Un corpus col·loquial i dialectal del valencià: PARLARS

Sandra Montserrat (Alacant)
Carles Segura (Alacant)

Summary: The corpus PARLARS aims to document the colloquial not interfered Catalan (Beltran / Segura, 2017) before it disappears, owing to both the pressure of standard and especially linguistic contact with Spanish (Segura, 2003). It also wants to offer researchers adequate materials to develop studies of description and analysis of the linguistic variation of Catalan, especially the one related to functional (colloquial) and dialectal varieties. Finally, using data provided by the corpus, we try to test the hypothesis that languages function as conventionalized structures, as explained by Cognitive Construction Grammar (Goldberg, 2003; Taylor, 2012; Hilpert, 2014). This article presents the first phase of the construction of this corpus which consists precisely of defining its main characteristics, the working methodology and the transcription system, as well as the results obtained so far.
Keywords: corpus, colloquial variety, variation, Catalan, Valencian, synchrony, Cognitive Construction Grammar

Received: 17-07-2020 · Accepted: 04-08-2020

1 Introducció

El català disposa de corpus lingüístics de tipologia diversa que han anat desenvolupant-se en els darrers anys, especialment dedicats a la llengua escrita. En primer lloc, cal destacar el *Corpus textual informatitzat de la llengua catalana* (CTILC) de l'Institut d'Estudis Catalans (IEC), un corpus de referència de la llengua contemporània (1833–2018), amb més de 52 milions de paraules. A més, existeixen diversos projectes de corpus amb finalitats molt específiques, com ara el *Corpus tècnic de l'IULA*, un corpus multilingüe amb textos d'especialitat cientificotècnica, elaborat per l'Institut de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra, o alguns corpus amb anotacions sintàctiques, com ara AnCora (Taulé *et al.*, 2008), de textos periodístics del català i de l'espanyol dissenyat a la Universitat de Barcelo-

ISSN 0932-2221 · eISSN 2199-7276
https://doi.org/10.46586/ZfK.2020.9-44

na, o el CuCWeb (Badia i Boleda, 2008), compilat automàticament a partir de webs, també de la Universitat Pompeu Fabra.[1]

En l'àmbit de l'oralitat també comptem amb diversos recursos. En primer lloc, cal destacar el *Corpus del Català Contemporani de la Universitat de Barcelona* (CCCUB) (Alturo *et al*., 2004; Boix-Fuster *et al*., 2007; Carrera-Sabaté i Vilaplana, s.a.; Payrató i Alturo, 2002; Pons i Vilaplana, 2009; Vilaplana i Perea, 2003; Vilaplana *et al*., 2007).[2] Aquest corpus està format per diversos subcorpus que abracen tots els eixos de variació en la nostra llengua. Així, el *Corpus oral dialectal* (COD) recull textos orals de tots els dialectes catalans de tipologia narrativa amb enregistraments ordinaris (no secrets) i entrevistes semidirigides amb la participació de l'investigador; el *Corpus oral de conversa col·loquial* (COC) està format per converses lliures de parlants de Barcelona i sense observació participant d'investigadors; el *Corpus de varietats socials* (COS) té textos orals de parlants de classe treballadora de diferents franges d'edat; i, finalment, el *Corpus oral de registres* (COR) que inclou converses i intervencions orals de diversos àmbits públics, privats i professionals.

En segon lloc, en fase primerenca d'elaboració, hi ha el projecte *Corpus oral de la llengua catalana* (COLC) de l'IEC que pretén la recollida, el tractament i l'anàlisi de mostres significatives de la llengua catalana parlada en totes les varietats i tots els dialectes majors de la llengua catalana. Pel fet de ser un projecte d'infraestructura normativa, cerca trobar anivellaments dialectals i tendències d'estandardització espontània i, per això mateix, en aquesta primera fase, ha començat a arreplegar materials de parla formal no espontània.

El CCCUB és, doncs, el referent més important del projecte del corpus PARLARS,[3] que presentem en aquest article, especialment el COD i el

1 Pel que fa al català en diacronia, diposem des de fa temps de diversos corpus de referència: el *Corpus informatitzat del català antic* (CICA), el *Corpus Informatitzat de la Gramàtica del Català Antic* (CIGCA) i, en fase d'elaboració, el *Corpus Informatitzat de la Gramàtica del Català Modern* (CIGCMod).

2 També tenim altres recopilacions per a l'estudi de l'oralitat com l'*Arxiu audiovisual dels dialectes catalans de les Illes Balears* o *l'Atles interactiu de l'entonació del català* (Prieto i Cabré 2007–2012) que té per objectiu l'estudi de l'entonació. Cal esmentar també el *Museu de la Paraula*, que, tot i que no té com a objectiu l'estudi lingüístic, pretén conservar el testimoniatge oral antropològic valencià.

3 El corpus PARLARS està elaborat en el marc del projecte «CorDiVal: Elaboració d'un corpus oral dialectal del valencià col·loquial» (GV/2017/094), finançat per la Generalitat Valenciana (<https://www.uv.es/corvalc/>), sota la direcció del Dr. Andreu Sentí, del Departament de Filologia Catalana de la Universitat de València. Hi participen especialistes del Departament de Filologia Catalana (Dr. Vicent Beltran, Dra. Maribel

COC, ja que, com veurem més avant (vg. 2.1), ens proposem construir un corpus dialectal i col·loquial. No obstant això, hi ha diversos motius que justifiquen la necessitat d'elaboració d'un corpus com PARLARS. En primer lloc, el CCCUB no està integrat en una interfície de consulta que facilite el treball al lingüista de corpus, sinó que els textos estan publicats en paper i/o en versió en PDF en el web (<http://www.ub.edu/cccub/>). Únicament, el COD té una interfície, que pren el nom de DialCat consultable en <http://stel.ub.edu/dialcat/> (Duran *et al.*, 2007) amb anotació morfològica i lematització. En segon lloc, el COD és poc representatiu de les varietats valencianes i, de més a més, no és prou extens per a estudis de tipus sintàctic, semàntic o pragmàtic. Finalment, cal assenyalar que el corpus col·loquial (COC) és molt reduït territorialment, ja que com hem assenyalat adés, només incorpora informants de Barcelona: no s'ocupa dels registres col·loquials conversacionals del català parlat al País Valencià.

Comptat i debatut, no existeix un corpus textual de grans dimensions, prou representatiu per a poder estudiar les varietats valencianes, tant la variació dialectal com la funcional (col·loquial). Així mateix, també és necessari un corpus que facilite l'estudi de la morfosintaxi, la semàntica i la pragmàtica de la llengua oral col·loquial (Fernández Ordóñez, 2011; Llop i Pineda, 2017), anotat morfosintàcticament i amb una bona interfície de consulta web. Provarem d'omplir aquest buit amb el corpus PARLARS.

Guardiola, Dra. Sandra Montserrat i Dr. Carles Segura) i del Departament de Llenguatges i Sistemes Informàtics (Dr. Miquel Esplà-Gomis), de la Universitat d'Alacant. Tots ells combinen la recerca en diversos aspectes de la llengua catalana amb l'experiència d'elaboració i explotació de corpus lingüístics diversos. – Agraïm la faena feta als informants i, molt especialment, als col·laboradors locals i transcriptors (Elena Verdú, Marta Soriano, Ares Llop, Anna Paradís, Anna Pineda, Aina Torres, Senén Magraner i Betlem Pallardó). També, al Dr. Esteve Clua (UPF) i a la Dra. Maria-Rosa Lloret (UB), amb qui n'hem discutit els criteris de transcripció. Així mateix, als col·legues de *Val.Es.Co.* com el Dr. Antonio Briz (UV) i el Dr. Adrián Cabedo (UV) i del Corpus *Coser*, com ara, a la Dra. Inés Fernández-Ordóñez (UAM). Finalment, als companys i companys del Servei d'audiovisuals de la Facultat de Filologia, Traducció i Interpretació de la Universitat de València, i al servei de gestió del Departament Filologia Catalana UV i de la UA.

2 Presentació del corpus

2.1 Objectius i característiques del corpus

Els objectius del corpus PARLARS es poden sintetitzar en els punts següents. Pretenem:

a) Documentar el català col·loquial poc interferit (Beltran i Segura, 2017). Com és ben sabut, a causa de la pressió de l'espanyol, la llengua catalana canvia ràpidament (Boix i Vila, 1998; Pradilla, 2004). També, la pressió de la varietat estàndard hi influeix (Segura, 2003; Beltran, Baldaquí 2006 i Segura, 2017). Fet i fet, doncs, volem fer la darrera fotografia –segurament– d'aquesta varietat genuïna de la llengua en totes les comarques valencianes, abans que desaparega per complet.
b) Aconseguir una representació massiva de tots els subdialectes del català parlat al País Valencià i a la comarca murciana del Carxe (Beltran i Segura, 2017), tal com els descriurem en el punt 2.2.
c) Oferir als investigadors materials adequats per a desenvolupar estudis de descripció i anàlisi de la variació lingüística del català oral, col·loquial i dialectal. Per tant, és fonamental dissenyar un corpus anotat morfosintàcticament, consultable en un web de manera unificada.
d) Elaborar un subcorpus de textos orals col·loquials mediatitzats per ordinador (WhatsApp, per exemple).
e) Unificar altres materials de converses orals preexistents però que no s'han constituït com a corpus i, fins i tot, incorporar textos d'altres corpus en el futur.
f) Avançar en la comprovació de la hipòtesi de la Gramàtica de Construccions Cognitiva que entén que les llengües funcionen com a estructures convencionalitzades (Goldberg, 2003; Taylor, 2012; Hilpert, 2014).

Per a aconseguir aquests objectius, PARLARS haurà de reunir tot un seguit de característiques: (1) la primera fase en l'arreplega de dades se centrarà en els entorns rurals o ciutats petites, així com en informants d'edats avançades (majors de 60 anys) per a assegurar que podem documentar la llengua en l'estat menys interferit possible. A mesura que avance el projecte s'hi incorporaran la resta d'informants; (2) els tipus de textos orals que documentarem seran col·loquials, en altres paraules, converses espontànies, sense intervenció de l'investigador, o almenys, amb la menor possible; (3) el corpus estarà anotat morfològicament i formarà part d'un sol repositori que permetrà cerques diverses (paraules, fragments de paraules, lemes, combinacions de paraules; comarca i població; característiques sociològi-

ques dels informants com ara edat, sexe, formació, etc.); (4) les eines informàtiques permetran que, en projectes futurs, es puga avançar en la revisió de l'anotació i en anotacions més desenvolupades, com ara a fonètiques, sintàctiques, semàntiques i pragmàtiques; finalment, els textos orals recopilats i les anotacions es difondran sota llicencia tipus *Creative Commons* amb atribució. Això permetrà que altres centres d'investigació puguen continuar la recerca, completar-la i ampliar-ne els recursos, així com també afinar-ne els etiquetatges per a finalitats específiques.

Explicarem al detall aquestes característiques en els apartats següents.

2.2 El registre col·loquial

El corpus que presentem vol documentar i caracteritzar el registre col·loquial dels dialectes valencians. Com és sabut, la variació diafàsica o funcional depèn de la situació comunicativa, mentre que la variació diatòpica, de l'usuari. En aquest sentit, els textos que pretenem recopilar seran una intersecció entre les varietats diatòpiques del valencià (vg. 2.3) en situacions d'informalitat i quotidianitat, és a dir, en la seva modalitat més corrent i primària: els registres col·loquials. Això ens permetrà entrar en contacte amb els textos més dialectals, ja que és en les situacions informals on apareixen naturalment els trets dialectals (Bibiloni, 1997). Per aquest motiu, deixem de costat situacions formals, professionals o presentacions orals en públic.

Tant la caracterització del registre col·loquial en el conjunt de la variació diafàsica com la del gènere conversacional en la variació textual ja han estat ben establertes. Ens basem en la caracterització del grup Val.Es.Co (Briz, 2002 i 2010)[4] per a definir el segment de variació que volem arreplegar en el corpus PARLARS (cf. Figura 1). Aquesta caracterització aplica la Teoria del Prototip de la Lingüística Cognitiva (Geeraerts, 1997) per a explicar els trets col·loquials i formals dels textos en termes de prototipicitat, conformant, per tant, un contínuum. Així, el col·loquial prototípic té les característiques nuclears següents: (a) relació d'igualtat entre els interlocutors, (b) relació vivencial de proximitat entre els interlocutors, (c) marc d'interacció familiar, (d) quotidianitat temàtica, (e) planificació sobre la marxa, (f) fi interpersonal i (g) to informal (Briz 2010).

4 Podeu consultar Briz (2010) per a aprofundir sobre la distinció conceptual i terminològica entre la col·loquialitat (registre dins de la variació diafàsica) i la conversa (gènere discursiu dins de la variació textual).

Variació funcional	Variació textual	
Registre col·loquial prototípic: + Relació d'igualtat entre interlocutors + Relació vivencial de proximitat entre els interlocutors + Marc d'interacció familiar + Quotidianitat temàtica + Planificació sobre la marxa + Fi interpersonal + To informal	Converses prototípiques	+ oral + immediat + dialogal + retroalimentació + cooperatiu + dinàmic + alternança de torns no per defecte + secreta presencial
	Converses no prototípiques (semidirigida)	+ oral + immediat + dialogal + dinàmic presencial
	Monòleg (seqüències narratives)	+ oral + immediat presencial
	Discurs oral mediatitzat per ordinador (converses no prototípiques)	+ oral no presencial

Figura 1. Variació, diafàsica i textual del català en el corpus PARLARS

L'oralitat col·loquial (o informal) es pot manifestar en diversos gèneres discursius (variació textual). La conversa espontània, secreta, com ara la que es pot donar al carrer entre dos coneguts que es troben, és, certament, la prototípica i, per tant, la més desitjable en un corpus que representa l'oralitat col·loquial. Tanmateix, aquest tipus de conversa és difícil de compilar. Per això mateix i atés que l'objectiu principal del corpus no és l'estudi del funcionament de la conversa sinó l'anàlisi del col·loquial (i la variació dialectal), el corpus PARLARS inclou, a banda de les converses prototípiques, quan siga possible, altres tipus de converses no prototípiques: concretament, converses semidirigides i monòlegs (vg. 3.1.3).

Les converses semidirigides estaran guiades per un investigador-entrevistador que conduirà els informants i, en conseqüència, seran menys espontànies (tot i que, a voltes, s'intentarà que també siguen secretes). Se'ls proposaran temes controvertits que inciten a l'argumentació. Aquest tipus d'entrevista manté alguns dels trets prototípics de la conversa (oral, immediata, dinàmica i dialogal), però, segons el funcionament de la conversa, pot haver-hi canvis en els trets de retroalimentació i cooperació entre els interlocutors i l'alternança de torns.

El corpus incorporarà també monòlegs: l'entrevistador demanarà a l'informant que conte vivències perquè construïsca un discurs essencialment narratiu al voltant de temes locals, del seu entorn o la seua vida. Encara que el diàleg s'associa de forma més directa amb la col·loquialitat que no el monòleg (Briz, 2010), també existeixen monòlegs col·loquials que poden ser una font d'informació per a completar les tipologies textuals que conformen el corpus i l'aparició de fenòmens lingüístics interessants. Fet i fet, doncs, en aquestes seqüències narratives podem trobar alguns trets de la conversa (oral, immediat, dinàmic) tot i que se n'alteren d'altres. D'una banda, no hi ha diàleg ni retroalimentació i cooperació. De més a més, tampoc hi ha alternança de torns lliure: aquests estan predeterminats per la figura de l'entrevistador i l'entrevistat.

Finalment, el corpus inclourà un tipus de textos orals que resulten de l'aparició i l'ús de les noves tecnologies: les converses orals mediatitzades per ordinador (missatges de veu de WhatsApp, per exemple). Aquest tipus de converses mantenen la característica essencial del gènere, que sigui oral, però la resta de trets canvien (no és tan immediat ni tan dinàmic, les intervencions són més llargues i no es retroalimenta ni coopera de la mateixa manera). A més, hi ha diferents graus d'immediatesa en aquest tipus de textos (des de notes de veu esporàdiques a diàlegs immediats per ordinador passant per converses amb notes de veu amb certa alternança de torns).

Ateses les característiques dels quatre tipus textuals que incorpora PARLARS, el nucli del nostre corpus estarà format per la conversa prototípica i la conversa semidirigida, que duraran entre 45 i 70 minuts: no debades, com hem explicat, aquests dos tipus textuals reuneixen més característiques del registre col·loquial prototípic.

2.3 Variació dialectal

Com hem indicat adés (vg. 2.1), el segon objectiu del corpus PARLARS és el d'aconseguir enregistrar una bona representació de la variació dialectal del

valencià i, per tant, servir de font per als estudis dialectals posteriors. Per decidir quina ha de ser la representativitat d'aquestes parles en el nostre corpus, hem partit dels estudis de dialectologia més recents. Així doncs, seguim la nova proposta de divisió dialectal de Beltran i Segura (2017: 31–32), que consisteix a establir cinc subdialectes: el valencià tortosí, el valencià septentrional, el valencià central, el valencià meridional i el valencià alacantí. En concret:

a) *El valencià tortosí* té trets propis de la varietat tortosina del català. Sumàriament, abraça les comarques del Baix Maestrat, l'Alt Maestrat, els Ports i les zones septentrionals de la Plana Alta i l'Alcalatén. Es pot considerar àrea marginal en el si dels parlars valencians.
b) *El valencià septentrional* té característiques que es desprenen de les varietats tortosines i d'altres que provenen de la influència històrica de València. La resta de la Plana Alta i l'Alcalatén i la Plana Baixa són les àrees que s'hi poden adscriure.
c) *El valencià central* correspon a la varietat que ensordeix les sibilants sonores i es troba clarament sota la influència del Cap i Casal. En concret, s'hi lliguen les comarques del Baix Palància, el Camp de Túria, l'Horta, gran part de la Ribera Alta i una àrea de la Ribera Baixa.
d) *El valencià meridional* és la darrera zona al sud del país on amb claredat arriba la influència de València. S'hi mantenen trets especialment conservadors, sobretot de caràcter fonètic. La resta de la Ribera del Xúquer, la Costera, la Vall d'Albaida, la Safor, l'Alcoià, la Marina Baixa i part de l'Alacantí són les comarques que s'hi poden adscriure. S'hi inclou la varietat del *valencià mallorquí*, això és, uns bon feix de trets de caràcter oriental, superposats a aquesta varietat, que hi foren empeltats amb l'expulsió dels moriscos i la subsegüent repoblació mallorquina. Fonamentalment es concentren a la major part de la Marina Alta i part de la Marina Baixa i la Safor.
e) *El valencià alacantí* és l'àrea marginal meridional dels parlars valencians, on es conserven trets que corresponen a l'antiga varietat perduda de l'oriolà. S'hi mantenen trets poc habituals en la resta del territori, especialment arcaismes, però també castellanismes i formes potser d'origen remotament catalanooriental.

Seguint les indicacions que ens proporcionen aquests investigadors, a banda d'aquestes grans àrees o subvarietats dialectals, convé no oblidar els parlars fronterers o bé les parles derivades de situacions històriques especials. S'encabeixen en aquesta categoria parlars com ara els de Vistabella, Suera, Tales, Torís, la Canyada, el Carxe, Crevillent o Guardamar.

3 Metodologia de treball

3.1 Recopilació de dades

3.1.1 Elecció de localitats

- **Els Ports** (Cinctorres, Morella/Herbers)
- **El Baix Maestrat** (la Jana/Xert, Rossell/Alcalà de Xibert)
- **L'Alt Maestrat** (Benassal, Ares)
- **L'Alcalatén** (les Useres/l'Alcora/Costur, Atzeneta)
- **La Plana Alta** (Cabanes, Albocàsser/Tírig)
- **La Plana Baixa** (Borriana, Tales/Suera)
- **El Camp de Morvedre** (Algar, una població de les Valls)
- **El Camp de Túria** (Llíria/Benaguasil, Nàquera)
- **L'Horta** (Massanassa, Torrent, Museros/Foios, Alfara del P./Aldaia)
- **La Ribera Alta** (Torís, Benimuslem)
- **La Ribera Baixa** (Sueca/Cullera, Sollana)
- **La Costera** (la Font de la Figuera, Xàtiva)
- **La Vall d'Albaida** (Ontinyent, Benigànim)
- **La Safor** (Gandia/Oliva, una població de la Valldigna)
- **El Comtat** (Muro, Benillup, l'Orxa)
- **La Marina Alta** (Gata/Pedreguer, Benissa, Castells/Ebo/Gallinera/Laguar)
- **La Marina Baixa** (la Vila, Callosa, Tàrbena)
- **L'Alcoià** (Alcoi, Castalla)
- **L'Alt Vinalopó** (la Canyada)
- **L'Alacantí** (Mutxamel, Agost, Xixona)
- **Les Valls del Vinalopó** (Monòver, Novelda, el Carxe)
- **El Baix Vinalopó** (Elx, Crevillent)
- **Baix Segura** (Guardamar)

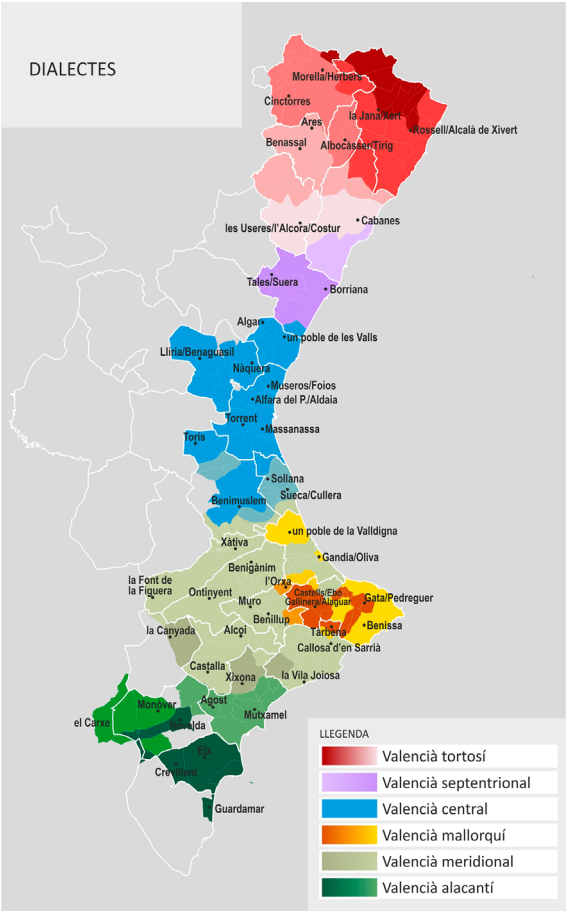

Figura 2. Pobles del País Valencià on s'arreplegaren dades per al corpus PARLARS

Cada comarca del País Valencià es veu representada per un, dos o tres pobles depenent de les característiques geogràfiques i lingüístiques de cada

comarca. Si seguim els criteris establerts per Beltran i Segura (2017), a banda de la representació dels cinc subdialectes que tot just acabem de descriure, fa falta representar models urbans i rurals, de mar i d'interior, i els casos particulars. Concretament, hem seleccionat 51 pobles de 23 comarques del País Valencià (cf. Figura 2 a la pàgina anterior).[5]

3.1.2 El perfil de l'informant

Hem dividit els informants en tres grups, atenent la cronologia dels enregistraments. En una primera fase, hem fet les entrevistes a majors de 60 anys. Durant la segona fase, enregistrarem parlants entre 35 i 60 anys i, finalment, durant la tercera fase, entrevistarem la població que no arribe als 35 anys d'edat. Seguint criteris habituals de la dialectologia, l'entrevistat o entrevistada ideal és qui té pare i mare, avi i àvia del mateix poble, que s'ha mogut poc del lloc on viu i amb poca formació. Tanmateix, això variarà a mesura que l'informant siga més jove.

Les dades sol·licitades a l'informant estan desplegades en una fitxa corresponent (cf. Figura 3 a la pàgina següent).

3.1.3 Modalitats d'entrevistes

El corpus PARLARS tindrà, bàsicament, tres modalitats de textos per a il·lustrar diferents tipologies del registre col·loquial, d'acord amb el que hem exposat en el punt 2.2: converses prototípiques, converses semidirigides i monòlegs.

En primer lloc, en les converses prototípiques secretes els interlocutors no saben que se'ls grava i l'investigador-entrevistador no forma part de la conversa.[6] La recollida d'aquest tipus de textos és l'objectiu principal del projecte, però presenta alguns problemes. En primer lloc, l'investigador encarregat de la gravació ha de ser una persona de confiança que puga estar amb els informants malgrat que no hi intervinga. En segon lloc, encara que siga secreta, cal demanar-ne permís previ: es pot demanar el permís unes

5 No estan representades les comarques de predomini lingüístic castellà. El símbol [/] indica que els pobles són equivalents quant a les característiques geogràfiques i lingüístiques i que s'ha enregistrat l'un o l'altre en relació amb la facilitat de trobar informants.

6 En algun cas, qui s'encarrega de l'enregistrament estarà present si forma part de l'entorn dels informants, però es mantindrà al marge de la conversa.

Fitxa de l'informant

Parlant:

Data de naixement:

Sexe: ☐ Home ☐ Dona

Població de naixement i comarca:

Població de naixement dels pares:

Població de residència i comarca:

Altres poblacions on ha viscut:

Parlar: ☐ v. tortosí ☐ v. septentrional ☐ v. central ☐ v. meridional
☐ v. mallorquinitzant ☐ v. meridional-alacantí

Grup social (professió):

Estudis: ☐ analfabet ☐ primaris (EGB)
☐ secundaris (ESO/BUP-COU/Batxillerat/FP)
☐ universitaris (especifica'ls: _____)

Coneixement de valencià: ☐ col·loquial i familiar ☐ escola i secundària
☐ títol EOI o JQCV o CIEACOVA (especifica'l: _____)
☐ altres: _____

Llengua familiar: ☐ valencià ☐ castellà ☐ una altra *(es pot fer més d'una creu)*

Llengua habitual: ☐ valencià ☐ castellà ☐ una altra *(es pot fer més d'una creu)*

Figura 3. Fitxa de l'informant

quantes setmanes abans sense que els entrevistats sàpiguen quan se'ls gravarà. D'altra banda, aquest tipus d'enregistrament secret comporta que el so no és de tanta qualitat atés que la gravadora està amagada i els interlocutors fan sorolls o es mouen. Finalment, el text resultant pot ser en moltes ocasions fragmentari.

La segona modalitat és la conversa semidirigida (no secreta), en la qual l'investigador està present, condueix la conversa amb dos o tres informants (fins i tot, quatre) i va introduint temes o demana l'opinió dels informants sobre aspectes que els puguen interessar. Intenta fer fluir la conversa i evita de participar-hi ni marcar-ne el rumb sempre que estiga avançant. D'aques-

ta manera, encara que no és una conversa totalment espontània, s'hi pot acostar, tal com hem descrit en el punt 2.2. Aquesta modalitat és més fàcil d'aconseguir i per això hem arreplegat mostres de converses de manera sistemàtica de totes les localitats descrites a 3.1.1, la qual cosa garanteix la representativitat diatòpica del corpus.

Figura 4. Entrevista semidirigida a Pinet (la Vall d'Albaida)

Figura 5. Entrevista semidirigida a Fleix (poble de la Marina Alta)

La tercera modalitat d'enregistrament és l'entrevista a un únic informant. En aquest cas, el text és un monòleg especialment narratiu. L'entrevistador demanarà a l'informant que narre històries personals o familiars, experiències, anècdotes, etc.

Figura 6. Monòleg a Cinctorres (els Ports)

Condicions de la gravació

Condicions de la gravació: ☐ conscient de la gravació ☐ no conscient

Tipus de conversa: ☐ espontània ☐ semidirigida ☐ monòleg

Nombre de participants:

Tipologia textual: ☐ narració ☐ exposició ☐ argumentació ☐ instrucció ☐ conversació

Temàtica (especifica-la)**:**

Durada:

Enregistrament: ☐ vídeo ☐ àudio

Marca i model de la gravadora i/o càmera:

Lloc de l'enregistrament:

Enquestador/a:

Data de l'enregistrament:

Codi de la conversa:

Observacions/Notes:

Figura 7. Fitxa que reflecteix les condicions de la gravació

La durada dels enregistraments oscil·la entre 45 i 70 minuts. Les converses es graven sempre amb gravadora i en alguns casos, en converses semidirigides i monòlegs, també hi ha un enregistrament en vídeo, quan el perfil de l'informant justifica l'interés antropològic de l'entrevista.

Com a complement a aquestes tres modalitats, elaborarem un subcorpus amb textos orals mediatitzats per ordinador, com per exemple notes de veu en missatgeria instantània. Aquesta arreplega tot just acaba de començar.

Cada enregistrament disposa d'una fitxa, amb les dades corresponents (cf. Figura 7, a la pàgina anterior).

3.2 Altres enregistraments

Com també hem indicat en el punt 2.1, a més del treball de camp, també ens proposem d'aplegar enregistraments de fons públics o privats d'interés per a l'estudi de la llengua oral i els objectius del projecte. Això ens permetrà concentrar el material oral al si del nostre projecte per tal de preservarlo i, també, perquè en el futur puga formar part del corpus PARLARS, si escau. En aquest sentit, fins al moment, hem documentat i aplegat –amb el permís dels autors i autores– una gran quantitat de gravacions d'arxius molt diferents. En acabar el treball de camp, intentarem que formen part del corpus, indicant-ne la procedència:

1) Textos orals de parlants valencians del corpus COD, subcorpus del CCCUB dirigit pel Dr. Esteve Clua (Universitat Pompeu Fabra), la Dra. Maria-Rosa Lloret (Universitat de Barcelona) i la Dra. Maria-Pilar Perea (Universitat de Barcelona).
2) Textos orals de parlants valencians del projecte *Atles interactiu de l'entonació del català* dirigit per la Dra. Pilar Prieto (Universitat Pompeu Fabra).
3) Arxiu personal del prof. Vicent-Josep Pérez Navarro. Està format de 14 monòlegs i tres converses gravades entre el 1997 i el 2018, a informants nascuts entre 1906 i 1942, nadius de les comarques del Baix Vinalopó, les Valls del Vinalopó, l'Alacantí i l'Alcoià.
4) Arxiu personal del Dr. Vicent Beltran (Universitat d'Alacant). Hem digitalitzat l'arxiu personal d'aquest dialectòleg i membre del projecte per tal de conservar les nombroses gravacions de què disposava.
5) Arxiu personal de la Dra. Maribel Guardiola (Universitat d'Alacant). Hem digitalitzat l'arxiu personal d'aquesta lexicòloga i membre del projecte per tal de conservar les nombroses gravacions de què disposava.

6) Arxiu personal del Dr. Miquel Àngel Pradilla (Universitat de Rovira i Virgili) que va usar en l'elaboració de la seua tesi doctoral (Pradilla, 1993) sobre Benicarló.
7) Arxiu oral de Benassal (Barreda i Ferrando, 2007), conservat al museu de Benassal.
8) Arxiu oral de les Useres, del cronista Josep Rubio Miguel (Beltran i Rubio, 2008).
9) Arxiu oral de l'Associació Pòrxens de Pedreguer. Hem seleccionat dues gravacions de monòlegs entre les gravacions d'aquest arxiu.

4 Transcripció

La transcripció «és un procediment de trasllat o transposició a una forma escrita d'unes dades que originalment s'han produït a través del canal oral» (Payrató, 2010: 208). Hi ha diversos tipus de transcripció: la transcripció pròpiament dita i la transliteració ortogràfica (Hidalgo i Sanmartín, 2005; Bladas, 2009). La transcripció és un model més detallat que pretén reflectir la llengua oral sense perdre gaire informació o el mínim possible, com ara la transcripció fonètica amb l'Alfabet Fonètic Internacional (AFI) o altres models amb símbols prosòdics sobre pauses, encavalcaments, gestualitats entre altres aspectes vocals (vg. COC; Alturo i Payrató, 2002; Val.Es.Co; Briz, 2002; Bladas, 2009). La transliteració, en canvi, és un model que usa les normes ortogràfiques convencionals i, per tant, perd part de la informació com ara elements prosòdics (vg. la transcripció fonoortogràfica del COD, Viaplana i Perea, 2003; *Atles entonatiu del català*, Prieto & Cabré, 2007–2012; el corpus COSER, De Benito, Pueyo i Fernández-Ordóñez, 2016; o el *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico* [MC-NLCH], Samper, Hernández i Troya, 1998).

Els criteris de transcripció que presentem s'ajusten als objectius del corpus PARLARS, és a dir, un corpus orientat a l'estudi de la variació dialectal i diafàsica de la llengua oral. També s'han tingut en compte els criteris d'altres projectes anteriors, especialment els del CCCUB (Payrató i Alturo, 2002; Viaplana et al., 2003; Viaplana i Perea, 2003; Foix-Fuster *et al.*, 2007; Alturo *et al.*, 2009; Carrera-Sabaté i Viaplana, s.a.; Pons i Viaplana, 2009; Blesa, 2009; Payrató, 2010) i els de l'*Atles interactiu de l'entonació del català* (Prieto i Cabré, 2007–2012). Finalment, cal tindre en compte que ens hem basat en les característiques lingüístiques dialectals generals (Veny, 1982; Solà *et al.* 2002; Beltran i Segura, 2017).

Val a dir que l'establiment dels criteris de transcripció definitius ha experimentat dues fases. En un primer moment, vam decidir que el corpus tindria una transcripció que reflectís al màxim possible les característiques dialectals i lingüístiques del discurs alhora que estandarditzés alguns aspectes fonètics per facilitar les tasques d'anotació (Beltran *et al.*, 2019a). Vam aplicar aquests criteris a un text de manera experimental per valorar-ne l'eficàcia quant al processament de dades. En aquesta versió inicial vam optar per una estratègia de doble anotació i alineació text-veu per parlant. Aquesta solució permetia disposar d'una transcripció molt pròxima a la pronunciació real –útil per a observar la producció oral– i una altra transcripció ortogràfica estandarditzada més pròxima a la llengua escrita codificada –útil per a l'anotació del corpus. Tanmateix, vam trobar algunes dificultats tecnològiques a l'hora d'alinear *token* a *token*[7] les dues transcripcions, per la qual cosa vam desistir d'aquest model (vg. Esplà-Gomis i Sentí, en preparació).

La segona proposta opta per una única transcripció. Abans de definir-ne els nous criteris, vam tokenitzar i lematitzar un text de prova seguint els criteris anteriors i vam observar quines formes no eren identificades automàticament pel lematitzador automàtic d'*Apertium* (Forcada *et al.*, 2011). D'acord amb aquesta informació, vam afinar els criteris de transcripció amb la intenció que el lematitzador automàtic pogués identificar el màxim de formes possibles, ja siga canviant el criteri de transcripció o bé pensant quins canvis introduiríem en el lematitzador. En definitiva, els nous criteris responen a la voluntat de facilitar la tasca de lematització i, per això, no reflecteixen la majoria dels fenòmens fonètics.

Comptat i debatut, els criteris de transcripció definitius, com veurem, responen a un model més pròxim a la llengua codificada ortografiada en els aspectes fonètics (transliteració), però que preserva els trets morfològics, sintàctics i lèxics propis de la conversa o monòleg (transcripció) (Beltran *et al.*, 2019b).

7 Un *token* és una cadena de caràcters i xifres que estan tancades entre espais (Procházková, 2006: 2). Aquest anglicisme equival a *ocurrència*, *cas*, *paraula* o *mot*. Usarem l'anglicisme quan ens referirem a la unitat que es deriva del procés pel qual se separen automàticament les ocurrències d'un text oral, usant *Apertium*. Denominarem aquest procés *tokenitzar*.

4.1 Criteris lingüístics de la transcripció

Com hem dit, els criteris de transcripció que proposem pretenen mostrar la realització dialectal del text oral utilitzant les convencions de l'ortografia catalana. Els aspectes fonètics no seran sempre representats en la transcripció, però sí que hi haurà molta fidelitat a les característiques morfosintàctiques.

4.1.1 Fonètica

Els criteris de transcripció intenten acostar-se a la fonètica real, però no pretenem fer una transcripció fonètica rigorosa, així com tampoc no usarem l'Alfabet Fonètic Internacional (AFI). Com a criteri general, la transcripció representa els fenòmens fonètics que impliquen un canvi sil·làbic (*vesprà*) i, de vegades, una elisió o addició d'un so (*prèmit*), però no les neutralitzacions vocàliques, l'elevació lingual o l'obertura de les vocals (no transcriurem, per exemple, les pronúncies *còsa*, *pén* o *què*), ni l'ensordiment, el ieisme o el betacisme. Vegem els criteris dels principals fenòmens.

La transcripció ha de representar alguns fenòmens fonètics que impliquen elisions o canvis molt significatius. Per tant, marcarem els fenòmens següents tal com els pronuncie el parlant:

a) En català occidental hi ha alguns mots, que provenen del llatí amb una Ē, que han evolucionat a una [e] mentre que en català oriental tenen una [ɛ] o [ə]. Quan aquests mots tenen aquesta vocal en posició que ha de ser accentuada, optarem per l'accentuació que reflecteix la pronúncia occidental: *francés* i no *francès*. Els contextos més habituals són: terminacions *-és* de gentilicis (*anglés*), participis (*admés*, *compromés*) i adjectius (*cortés*); numerals ordinals acabats en *-é* (*cinquè*) i alguns substantius (*cafè*); la terminació *-én* de la tercera persona del plural del present d'indicatiu d'alguns verbs de la II conjugació (*aprén, comprén, depén*); els infinitius acabats en *-éixer* (*conéixer*) i *-éncer* (*convéncer*); la segona i tercera persona del plural dels imperfets d'indicatiu amb accent al radical (*féiem, féieu*). Ara bé, escriurem els mots *què*, *perquè*, *València*, *sèrie*, *època*, entre d'altres, amb accent greu, independentment de la pronúncia del parlant, tal com estipula la norma ortogràfica.
b) Canvi de síl·laba accentuada. Transcriurem la pronúncia esdrúixola de mots com *cànvie* 'canvie', *òdie* 'odie', etc.

c) La pronúncia castellanitzant amb /u/ tònica de paraules com *pluma* (i no escriurem *ploma*). La pronúncia castellanitzant amb /o/ tònica o àtona de paraules com *montar* o *monta* (i no escriurem *muntar* o *munta*).
d) La caiguda de l'auxiliar de perfet *haver*: *li dit* 'li he dit', *li's fet* 'li has fet', *li'm dit* 'li hem dit'.
e) La preposició *a* prendrà les formes ortogràfiques *a*, *ad*, *an* segons com la pronuncie el parlant: *Li ho donaré ad ell.*
f) La caiguda de la *d* o *g* quan suposen una reducció vocàlica: *vesprà* 'vesprada', *Nal* 'Nadal', *palar* 'paladar'.
g) En general, només marcarem les caigudes de vocals en paraules gramaticals com els demostratius. Per tant, transcriurem *ixe* 'eixe'. No transcriurem, però, la caiguda de la *a* en mots com *allà*, *allí*, *ahí* perquè aquesta elisió es deu a motius prosòdics i del context fonètic. Tampoc marcarem altres afèresis (caiguda de vocal inicial) perquè, tot i que és un fenomen oral habitual, hi ha molta variació fins i tot intraparlant: en lloc de *nar*, *via* o *metla*, escriurem *anar*, *havia* i *ametla*.
h) Fenòmens fonètics diversos com ara *quidrar/quirdar* 'cridar', *auia* 'aigua', *bragó* 'braó', *a von* 'a on'; vocals de suport com ara *vos e les duc*; transcriurem els casos de variants fonètiques incorporades en diccionaris normatius: *flare* 'frare', *almorzar* 'esmorzar'.
i) La contracció de la preposició *per (a)* que no coincidisca amb l'ortografia: *pa tu* ('per a tu'), *pa fer-ho* ('per a fer-ho'), *pa asfaltar* ('per a asfaltar'), *pa ahí* ('per ahí'). També transcriurem la reducció de *per (a)* a *pe*: *pe tu* ('per a tu'), *pe hí* ('per ahí').
j) La palatalització d'*haver-hi*: *no ny'ha* 'no n'hi ha'.

Hi ha una sèrie de trets propis de certs dialectes (o idiolectes) que no representem amb una grafia diferent de l'ortografia estàndard; és a dir, regularitzarem aquests casos d'acord amb la norma independentment de la realització del parlant per facilitar l'anotació posterior:

a) L'elevació de les vocals *e* i *o* quan porten accent gràfic. Seguirem l'ortografia: transcriurem *telèfon*, independentment de la pronúncia [teˈlɛfon] o [teˈlefon].
b) Les neutralitzacions de vocals *e/a* i *o/u* àtones: *esperar*, *calendari*, *llençol*, *Josep*, *Joan*, *joventut*, *cosir*. Tampoc marcarem la pronúncia amb /e/ de mots com *demunt* 'damunt', *trevessar* 'travessar'.
c) Les diverses realitzacions vocàliques de l'auxiliar *haver*: no transcriurem *Jo ha fet*, sinó *Jo he fet*.
d) L'harmonia vocàlica: *terre* 'terra'; *porto* 'porta'. Escriurem *terra* i *porta*.

e) El tancament d'una /e/ > /i/. No transcriurem *ginoll*, *sinyor*, sinó *genoll* i *senyor*.
f) Les formes dialectals amb *e* com ara *vengut*, *tengut*, *tendria*, etc. es transcriuran amb la forma estàndard *vingut*, *tingut*, *tindria*, etc.
g) Les diverses pronúncies possibles del mot *diumenge* (*dumenge*, *domenge*...).
a) La tonicitat o atonicitat de l'auxiliar de perfet: *ha fet* [ˈa ˈfet] o [a ˈfet].
i) Les diverses realitzacions fonètiques del pronom *ho* [o], [w], [ew] o [aw] (com ara *hau sé*, *heu sé*, *t'heu dic*, *heu agarre*, *hau agarre*, *[w] agarre*, *compra-[w]*, *comprar-[o]*): *ho sé*, *t'ho dic*, *ho agarre*, *compra-ho*, *comprar-ho*. Sempre transcriurem ortogràficament *ho*.
j) No indicarem la reducció de la vocal de suport de l'article definit [l] i, per tant, escriurem *ara els ulls* o *arribarà el dia*. També escriurem la *l* de l'article definit en plural encara que en la pronúncia caiga (*es troncs*): *els troncs*.
k) Com hem dit més amunt, en general no transcrivim l'elisió de les vocals inicials: en lloc de *llí*, *nar*, *via*, *metla* o *scola*, escriurem *allí*, *havia*, *ametla*, *escola*.
l) Quan es combina la preposició *de* i una paraula començada per *a*, sovint, en alguns parlars, hi ha una afèresi d'aqueixa *a* i, per tant, la preposició *de* no elideix la vocal: *de xò* 'd'això', *de llí* 'd'allí', *de nar* 'd'anar'. Transcriurem aquests casos amb *de* i sense representar l'afèresi: *de això*, *de allí*, *de anar*. Apliquem el mateix criteri en casos similars com *El peix es via de dur* 's'havia de dur', construcció que transcriurem així: *El peix es havia de dur*.
m) No indicarem les elisions per contacte vocàlic: *onze anys*, *una hora*, *la mateixa hora*, *sense ou*, *eixe home*…
n) No marcarem les sinalefes perquè és un fenomen molt general: *una amiga*.
o) Tampoc no marcarem les formes *sixanta*, *ixim*, *ixiu*... sinó que les transcriurem segons la norma *seixanta*, *eixim, eixiu*. Tampoc no escriurem *coranta*, sinó *quaranta*.
p) No marcarem la caiguda de la *d* intervocàlica o altres sons quan no implique una reducció sil·làbica, és a dir, la pronúncia de *maür, llauraor, juar*, *iual* o *aiua* la transcriurem ortogràficament *madur*, *llaurador*, *jugar*, *igual* o *aigua*. En canvi, si hi ha metàtesi, sí que es transcriu *auia* 'aigua'.
q) No anotarem i, doncs, usarem la forma normativa en elisions com *nessitar* 'necessitar', *tallaïns* 'tallarins', *carreó* 'carreró' o canvis fonètics com *ottubre* 'octubre', *moixca* o *moxca* 'mosca', *almari* o *asmari* 'armari', *ambercoc* o *asbercoc* 'albercoc'.

r) L'emmudiment de la *-r* final. Escriurem *dir*, *comprar*, *corredor*.
s) Emmudiment de la *l*: *atre* 'altre'.
t) L'emmudiment de la *r* en mots com *perdre*, *prendre* (i *sorprendre*, *mamprendre*, *emprendre*...) *arbre*, *marbre*, *dimarts*, *diners*, entre altres.
u) La pronunciació amb *n* de mots com *contar* 'comptar', *prensa* 'premsa', *pronte* 'prompte', etc. Per tant, els transcriurem d'acord amb la llengua normativa.
v) L'ensordiment de les sibilants sonores: *casa*, *dotze*, *metge*, *pluja*. Tampoc marcarem els canvis produïts en les sibilants per la fonètica sintàctica: *peix i carn*, *cases altes*.
w) L'ensordiment de la [z] del mot *zero* (no transcriurem *sero* sinó *zero*). El mateix per a altres mots que experimenten una sonorització de la [s] com *zenzillo* 'senzill', que transcriurem *senzillo*.
x) La sonorització en fonosintaxi de les oclusives sordes: *cin*[ɡ] *o sis*; *se*[ð] *o huit*. És a dir, escriurem *cinc o sis* i *set o huit*.
y) No marcarem la palatalització de la sibilant alveolar en contacte amb una velar: no escriurem *servixca*, *crexca* o *vixcut* sinó *servisca*, *cresca* i *viscut*.
z) La presència/absència de la iod davant de la palatal fricativa sorda i sonora (no transcriurem *caxa*, *baijoca*, *correija*, *puijar* o *meige* sinó *caixa*, *bajoca*, *corretja*, *pujar*, *metge*).
aa) Les aspiracions de la *s*: *és que*.
bb) El ieisme: *llavi*, *palla*, *pell*
cc) El betacisme: *vinc*, *vols*.
a) La distinció entre fricatives [ʒ] i africades [d͡ʒ]: *Jaume*.
ee) Les diferents realitzacions fonètiques del pronom *jo* ['jo], ['d͡ʒo]... Sempre transcriurem *jo*.
ff) La realització de les oclusives fricatives: *ceba*, *pagar*...
gg) El reforç de les consonants finals: [pontᵊ] es transcriurà *pont*.
hh) La caiguda de les oclusives finals: *pont*, *alt*, *camp*, *alts*, *malalts*, *quant*. En canvi, sí que marcarem l'afegit d'una *t* a la paraula *quan*: *Quant vindràs, demà?*
ii) La palatalització del grup *-ts* final: *acabats*, *tots*.
jj) Escriurem d'acord amb l'ortografia les paraules que contenen sons geminats: *col·legi*, *guatla*, *ratlla*, *setmana*.

4.1.2 Morfosintaxi

En general, transcrivim tots els trets morfosintàctics del discurs del parlant. Alguns exemples de morfologia nominal i verbal són:

a) L'article *lo*: *lo llibre*. També l'article neutre *lo*: *lo que dius*, *lo bonic que és*.
b) L'article salat: *sa llibreta*.
c) Transcriurem els demostratius *este*, *eixe* i *aquell* mantenint la forma no reforçada i reflectint la realització fonètica: *ixe xic*, *esta dona*. No reflectirem la reducció vocàlica *est' home*, que transcriurem *este home*.
d) La forma femenina del numeral *dos*: transcriurem *dos cases* i no ho modificarem per *dues cases*.
e) Les realitzacions morfològiques dialectals dels pronoms forts com *voatros* 'vosaltres', i dels pronoms febles com *mos*, *nos*, etc.: *me fa* 'em fa', *mos diu* 'ens diu', *se n'anem o mo n'anem* 'ens n'anem', *comprar-mo'n* 'comprar-nos-en', *compra-lo* 'compra'l'.
f) La preposició *amb* es transcriurà *en* d'acord amb la pronúncia: *en ell*, *en Maria*. No marcarem altres canvis fonètics que afecten aquesta preposició, com ara la neutralització de la *e* en *a* o l'assimilació de la nasal. Transcriurem doncs *Estic en Toni*, *Estic en Pere* i evitarem *Estic an Toni*, *Estic em Pere*.
g) La flexió verbal mostrarà la forma usada: *perc* 'perd', *vega* 'veja', *cantave* 'ell cantava', *porte* 'ell porta',
h) Independentment que les formes verbals siguen normatives o no, reflectirem la forma usada pel parlant. Per exemple, en alguns verbs de la II conjugació, els parlars valencians no tenen la [j] antihiàtica: *féem*, *caem*, *veeu*… També marcarem si el parlant pronuncia *veguem* o *vegem*.
i) El morfema de passat imperfet /va/ o /ve/ sense la *v*, *cantàem*, *cantàeu*, *cantaen*, *vàem*.
j) La terminació de verbs de la II conjugació com *víndrer* que en alguns dialectes per analogia poden acabar en *-er* en lloc de *-re*.
k) El sufix *-ea*: *vellea*.
l) Duplicació del clític de datiu: *Li vaig dir a ma mare*.
m) Altres pleonasmes: *N'hi havia dos cases*.
n) La preposició *a* davant complement directe: *He vist a ma mare al mercat*.
o) No corregirem la sintaxi dels parlants i, per tant, si un parlant no usa un pronom que en la llengua normativa seria obligatori, no l'afegirem: si un parlant diu, per exemple, *Volia un*, no ho canviarem per *En volia un*. Tampoc en el cas de formes lexicalitzades: si el parlant diu *Si havien persones*, no transcriurem *Si hi havien persones*.
p) La concordança del verb *haver-hi* amb el SN: *Hi han dos xiquets*.
q) No canviarem els mots normatius com *tindre*, *vindre*, *calfar*... per *tenir*, *venir*, *escalfar*...

4.1.3 Lèxic

Independentment de l'origen del mot o de les recomanacions prescriptives, escriurem tots els mots sense modificar-los: *mensatge* 'missatge'. Si són mots dialectals, utilitzarem com a referència el DCVB: *setiet* 'estalvis', *xicotiu* 'molt petit'. Si no apareixen en el DCVB (p. ex., *xicorrotiniu*) els transcriurem seguint els criteris fonètics i morfològics descrits.

Transcriurem els manlleus usant l'ortografia catalana, independentment de si ha estat normalitzat pel TERMCAT, el Porterval, els diccionaris normatius (DIEC, DNV) o el GDLC: *tetxo* (esp. *techo*), *unya* (esp. *uña*), *sumo* (esp. *zumo*), *calsonsillos* (esp. *calzoncillos*), *uàssap*, *uassap*, *uasap*, *uatzap*... (ang. *whatsapp*), *quesso* (esp. *queso*), *trage* (esp. *traje*), *entonces* (esp. *entonces*), *límpio* (esp. *limpio*), *tuit* (ang. *twit*). Les paraules més consolidades en la llengua col·loquial presenten els trets de la flexió i la derivació catalanes (*unyes*, *sumet*, *tetxar*, *quesset*). Aquests casos, doncs, els transcriurem amb l'ortografia catalana i els considerem manlleus adaptats.

En el cas de manlleus que tinguen sons que són aliens a la fonètica catalana, és a dir que no han estat adaptats, utilitzarem la grafia de l'idioma original (castellà, anglès, francès, etc.), tot i que no coincidisca amb la fonètica catalana: *jauja*, *zumo*, *jefe*, *jamón* o *jamó*, *heavy*, *traje*, *prêt-à-porter*, *voilà*...

En aquests casos de manlleus no adaptats haurem d'afegir una anotació, és a dir, escriurem una etiqueta abans de la paraula afectada (<ManlleuNoAdaptat>) i una després (</ManlleuNoAdaptat>), sense deixar-hi espais en blanc:

<ManlleuNoAdaptat>botellón</ManlleuNoAdaptat>
<ManlleuNoAdaptat>piscina</ManlleuNoAdaptat> [pronúncia castellana]
<ManlleuNoAdaptat>iPad</ManlleuNoAdaptat>
<ManlleuNoAdaptat>gauche divine</ManlleuNoAdaptat>

En general, doncs, mantindrem la transcripció següent en els mots: *casi*, *inglés*, *endespués*, *después*, *bueno*, *disfràs*, *pues* o *pos*, *algo*, *aixina*, *entonces*, *uelo*, *uela*, *sombrero*, *madera*, *txiringuito*, *màrmol*, *assentar*, *encontrar*, *sin embargo*, etc.

Les xifres s'escriuran seguint l'ortografia convencional i respectant la fonètica del parlant (i els criteris de transcripció fonètica que hem establert): *vuit-cents*, *seixanta-huit*. Segueix aquest criteri, la transcripció del nom de les hores (*són les onze i quart*) i el nom de les lletres, siguen manlleus o no (*ve*, *uve*, *jota*, *hatxe*, *bè*)...

4.2 Puntuació

La puntuació és un sistema gràfic molt ric per a la llengua escrita que vol organitzar el discurs, separar les oracions o constituents oracionals i facilitar-ne la lectura. Per tant, la puntuació no serveix per a marcar les pauses, sinó que fa altres funcions. Usar la puntuació per a la transcripció de la llengua oral és, doncs, aplicar-la a un àmbit que li és aliè. Tot i això, en algun cas usarem algun símbol de puntuació.

4.2.1 Pauses

En principi, les pauses s'indiquen amb la segmentació (vg. Gomis-Esplà i Sentí (en preparació)). Ara bé, pot passar que dins d'un segment hi haja una pausa i es preferisca marcar-la d'alguna manera que recórrer a tancar el segment. En aquest cas, usarem els símbols:

/ pausa breu
// pausa llarga

Per tant, en la transcripció no marcarem les pauses amb els símbols que utilitza la fonètica –com ara (|) o (| |)– ni amb els signes de puntuació, com el punt (.), la coma (,), els dos punts (:), els guions (–), el punt i coma (;) o els parèntesis.

Com que la fi de segment ja marca que hi ha pausa final, no cal acabar els segments amb / ni //.

4.2.2 Altres signes de puntuació

A més a més de marcar les pauses, usarem alguns signes de puntuació per a indicar alguns aspectes prosòdics: la interrogació (?), l'exclamació (!) i els punts suspensius (...). Vegem-ne els usos:

a) *Interrogació (?)*
Usarem el signe d'interrogació al final d'una oració per indicar que fem una pregunta directa: *Què passa? Anem a comprar?*

b) *Exclamació (!)*
Usarem el signe d'exclamació per expressar gràficament sorpresa o èmfasi: *Au! Ves-te'n a pastar fang!*

c) *Punts suspensius (...)*
Usarem els punts suspensius en el mateix sentit que la llengua escrita: per marcar que una oració és inacabada, que hi ha una interrupció, que l'oració es deixa en suspens o que una enumeració no s'ha acabat: *És que ell és molt...*

d) *Cometes (“”)*
No usarem les cometes. No indicarem l'estil directe o les citacions amb la puntuació, sinó amb un etiquetatge específic (vg. §5.5).

4.3 Criteris tipogràfics

Com ja hem dit, no utilitzarem la negreta, la cursiva, les versaletes, el subratllat ni cap altre recurs tipogràfic.

4.3.1 Majúscules i minúscules

A diferència de la llengua escrita, restringirem l'ús de la majúscula a casos molt clars i molt concrets:

- La primera lletra de l'inici del discurs de cada parlant.
- Usarem la majúscula després d'una pausa llarga amb canvi de tema.
- Per a iniciar un discurs reportat, encara que abans hi haja una barra (que indica pausa breu): *Va agafar el micròfon i va dir / He decidit que esta nit esteu tots convidats.*
- Els noms propis: antropònims (*Enric Valor*), topònims (*el Montgó*), els noms d'institucions, organismes o empreses (*Generalitat*, *la Caixa*). També quan siguen acrònims: *Renfe*, *Unicef.*
- Entitats religioses: *Déu*, *Esperit Sant.*

Així doncs, escriurem en minúscula els altres contextos i, sobretot, en cas de dubte, prioritzarem l'ús de la minúscula: *ajuntament*, *diputació*, *estat*…

4.4 Aspectes prosòdics i altres

4.4.1 Entonació

PARLARS vol ser un corpus per a estudis de diversos tipus, especialment gramaticals. Per això, l'estudi de la prosòdia no n'és un objectiu prioritari. També volem que siga un corpus extens i robust. Així doncs, el detall en la transcripció prosòdica ha de reduir-se al mínim. Tot i això, tal com hem indicat adés, marcarem mínimament els aspectes prosòdics, ja que usarem les exclamacions, interrogacions ortogràfiques i els punts suspensius per a marcar l'entonació descendent (*No tens raó!*), l'ascendent (*Vindràs demà?*) i el manteniment (*El que em vas explicar ahir…*), respectivament.

4.4.2 Paraules repetides

Les paraules repetides es transcriuran tantes vegades com s'hagen pronunciat: *que que que que vingues / dic!*

4.4.3 Paraules inacabades

Les paraules inacabades aniran seguides d'un guionet (-), sense espai: *escol-* 'escolta', *t'ana-* 't'anava'.

4.4.4 Paraules reduides

Anotarem les paraules que tenen una pronúncia reduïda amb l'etiqueta següent:

<PronunciaReduida t="a">ara</PronunciaReduida>
<PronunciaReduida t="en sia">en seguida</PronunciaReduida>

4.4.5 Interjeccions i sons paralingüístics

En el registre col·loquial solem trobar diversos recursos lingüístics discursius i expressius que doten d'eficàcia el missatge i que podem englobar dins de l'etiqueta general de les interjeccions (Cuenca, 2002). La transcripció d'aquests recursos és, sovint, complexa. Per això, seguirem l'ortografia adaptada que proposa Cuenca (2002) i la completarem amb la proposta de Riera-Eures i Sanjaume (2002 i 2010) i el llistat de 163 interjeccions que podem recuperar en la cerca avançada del DNV i 173 del DIEC2 (incloses les onomatopeies). En cas de ser necessari, emprarem la grafia convencional *h* per a marcar aspiracions tant en els casos més estandarditzats, com *ha-ha-ha* o *ehem* com els que no ho són tant, com *ahà* o *ha*.

D'altra banda, transcriurem les altres interjeccions pròpies o impròpies (normalment localismes), siguen catalanes o manlleus, que no apareixen en els reculls esmentats adés, seguint els criteris descrits: *arrea*, *ira, baia*, *equiliquà*, *ió*, *voilà*, etc.

4.4.6 Onomatopeies

Per a la transcripció de les onomatopeies seguirem el llistat de 92 onomatopeies presents en el DNV i de les que figuren entre els interjeccions del DIEC2. Completarem aquest llistat, en cas de necessitat, amb el que conté Riera-Eures i Sanjaume (2002 i 2010). Si apareix en el discurs algun ele-

ment onomatopeic no present en els repertoris assenyalats, el transcriurem seguint els criteris descrits adés.

Les onomatopeies han d'anar marcades amb una etiqueta específica abans i després de la paraula, com en l'exemple següent:

<Onomatopeia>bub-bub</Onomatopeia>

4.4.7 Paraules o fragments incomprensibles

Marcarem les paraules o fragments incomprensibles amb l'etiqueta <Incomprensible/>. Aquest marcatge no té una etiqueta d'inici i una de final, sinó que és una etiqueta única que indica que hi ha una paraula o fragment que no podem desxifrar.

Si la paraula o fragment és difícil d'entendre però tenim una hipòtesi d'interpretació, ho marcarem així:

<Dubte>paraula o paraules dubtoses</Dubte>

4.4.8 Estil directe

Els fragments en estil directe o les citacions tindran les etiquetes següents: <EDirecte> i </EDirecte>. Exemple:

Vam arribar a la plaça i fa <EDirecte>És que mai ve ningú. Quin desastre!</EDirecte>

4.4.9 Topònims

Marcarem els topònims menors (partides, muntanyes locals, noms de carrers...) amb les etiquetes següents:

<Lloc>Segària</Lloc>
Carrer <Lloc>Major</Lloc>

4.4.10 Fragments en altres llengües

Marcarem els fragments en altres llengües amb les etiquetes següents:

<LlenguaCastella>Qué pasó?</LlenguaCastella>
<LlenguaAngles>the best</LlenguaAngles>

5 Participants i anonimització

Per motius ètics i de protecció de dades, els textos han d'anonimitzar els informants i altres referents humans que s'esmenten abans que es publiquen al corpus. De fet, l'enregistrament també amagarà aquests fragments. Per a fer això, seguirem dues fases:

a) En la transcripció inicial, el transcriptor sí que conservarà tots els noms, però caldrà etiquetar-los. Per exemple, quan un parlant faça referència al nom d'un dels interlocutors de la conversa, caldrà etiquetar el nom amb aquesta etiqueta en què identifiquem amb un número cada parlant:

"Tu/ <NomInterlocutor ID="001">Vicent</NomInterlocutor ID="001"> / què trobes?
<NomEntrevistador ID="001">Pepa</NomEntrevistador ID="001">

També marcarem els noms o els malnoms referits a altres persones alienes als interlocutors de la conversa amb unes etiquetes. Vegem-ne uns exemples:

M'ha dit <NomExtern>Batiste</NomExtern> que no vindrà
Com em va dir <Malnom>Quico el del Pla</Malnom>
Mira per ací ve <Malnom>Teresa la de la plaça del rellotge</Malnom>

b) Fase de revisió i anonimització. En aquesta fase, crearem un fitxer nou anonimitzat en què substituirem els noms d'interlocutors, noms aliens a la conversa i malnoms per inicials, com en aquests exemples:

"Tu/ <NomInterlocutor ID="001">V</NomInterlocutor ID="001"> / què trobes?
M'ha dit <NomExtern>B</NomExtern> que no vindrà
Com em va dir <Malnom>Q</Malnom>
Mira per ací ve <Malnom>T</Malnom>

Si hi hagués lletres repetides que pogueren provocar una mala interpretació de la referencialitat, optaríem per una segona lletra, sempre que no facilités la identificació de la persona. Per exemple:

I també vindrà <NomExtern>Bernat</NomExtern>
I també vindrà <NomExtern>Be</NomExtern>

6 Disseny i criteris d'anotació

El corpus PARLARS ha estat construït fent servir una estratègia *stand-off*, és a dir, té diverses capes, una amb el senyal acústic dels parlants i d'altres amb les diverses capes d'anotació per a cada participant en la conversa (transcripció, tokenització, lema i categoria gramatical). A més, aquest model també fa possible la incorporació de noves capes d'anotació en el futur, segons els interessos d'investigació.

Per a codificar aquesta informació s'ha utilitzat el format EAF, [8] definit per a la ferramenta de transcripció ELAN (Wittenburg *et al.*, 2006), i que es basa en l'estàndard XML.

El disseny del corpus, els criteris d'anotació i l'ús i adaptació de la ferramenta *Apertium* (Forcada *et al.*, 2011) per a la tokenització i l'anàlisi morfològica de la llengua oral dialectal són explicats a Esplà-Gomis i Sentí (en preparació).

7 Primers resultats

Tot i que el projecte va començar l'any 2018, es troba avançat quant a la fase de gravació, sobretot. Hem dut a terme 86 gravacions en àudio (en alguns casos també en vídeo) en totes les comarques valencianes, llevat de la Safor i el Camp de Morvedre. El treball de camp que hem realitzat ha superat totes les expectatives ja que hem aconseguit moltes mostres d'interaccions orals (en forma de monòleg, conversa semidirigida i conversa secreta) de moltes poblacions i, per tant, gairebé totes les comarques estan més ben representades del que havíem previst inicialment. De més a més, la fase de transcripció ha començat i ja hi ha 16 transcripcions definitives.

A la fi, aquestes transcripcions ens han permés ja avançar en el darrer objectiu del projecte, és a dir, encetar investigacions sobre construccions lingüístiques en la conversa col·loquial. En aquest sentit són interessants els estudis d'Antolí i Sentí (2020) sobre les marques evidencials gramaticalitzades tipus *diu que* que només és possible trobar en aquest tipus de textos:

(1) La rabera sabia on havia d'anar. *Diu que* els animals tenen més coneiximent que les persones
(PARLARS, Xert, conversa semidirigida)

8 <http://www.mpi.nl/tools/elan/EAF_Annotation_Format.pdf>.

(2) A: Ací havien moltes sabateries i ara no sé si en quedarà una o dos, si en queden.
B: una va tancar ja
A: eh, per això dic jo que...
B: *diu que* si ja s'ho deixen

(PARLARS, Benissa, conversa semidirigida)

Comptat i debatut, el corpus parlars comença a ser una realitat. Actualment, disposem ja de molt de material enregistrat i, també, transcrit i anotat. Confiem que, aviat, amb la interfície de consulta –en fase d'elaboració–, es puga posar a l'abast de tots els investigadors i investigadores.

Referències bibliogràfiques

Albelda, Marta (2005): «Sistemas de transcripción de los corpus orales del español», in: Carrió Pastor, M. Luisa (coord.): *Perspectivas interdisciplinares de la lingüística aplicada*, València: AESLA, vol. 2, 381–387.

Alturo, Núria / Bladas, Òscar / Payà, Marta / Lluís Payrató (ed.) (2004): *Corpus oral de registres. Materials de treball*, Barcelona: Publicacions i Edicions de la Universitat de Barcelona.

Antolí, Jordi / Sentí, Andreu (2020): «Evidentiality in spoken Catalan. The evidential marker *diu que*», *Anuari de Filologia. Estudis de lingüística* 10.

Badia, Toni / Boleda, Gemma (2008): «CuCWeb: un corpus de la llengua catalana construït a partir de la web», *Estudis Romànics* 30, 291–293.

Baldaquí Escandell, Josep M. (2006): *El model de llengua i la seguretat lingüística dels jóvens valencians*, València / Barcelona: IIFV / Publicacions de l'Abadia de Montserrat.

Barreda Edo, Pere Enric / Ferrando Puig, Emili (coords.) (2007): *Benassal. Segle XX. Estudi d'un poble rural realitzat amb fonts orals. Obra completa*, 5 vols., Castelló: Grup de la Recuperació de la Memòria Històrica.

Beltran Zaragozà, Andreu / Rubio Miguel, José (2008): *Les Useres. Història d'un parlar*, València: Acadèmia Valenciana de la Llengua.

Beltran, Vicent / Segura-Llopes, Carles (2017): *Els parlars valencians*, València: Publicacions Universitat de València.

— / Esplà, Miquel / Guardiola, M. Isabel / Montserrat, Sandra / Segura, Carles / Sentí, Andreu (2019a): *Criteris per a la transcripció del corpus* PARLARS, València: Universitat de València, <http://roderic.uv.es/handle/10550/69633> [14.07.2020].

— / — / — / — / — / — (2019b): *Criteris per a la transcripció del corpus* PARLARS, València: Universitat de València, <http://roderic.uv.es/handle/10550/71244> [14.07.2020].

Bibiloni, Gabriel (1997): *Llengua estàndard i variació lingüística*, València: Tres i Quatre.

Bladas, Òscar (2009): *Manual de transcripció del discurs oral. Materials de treball*, Barcelona: Universitat de Barcelona.

Boix-Fuster, Emili / Àlamo Sala, Marina/ Galindo Solé, Mireia/ Vila i Moreno, Francesc X. (ed.) (2007): *Corpus de Varietats Socials. Materials de treball*, Barcelona: Publicacions i Edicions de la Universitat de Barcelona.

— / Vila, F. Xavier (ed.): (1998): *Sociolingüística de la llengua catalana*, Barcelona: Ariel Lingüística.

Briz Gómez, Antonio (2010): «Lo coloquial y lo formal, el eje de la variedad lingüística», in: Castañer, Rosa Ma. / Lagüens, Vicente (eds.): *De moneda nunca usada. Estudios dedicados a José Ma Enguita*, Zaragoza: Universidad Autónoma de Nuevo León, 125–133, <https://ifc.dpz.es/recursos/publicaciones/29/95/11briz.pdf> [14.07.2020].

— / Grupo Val.Es.Co. (2002): «Corpus de conversaciones coloquiales», *Anejo de la revista Oralia*, Madrid: Arco/Libros.

Brugman, Hennie / Russel, Albert (2004): «Annotating multimedia/multimodal resources with *ELAN*», in: Lino, M. Teresa et al. (eds.): *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisboa: European Language Resources Association (ELRA), 2065–2068, <http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf> [14.07.2020].

Carrera-Sabaté, Josefina / Viaplana, Joaquim (eds.) (s.a.): *Corpus Oral Dialectal (COD). Textos orals del nord-occidental*, Barcelona: Dipòsit Digital de la UB, <http://www.ub.edu/cccub/cod-nord-occidental-C&V.html> [14.07.2020]

Cuenca, Maria Josep (2002): «Els connectors textuals i els interjeccions», in: Solà, Joan (dir.): *Gramàtica del català contemporani*, vol. 3, Barcelona: Empúries, 3173–3237.

DCVB = Alcover, Antoni M. / Moll, Francesc de B. (1928–1962): *Diccionari Català-Valencià-Balear*, Palma: Editorial Moll, 10 vol., <http://dcvb.iecat.net> [14.07.2020].

De Benito Moreno, Carlota / Pueyo, Javier / Fernández-Ordóñez, Inés (2016): «Creating and designing a corpus of rural Spanish», in: Misra

Sharma, Dipti / Sangal, Rajeev / Kumar Singh, Anil (eds.): *Procceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, Varanasi: NLP Association of India, 78–83, <https://www.linguistics.rub.de/konvens16/pub/10_konvensproc.pdf> [14.07.2020].

DIEC2 = Institut d'Estudis Catalans, *Diccionari de la llengua catalana*, 2a edició, <https://dlc.iec.cat/> [14.07.2020].

DNV = Acadèmia Valenciana de la Llengua, *Diccionari Normatiu Valencià*, <http://www.avl.gva.es/lexicval/> [14.07.2020].

Duran Cals, Jordi / Martí Antonín, Maria Antònia / Perea Sabater, Maria Pilar (2007): «HistoCat y DialCat: extensiones de un analizador morfológico para tratar textos históricos y dialectales del catalán», *Procesamiento del lenguaje natural* 39, 279–280, <https://rua.ua.es/dspace/bitstream/10045/3055/1/PLN_39_35.pdf> [14.07.2020].

Esplà, Miquel / Beltran, Vicent / Guardiola, M. Isabel / Montserrat, Sandra / Segura, Carles / Sentí, Andreu (2018): *Tutorial d'ELAN per a la transcripció del corpus* PARLARS, València: MMedia, Universitat de València, <https://mmedia.uv.es/buildhtml/52147> [14.07.2020].

Esplà-Gomis, Miquel / Sentí, Andreu (en preparació): *The spoken corpus PARLARS: design, transcription and annotation of an informal corpus for Valencian Catalan*.

Fernández Ordóñez, Inés (2011): «Nuevos horizontes en el estudio de la variación gramatical del español: el *Corpus Oral y Sonoro del Español Rural*», in: Colón, Germà / Gimeno, Lluís (eds.): *Noves tendències en la dialectologia contemporània*, Castelló de la Plana: Universitat Jaume I, 173–203.

Forcada, Mikel L. / Garcia, Marinel / Iturraspe, Amaia / Gilabert, Patrícia / Montserrat, Sandra (2005): «Desenvolupament guiat per corpus d'un analitzador morfològic de català antic», in: Pusch, Claus D. / Kabatek, Johannes / Raible, Wolfgang (eds.): *Romanistische Korpuslinguistik / Romance Corpus Linguistics: Korpora und gesprochene Sprache / Corpora and Spoken Language*, Tübingen: Narr, 243–252.

— / Ginestí-Rosell, Mireia / Nordfalk, Jacob / O'Regan, Jim / Ortiz-Rojas, Sergio / Pérez-Ortiz, Juan Antonio / Sánchez-Martínez, Felipe / Ramírez-Sánchez, Gema / Tyers, Francis M. (2011): «*Apertium*: a free/open-source platform for rule-based machine translation», *Machine translation* 25:2 (Special Issue on *Free/Open-Source Machine Translation*), 127–144.

GDLC = *Gran Diccionari de la Llengua Catalana*, Enciclopèdia Catalana, <https://www.enciclopedia.cat/gran-diccionari-de-la-llengua-catalana> [14.07.2020].

Geeraerts, Dirk (1997): *Diachronic Prototype Semantics. A Contribution to Historical Lexicology*, Oxford: Clarendon Press.

GNV = Acadèmia Valenciana de la Llengua, *Gramàtica normativa valenciana*, <http://www.avl.gva.es/documents/31987/65233/GNV> [14.07.2020].

Goldberg, Adele E. (2003): «Constructions: A new theoretical approach to language», *Trends in Cognitive Sciences* 2:5, 219–224.

Hidalgo, Antonio / Sanmartín, Julia (2005): «Los sistemas de trascripción de la lengua hablada», *Oralia* 8, 13–36.

Hilpert, Martin (2014): *Construction Grammar and its Application to English*, Edimburgh: Edimburgh University Press.

Ide, Nancy / Pustejovsky, James (eds.) (2017): *Handbook of Linguistic Annotation*, Dordrecht: Springer.

Koch, Peter / Oesterreicher, Wulf (1990): *Gesprochene Sprache in der Romania: Französisch, Italienisch, Spanisch*, Tubinga: Niemeyer.

Llop, Ares / Pineda, Anna (2017): «L'estudi de la variació sintàctica en català: On som i cap on anem?», in: Pérez Saldanya, Manuel / Roca i Ricart, Rafael (eds.): *Actes del XVIIè Col·loqui Internacional de Llengua i Literatura Catalans. Universitat de València, 7–10 de juliol de 2015*, Barcelona: IEC, 527–542.

Lüdeling, Anke / Kytö, Merja (eds.) (2008): *Corpus Linguistics. An International Handbook*, Berlin / New York: De Gruyter.

Luong, Thang / Socher, Richard / Manning, Christopher (2013): «Better word representations with recursive neural networks for morphology», in: Hockenmaier, Julia / Riedel, Sebastian (eds.): *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, Sofia: Association for Computational Linguistics, <https://nlp.stanford.edu/~lmthang/data/papers/conll13_morpho.pdf> [14.07.2020].

Nissim, Malvina / Pietrandrea, Paola (2017): «MODAL: A multilingual corpus annotated for modality», in: Basili, Roberto / Nissim, Malvina / Satta, Giorgio (dir.): *Proceedings of the Fourth Italian Conference on Computational Linguistics (Clic-it 2017)*, Roma: Collana dell'Associazione Italiana di Linguistica Computazionale, <http://ceur-ws.org/Vol-2006/paper068.pdf> [14.07.2020].

Nivre, Joakim (2008): «Treebanks», in Lüdeling / Kytö (eds.), 225–241.

Observatori de Neologia (2004): *Metodologia del treball en neologia: criteris, materials i processos*, Barcelona: Universitat Pompeu Fabra / Institut Universitari de Lingüística Aplicada.

OIEC = Institut d'Estudis Catalans, *Ortografia catalana*, <https://www.iec.cat/llengua/documents/ortigrafia_catalana_versio_digital.pdf> [14.07.2020].

Payrató, Lluís (2010): *Pragmàtica, discurs i llengua oral*, Barcelona: Editorial UOC.

— / Alturo, Núria (ed.) (2002): *Corpus oral de conversa col·loquial. Materials de treball*, Barcelona: Publicacions de la Universitat de Barcelona.

Perea, Maria-Pilar / Viaplana, Joaquim: *Corpus Oral Dialectal (COD). Selecció de textos*, Dipòsit Digital de la Universitat de Barcelona, <http://www.ub.edu/cccub/cod-seleccio.html> [14.07.2020].

Pons, Clàudia / Viaplana, Joaquim (ed.) (2009): *Corpus oral dialectal (COD). Textos orals del balear*, Dipòsit Digital de la Universitat de Barcelona, <http://www.ub.edu/cccub/cod-balear2009.html> [14.07.2020].

Porterval = Acadèmia Valenciana de la Llengua, *Portal Terminològic Valencià*, <https://www.avl.gva.es/lexicval/ptv> [14.07.2020].

Pradilla Cardona, Miquel Angel (1993): *Variació i canvi lingüístic en curs al català de transició nord-occidental valencià*, Tarragona: Universitat Rovira i Virgili, Departament de Filologies Romàniques (tesi doctoral).

— (2004): *El laberint valencià: apunts per a una sociolinguística del conflicte*, Benicarló: Onada.

Prieto, Pilar / Cabré, Teresa (coords.) (2007–2012): «Criteris bàsics de transcripció ortogràfica de l'*Atles interactiu de l'entonació del català*», in: *Atles interactiu de l'entonació del català*, <http://prosodia.upf.edu/atlesentonacio/> [14.07.2020].

— / — (coords.) (2013): *L'entonació dels dialectes catalans*, Barcelona: Publicacions de l'Abadia de Montserrat.

Procházková, Petra (2006): *Fundamentos de la lingüística de corpus. Concepción de los corpus y métodos de investigación con corpus*, <http://www.prochazkova.de/fundamentos_de_la_lingüística_de_corpus.pdf> [14.07.2020].

Riera-Eures, Manel / Sanjaume, Margarida (2002): *Diccionari d'onomatopeies i mots de creació expressiva*, Barcelona: Edicions 62.

— / — (2010): *Diccionari d'onomatopeies i altres interjeccions*, Vic: Eumo.

Samper Padilla, Jose A. / Hernández, Clara Eugenia / Troya, Magnolia (1998): *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico (MC-NLCH)*, Las Palmas: Servicio de Publicaciones de la Universidad de Las Palmas de Gran Canaria.

Sauer, Simon / Lüdeling, Anke (2016): «Flexible multi-layer spoken dialogue corpora», *International Journal of Corpus Linguistics* 21:3, 419–438.

Segura, Carles (2003): *Variació dialectal i estandardització al Baix Vinalopó*, Alacant / Barcelona: Institut Interuniversitari de Filologia Valenciana / Publicacions de l'Abadia de Montserrat.

Sentí, Andreu (2019): «Les perífrasis modals catalanes: prescripció i descripció», in: Escartí, Vicent J. (ed.): *Nunc dimittis. Estudis dedicats al professor Antoni Ferrando*, València: Publicacions de la Universitat de València, 525–562.

Solà, Joan (dir.) (2002): *Gramàtica del català contemporani*, 3 vols., Barcelona: Empúries.

Taulé, Mariona / Martí, M. Antònia / Recasens, Marta (2008): «*AnCora*: Multilevel annotated corpora for Catalan and Spanish», in: Calzolari, Nicoletta et al. (eds.): *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech: European Language Resources Association (ELRA), 96–101, <http://www.lrec-conf.org/proceedings/lrec2008/pdf/35_paper.pdf> [14.07.2020].

Taylor, John R. (2012): *The Mental Corpus*, Oxford: Oxford University Press.

TERMCAT = *Centre de Terminologia*, <http://www.termcat.cat> [14.07.2020].

Veny, Joan (2002 [1982]): *Els parlars catalans*, Palma: Editorial Moll.

Viaplana, Joaquim / Perea, Maria Pilar (ed.) (2003): *Textos orals dialectals del català sincronitzats. Una selecció*, Barcelona: Promocions i Publicacions Universitàries (PPU), <http://www.ub.edu/cccub/cod-textos2003.html> [14.07.2020].

— / Lloret, Maria-Rosa / Perea, Maria-Pilar / Clua, Esteve (2007): *COD. Corpus Oral Dialectal*, Barcelona: Promocions i Publicacions Universitàries (PPU).

Wichmann, Anne (2008): «Speech corpora and spoken corpora», in Lüdeling / Kytö (eds.), 187–207.

Wittenburg, Peter / Brugman, Hennie / Russel, Albert / Klassmann, Alex, Sloetjes, Han (2006): «ELAN: a Professional Framework for Multimodality Research», in: Calzolari, Nicoletta et al. (eds.): *Proceedings of the Fifth International Conference on Language Resources and Evaluation*

(LREC'06), Genoa: European Language Resources Association (ELRA), 1555–1559, <http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf> [14.07.2020].

Corpus lingüístics en línia

AnCora, <http://clic.ub.edu/corpus/es/ancora> [14.07.2020].

Arxiu audiovisual dels dialectes catalans de les Illes Balears, <http://www.uib.cat/catedra/camv/arxiu.html#inici> [14.07.2020].

CCCUB = *Corpus de Català Contemporani de la Universitat de Barcelona*, Universitat de Barcelona, <http://www.ub.edu/cccub/> [14.07.2020].

CICA = Torruella, Joan (dir.) / Pérez Saldanya, Manuel / Martines, Josep / Martines, Vicent: *Corpus Informatitzat del Català Antic*, <http://cica.cat> [14.07.2020].

CIGCA = Martines, Josep / Martines, Vicent (dirs.): *Corpus Informatitzat de la Gramàtica del Català Antic*, Alacant: ISIC-IVITRA, Universitat d'Alacant.

CIGCMod = Martines, Josep / Martines, Vicent (dirs.): *Corpus Informatitzat de la Gramàtica del Català Modern*, Alacant: ISIC-IVITRA, Universitat d'Alacant.

COC = *Corpus oral de conversa col·loquial*, Barcelona: Universitat de Barcelona, <http://www.ub.edu/cccub/corpusoraldeconversacolloquial-coc.html> [14.07.2020].

COD = *Corpus oral dialectal*, Barcelona: Universitat de Barcelona, <http://www.ub.edu/cccub/corpusoraldialectal-cod.html> [14.07.2020].

COLC = Dols, Nicolau (dir. / Martines, Vicent / Yzaguirre, Lluís de: *Corpus Oral de la Llengua Catalana*, Barcelona / Palma de Mallorca: IEC / Govern Balear, <https://www.iec.cat/recerca/projecte1.asp?codi=PR2017-S04-DOLS> [14.07.2020].

COR = *Corpus oral de registres*, Barcelona: Universitat de Barcelona, <http://www.ub.edu/cccub/corpusoralderegistres-cor.html> [14.07.2020].

Corpus tècnic de l'IULA, <http:((bwananet.upf.edu/> [14.07.2020].

COS = *Corpus oral de varietats socials*, Barcelona: Universitat de Barcelona, <http://www.ub.edu/cccub/corpusdevarietatssocials-cos.html> [14.07.2020].

COSER = *Corpus Oral y Sonoro del Español Rural*, <http://www.corpusrural.es/> [14.07.2020].

CTILC = *Corpus textual informatitzat de la llengua catalana*, Barcelona: Institut d'Estudis Catalans, <https://ctilc.iec.cat/> [14.07.2020].

Museu de la paraula. Arxiu de la memòria oral valenciana, <http://www.museudelaparaula.es/web/home/info.php> [14.07.2020].

Val.Es.Co. = *Corpus Val.Es.Co. 2.1*, <http://www.valesco.es/?q=consulta> [14.07.2020].

- Sandra Montserrat, Universitat d'Alacant, Departament de Filologia Catalana, Campus de Sant Vicent del Raspeig s/n, E-03690 Alacant, <sandra.montserrat@ua.es>.
- Carles Segura, Universitat d'Alacant, Departament de Filologia Catalana, Campus de Sant Vicent del Raspeig s/n, E-03690 Alacant, <carles.segura@ua.es>.