Treballs de Sociolingüística Catalana, núm. 25 (2015), p. 287-302 DOI: 10.2436/20.2504.01.103
http://revistes.iec.cat/index.php/TSC ISSN (ed. impresa): 0211-0784 ISSN (ed. electrònica): 2013-9136

Primers passos cap a la documentació de discurs signat: el projecte pilot de constitució del corpus de la llengua de signes catalana

First steps towards the documentation of signed discourse: the pilot project for the creation of the Catalan Sign Language corpus

Gemma Barberà,* Josep Quer** i Santiago Frigola***
* Centre National de Recerca Científica (CNRS) - Universitat de París VIII
** Institució Catalana de Recerca i Estudis Avançats - Universitat Pompeu Fabra
*** Universitat Pompeu Fabra

Data de recepció: 28 de maig de 2014
Data d'acceptació: 25 d'agost de 2014

Resum
Aquest article presenta la creació del projecte pilot del primer corpus de referència de la llengua de signes catalana (LSC). El corpus d'una llengua consisteix en una col·lecció representativa d'exemples en format llegible per una màquina i té com a objectiu oferir una visió del conjunt d'un domini lingüístic, ja que constitueix una representació àmplia de la llengua i de les seves varietats geogràfiques, de registre, generacionals i discursives. Pel fet de tractar-se d'un corpus de llengua de signes, les col·leccions de vídeos anotats contenen material sincronitzat amb les dades visuals enregistrades. Aquestes dades contribueixen a millorar les descripcions gramaticals i sociolingüístiques de la LSC, tot proporcionant una base sòlida per a l'anàlisi posterior, així com a fer possible una anàlisi basada en dades tant espontànies com induïdes. Aquí presentem la feina feta durant el procés d'elaboració d'aquest projecte, tant pel que fa a la metodologia d'obtenció de dades, a les filmacions i als materials que s'han utilitzat, com a l'anotació que s'ha dut a terme de les dades signades. Aquesta feina constitueix el pas previ necessari per al projecte de constitució del corpus de referència de la LSC.

Paraules clau: anotació, documentació, corpus, llengua de signes catalana (LSC), modalitat gestovisual, obtenció de dades.

Abstract
This article presents the creation process of the pilot project of the first reference corpus of Catalan Sign Language (LSC). A corpus of a particular language consists of a representative collection of language samples in a machine-readable format and aims to provide an overview of a linguistic domain, as it is a broad representation of the language and its varieties depending

Correspondència: Josep Quer. Universitat Pompeu Fabra. Departament de Traducció i Ciències del Llenguatge. Carrer de Roc Boronat, 18. 08018 Barcelona. A/e: *lsc@iec.cat.* A/I: *http://blogs.iec.cat/lsc/corpus/.* Tel.: 935 421 136.

on the geographical area, register, age and different discourse situations. As a sign language project, the video collections contain synchronised annotations aligned with the recorded visual data. These data contribute to improving the grammatical and sociolinguistic descriptions of LSC, providing a solid basis for further analysis based on spontaneous and elicited data. Here we present the work done during the process of preparation of this project, both in terms of the methodology of data collection, recordings and materials used, as well as the annotation procedure carried out. This work constitutes the first step necessary for the creation of the reference corpus project of LSC.

KEYWORDS: annotation, Catalan Sign Language (LSC), corpus, data collection, documentation, gestural-visual modality.

1. INTRODUCCIÓ

Malgrat l'important creixement de la recerca durant els darrers anys, el nostre coneixement lingüístic de la llengua de signes catalana (LSC) és encara limitat. Des que als anys cinquanta les llengües de signes van començar a ser reconegudes com a llengües de ple dret, tot un seguit d'estudis científics s'han esforçat a descriure'n i a analitzar-ne aspectes concrets, tant gramaticals com lèxics. Aquests estudis no només són importants pel valor intrínsec de presentar una descripció i anàlisi d'aquestes llengües gestovisuals pràcticament no documentades, sinó també perquè contribueixen a la difusió de llengües que formen part del patrimoni lingüístic i cultural de la societat en general. Tanmateix, aquests estudis lingüístics i sociolingüístics se centren generalment a descriure un fenomen particular i ho fan a partir de reculls restringits de dades que no poden donar compte de la variació existent en la llengua. En el cas de la LSC, no se n'ha pogut donar encara una visió extensa i global per la manca d'una mostra àmplia, representativa i general de la llengua. Aquest article presenta una descripció del projecte pilot de constitució del corpus de la LSC dut a terme per l'Institut d'Estudis Catalans (IEC) en el període 2012-2013. Explica la feina feta durant el procés d'elaboració d'aquest projecte, tant pel que fa a la metodologia d'obtenció de dades, a les filmacions i als materials que s'han utilitzat, com a l'anotació que s'ha dut a terme de les dades signades. Aquesta feina constitueix el pas previ necessari per a la constitució i l'establiment del projecte de corpus de referència.

2. LLENGÜES DE SIGNES I COMUNITAT SORDA

Les llengües de signes són els sistemes lingüístics propis de les persones sordes i sordcegues signants. Com a llengües naturals que són, emergeixen i evolucionen de manera espontània en el si de les comunitats de persones sordes i oients i ho fan amb independència de la llengua oral que es parla a l'entorn, si deixem de banda els fenò-

mens de contacte, típics de contextos multilingües. L'evolució natural i espontània fa que no n'existeixi una d'utilitzada universalment per tothom, sinó que existeixin innombrables llengües diferenciades, que també presenten variació dialectal. La majoria de llengües de signes són llengües minoritàries i, a més, sovint estan minoritzades. D'una banda, es consideren minoritàries perquè acostumen a tenir pocs signants, conviuen amb llengües orals majoritàries i presenten una transmissió discontínua entre generacions. D'altra banda, són llengües minoritzades perquè tradicionalment s'han associat a una discapacitat invisible, la sordesa, i això fa que sovint no hagin rebut el reconeixement lingüístic que mereixen ni la intervenció de les administracions per tal de normalitzar-les. Tot i que això està començant a canviar, molta gent encara veu només la patologia de la sordesa i desconeix la riquesa de la llengua de signes.

La LSC és una llengua viva més del conglomerat lingüístic de Catalunya. Tot i que les dades són aproximatives, segons la Federació de Persones Sordes de Catalunya (FESOCA), es calcula que actualment uns vint-i-cinc mil signants la utilitzen en la seva quotidianitat personal i laboral. Entre aquests signants, s'hi inclouen persones sordes i persones oients que per raons professionals i personals la utilitzen com a mitjà de comunicació preferent. La LSC té un abast territorial que coincideix a grans trets amb el del Principat de Catalunya, de manera que a la resta de l'Estat espanyol se'n signa una altra, la llengua de signes espanyola (LSE). Respecte de l'ús d'aquestes dues llengües, és important remarcar que a Catalunya no existeix un bilingüisme LSC-LSE, com sí que ocorre amb el català i el castellà. Així doncs, si una persona sorda signant de fora de Catalunya ve a viure aquí, adopta la LSC com a llengua exclusiva de comunicació, deixant de banda la seva llengua de signes d'origen.

Els grups de signants formen el que es coneix com a comunitat sorda, és a dir, aquella «comunitat de parla signant que engloba un conjunt d'individus que comparteixen una llengua i uns valors sociolingüístics comuns, com regles d'ús i actituds lingüístiques, així com també una ideologia i una actitud vers aquests usos lingüístics» (Gras, 2006: 5). Tot i així, l'edat de la primera exposició a la llengua, la llengua materna de casa, la varietat social i regional i l'escolarització són elements que contribueixen a donar una gran heterogeneïtat a aquesta comunitat. Un aspecte comú a les comunitats sordes és que la gran majoria dels seus membres van entrar en contacte amb la llengua durant l'escolarització, ja que gran part dels infants sords procedeixen de famílies oients (de manera estàndard es calcula que representen entre el 90 % i el 95 %, però aquest percentatge és probablement superior). A més, cal tenir també en compte que estem parlant d'una llengua que va estar prohibida a l'escola durant molt de temps i relegada, per tant, a l'ús familiar i informal. Les distàncies territorials i les polítiques educatives aplicades en diferents èpoques també han fet que la llengua evolucionés de manera diferent a cada regió i això fa que ens trobem davant d'una variació regional considerable. Tots aquests aspectes, cal tenir-los en compte a l'hora d'engegar qualsevol estudi lingüístic de la llengua. Coneixedors d'aquests fets i conscients que la llengua és un mosaic representatiu de tot un territori, els lingüistes hem de prendre les mesures necessàries per tal que la base empírica de la nostra recerca inclogui la realitat heterogènia de la llengua.

3. Importància dels corpus de llengües de signes

Un corpus de referència és una col·lecció representativa d'exemples d'una llengua, en format llegible per una màquina, que s'utilitza per a estudiar el tipus i la freqüència d'unitats i trets lingüístics. A més, en la mesura que es planteja oferir una visió del conjunt d'un domini lingüístic, constitueix una representació àmplia de la llengua i de les seves varietats geogràfiques, de registres i generacionals. En el cas particular de les llengües de signes, els corpus es caracteritzen per ser col·leccions de vídeos anotats que contenen material escrit alineat, és a dir, sincronitzat amb les dades principals de la llengua de signes en vídeo (Schembri i Crasborn, 2010). Els beneficis que aquests tipus de corpus presenten són, d'una banda, posar a l'abast de les comunitats científica i educativa un conjunt de dades que contenen una mostra àmplia i representativa de la llengua de signes en qüestió, llengües que en la gran majoria es caracteritzen per presentar una important variació, i, de l'altra, preservar-la com a part important del patrimoni social i lingüístic d'una societat.

De manera general, els corpus contribueixen a millorar les descripcions d'una llengua signada, tot proporcionant una base sòlida per a l'anàlisi posterior, així com a fer possible una anàlisi basada en dades tant espontànies com induïdes. Aquest conjunt de dades permet formular hipòtesis i preguntes de recerca relacionades amb cada un dels components gramaticals, com ara la fonologia, la morfologia, el lèxic, la sintaxi i el discurs, així com amb alguns aspectes de la variació dialectal i sociolingüística. Tanmateix, hi ha tres motius principals pels quals és important treballar amb dades signades provinents d'un corpus. En primer lloc, les llengües de signes són llengües joves de comunitats minoritàries que no disposen d'un sistema d'escriptura estàndard establert, així com tampoc dels estàndards de correcció desenvolupats que sovint acompanyen l'alfabetització (Johnston, 2010). En segon lloc, presenten una transmissió discontínua entre generacions a causa dels diferents estigmes relacionats amb la sordesa al llarg dels anys i del fet que la llengua té un nombre molt reduït de signants nadius (Costello, Fernández i Landa, 2008). En darrer lloc, l'anotació tradicional d'exemples signats esdevé sovint inaccessible per a alguns investigadors, ja que no tots els investigadors comparteixen les mateixes convencions d'anotació i perquè en la majoria dels casos les dades enregistrades no acompanyen els treballs publicats. Per aquests motius i amb l'objectiu de basar la recerca lingüística en dades de corpus, en els darrers deu anys, s'han iniciat i s'han desenvolupat diferents projectes de corpus per a llengües de signes.

La iniciativa del projecte de corpus de la LSC compta amb el valuós precedent de projectes europeus semblants que es troben davant de mancances comparables i que estan en fase de construcció, anotació o finalització, depenent del cas. L'experiència acumulada en aquests projectes ens permet avançar encara amb més solidesa i eficiència en la constitució del corpus de la LSC, sobre la base de criteris fiables. Així, el corpus de la llengua de signes australiana (AUSLAN) va ser el primer gran projecte de corpus d'una llengua de modalitat gestovisual. Es tracta d'una gran col·lecció de dades signades amb l'objectiu de disposar d'un arxiu d'una llengua amenaçada per

l'accelerat decreixement del nombre de signants (Johnston, 2010). El corpus de la llengua de signes holandesa (NGT) recull dades de signants de diferents regions del país i actualment disposa d'una anotació bàsica parcialment accessible en línia (Crasborn i Zwitserlood, 2008). El corpus de la llengua de signes britànica (BSL) té com a objectiu crear una col·lecció de vídeos signats en BSL per sords nadius d'arreu del Regne Unit i posar-los a disposició en línia per tal d'estudiar-ne la variació gramatical i lèxica (Schembri *et al.*, 2013). El corpus de la llengua de signes alemanya (DGS) té com a objectiu recollir dades signades i fer-ne una compilació per a l'elaboració d'un diccionari DGS-alemany (Blanck *et al.*, 2010). El corpus de la llengua de signes sueca (STS) es proposa, sobretot, posar a l'abast material per a l'ensenyament i l'aprenentatge d'aquesta llengua (Mesch i Wallin, 2008). Altres projectes en curs són el corpus de la llengua de signes francesa (LSF) (Balvet *et al.*, 2010), el corpus de la llengua de signes italiana (LIS) (Geraci *et al.*, 2011), el corpus de la llengua de signes eslovena (SZJ) i el corpus de les variants col·loquials de la llengua de signes japonesa (JSL), entre d'altres.

4. Projecte pilot de constitució del corpus de la LSC

4.1. *Objectius*

Els objectius principals del projecte de constitució de corpus de la LSC són, d'una banda, disposar d'un conjunt de dades representatives i naturals de la llengua, tot documentant-ne l'estat actual mitjançant una mostra àmplia i representativa de diferents tipus de discurs signat de signants nadius o quasi nadius; i de l'altra, desenvolupar una anotació bàsica del corpus com a punt de partida per a la recerca lingüística que ha de permetre avançar en el coneixement de la gramàtica i el lèxic de la LSC. Aquestes dues finalitats es proposen oferir una eina útil per a la recerca en general, no només de caire teòric, ja que permetrà disposar d'un conjunt de dades anotades sobre les quals basar descripcions i anàlisis per a conèixer millor la LSC, i també fer recerca aplicada, perquè ha de servir com a instrument de referència en l'elaboració d'obres lexicogràfiques i bases de dades lèxiques, en els treballs de síntesi de llengua de signes o en els programaris de traducció automàtica.

4.2. *Metodologia*

En col·laboració amb Política Lingüística de la Federació de Persones Sordes de Catalunya, en aquesta fase inicial es va determinar la ciutat del territori català on es durien a terme les primeres gravacions a partir dels perfils socials i lingüístics dels signants que les característiques del projecte requerien, a fi d'obtenir unes dades representatives. La selecció dels sis signants sords va ser feta a partir de criteris de gènere i edat, i tenint en compte que tinguessin familiars directes sords (pares i/o germans) o bé que haguessin estat internats en una escola específica d'infants sords. L'Associació

de Persones Sordes de Terrassa complia els criteris necessaris per a enregistrar la parella de signants home-dona del grup d'edat mitjana (30-50 anys) i la parella home-dona del grup d'edat adulta-gran (50-80 anys). La parella home-dona del grup de joves (18-30 anys) va ser escollida amb l'ajuda de la Comissió de Joventut de la FESOCA.

Seguint una pràctica comuna i reconeguda com a fonamental en la metodologia d'obtenció de dades de les llengües signades (Neidle *et al.*, 2000), els enregistraments van ser guiats en tot moment per la figura de l'entrevistador sord expert. El fet que l'entrevistador sigui una persona sorda aporta naturalitat a la sessió de filmació i evita que la llengua de signes que espontàniament signen els informants tingui influència estructural, lèxica o estilística de la llengua oral de l'entorn.

Prèviament a l'inici de la sessió d'enregistrament, l'entrevistador sord i la coordinadora del projecte van fer una presentació curta i explicativa sobre els objectius, les finalitats i els beneficis d'un projecte de corpus, de manera que els informants signants captessin l'objectiu primordial: intentar signar de la manera més natural possible i evitar qualsevol influència de nocions de correcció o gramaticalitat i de la llengua oral.

L'escenari de les filmacions estava disposat de la següent manera. Un fons de color llis i uniforme, utilitzat a manera de croma, feia de capçalera just darrere de l'esquena dels signants. Aquest fons llis dóna una bona qualitat d'imatge en l'edició dels vídeos i, així, a l'hora de manipular-los, la imatge de la configuració de les mans i del moviment del cos és més nítida. Els informants estaven asseguts de costat i una mica en diagonal, per tal que les dues càmeres (ubicades de manera creuada) poguessin filmar de cara el signant (figura 1). L'entrevistador sord estava situat entre les càmeres i darrere seu hi havia col·locats els focus de llum. Aquesta disposició és molt adequada, ja que evita que els informants mirin frontalment a la càmera i permet una col·locació molt adient a l'hora d'obtenir un entorn natural de conversa.

FIGURA 1
Disposició de l'escenari de les filmacions

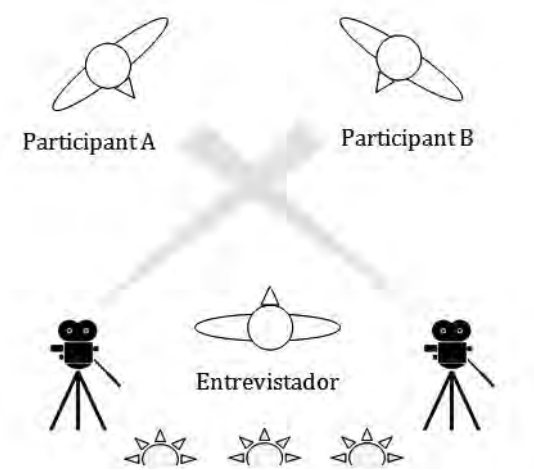


FIGURA 2
Instantànies de les filmacions

Per a les filmacions es van utilitzar dues càmeres panoràmiques d'alta definició que permeten fer una captura no només de l'activitat manual, sinó també dels moviments no manuals i dels gestos facials, de manera que per a l'anotació es puguin utilitzar dos arxius. D'una banda, l'arxiu de pla mitjà permet visualitzar el tors, els braços i el cap del signant en la seva extensió (figura 3). De l'altra, el pla de detall permet apreciar amb més precisió tots els moviments de l'activitat facial (figura 4).

FIGURA 3
Pla mitjà

FIGURA 4
Pla de detall

La preparació de les mostres de discurs signat que vam enregistrar es va fer seguint els protocols estàndard de qualitat tècnica i lingüística. Per tal de complir els requeriments ètics establerts per al treball lingüístic amb signants i per a l'arxivament i la difusió de les dades recopilades, vam seguir un protocol ètic que consistia en l'elaboració d'un document de protecció de dades i un document de metadades. El document de protecció de dades es va presentar de manera escrita i signada als participants perquè poguessin donar l'autorització per a l'enregistrament i la utilització de les dades recollides. D'altra banda, el document de metadades consigna informació personal per tal de determinar el perfil lingüístic del signant, que podrà ser objecte de cerques. Aquesta recollida de metadades s'ajusta al protocol estandarditzat ISCLE Meta Data Initiative (IMDI), que s'ha utilitzat en altres projectes de bases de dades lingüís-

tiques multimèdia i multimodals. El protocol permet l'assignació d'un codi anònim per a cada signant, necessari a causa de l'especificitat que suposa per a la privacitat dels participants no poder deslligar les dades lingüístiques en vídeo de la identitat del signant. Així doncs, les metadades personals de cada signant estan identificades amb un codi secret, que no en permet l'associació amb les dades reals del signant.

4.3. *Materials i tècniques d'obtenció de dades*

Cada sessió de gravació va incloure vuit tasques diferents, les quals tenien objectius diversos i utilitzaven, per tant, materials i tècniques d'obtenció de dades (inducció; en anglès, *elicitation*) diferents. A continuació, detallem cada tasca, per l'ordre d'ús que se'n va fer. La primera tasca consistia en una presentació individual, en què el signant expressava el seu nom i signom. A banda de servir d'escalfament per a signar davant les càmeres, aquesta tasca té com a objectiu documentar signoms i recollir i conservar aspectes característics de la cultura de la comunitat sorda.

La segona tasca consistia a explicar les vinyetes del còmic *La història de la granota* (Mayer, 1969). Es tracta d'un material estàtic basat en vinyetes, sense cap mena de text, que ha estat utilitzat extensament en recerca en llengües de signes i orals i, concretament, en altres corpus, tant signats com de llengües orals, per a fer anàlisis posteriors comparatives. L'objectiu d'aquesta tasca era disposar de dades de discurs narratiu, que s'esperava que contingués un important ús de l'espai, d'estructures de canvi de rol i de classificadors. Es pretenia també obtenir dades que permetessin una recerca comparativa amb altres llengües de signes i amb llengües orals. Per tal que no hi hagués influència a l'hora de narrar la historieta, cada signant va signar la història a l'entrevistador per separat.

La tercera tasca va consistir en la visualització d'un fragment del vídeo *Silvestre i Piuet* i la posterior explicació també a l'entrevistador d'aquest material dinàmic. La finalitat última era disposar d'una narració induïda que s'esperava que contingués una quantitat important de verbs de moviment i localització, classificadors i estructures de canvi de rol. El fet que aquest material s'hagi utilitzat en altres corpus signats i en corpus de gestualitat en la llengua parlada permetrà una recerca comparativa amb altres llengües de signes i amb dades de producció gestual.

La quarta tasca consistia a explicar una anècdota personal del passat relacionada amb el fet de ser sord. Aquest discurs narratiu/descriptiu obtingut sense cap material d'estímul serveix per a disposar de dades espontànies i per a documentar fets de la cultura de la comunitat sorda. Cada participant ho explicava a l'altre i l'entrevistador intervenia en els moments necessaris, tot creant una conversa espontània i distesa.

La cinquena i la sisena tasca van consistir en unes activitats relacionades amb materials visuals, en què els signants havien d'interaccionar directament. En el primer cas, l'informant *A* havia de reproduir el dibuix de l'informant *B* a partir de les seves explicacions. Els dibuixos se centraven en aliments, parts del cos i colors, per obtenir, així, lèxic d'aquests àmbits diferents en què generalment es considera que hi ha més

variació, en tractar-se d'àmbits quotidians on cada unitat familiar, escola o comunitat utilitza signes idiosincràtics. A més, aquesta tècnica serveix també per a induir preguntes.

La setena tasca consistia a explicar un vídeo mut de gènere dramàtic, per a un informant, i romàntic, per a l'altre. L'objectiu principal era obtenir dades per a analitzar l'expressió dels sentiments, l'ús de l'expressió facial en general, així com també les pressuposicions fetes a partir d'un material dinàmic mut.

Finalment, la darrera tasca consistia a discutir sobre un tema polèmic i força controvertit en la comunitat sorda actual. L'entrevistador obria el debat amb la pregunta de si calia mantenir les associacions de persones sordes. Amb l'objectiu de donar pas a una discussió emotiva i d'obtenir dades de tipus argumentatiu, els informants eren lliures d'opinar i exposar els seus arguments a favor i en contra de la pregunta d'obertura. En tot moment, l'entrevistador sord es va encarregar de controlar i gestionar la discussió, així com en totes les altres tasques que acabem de descriure breument. Tant en les tasques individuals com en les conjuntes, els informants sempre tenien l'entrevistador sord i el company com a interlocutors i en cap cas signaven a la càmera.

4.4. *Anotació*

Com ja hem esmentat a l'inici de l'article, la recerca lingüística de les llengües de signes no disposa encara d'uns criteris comuns d'anotació de les dades d'ús estès i generalitzat. A més, molts fenòmens de la LSC encara continuen inexplorats o amb un coneixement limitat. Amb aquest projecte hem pretès, doncs, establir unes convencions d'anotació generals i neutres que només es basin en els trets formals de la llengua i que, en la mesura del possible, evitin introduir qualsevol anàlisi lingüística, ja que idealment aquesta només s'hauria d'introduir posteriorment. Així doncs, les convencions d'anotació utilitzades es proposen servir com un sistema neutre amb un objectiu descriptiu (Dryer, 2006). Un cop l'anotació descriptiva està feta, cada lingüista, a partir del marc teòric que segueixi i els objectius concrets que es plantegi en relació amb les dades del corpus, podrà fer-ne una proposta d'anàlisi. Així, per exemple, s'ha fet poca recerca per a conèixer en detall quins són els criteris fonològics i morfosintàctics que diferencien els noms, els verbs i els adjectius. Per aquest motiu, hem optat per anotar la glossa a partir del nom i, així, evitar possibles errors d'etiquetatge gramatical induïts per la traducció cap a la llengua oral. Amb aquesta etiqueta, permetem que les cerques al corpus mostrin tots els exemples i que la recerca posterior estableixi les diferències de les categories lèxiques a partir d'una anàlisi exhaustiva de la major quantitat possible d'ocurrències. De la mateixa manera, per exemple, no s'ha volgut distingir entre els diferents subtipus de classificadors (semàntics, descriptius i de manipulació; vegeu Quer *et al.*, 2005) i només els hem marcat amb l'etiqueta *CL* i una descripció aproximada del classificador. Les diferències fonètiques també s'han identificat en cada anotació. Així, les variants dels signes que es distingeixen per un tret fonològic diferenciat apareixen marcades. Per exemple, el signe MARE es pot

articular amb la configuració que correspon a la lletra *N* de l'alfabet dactilològic (figura 5) o bé amb la configuració que correspon a la lletra *B* (figura 6). Les anotacions corresponents, per tant, s'han diferenciat amb les anotacions MARE(*N*) o MARE(*B*), respectivament. Amb la finalitat de facilitar les cerques posteriors, tots els moviments que pertanyen a l'àmbit gestual s'han marcat amb una *g* seguida d'una descripció entre cometes del significat del gest. Aquesta etiqueta facilita que qualsevol persona interessada en la gestualitat dins d'un discurs signat pugui discriminar tots els gestos que van més enllà del lèxic de la llengua.

FIGURA 5
Signe MARE amb la configuració N

FIGURA 6
Signe MARE amb la configuració B

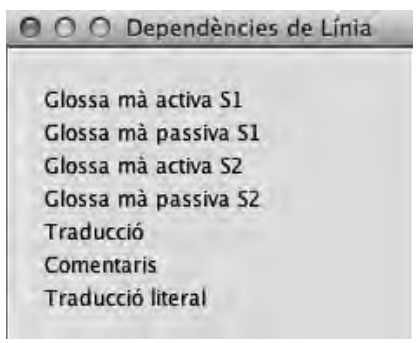
Per a les anotacions hem utilitzat el programa ELAN, un programa desenvolupat a l'Institut de Psicolingüística Max Planck Institute de Nimega (Països Baixos). Aquesta interfície, creada inicialment per a la transcripció del discurs parlat i de la gestualitat en les llengües orals, permet treballar les dades a partir de la transcripció sincronitzada amb la pantalla del vídeo signat. Es tracta d'una eina flexible i sofisticada que permet alinear anotacions complexes amb les dades originals de vídeo i fer-ne cerques a partir de les anotacions. A banda de les seves qualitats inherents, la selecció d'aquesta eina permet la intercanviabilitat amb altres projectes internacionals de corpus que estan emprant el mateix programari. Un avantatge addicional és la facilitat d'ús, un aspecte que simplifica l'entrenament dels anotadors pel que fa al vessant tècnic. El fet que la interfície estigui disponible en català i en castellà, entre altres llengües, facilita que puguin treballar-hi anotadors d'aquí i que s'afavoreixi, així, la participació de col·laboradors sords. Actualment l'ELAN és el programa d'anotació de referència en la comunitat d'investigadors de les llengües de signes, fet que ofereix l'avantatge d'inter-

canviar arxius amb facilitat. La figura 7 mostra una captura de pantalla de la interfície ELAN. Al marge superior esquerre es visualitzen els vídeos signats. A l'esquerra, veiem el vídeo amb el pla mitjà i a la dreta, el vídeo amb el pla de detall. A la meitat inferior, apareixen per separat cada una de les línies d'informació lingüística que hem establert, on transcrivim les glosses o bé fem una descripció en el cas del component no manual. Al costat superior dret del vídeo, podem escollir quina línia volem visualitzar.

FIGURA 7
Exemple d'anotació sincronitzada amb el vídeo signat

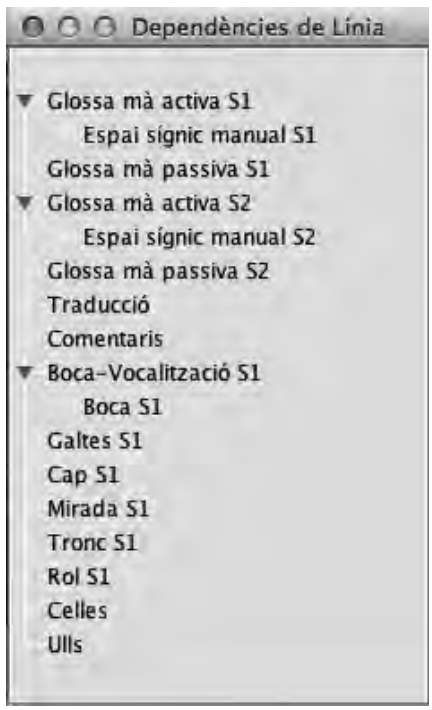
En la fase pilot hem volgut comparar el cost de recursos humans que suposa fer una anotació detallada de les dades amb una de bàsica. Les següents figures recullen les tires en cada versió. L'anotació bàsica (figura 8) inclou les glosses per a la mà activa (és a dir, la mà dominant del signe) i la mà passiva per a cada un dels signants, la traducció al català i els comentaris corresponents.

FIGURA 8
Tires lingüístiques de l'anotació bàsica



L'anotació complexa (figura 9) inclou aspectes molt més detallats, no només dels signes manuals, sinó també del component no manual (expressió facial, moviment del tors, etc.), l'ús de l'espai sígnic, així com també l'abast de les estructures de rol. Com era d'esperar, l'anotació complexa requereix un temps d'anotació molt més llarg que no pas l'anotació bàsica. De totes maneres, a l'hora de calcular la ràtio d'anotació no només cal tenir en compte el tipus d'anotació utilitzat, sinó també factors relacionats amb l'expertesa dels anotadors i amb el perfil del signants. Així, d'una banda, l'experiència i la formació prèvia en l'àmbit de l'anotació en la recerca contribueixen a un quocient menor entre el temps del vídeo i el temps necessari per a l'anotació. D'altra banda, els signants de la franja d'edat d'entre divuit i vint-i-cinc anys tendeixen a utilitzar unes estructures més aglutinadores i a fer més ús dels diferents articuladors de la llengua de signes. És a dir, els seus enunciats exploten més l'articulació simultània de les dues mans i dels diferents components no manuals: el cap, el tors, el moviment de les celles, el moviment de les galtes, el component bucal (moviments amb la boca, els llavis i la llengua) i el component oral (moviments que tenen l'origen en el gest de la pronunciació de la paraula corresponent al signe en la llengua parlada). Aquestes estructures tan complexes fan que la ràtio entre el temps d'anotació i la durada del vídeo sigui molt major en l'anotació dels vídeos del grup de joves en comparació dels vídeos dels grups d'edat mitjana i d'adulta-gran.

FIGURA 9
Tires lingüístiques de l'anotació complexa



Tal com havíem previst, en l'etapa de constitució del corpus només serà possible oferir una transcripció neutra i general que pugui ser útil per a objectius diversos de recerca. Aquesta transcripció bàsica del corpus podrà ser ampliada en el futur amb projectes dedicats a enriquir algun aspecte concret de la transcripció o bé per investigadors individuals que treballin en projectes específics. En la mesura que es consideri convenient, aquestes transcripcions podran incorporar-se posteriorment al corpus després d'haver estat supervisades.

5. Conclusions

Aquest corpus representa una contribució important en el camp de la recerca en la lingüística i la sociolingüística de les llengües de signes i de la comunitat de signants de Catalunya. En primer lloc, esdevé una eina necessària per a l'establiment de les bases per a l'estandardització de la LSC, tant des d'un punt de vista lingüístic, ja que intenta oferir una mostra de diferents varietats de la llengua que contribueixin a l'estudi de la gramàtica i del lèxic propis de la LSC, com des del punt de vista metodològic, ja que estableix uns criteris compartits, tot fixant unes convencions d'anotació. Contribueix, per tant, als objectius específics plantejats per a la normalització de la llengua en si, en el coneixement tant gramatical com lèxic de la llengua. En segon lloc, és una eina molt útil per a la recerca, tant teòrica, ja que proporciona una quantitat important i representativa de dades anotades, com aplicada, atès que ofereix una base sòlida per a la creació d'obres lexicogràfiques, per a l'activitat neològica i per a la traducció automàtica. A més, aquest corpus garanteix l'accés a materials de referència útils com a estris d'aprenentatge com a primera i com a segona llengua. Un aspecte valuós d'utilitzar-lo com a material d'aprenentatge és que conté diferents varietats de la LSC, així com també models de signants diversos, pel que fa a l'edat, al gènere, a les escoles on han anat o a l'associació de persones sordes a què estan afiliats. Finalment, el corpus fa visible la LSC d'una manera més general al conjunt de la societat i en millora l'estatus.

L'objectiu final de tots els corpus moderns de llengües, tant orals com signades, és que es construeixin i estiguin disponibles en un format llegible per una màquina. Per aquest motiu, les anotacions cal que siguin sistemàtiques i coherents en el seu conjunt (vegeu l'apartat 5), que els materials utilitzats siguin els que s'utilitzen en altres projectes per tal d'emprendre estudis comparatius (vegeu l'apartat 4), així com que estigui ben documentat, tant pel que fa al tractament de les dades de caire sociolingüístic, com a la tria de les diferents regions del país on s'obtenen mostres representativament àmplies de la llengua (vegeu l'apartat 3). Això requereix una tecnologia especialitzada, fornida especialment pel programa d'anotació ELAN. Tanmateix, també és del tot necessari que les etiquetes dels signes identifiquin cada un dels lemes de manera coherent i sistemàtica (Johnston, 2008). La identificació determinada de cada glossa passa per associar l'anotació a una base de dades lèxica que contingui cada lema concret. L'associació a una base de dades amb les glosses que identifiquen cada peça lèxica

(conegudes com a *ID-glosses*) és determinant per a diferenciar cada tipus de signe (*types*) de les diferents ocurrències (*tokens*), amb les corresponents variants fonètiques o variants produïdes per processos d'assimilació fonològica (és a dir, assimilació d'algun paràmetre formatiu del signe pel contacte amb el signe previ o posterior). L'associació a la base de dades lèxica és un aspecte necessari que caldrà incloure en el projecte definitiu de corpus de la LSC.

El nou marc legal, un cop aprovada la Llei de la llengua de signes catalana pel Parlament de Catalunya el 2010, estableix l'inici d'un procés de normalització en el qual s'atribueix a l'IEC un paper decisiu, no solament com a autoritat normativa sinó també com a entitat vertebradora de la recerca en LSC. Una condició indispensable per a dur a terme accions normalitzadores ben fonamentades és garantir la descripció i l'anàlisi bàsica de la llengua (Quer, 2010). Ara bé, de cara a desplegar les accions normalitzadores previstes en la Llei de la llengua de signes catalana del 2010, sembla ineludible crear un corpus de la llengua que sigui representatiu de tota la seva variació geogràfica, de registres i generacional (Gras, 2006; Jarque, 2012; Quer, 2010). Sense aquesta eina, es perpetuaria la situació actual, en què tenim una visió fragmentada i només parcialment representativa de la llengua emprada pel conjunt de la comunitat de signants. Això no solament suposa un greuge respecte a aquesta llengua pròpia de Catalunya i als seus usuaris, sinó que també ens impedeix donar la base científica necessària per al desenvolupament de materials destinats, per exemple, a l'educació d'infants signants o a la formació dels intèrprets de llengua de signes.

Agraïments

Aquest projecte de l'Institut d'Estudis Catalans ha rebut el suport logístic del Laboratori de Llengua de Signes Catalana del Departament de Traducció i de Ciències del Llenguatge de la Universitat Pompeu Fabra, així com també la col·laboració de la Federació de Persones Sordes de Catalunya. Volem agrair especialment la col·laboració i el suport que hem rebut de Política Lingüística de la FESOCA, l'Associació de Persones Sordes de Terrassa i la Comissió de Joventut de la FESOCA. Encara més especialment, volem agrair la col·laboració dels informants sords nadius que han participat en l'elaboració d'aquest projecte de corpus, sense els quals la feina d'obtenció de dades i de filmació no hauria estat possible. L'equip executiu del projecte pilot de constitució de corpus de la LSC ha estat format per les anotadores Delfina Aliaga i Noelia Hernández, l'expert en eines d'anotació Guillem Massó i la traductora i intèrpret Sara Costa. Aquest projecte ha rebut el finançament de l'Obra Social "la Caixa" a través del Departament de Cultura de la Generalitat de Catalunya.

Bibliografia de referència

Balvet, Antonio; Courtin, Cyril; Boutet, Dominique; Cuxac, Christian; Fusellier-Souza, Ivani; Garcia, Brigitte; L'Huillier, Marie-Thérèse; Sallandre, Marie-Anne (2010).

«The Creagest Project: a digitized and annotated corpus for French sign language (LSF) and natural gestural languages». A: *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*. París: ELDA, p. 469-475.

BLANCK, Dolly; HANKE, Thomas; HOFMANN, Ilona; HONG, Sung-Eun; JEZIORSKI, Olga; KLEYBOLDT, Thimo; KÖNIG, Lutz; KÖNIG, Susanne; KONRAD, Reiner; LANGER, Gabriele; NISHIO, Rie; RATHMANN, Christian; VORWERK, Stephanie; WAGNER, Sven (2010). «The DGS Corpus Project. Development of a corpus based electronic dictionary *German sign language – German*». Pòster presentat al congrés 10th Theoretical Issues in Sign Language Research (TISLR), Universitat Purdue, Indiana, Estats Units.

COSTELLO, Brendan; FERNÁNDEZ, Javier; LANDA, Alazne (2008). «The non-(existent) native signer: sign language research in a small population». A: QUADROS, Ronice (ed.). *Sign languages*: *Spinning and unravelling the past, present and future*: *TISLR 9*: *Papers from the 9th Theoretical Issues in Sign Language Research Conference, Florianopolis, Brazil, December 2006*. Petrópolis: Arara Azul, p. 77-94.

CRASBORN, Onno; ZWITSERLOOD, Inge. (2008). «The Corpus NGT: an online corpus for professionals and laymen». A: CRASBORN, Onno; HANKE, Thomas; EFTHIMIOU, Eleni; ZWITSERLOOD, Inge; THOUTENHOOFD, Ernst (ed.). *Construction and exploitation of sign language corpora*: *3rd Workshop on the Representation and Processing of Sign Languages*. París: ELDA, p. 44-49.

DRYER, Matthew (2006). «Descriptive theories, explanatory theories, and basic linguistic theory». A: AMEKA, Felix; DENCH, Alan; EVANS, Nicholas (ed.). *Catching language*: *Issues in grammar writing*. Berlín: Mouton de Gruyter, p. 207-234.

GERACI, Carlo; BATTAGLIA, Katia; CARDINALETTI, Anna; CECCHETTO, Carlo; DONATI, Caterina; GIUDICE, Serena; MEREGHETTI, Emiliano (2011). «The LIS Corpus Project. A discussion of sociolinguistic variation in the Lexicon». *Sign Language Studies*, vol. 11 (4), p. 528-574.

GRAS, Victòria (2006). «La comunidad sorda como comunidad lingüística: panorama sociolingüístico de la/s lengua/s de signos en España». Tesi doctoral. Universitat de Barcelona.

JARQUE, Maria-Josep (2012). «Las lenguas de signos: su estudio científico y reconocimiento legal». *Anuari de Filologia. Estudis de Lingüística*, núm. 2, p. 33-48.

JOHNSTON, Trevor (2008). «Corpus linguistics and signed languages: no lemmata, no corpus». A: CRASBORN, Onno; HANKE, Thomas; EFTHIMIOU, Eleni; ZWITSERLOOD, Inge; THOUTENHOOFD, Ernst (ed.). *Construction and exploitation of sign language corpora*: *3rd Workshop on the Representation and Processing of Sign Languages*. París: ELDA, p. 82-87.

— (2010). «From archive to corpus: transcription and annotation in the creation of signed language corpora». *International Journal of Corpus Linguistics*, vol. 15 (1), p. 104-129.

MAYER, Mercer. (2009*). Rana, ¿dónde estás?* Madrid: Los Cuatro Azules. [Edició original en anglès: 1969]

MESCH, Johanna; WALLIN, Lars (2008). «Use of sign language materials in teaching». A: CRASBORN, Onno; HANKE, Thomas; EFTHIMIOU, Eleni; ZWITSERLOOD, Inge; THOUTENHOOFD, Ernst (ed.). *Construction and exploitation of sign language corpora*: *3rd Workshop on the Representation and Processing of Sign Languages*. París: ELDA, p. 134-137.

NEIDLE, Carol; KEGL, Judy; MACLAUGHLIN, Dawn; BAHAN, Benjamin; LEE, Robert (2000). *The syntax of American sign language*. Cambridge: The MIT Press.

QUER, Josep (2010). «La normalització de les llengües de signes». A: MARTÍ, Joan; MESTRES, Josep Maria (ed.). *Les llengües de signes com a llengües minoritàries*: *Perspectives lingüístiques, socials i polítiques*. Barcelona: Institut d'Estudis Catalans, p. 239-255.

QUER, Josep; RONDONI, Eva-Maria; BARBERÀ, Gemma; FRIGOLA, Santiago; ALIAGA, Delfina; BORONAT, Josep; GIL, Joan; IGLESIAS, Pilar; MARTÍNEZ, Marina. (2005). *Gramàtica bàsica LSC*. Barcelona: FESOCA: DOMAD. També disponible en línia a: <http://blogs.iec.cat/lsc/gramatica/>.

SCHEMBRI, Adam; CRASBORN, Onno (2010). «Issues in creating annotation standards for sign language description». A: DREW, Philippe; EFTHIMIOU, Eleni; HANKE, Thomas; JOHNSTON, Trevor; MARTINEZ-RUIZ, Gregorio; SCHEMBRI, Adam (ed.). *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. París: ELDA, p. 212-216.

SCHEMBRI, Adam; FENLON, Jordan; RENTELIS, Ramas; REYNOLDS, Sarah; CORMIER, Kearsy (2013). «Building the British sign language corpus». *Language Documentation and Conservation*, núm. 7, p. 136-154.