

# Introducció

Vicent Beltran (Alacant)

Maribel Guardiola (Alacant)

El dossier que teniu a les mans<sup>1</sup> sorgeix de la celebració del I i II Col·loqui Internacional sobre Lingüística de Corpus Orals (LingCor) organitzats en el marc del projecte de recerca «Elaboració d'un corpus oral dialectal del valencià col·loquial (CorDiVal)», dirigit pel Dr. Andreu Sentí i finançat per la Generalitat Valenciana (ref. GV/2017/094).

La primera edició de LingCor (Universitat de València, 2018) es va centrar tant en la construcció de corpus textuais (arreplega de dades, selecció d'informants, divisions dialectals i sociolectals, etiquetatge i anotació de corpus) com en l'explotació i els estudis basats en corpus (fonètica, morfosintaxi, semàntica i pragmàtica). Tot i que s'acceptaven propostes amb tots els enfocaments teòrics i totes les temàtiques al voltant de la lingüística de corpus orals, els temes principals van ser bàsicament els corpus sincrònics orals del català i de llengües europees minoritzades, els corpus dialectals de les llengües romàniques i les llengües europees, la planificació de corpus orals, el processament de les dades dels corpus orals, els estudis sobre el canvi lingüístic i la variació dialectal i funcional a partir de corpus orals, la lingüística cognitiva, la gramàtica de construccions i la pragmàtica en relació amb els corpus orals. Hi van participar com a ponents convidats Anke Lüdeling (Humboldt-Universität zu Berlin), Martin Hilpert (Université de Neuchâtel), Antonio Briz (Universitat de València) i Xosé Luis Regueira (Universidade de Santiago de Compostela).

D'altra banda, la segona edició de LingCor (Universitat d'Alacant, 2019) pretenia ser un fòrum de discussió entre lingüistes que treballen en una metodologia de corpus orals. Com en l'edició anterior, aquesta jornada se centrava tant en la construcció de corpus textuais (arreplega de dades, selecció d'informants, divisions dialectals i sociolectals, etiquetatge i anotació de corpus) com en l'explotació i els estudis basats en corpus (fonètica,

---

1 Aquest treball ha rebut el suport del projecte «Elaboració d'un corpus oral dialectal del valencià col·loquial (CorDiVal)», finançat per la Generalitat Valenciana (Ref. GV/2017/094).



morfosintaxi, semàntica i pragmàtica). Hi van participar com a ponents convidats Carlota de Benito (Universität Zürich), Peter Herbeck (Universität zu Köln), Salvador Pons (Universitat de València), Juanma Garrido (UNED), Mario Barcala (NLPgo) i també Jaume Corbera (Universitat de les Illes Balears) que, amb Vicent Beltran (Universitat d'Alacant) i Carles Segura (Universitat d'Alacant) van mantenir un diàleg sobre els corpus orals actuals del català.

Les pàgines següents arpleguen cinc treballs d'aquells participants que han disposat del temps i l'energia per a poder-los enllestir.

El treball de Sandra Montserrat i Carles Segura (Universitat d'Alacant) presenta la primera fase de la construcció del corpus PARLARS, un corpus oral del català col·loquial no interferit parlat al País Valencià que tracta de documentar aquesta varietat abans que no desaparega. L'interès d'aquest corpus, que té pretensions de continuar actiu, és, a més de permetre estudiar la llengua col·loquial des d'un punt de vista pragmàtic i funcional, estudiar la variació diatòpica en els termes tradicionals de la dialectologia i dotar la lingüística catalana d'una eina útil i representativa del valencià actual per a l'estudi de la variació (fonètica, morfosintàctica, semàntica i lèxica), a més de comprovar si les llengües funcionen com a estructures convencionalitzades, tal com explica la Gramàtica de Construccions. En el marc del projecte de recerca «Elaboració d'un corpus oral dialectal del valencià col·loquial (CorDiVal)» s'han dut a terme les tasques següents durant el 2017-2019: a) planificació del corpus; b) realització del treball de camp; c) transcripció i anotació del corpus; d) construcció de la interfície d'accés i cerca. En aquest article, però, se'n defineixen els objectius i les característiques principals, la metodologia de treball i el sistema de transcripció, anotació i consulta, així com també els resultats obtinguts fins al moment.

Per part seua, Renata Enghels, Fien De Latte i Linde Roels (Ghent University) presenten materials per a l'estudi sociolingüístic de l'espanyol col·loquial arplegats al CORMA (*Corpus Oral de Madrid*). El CORMA és un corpus d'espanyol peninsular que inclou converses espontànies de parlants madrilenys de diferents gèneres, edats (amb atenció especial al sociolèctes dels jòvens) i nivells socioculturals, enregistrades a Madrid entre el 2016 i el 2019, tenint en compte una àmplia gamma de trets sociolingüístics i situacionals. L'article fa un recorregut pels diferents passos realitzats per a crear el conjunt de dades i presenta les característiques principals del corpus final. També s'hi inclou una il·lustració de com es poden explotar amb èxit les dades mitjançant un estudi pilot de l'augment i la difusió del marcador d'intensificació *mazo* (*de*) en la parla jove de l'espanyol des de

principis del segle XXI, que ha disminuït la freqüència d'ús darrerament i s'ha convertit en un intensificador multifuncional que cada vegada acompleix més funcions pragmàtiques a mode de falca.

Des del País Basc, l'article d'Irantzu Epelde (Euskal Herriko Unibertsitatea) i Bernard Oyharçabal (Université Michel de Montaigne – Bordeaux III) se centra en la marca d'ergatiu en el canvi de codis en la parla espontània dels bilingües bascos dels dos costats de la frontera administrativa. En eusquera el subjecte d'un verb transitiu incorpora la marca d'ergatiu, de manera que aquest informa o avança que el verb que apareixerà tot seguit ha de regir un sintagma nominal que fa funció de complement directe. Així doncs, el basc és una llengua ergativa, on el subjecte de les construccions intransitives coincideix amb el complement directe de les transitives. D'acord amb les dades arreplegades al País Basc, amb 200 hores de gravació i 120 informants, els bascos, en canviar de llengua durant una conversa espontània, mostren alguns patrons clars en l'ús de l'ergatiu: utilitzen el determinant com a desencadenant principal en la marca ergativa, ja que és ací on es pot observar clarament quan el determinant és francès / espanyol. Hi ha quatre exemples amb verbs en francès / castellà que porten clarament el sufix ergatiu *-k*; en tots, la frase es dona íntegrament en basc; en els quatre, el subjecte es troba a l'esquerra del verb mentre que la resta d'elements estan en francès / espanyol. S'han tingut en compte sis perfils sociolingüístics a l'hora de seleccionar els informadors que s'han entrevistat: els majors de 55 anys (tant a França com a Espanya), els informadors d'entre 18 i 35 anys, i els menors de 18 anys. Gairebé tots els que s'han entrevistats eren bastant capaços de comunicar-se en basc i gairebé totes les converses eren en aquesta llengua.

Shima Salameh Jiménez (Universitat de València) focalitza el seu estudi en la construcció *pues eso* des del model Val.Es.Co., entesa com a construcció interactiva dins del marc de la Gramàtica de Construccions. Segons l'autora, *pues eso* no funciona de manera composicional, com si fora la unió del marcador discursiu *pues* i el dític *eso*, tot i que conserva propietats dels components heretats d'origen, sinó que ha adquirit nous trets i usos que demanen descripcions addicionals. Per provar aquesta idea, s'ha aplicat de manera sistemàtica el model de segmentació d'unitats i posicions Val.Es.Co. a les diferents característiques prosòdiques, sintàctiques i semàntiques connectades a les cinc funcions de *pues eso*. Aquesta combinació d'unitats, posicions i característiques explica les diferències entre *pues eso* i les formes originals a partir de les quals va evolucionar. Aquesta anàlisi sistemàtica sincrònica es completa amb un tractament diacrònic que explica com va desen-

volupar la seua trajectòria compositiva i la seua relació amb altres construccions semblants. La inclusió d'aquest model d'unitats de segmentació revela el motiu d'aquesta adquisició de trets que distingeixen *pues eso* dels seus components originals, posició i unitat, entesos com a trets de l'estructura externa de la construcció que perfilen les funcions de la construcció i concreten el caràcter continuatiu i díctic. Finalment, l'article permet plantejar-se si aquesta construcció és única o forma part d'una xarxa de construccions més ampla, una xarxa metadiscursiva en la qual es poden incloure altres estructures de l'espanyol, com ara *y nada, y eso, pues nada, y tal o no sé qué*.

Des d'una altra perspectiva, l'article de Juan María Garrido (UNED) fa un repàs de l'inventari d'eines automàtiques disponibles actualment, de pagament o de lliure distribució, que poden fer-se servir per a les diferents fases de desenvolupament i explotació de corpus orals en català, amb especial atenció a aquelles que poden ser més útils en l'àmbit de la fonètica; a més, presenta també un exemple concret d'ús d'aquestes eines per al desenvolupament d'un corpus oral en català. S'hi dona compte que el procés de construcció d'un corpus implica habitualment un esforç important que requereix tot un equip de persones i un treball de coordinació i formació importants si es fa de forma manual; d'altra banda, adverteix que actualment existeixen eines computacionals que permeten l'automatització total o parcial de determinades tasques en les diferents fases del procés, que redueixen la faena i els temps de desenvolupament. La revisió feta en aquest treball permet concloure que, tot i que poques, hi ha eines automàtiques per a determinades fases del desenvolupament i l'explotació de corpus orals en català, com ara la transcripció fonètica, l'annotació prosòdica, l'annotació lingüística o l'alineació amb el senyal acústic. Hi ha, però, altres àmbits, com la transcripció automàtica, on les eines no estan disponibles per al gran públic. També hi ha eines per a l'anàlisi de dades de corpus orals que permeten la realització d'estudis fonètics basats en gran quantitat de dades, al menys en àmbits com l'anàlisi prosòdica. Aquestes eines permeten alleugerir moltes fases del procés, però en el seu estat de desenvolupament encara requereixen una fase posterior de revisió manual per obtenir un resultat de qualitat. Cal encara, doncs, molts esforços de recerca per millorar el rendiment de les eines existents si es vol arribar a eliminar les tasques de revisió manual, i també per desenvolupar noves eines que permeten realitzar tasques encara no cobertes, com en el cas de l'annotació lingüística.

Tot plegat, aquesta publicació pretén oferir cinc perspectives diferents sobre la lingüística de Corpus Orals que permeten actualitzar la informació existent fins ara, que mostren les principals línies de treball dels investigadors contemporanis i que inciten a reflexionar sobre la manera d'afrontar els reptes que aquesta disciplina ens brinda en uns moments de canvis socials sobtats, davant la necessitat urgent d'arreplegar un material inèdit, en molts casos lligat a un estil de vida efímer, i d'adaptar els recursos existents a les noves tecnologies. ■

- Vicent Beltran, Universitat d'Alacant, Departament de Filologia Catalana, Campus de Sant Vicent del Raspeig s/n, E-03690 Alacant, <vicent.beltran@ua.es>.
- Maribel Guardiola, Universitat d'Alacant, Departament de Filologia Catalana, Campus de Sant Vicent del Raspeig s/n, E-03690 Alacant, <maribel.guardiola@ua.es>.