# Criteris per a la transcripció del corpus *Parlars*. Segona versió

## 1. Introducció

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 (vg. Hidalgo & 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 AFI o altres models amb símbols prosòdics sobre pauses, encavalcaments, gestualitats entre altres aspectes vocals (vg. COC, Alturo & 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 & Perea 2003; *l'Atles entonatiu*, Prieto & Cabré 2007-2012; el corpus COSER, de Benito, Pueyo & Fernández-Ordóñez 2016; o el MC-NLCH, Samper, Hernández & Troya 1998).

Els criteris de transcripció que presentem s'ajusten als objectius del corpus *Parlars* (projecte CorDiVal: *Elaboració d'un corpus oral dialectal del valencià col·loquial*, Ref. GV/2017/094, finançat per la Generalitat Valenciana; https://www.uv.es/corvalc/), 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 *Corpus de Català Contemporani de la Universitat de Barcelona* (Payrató & Alturo 2002, Viaplana et al. 2003, Viaplana & Perea 2003, Foix-Fuster *et al.* 2007, Alturo *et al.* 2009, Carrera-Sabaté & Viaplana, Perea & Viaplana, Pons & Viaplana 2009, Blesa 2009, Payrató 2010) i els de l'*Atles interactiu de l'entonació del català* (Prieto & Cabré 2007-2012). Finalment, també ens hem basat en les característiques lingüístiques dialectals generals (Veny 1982, Solà *et al.* 2002, Beltran & Segura 2017).

El corpus *Parlars* tindrà una transcripció que provarà de reflectir al màxim possible les característiques dialectals i lingüístiques del discurs alhora que estandarditzarà alguns aspectes fonètics per facilitar les tasques d'anotació. Optem per una estratègia d'anotació multicapes (*stand-off*) que permet la superposició de diversos nivells de transcripció i anotació alineats amb el document audiovisual original. La font primària del corpus és l'arxiu de so (o vídeo) dividit en segments breus que faciliten la cerca d'elements lingüístics transcrits i el fragment corresponent del document multimèdia on hi apareixen. La transcripció es realitza amb l'eina ELAN (Brugman & Russel 2014) (vg. les instruccions del funcionament del programa amb el tutorial que hem preparat per a l'elaboració del corpus *Parlars*, Esplà *et al.* 2018). La tokenització i lematització es durà a terme amb l'analitzador morfològic proporcionat per la col·lecció d'eines Apertium (Forcada *et al.* 2011).

L'estratègia d'anotació multicapa permet afegir diverses capes de transcripció o anotació segons les necessitats. Les diverses capes que s'introduïsquen dependran de cada text, però previsiblement hi ha almenys les següents:

a) Una capa amb els silencis més llargs d'un determinat llindar (vg. la primera part del tutorial, Esplà *et al.* 2018).
b) Una capa amb la transcripció per cada parlant.
c) Una capa amb la segmentació per tokens.

d) Una capa amb el lema.
e) Una capa amb la categoria gramatical.
f) Una capa per marcar els fragments no transcrits. Els fragments que decidim no transcriure perquè no tenen interès o perquè hi ha massa soroll o massa frases que no es poden entendre aniran marcats amb un segment en aquesta capa.

En la mesura que avance el projecte, podrem afegir-hi noves capes amb anotació de fenòmens diversos (fonètica i prosòdia, sintaxi, semàntica, pragmàtica, etc.) (Nivre 2008; Nissim & Pietrandrea 2017), així com també altres investigadors aliens al projecte podran descarregar-se els fitxers i afegir les capes i anotacions que necessiten les seues investigacions.

En definitiva, el model de transcripció que presentem no tindrà un fitxer en format .doc o .txt, és a dir, no serà un text lineal de l'àudio en forma de text escrit. Per tant, no hi haurà trets tipogràfics (negreta, cursiva, subratllat, versaleta…), ni caldran símbols per a marcar encavalcaments o una disposició del text determinada dins la pàgina. El fitxer resultant de la transcripció, sincronització i segmentació amb ELAN tindrà un format estàndard XML TEI (Text Encoding Initiative - TEI Consortium, 2017) (vg. Esplà *et al.* 2018).

La transcripció que proposem és a mig camí entre la *transcripció discursiva* que incorpora molts elements del context (Du Bois *et al*. 1991, Bladas 2009: §4) i la *transliteració ortogràfica*. Aquesta transcripció permet tenir una versió aproximada a la fonètica del text oral i molt fidel a les característiques morfosintàctiques, però no entrarà en detall a marcar aspectes prosòdics com l'entonació. Amb tot, sí que incorporarà alguns aspectes discursius que es puguen marcar amb la puntuació o altres signes, és a dir, s'indicaran les pauses, les interjeccions o altres sons (vg. §6). És a dir, es tracta d'una transcripció intermèdia-estreta.

En aquest protocol de transcripció presentem la segmentació de la llengua oral (§2), els criteris lingüístics de la transcripció (§3), els criteris de puntuació (§4), els criteris tipogràfics (§5), els aspectes prosòdics (§6) i el tractament de la privacitat (§7).

## 2. Segmentació de la llengua oral

La segmentació del discurs en unitats de transcripció l'elaborarem semiautomàticament. Primer, amb el programa ELAN (vg. Esplà *et al*. 2018) obtindrem una segmentació automàtica en què s'aprofiten les pauses per a marcar una unitat o segment. Aquesta segmentació serà reajustada pel transcriptor per assegurar-nos que el tall es produeix coincidint en una pausa. El transcriptor també ajustarà manualment cada segment a la línia del parlant que té la paraula, és a dir, cada segment estarà situat a la línia del parlant en qüestió. Evidentment, hi haurà segments solapats sempre que hi haja més d'un parlant intervenint al mateix moment.

Cada segment serà més o menys equivalent a la unitat tonal o unitat prosòdica (Prieto & Cabré 2013). La segmentació de cada unitat tonal no serà del tot exhaustiva perquè és molt difícil determinar les característiques tonals amb una o dues audicions (Edwards 1993 *apud* Bladas 2009) (vg. §6). Tanmateix, el transcriptor intentarà acostar-s'hi i aprofitar el silenci de cada parlant per a acabar un segment.

## 3. Criteris lingüístics de la transcripció

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.

Els criteris que presentem a continuació han estat discutits col·lectivament i són el resultat d'unes proves pilot. En aquestes proves, hem lematitzat automàticament uns quants textos orals amb Apertium (Forcada *et al.* 2011) seguint uns criteris fonoortogràfics molt pròxims a la pronunciació (hem seguit uns criteris provisionals, vg. Beltran *et al.* 2019) i hem observat quins mots que no han estat reconeguts pel lematitzador automàtic. Després, hem valorat quins són els fenòmens més freqüents que dificulten la lematització i quins fenòmens podrien ser reconeguts en el futur si afegim nous lemes o noves regles a Apertium. En definitiva, els criteris que presentem responen a la voluntat de facilitar la tasca de lematització i, per això, no reflecteixen la majoria dels fenòmens fonètics.

### Fonètica

Els criteris de transcripció intenten acostar-se a la fonètica real, però no pretenen 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 (*premit*), però no les neutralitzacions vocàliques, l'elevació lingual o obertura de les vocals (no transcriurem les pronúncies *còsa, péu* o *qué*), 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, els fenòmens següents els marcarem 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 *-es* de gentilicis (*anglés*), participis (*admés*, *compromés*) i adjectius (*cortés*); numerals ordinals acabats en *-e* (*cinqué*) i alguns substantius (*café*); la terminació *-en* de la tercera persona del plural del present d'indicatiu d'alguns verbs de la II conjugació (*aprén, comprén, depén*); els infinitius acabats en *-eixer* (*conéixer*) i *-encer* (*convéncer*); la segona i tercera persona del plural dels imperfets d'indicatiu amb accent al radical (*féiem, féieu*).
b) Canvi de síl·laba accentuada. Transcriurem la pronúncia esdrúixola de mots com *cànvie* 'canvie', *ódie* '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'. No deixem espai entre el pronom i l'apòstrof perquè considerem que l'auxiliar és un clític.
e) La preposició *a* l'escriurem *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', *Naal* 'Nadal', *palar* 'paladar'. En canvi, no marcarem la caiguda de la *d* quan no hi ha reducció sil·làbica: en lloc de *llauraor* transcriurem *llaurador*.
g) En general, només marcarem les caigudes de vocals en paraules gramaticals com els demostratius. Per tant, transcriurem *ixe* 'eixe', *ixò* 'això'. 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 vocal inicial) perquè, tot i que és un fenomen oral habitual, hi ha molta variació fins i tot intraparlant: en lloc de *nar* escriurem *anar*, en lloc de *via* escriurem *havia*, en lloc de *metla* escriurem *ametla*.
h) Fenòmens fonètics diversos com ara *quidrar/quirdar* 'cridar', *auia* 'aigua', *bragó* 'braó', *a von* 'a on'; 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'). La reducció de *per (a)* a *pe*, també la transcriurem: *pe tu* ('per a tu'), *pe hí* ('per ahí').
j) La palatització d'*haver-hi*: *no ny'ha* 'no n'hi ha' (apòstrof sense espai ja que té un únic accent).

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, trendria*, etc., les transcriurem amb la forma estàndard *vingut, tingut, tindria*, etc.
g) Les diverses pronúncies possibles del mot *diumenge* (*dumenge, domenge*...) no les representarem.
h) 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 les transcriurem ortogràficament amb *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* escriurem *allí*, en lloc de *nar* escriurem *anar*, en lloc de *via* escriurem *havia*, en lloc de *metla* escriurem *ametla*, en lloc de *scola* escriurem *escola*.
l) No indicarem les elisions per contacte vocàlic: *onze anys, una hora, la mateixa hora, sense ou, eixe home…*
m) No marcarem les sinalefes perquè és un fenomen molt general: *una amiga*.
n) Tampoc no marcarem les formes *sixanta*, *ixim, ixiu*... sinó que les transcriurem segons la norma *seixanta*, *eixim, eixiu*. Tampoc no escriurem *coranta*, sinó *quaranta*.
o) 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 ho transcrivim. Per això, sí que escriuríem la forma *auia*.
p) 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' no les anotarem i usarem la forma normativa.
q) L'emmudiment de la *-r* final. Escriurem *dir, comprar, corredor.*
r) L'emmudiment de la r en mots com *perdre, prendre* (i *sorprendre*, *mamprendre, emprendre*...) *arbre, marbre, dimarts, diners*...
s) La pronunciació amb *n* de mots com *contar* 'comptar', *prensa* 'premsa', *pronte* 'prompte', etc. Per tant, els transcriurem d'acord amb la llengua normativa.
t) L'ensordiment de les sibilants sonores: *ca**s**a, do**tz**e, me**tg**e, plu**j**a.* Tampoc marcarem els canvis produïts per la fonètica sintàctica: *pe**ix** i carn*, *case**s** altes*.
u) L'ensordiment de la [z] del mot *zero* no la transcriurem així *sero*, sinó amb l'ortografia corrent, zero. El mateix per a altres mots com *zenzillo* 'senzill', el transcriurem *senzillo*.
v) La sonorització en fonosintaxi de les oclusives sordes: *cin[g] o sis; se[d] o huit*. És a dir, escriurem *cinc o sis* i *set o huit*.
w) La paletització de la sibilant alveolar en contacte amb una velar no la marcarem: no escriurem *servixca, crexca* o *vixcut* sinó *servisca*, *cresca* i *viscut*.
x) La presència/absència de la iod davant de la palatal fricativa sorda i sonor (*caixa/caxa, bajoca/baijoca, corretja/correija, pujar/puijar, metge/meige*).
y) Les aspiracions de la *s*: *és que*.
z) El ieisme: *llavi, palla, pell*
aa) El betacisme: *vinc, vols*.
bb) La distinció entre fricatives [ʒ] i africades [d͡ʒ]: *Jaume*.
cc) Les diferents realitzacions fonètiques del pronom *jo* [io], [d͡ʒo]... Sempre transcriurem *jo*.
dd) La realització de les oclusives fricatives: *ceba, pagar*...
ee) El reforç de les consonants finals: [pontᵊ] ho transcriurem *pont.*
ff) 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?*
gg) La palatalització del grup *-ts* final: *acabats, tots*.

hh) Les paraules geminades les escriurem d'acord amb l'ortografia: *col·legi, guatla, ratlla, setmana*.

**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) Els demostratius *este*, *eixe* i *aquell* es transcriuen mantenint la forma no reforçada i reflectint la realització fonètica: *ixe xic, esta dona.* La pronúnica *est' home* no la marquem, escriurem *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* la transcrivim *en* d'acord amb la pronúncia: *en ell*, *en Maria*. No marcarem si hi ha una neutralització de la *e* en *a*: *estic an Pep* ho transcriurem per *estic en Pep*. Tampoc no marcarem si la *n* s'assimila a la bilabial de la paraula següent: transcriurem *en pa* 'amb pa' i no, *em pa*.
g) La flexió verbal mostrarà la forma usada: *perc* 'perd', *vega* 'veja', *ell cantave* 'cantava', *ell porte* '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 *vindrer* 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 *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...*

**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.

Els manlleus els transcriurem 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. *calsonsillos*), *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 que 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...* Aquests exemples són manlleus no adaptats.

En el cas dels manlleus no adaptats haurem d'afegir una anotació, és a dir, escriurem una etiqueta[1] abans de la paraula afectada (<ManlleuNoAdaptat>) i una després (</ManlleuNoAdaptat>), sense deixar-hi espais en blanc. Vegeu l'exemple següent:

<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, domés, 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: *quatre-cents sixanta-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, be*)...

## 4. 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.

**Pauses**

En principi, les pauses s'indiquen amb la segmentació (vg. §2). 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 el símbols:

[1] Advertència sobre l'ús de les etiquetes: És millor començar i tancar una etiqueta dins d'un segment i no que hi haja diversos segments entre l'inici i el final de l'etiqueta. També és molt important que estiguen escrites correctament (fixeu-vos-hi que no hi ha espais, que hi ha majúscules, que no hi ha accents, etc.).

/ 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 //.

**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:

*Interrogació (?)*

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

*Exclamació (!)*

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

*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...*

*Cometes ("")*

No usarem les cometes. L'estil directe o les citacions no les indicarem amb la puntuació, sinó amb un etiquetatge específic (vg. §6).

## 5. Criteris tipogràfics

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

**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 coma (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 casos i, sobretot, en cas de dubte, prioritzarem l'ús de la minúscula: *ajuntament, diputació, estat*…

## 6. Aspectes prosòdics i altres

### Entonació

El corpus *Parlars* vol ser un corpus per a estudis de diversos tipus, especialment gramatical. Per això, l'estudi de la prosòdia no és l'objectiu principal. 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 vist en §4, 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.

### Allargament de vocals

No marquem els allargaments de vocals finals (o no finals) amb la repetició de les vocals pertinents que sovint es donen en el discurs oral per a manifestar sorpresa i exclamació:

- *Què?* i no *Quèe?, Quèee? o Quèeee?*
- *Sí i no síiii.*
- *Vols... vindre?* i no *Volss... vindre?*

### Paraules repetides

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

### Paraules inacabades

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

### Pauses i encavalcaments

Ja hem vist com les pauses es marquen (vg. §3). Si les pauses són molt llargues o molt breus es podrà observar en l'enregistrament i no amb la transcripció. No es marcaran tampoc els encavalcaments conversacionals perquè cada parlant té una capa d'anotació individual i, per tant, hi poden haver superposicions en el mateix moment temporal (vg. Esplà et al. 2018).

### 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 & 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*.

Les interjeccions més generals són:

| **Valor** | **Forma** |
|---|---|
| alegria, enuig | *xe, xi* |
| apel·lació | *txit, psit* |
| assentiment, corroboració o comunicació que se segueix la conversa | *mhm, hu-hu, sè* |
| avís o prevenció | *ui-ui-ui* |
| crida, avís o salutació | *ei, ep, ie* |
| demanda de conformitat | *eh, tat, vitat* |
| desacord | *ps* |
| dolor | *ai, au, oi, ui* |
| èmfasi | *fu, buf* |
| estranyesa | *ai* |
| fàstic | *ec, ecs, uà, uf, uix* |
| fruïció, desig | *hum* |
| frustració o desacord respecte al que s'ha dit (és un clic) | *ntx* |
| indiferència o menyspreu | *pse* |
| inici o canvi de tema de conversa i altres usos | *ah* |
| menyspreu (és un clic) | *tx* |
| objecció o rectificació | *ep, uep* |
| omplidor de la parla que permet conservar el torn mentre es pensa què dirà a continuació | *e…, mm* |
| opressió, aclaparament | *uf* |
| per a saludar, cridar l'atenció o demanar silenci | *ts* |
| sorpresa, desil·lusió o contrarietat | *uei* |
| sorpresa, incredulitat | *ai, ca, ha, hu* |

**Taula 1**. *Interjeccions* (basada en Bladas 2009, DNV, DIEC2, Riera-Eures & Sanjaume 2002 i 2010)

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 (vg. §3): *arrea*, *ira, baia*, *equiliquà*, *ió*, *voilà*, etc.

Podrem afegir una capa d'anotació dedicada al lèxic on podrem anotar el sentit d'aquestes formes, si no coincideixen amb les previstes i el transcriptor considera que cal aclarir-ho.

Les interjeccions, les marquem amb una marca morfològica en la capa de la categoria gramatical.

**Onomatopeies**

Per a la transcripció de les onomatopeies seguirem el llistat de 92 onomatopeies present en el DNV, les que figuren entre els interjeccions del DIEC2 i el completarem, en cas de necessitat, amb el que conté Riera-Eures & Sanjaume (2002 i 2010). En cas que aparega en el discurs algun element onomatopeic no present en els repertoris assenyalats, el transcriurem seguint els criteris descrits adés (vg. §3).

Les onomatopeies han d'anar marcades amb una etiqueta específica abans i després de la paraula, com en l'exemple següent (sense deixar-hi espais):

<Onomatopeia>bub-bub</Onomatopeia>

**Riures i altres sons**

Els sons provocats pels parlants (riures, esternuts, tos, respiració, sospirs, xiulet...) o sons externs (portes, cotxes...) no els marquem en les transcripcions, tot i que se sentiran en l'àudio.

**Paraules o fragments incomprensibles**

Les paraules o fragments incomprensibles els marcarem així: <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 raonable ho marcarem així:

<Dubte>paraula o paraules dubtoses</Dubte>

**Estil directe**

Els fragments en estil directe o les citacions els marcarem així amb les següents etiquetes: <EDirecte> i </EDirecte>. Exemple:

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

**Topònims**

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

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

**Fragments en altres llengües**

Els fragments en altres llengües els marcarem amb les etiquetes següents:

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

Si són paraules soltes, no ho marcarem amb aquestes etiquetes.

## 7. Participants i anonimització

L'entrevistador és un col·laborador del projecte que coneix els informants, els enregistra i intenta fer fluir la conversa. Quan es tracta d'una entrevista-monòleg, la seua intervenció tendeix a ser mínima. Tot i això, quan intervé hi haurà una capa amb la transcripció de les seues intervencions. L'abreviació serà E. Si n'hi ha més d'un, els indicarem així: E1, E2, E3...

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?

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 haguera 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>

## Referències bibliogràfiques

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

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

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

Beltran, Vicent; Esplà, Miquel; Guardiola, M. Isabel; Montserrat, Sandra; Segura, Carles; Sentí, Andreu (2019): *Criteris per a la transcripció del corpus* Parlars. València: Universitat de València. Roderic. <http://roderic.uv.es/handle/10550/69633>

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

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

Briz Gómez, Antonio & Grupo Val.Es.Co. (2002): "Corpus de conversaciones coloquiales", *Anejo de la revista Oralia*. Madrid: Arco-Libros.

Brugman, H.; Russel, A. (2004): *Annotating Multimedia/Multi-modal resources with ELAN*, Proceedings of the Fourth International Conference on Language Resources and Evaluation. Lisboa: Portugal, p. 2065-2068,

CCCUB = *Corpus de Català Contemporani de la Universitat de Barcelona*. (<http://www.ub.edu/cccub/>)

DCVB = A. M. Alcover & F. de B. Moll (1928-1962) Dicci*onari Català-Valencià-Balear*, Palma, Editorial Moll, 10 vol. [http://dcvb.iecat.net.]

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

DIEC = Institut d'Estudis Catalans, *Diccionari de la llengua catalana*, 2a edició.

GDLC = *Gran Diccionari de la Llengua Catalana*, Enciclopèdia Catalana.

Carrera-Sabaté, Josefina & Joaquim Viaplana (ed.): *Corpus Oral Dialectal (COD). Textos orals del nord-occidental*. Dipòsit Digital de la UB.

COC = Payrató, Lluís & Núria Alturo (ed.) (2002): *Corpus oral de conversa col·loquial. Materials de treball*. Barcelona: Publicacions de la Universitat de Barcelona.

Cuenca, Maria Josep (2002): "Els connectors textuals i les interjeccions" dins Solà, Joan (dir.) *Gramàtica del català contemp*orani. Barcelona: Empúries, p. 3173-3237.

De Benito Moreno, Carlota, Javier Pueyo & Inés Fernández-Ordóñez (2016): "Creating and designing a corpus of rural Spanish", *Procceedings of the 13th Conference on Natural Language Processing* (KONVENS 2016), p. 78-83. (<https://www.linguistics.rub.de/konvens16/pub/10_konvensproc.pdf>)

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

Forcada, Mikel L.; 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 (Special Issue on Free/Open-Source Machine Translation) 25:2, 127-144.

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

Nissim, Malvina & Paola Pietrandrea (2017): "MODAL: A multilingual corpus annotated for modality". In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*.

Nivre, Joakim (2008): "Treebanks", dins Lüdeling, Anke & Kytö, Merja (ed.) *Corpus linguistics. An International Handbook*. Berlin / New York: Walter de Gruyter.

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/ortografia_catalana_versio_digital.pdf>)

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

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

Perea, Maria-Pilar & Joaquim Viaplana: *Corpus Oral Dialectal (COD). Selecció de textos*. Dipòsit Digital de la UB.

Pons, Clàudia & Joaquim Viaplana (ed.) (2009): *Corpus oral dialectal (COD)*. Textos orals del balear. Dipòsit Digital de la UB.

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

Prieto, Pilar & Cabré, Teresa (coords.) (2007-2012): "Criteris bàsics de transcripció ortogràfica de l'Atles interactiu de l'entonació del català", dins *Atles interactiu de l'entonació del català*. Pàgina web: <http://prosodia.upf.edu/atlesentonacio/>.

Prieto, Pilar & Cabré, Teresa (coords.) (2013): *L'entonació dels dialectes catalans*. Barcelona: Publicacions de l'Abadia de Montserrat.

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

Riera-Eures Manel & Margarida Sanjaume (2010): *Diccionari d'onomatopeies i altres interjeccions*. Vic: Eumo.

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

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

TEI Consortium, eds. *Guidelines for Electronic Text Encoding and Interchange*. [2017-11-29]. <http://www.tei-c.org/P5/>.

TERMCAT = *Terminologia Catalana* (<http://www.termcat.cat/ca/>)

Viaplana, Joaquim & Maria Pilar Perea (ed.) (2003): *Textos orals dialectals del català sincronitzats. Una selecció*. Barcelona: Promociones y Publicaciones Universitarias (PPU).

Viaplana, Joaquim; Maria-Rosa Lloret, Maria-Pilar Perea & Esteve Clua (2007): *COD. Corpus Oral Dialectal*. Barcelona: Promociones y Publicaciones Universitarias (PPU).

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

**Citació bibliogràfica d'aquest treball:**

Beltran, Vicent; Esplà, Miquel; Guardiola, M. Isabel; Montserrat, Sandra; Segura, Carles; Sentí, Andreu (2019b): *Criteris per a la transcripció del corpus* Parlars. Segona versió. València: Universitat de València. Roderic.