şi/sau amplificarea artificială a undei sonore pentru a limita, pe cât posibil, numărul transcrierilor incerte. Aceeaşi operaţie poate reprezenta o soluţie şi pentru unele dintre secvenţele indescifrabile. Cu toate acestea, în cazul în care informaţia nu se poate recupera, merită menţionat că se poate nota, de cele mai multe ori, conturul intonaţional şi pentru aceste secvenţe. Sugestia noastră ar fi adoptarea convenţiei folosite de Du Bois *et al.* (1988) şi, ulterior, de Du Bois (1991), potrivit căreia fiecare semn x ar nota o silabă din porţiunea indescifrabilă, iar nu un sunet. Segmentarea în cuvinte este aproape imposibilă în absenţa 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 doreşte eliminarea parantezelor rotunde simple şi se păstrează transcrierea incertă şi semnalarea cu x a fiecărei silabe indescifrabile, acest lucru este posibil.

6.4. În cadrul reorganizării unor elemente definite şi a unor notaţii pentru acestea, considerăm că ar fi un câştig dacă am exploata obişnuinţele de lectură ale utilizatorului, şi anume folosirea parantezelor. Folosirea parantezelor rotunde pentru secvenţe de text nesigure sau indescifrabile ca în *IVR* prezintă câteva inconveniente, întrucât aceste paranteze reflectă în general în ortografia curentă raportul informaţie principală – informaţie secundară. Acele cuvinte care nu au putut fi transcrise cu certitudine nu sunt mai puţin importante pentru construirea enunţului, ci doar accidental au ajuns să fie o informaţie nesigură. Optăm, în acest caz, pentru notaţiile din *CORV*, unde parantezele unghiulare notează numai cuvinte rostite în interacţiunea verbală (deci informaţie 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 secvenţe, este probabilă.

Suntem de părere că ar contribui la o mai bună organizare a transcrierii şi la o asimilare mai uşoară a convenţiilor de transcriere dacă s-ar nota diferit elementele nonverbale faţă de cele paraverbale. Pentru cele din urmă propunem parantezele simple (marcarea calităţii vocii păstrând paranteze unghiulare simple), iar pentru nonverbal parantezele duble. În acest fel, atenţia acordată de utilizator informaţiilor din interiorul parantezelor poate fi de acelaşi tip cu extragerea informaţiei la o lectură obişnuită: textul astfel izolat este parte integrantă din textul per ansamblu, dar de ordin secundar. Nonverbalul şi 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 expiraţia audibilă, să se folosească convenţiile lansate de Du Bois (1991): *(H)* pentru „inspiră adânc", *(Hx)* pentru „expiră", întrucât acestea izolează fenomenele vocale nonverbale de transcrierea rostirii propriu-zise (şi anume folosind constant acelaşi tip de paranteze, cele rotunde simple). Semnificaţia 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* şi *IVR*, fenomenele paraverbale discutate mai sus se notează astfel: *inspiră adânc* între paranteze

pătrate şi, respectiv, paranteze rotunde duble (dar astfel vor fi trecute laolaltă cu observaţii precum defectarea microfonului etc.). Pledăm aşadar pentru surprinderea acestor fenomene în transcriere, dar cu ajutorul unor convenţii cât mai simple, care să ocupe puţin spaţiu grafic şi să fie în concordanţă cu notaţiile pentru fenomene similare.

Propunem, de asemenea, notarea râsului ca în sistemele Du Bois *et al*. (1988) şi Du Bois (1991), adică inserarea câte unui semn @ pentru fiecare „silabă" de râs. Acest lucru ne va permite să marcăm durata relativă a secvenţei respective (faţă de notaţia din *CORV*, unde trecerea între paranteze, în text, a cuvântului *râde* nu oferea informaţii de acest tip), dar fără a introduce noi „cuvinte grafice" (adică nişte unităţi care nu corespund de fapt cuvintelor din rostirea participanţilor). 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) şi chiar <*@ hî hî hî*> (*IVR*, 172).

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

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 soluţie 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 întrebuinţat în *CORV* este, în prelucrarea computerizată, identic cu simbolul pentru contur descendent terminal; diferenţa dintre. şi. este aldin ~ alb, inoperantă pentru inteligenţa artificială). În acest caz, este important ca între # şi ((*X*s)) să nu fie introdus blancul.

O altă problemă este plasarea notaţiei pentru pauză în interiorul intervenţiei unui participant sau între intervenţii (între rânduri). Uneori distincţia între goluri, discontinuităţi şi tăceri semnificative (Ionescu-Ruxăndoiu, 1999, 36) nu este uşor de aplicat (v. Orletti / Testa, 1991, 273). Soluţia propusă de Jefferson (1978, xiii) pentru asemenea situaţii este, în opinia noastră, preferabilă, întrucât limitează interpretările din etapa transcrierii.

6.6. În ceea ce priveşte comentariile cercetătorului (glosări, observaţii privind înregistrarea etc.), dar şi marcarea întreruperii pasajului transcris, considerăm că soluţia folosirii barelor oblice */text/* este preferabilă aceleia de a combina tipuri de paranteze: ([, {[etc. De asemenea, folosirea notaţiilor specializate care includ litere *sic!*, *AK*, *K* ar putea fi izolată cu acelaşi tip de semne, */text/*, permiţând o lectură mai uşoară, dar şi excluderea lor, în funcţie de interesele celui care utilizează transcrierea. Menţionăm că există şi alte aşa-numite „notaţii specializate", pentru fenomene precum *false start,* semnul ⊥, sau *latching*, notat cu =, dar simbolurile nu sunt caractere alfanumerice (litere sau cifre) şi 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 convenţia din ortografia standard (majuscula). În *IVR* numele proprii nu sunt marcate. Pe de-o parte, transcrierea urmăreşte redarea rostirii şi din acest punct de vedere nu se justifică simboluri suplimentare pentru semnalarea numelor proprii. În plus, convenţia din ortografia curentă se suprapune cu notarea emfazei (care se face folosind majusculele), ducând uneori la ambiguitate, în cazul vocalelor iniţiale (v. *procesul de integrare-n Uniunea EuroPEAnă, CORV*, 228). Pe de altă parte, nemarcarea numelor proprii poate crea dificultăţi în înţelegerea textului. Un exemplu ar fi secvenţa: *o să văd codru* (*IVR*, 177), în care nu este vorba de o excursie în pădure, ci de o persoană („O să văd, Codru[ţa]".), fapt care reiese din lectura atentă a textului transcris: *nu ştiu codru↓ oricum↓ mai mă hotărăsc↓ şi: ((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 cunoştinţe comune locutorului şi interlocutorului (Bidu-Vrănceanu *et al.*, 2001, 415). Este vorba, în acest caz, de alt act de comunicare decât dialogul transcris; emiţătorul este autorul transcrierii, iar receptorul este cel care citeşte şi, eventual, utilizează transcrierile. Aşadar este greu de anticipat care dintre informaţii sunt cunoscute, mai ales atunci când nu este un antroponim, ci un titlu de lucrare, numele unei instituţii 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 informaţiei din comunicarea orală.

Din punct de vedere gramatical, clasa numelor proprii prezintă o serie de particularităţi, 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 consistenţă internă a sistemului de transcriere, ci cu încadrarea între bare oblice (*backslash*) a numelui:

CE legătură avem noi cu \uşa 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 şirurile de caractere (şi, 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 şirul de caractere dorit. Nu este suficient să copiem exact secvența grafică din text şi 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 şi această variantă în funcția de căutare apelăm la secvența *?@*, unde semnul *?* înseamnă orice caracter (unul şi 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 totuşi în şirurile 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	[ ] şi IPA	rostirea cuvintelor străine	*['ædvătaiziŋ]*	☐ precizia notației ☐ valorificarea unei convenții anterioare şi de largă circulație
	Tip de paranteze	Clasă de fenomene	Exemple	Avantaje
2	{ }	suprapuneri (eventual şi întreruperi)	***A;*** *unde# {am fost 1} eu vara trecută.* ***B;*** *{ai fost 1}*	☐ precizia notației ☐ eliminarea dificultăților tehnice în transferul de date ☐ prelucrare computerizată

				eficientă
3	< >	secvență grafică corespunzătoare rostirii: (1)mărci paradiscursive, (2) secvență neclară şi transcriere incertă şi (3) secvență indescifrabilă (fiecare **x** corespunde unei silabe rostite)	*(1) \transil<ʃvaniaʃ>* *(2) <ʔacoloʔ>* *(3) <xx>*	☐ păstrarea unității grafice a cuvântului ☐ precizia notației în cazul marcării a două sau mai multe mărci pentru aceeaşi secvență rostită ☐ posibilitatea realizării unor statistici computerizate ☐ posibilitatea selectării automate a uneia sau mai multor secvențe marcate ☐ posibilitatea eliminării automate a parantezelor de acest tip, păstrâdu-se doar textul corespunzător rostirii. ☐ înlesnirea lecturii
4	( )	elemente paraverbale: (1) descrierea în cuvinte a fenomenului şi (2) convenții pentru fenomenele mai frecvente: (@@), (H), (Hx) etc.	*(1)* ***A;*** *cred că noi (tuşeşte)* *(2)* ***A;*** *(H) domnule \pleşu\↓*	☐ posibilitatea eliminării automate a notațiilor respective dacă nu corespund obiectivelor utilizatorului ☐ posibilitatea realizării unor statistici ☐ înlesnirea lecturii
5	(( ))	elemente nonverbale; tăcerea, cu #((durata în secunde))	***A;*** *((se apropie de microfon))* *<Fstimați colegiF> #((3s))*	☐ posibilitatea eliminării automate a notațiilor respective ☐ posibilitatea realizării unor statistici ☐ înlesnirea lecturii
6	/ /	metatranscriere: (1) comentariile cercetătorului, (2) notații specializate: /K/, /AK/, /sic!/, (3) întreruperea pasajului transcris	*(1)* ***A;*** *convietuirea a fost posibilă /sună telefonul/ /.../* ***A;*** *regele lor \ştefan\ i-a creştinat. /14,5 sec./* *(2)* ***A;*** *v-ar place (sic!)*	☐ posibilitatea eliminării automate a notațiilor respective ☐ posibilitatea realizării unor statistici ☐ înlesnirea lecturii
7	\ \	marcarea numelor proprii	***A;*** *CE legătură avem noi cu \uşa interzisă\.*	☐ notarea unei informații importante fără a periclita consistența internă a sistemului (v. utilizarea majusculelor 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 şirul 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 *\1,* adică orice expresie rezultată în urma căutării automate va fi înlocuită cu ea însăşi) şi excluderea unor fenomene care nu prezintă interes pentru o anumită cercetare bazată pe transcriere (se tastează un blanc în *Replace with*), cum ar fi elementele nonerbale, spre exemplu.

7.1. Pentru a căuta în text numai cuvintele străine, notate cu IPA, folosim secvența *\[?@\]*.

7.2. Suprapunerile şi 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 \<*ŞOP?@ŞOP*\>. Secvența *ŞOP* 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 şi 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 *\/?@\/*, notațiile specializate cu *\/K\/*, *\/AK\/* şi, respectiv, *\/sic\!\/*, iar întreruperea pasajului transcris cu *\/...\/*.

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 şi eficient, inclusiv în domeniul ştiințelor umaniste, este un loc comun astăzi. Fişarea materialului după parametri bine stabiliți (eventual corelați) şi realizarea statisticilor pot prelua deja o parte migăloasă şi consumatoare de timp din munca specialiştilor. 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 – maşină.

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 priveşte 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 *(tuşeşte)*, 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 <*F*text*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}* şi 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*, Bucureşti, Editura Universităţii din Bucureşti, 2002, p. 337-348.
O'CONNELL, Daniel C. şi Sabine KOWAL, „Some Current Transcription Systems for Spoken Discourse: A Critical Analysis", în *Pragmatics*, 1994, 4, p. 81-107.
ORLETTI, Franca şi Renata TESTA 1991. „La transcrizione 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 intrigued 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 şi Unicode, standarde de reprezentare a informaţiei textuale, permit transferul datelor în computer, indiferent de platformă. Prin platformă se înţelege orice combinaţie posibilă de sisteme de operare (cum ar

fi Windows 98, Windows 2000, Linux, Mac-OS etc.) şi tipul de computer (IBM-PC, Macintosh etc.). ASCII are un inventar de 256 (2^8) unități. 128 dintre acestea (ASCII redus) codifică alfabetul englez şi un set limitat de semne de punctuație: a) valorile numerice cuprinse în intervalul 0-31 şi 127 codifică semne non-grafice (cum ar fi trecerea pe un rând nou, de pildă), b) 32 - pauza dintre cuvinte sau blancul şi c) valorile de la 33 la 126 codifică semne grafice: semnele de punctuație, cifrele şi literele (minuscule şi 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 şi fiecare simbol are o valoare numerică unică (deci poate fi transferat şi recunoscut de la o platformă la alta, dar şi 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, şi diacriticele), inclusiv ideografe. Toate simbolurile incluse de *Microsoft Word* (folosind comanda *Insert*, opțiunea *Symbol* şi 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-VRĂNCEANU, Angela, Cristina CĂLĂRAŞU, Liliana IONESCU-RUXĂNDOIU, Mihaela Mancaş, Gabriela PANĂ DINDELEGAN, *Dicționar de ştiințe ale limbii,* Bucureşti, Nemira, 2001.

CRESTI, Emanuela, *Corpus di italiano parlato*. Vol. I, II, Firenze, 2000.

DASCĂLU JINGA, Laurența, *Corpus de română vorbită (CORV). Eşantioane*, Bucureşti, 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*, Bucureşti, Editura Alcris, 1999.

HUANG, Xuedong, Alexandro ACERO şi Hsiao-Wuen HON, *Speech Processing*, www.clsp.jhu.edu/courses/zilla, 2000.

IONESCU-RUXĂNDOIU, Liliana, *Conversația: structuri şi strategii. Sugestii pentru o pragmatică a românei vorbite,* ediția a II-a, Bucureşti, ALL, 1999.

IONESCU-RUXĂNDOIU, Liliana (coord.) *Interacțiunea verbală în limba română actuală. Corpus (selectiv). Schiță de tipologie*, Bucureşti, Editura Universității din Bucureşti, 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.