CORILSE: a Spanish Sign Language Repository for Linguistic Analysis

Carmen Cabeza*, José M. Garcia-Miguel*, Carmen García-Mateo, José Luis Alba Castro****
Universidade de Vigo
*Facultade de Filoloxía e Tradución, Universidade de Vigo, 36310 Vigo (Spain)
**AtlantTIC Research Center, Escola de Enxeñaría de Telecomunicación, Universidade de Vigo, 36310 Vigo (Spain)
E-mail: cabeza@uvigo.es, gallego@uvigo.es, carmen.garcia@uvigo.es, jalba@gts.uvigo.es

Abstract

CORILSE is a computerized corpus of Spanish Sign Language (Lengua de Signos Española, LSE). It consists of a set of recordings from different discourse genres by Galician signers living in the city of Vigo. In this paper we describe its annotation system, developed on the basis of pre-existing ones (mostly the model of Auslan corpus). This includes primary annotation of id-glosses for manual signs, annotation of non-manual component, and secondary annotation of grammatical categories and relations, because this corpus is been built for grammatical analysis, in particular argument structures in LSE.
Up until this moment the annotation has been basically made by hand, which is a slow and time-consuming task. The need to facilitate this process leads us to engage in the development of automatic or semi-automatic tools for manual and facial recognition. Finally, we also present the web repository that will make the corpus available to different types of users, and will allow its exploitation for research purposes and other applications (e.g. teaching of LSE or design of tasks for signed language assessment).

Keywords: Spanish sign language, corpus development, linguistic annotation

1. Introduction

Spanish Sign Language (Lengua de Signos Española, LSE) can be considered an under-resourced language, which poses a major drawback for conducting linguistic research and for developing high-quality language technology (López-Ludeña, 2013). There are a number of limiting factors, among them standardization and difficulties in manual annotation. Therefore, the creation of modern language corpora builds on the development of proper tools for automatic sign language annotation.
In this work we discuss some strategies for the annotation of video clips which contribute to make our corpus usable for linguistic analysis, as well as for other applications, and describe the automatic tools we are developing to detect the regions of interest in the video, particularly eyes, lips and hands. All this has been made within the framework of the CORILSE ("Corpus informatizado para LSE", LSE computerized corpus) project that aims at building a high-quality software repository of LSE. Once completed, the corpus will be made publicly available for the benefit of researchers.
In this article, we present the two major components of the CORILSE project: (i) the video recordings and annotation files tailored to the particularities of LSE, and (ii) the software repository.

2. About Sign Language Corpora

The current research on sign language corpus is based on the widespread use of annotation tools as ELAN (Crasborn & Sloetjes, 2008), that enable to simultaneously visualize and annotate signing discourse. Some research teams have built sign language corpora from a set of video files annotated on ELAN and a lexical database. These previous works are our starting point: Auslan corpus (Johnston, 2010, 2013), NGT corpus (Crasborn & Meijer, 2012), and BSL corpus (Cormier et al., 2012)

Given that sign languages are supported on a visual-gestural modality, several articulators are simultaneously involved in their production (Meier, 2002; Aronoff et al., 2005). This fact has led researchers to make particular choices while glossing these languages. Sign language literature has generally favoured the use of capital letters for representing lexical units, highlighting the fact that orthographic words do not represent sound units, but lexical pieces of information in a visual modality language. Moreover, according to Johnston (2010), these capital letters words are interpreted as identifying glosses (ID-glosses) as they represent lemmas.

3. The Corpus

The CORILSE corpus is an output of the project RADIS, whose main goal is the study of predicates and argument structures in the grammar and discourse of LSE.
Although elaborating a corpus with representativeness and stratified criteria is not a specific goal in our project, we count on a growing body of recordings of LSE signing, produced by signers living in Vigo (Galicia, Spain). Our sample contains different discourse types (see table 1): most recordings consist of narratives, but there are also conversations or interviews. Elicited predicative expressions and "other" complete the corpus (see Table 1 for genres distribution and duration at the moment of writing this paper).

Genre	No_recordings	Duration
Conversation	3	00:19:03
Elicited examples	4	01:49:24
Interview	7	00:37:30
Narratives	18	01:20:02
Other	3	00:01:11
TOTAL	35	04:07:10

Table 1: Genre distribution in LSE Annotated Corpus.

A relevant part of the narratives has been recorded taking the picture book *Frog, where are you?* (Mayer 1969) and the video *The Pear Film* (Chafe 1980) as prompting resources for obtaining natural LSE narrative discourse. Both materials have been widely used in comparative linguistic research.

Other narratives included in the corpus are not the result of these elicitation resources. They can rather be considered as an example of deaf people's discourse and cultural practices. All interviews have been made by a deaf collaborator.

Elicited predicative expressions constitute a sample recorded with the aim of collecting data for a specific purpose of LSE grammar research: figuring out how different verb meanings are codified in LSE. Examples were recorded by presenting verb + arguments combinations to signers, as in (1). Participants were asked to verbalize them as LSE phrases:

(1) FOLLOW
Men → Women

Three sets of one hundred examples corresponding to the core verb meanings of the ValPaL project (Hartman et al. 2013) have been elicited. Other verb meanings and argument structures are to be found as contextualized items within the narratives. One of the main goals of our project is the annotation of predicate-argument relations and the compilation of a LSE lexicon of predicates and their semantic arguments (vid. infra §4.3).

4. Linguistic Annotation

We aim at enriching the whole corpus with a detailed annotation. Until now, the existing annotation has been carried out manually, which is a time-consuming task. An average of about 9 hours is spent for annotating 1 minute of recorded signed discourse –just primary annotation, non-manual components and secondary annotation not included. Therefore, at the moment we are we are exploring methods of automatic or semiautomatic recognition that can assist in the annotation of the outstanding recordings or in future stages.

The annotation process is performed in several phases and levels. A linguist with a thorough knowledge of LSE is responsible for the first phase, in collaboration with a Deaf person with extensive experience in tasks supporting research on LSE. With this assistance, the linguist solves the difficulties of interpretation. The Deaf collaborator has previously participated in the recording, conducting the interviews and explaining the objectives of research to the signing people being recorded. The outcome of the first annotation phase is primary annotation (described in section 4.1).

As to the second phase (see 4.2 below), the Deaf expert is responsible for completing the primary annotation by identifying and codifying the non-manual components (NMC)

In the third phase of the manual process of annotation, another linguist tags grammatical categories and argument structure as well as other important features for identifying referents in signing discourse, e.g. role shift or the location in signing space of entities holding some semantic role.

In order to ease this job, we have defined some preprocessing tools which will allow us to segment the regions of interest in the video as in Figure 1.

Automatic facial and hands detection will be superimposed in the video clip. The current version already includes the automatic detection of facial parts, as seen in Figure 1. Hence, zooming on specific parts will help to disambiguate some difficult annotations.

Figure 1: Screenshot of the automatic detection of facial elements

4.1 Primary Annotation of Manual Components

The primary level of annotation consists in a segmentation of discourse, followed by the identification and glossing of lexical and semi-lexical units, and a literal translation.

The manual annotation task, based on the Auslan project guidelines (Johnston, 2013), is being implemented in ELAN (Brugman & Russel, 2004). In addition, a lexical dictionary is being constructed with the aim of relating ID-glosses (Johnston, 2010) with their specific LSE articulation. In the medium term, this will be used to develop an automatic sign recognition system based on image processing.

Primary annotation requires the segmentation and lexical identification of tokens LSE discourse and involves the tiers *Ref* (Reference), *MD_Glosa* (Gloss for right hand), *MI_Glosa* (Gloss for left hand), and Trad (free translations into Spanish), as it is shown in Figure 2.

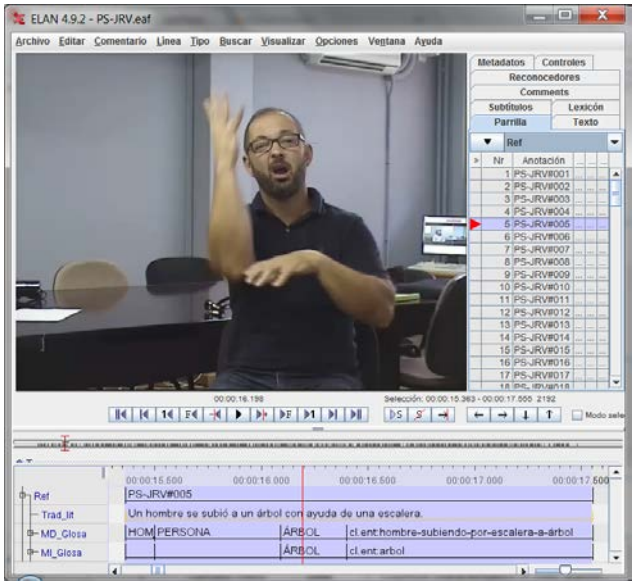

Figure 2: Primary annotation

A tier for each hand is needed since there are both monomanual and bimanual signs, and because in several cases each hand is articulating a different meaning.

The obtained tokens are tagged with ID-glosses, i.e. they are assigned to a type or lemma. Not all the elements resulting from LSE discourse segmentation are lexical units, and a set of particular annotation conventions is required for non-fully lexical signs, such as classifying or depicting signs, indexes, buoys, etc.

At the moment of writing up this text, we have completed the primary annotation of 17 recordings of the corpus, which correspond to duration of 3h 17m 50s. That means half the recordings currently available and 80% of the time that has been recorded so far. However, several annotation files still need to be thoroughly revised.

In this process of primary annotation we have already identified 4560 tokens and 1343 different types of signs articulated with the strong hand (the right hand), bimanual signs included. Among them, 637 types have been recognized as conventional lexical units and have become part of the lexical database. The remainder set of sign types is constituted by semi-lexical units, mainly depictive signs, also known as classifiers, as well as also indexes, buoys, and non-lexical gestures. For those semi-lexical units, the gloss consist of an identifier of unit type followed by a short description of its meaning (for example 'cl.ent:hombre-subiendo-por-escalera-a-árbol' = entity classifier:"man climbing a tree with a ladder").

The lexical database, in turn, contains a record of each ID-gloss and fields for its alternate meanings, its lexical category, and references to the corpus and to the LSE vocabulary repositories (DILSE, Sematos, Spreadthesign, etc.).

4.2 Annotation of Non-manual Components

The contribution of non-manual components (NMC) to sign language grammar and discourse has been widely recognized in the specialized literature (Nespor & Sandler 1999, Herrmann & Steinbach 2011). For example, while not aiming to be exhaustive, some of these NMC regulate the tracking of referents (eye gaze and body movement) and other identify the topic (eyebrows movement).

NMC are included as primary annotation devices that provide relevant information for secondary annotation. The following specific components are taken into account and annotated in different tiers: 1) eye gaze, 2) eyebrows-eyelids-nose movements, 3) oral gestures[1], 4) head, 5) shoulders and 6) body movements.

We have adopted an annotation system based on HamNoSys and the subsequent development by ViSiCast project (Hanke 2001).

This system consists of a set of easily usable alphanumeric codes. New tags are added to the original ViSiCast ones as new needs appear during the annotation process of LSE. For example, for the codification of eyebrow movements (annotated in the same tier as eyelids and nose), ViSiCast provided four codes:

RB (both eyebrows raised[2]):

Figure 3

RR (right eyebrow raised):

Figure 4

RL (left eyebrow raised):

Figure 5

FU (eyebrows furrowed):

Figure 6

New codes have been added to the original ones. They are as follow:

[1] Two types of mouth patterns are generally considered in signed language research: those related to the spoken contact language ('oralizations') and specific mouth gestures formed from within the sign language (Sutton-Spence & Boyes Braem 2001). Both are considered in our analysis.

[2] Drawings by Juan-Ramon Valiño (with consent ~~permission~~).

BA (basic neutral position):

Figure 7

CF (raised eyebrows and furrowed brow):

Figure 8

CFR (right eyebrow raised and furrowed):

Figure 9

CFL (left eyebrow raised and furrowed):

Figure 10

Figure 3 provides an example of NMC annotation in ELAN. Tiers for initial segmentation (*Ref*), translation into Spanish (*Trad*), and glosses for each hand (MD_Glosa, MI_Glosa) have been completed with others for eye-gaze (*Mirada*), eyebrows, lids, and nose (*Ojos_cejas*), oral gestures (*Gesto_boca*), torso movement (*Cuerpo*), shoulders (*Hombro*), and head movement (*Cabeza*), which include annotations similar to those exemplified for eyebrows.

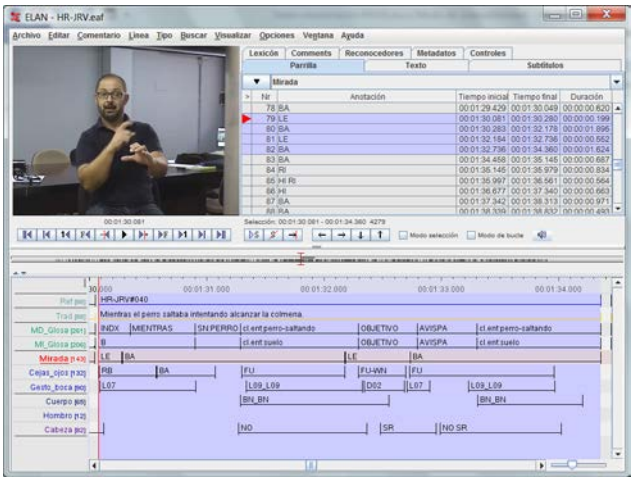

Figure 3: Annotation of non-manual components

4.3. Secondary Annotation

Secondary annotation aims at grammar description and linguistic typology research about predicates and argument structure, and it is consequently designed according to the specific object of study.

The ELAN tiers that are relevant for secondary annotation involve the following parameters:

- Lexical category (N, V, Indx…)
- Locus (relevant signing space locations for the articulation of arguments)
- Role shift (also referred as *constructed action* / *constructed dialogue*, *enactment* or *perspective shift* in sign language literature)
- Argument structure (predicates and arguments expressed by hand articulation, locus, and/or role-shift)

The annotation of lexical category is made semiautomatically using a simple script that chooses the category or categories of each item from the lexical database. This initial attribution is revised manually at the same time that the whole set of secondary annotation is applied manually and the primary annotation and segmentation are also revised.

Table 2 is a partial example of secondary annotation of a segment identified as predicative expression (a 'clause-like unit' or CLU, to borrow a term used by Johnston 2013 and Hodge 2013). The table includes glosses for right-hand and left-hand articulations, lexical category (only for right hand) and argument role for each hand.

Ref	PS-JRV#005			
	A man is climbing a tree (with a ladder)			
RH	**MAN**	**PERSON**	**TREE**	**cl.ent: man-climbing-tree**
	N.P	N.Loc	N	V.D
	A1		A2	V_A1
LH			**TREE'**	**cl.ent: tree**
			A2	A2

Table 2. A partial sample of secondary annotation

In the example, the predicate is a depicting verb (V.D) with two arguments. The first argument (A1) is lexically expressed by two signs articulated with the right hand (a plain noun –N.P– and a locatable noun –N.P–) and by a classifier also articulated by the right hand and incorporated into the predicate. The second argument (A2) is expressed by a lexical bimanual sign, and also by a classifier articulated with the left hand which provides a reference for the movement of the classifier articulated with the right hand.

The numbering of the arguments does not follow the order of appearance (as in Hodge 2013), but it is intended to be consistent across different uses of the same predicate. In the example, A1 should correspond to the climber and A2 to the climbing goal. In this sense, A1 and A2 represent verb-specific roles or 'micro-roles', but they can also be understood in a more general sense, because for each verb A1 corresponds to the more agent-like or subject-like argument, and A2 to the more undergoer-like or object-like argument. On the order hand, the annotation of argument roles is not limited to overt –nominal- arguments (as in Hodge 2013). In LSE, as in other sign languages, indicating directional predicates, depicting predicates and role-shift are usual resources used to point to a participant in an event, and our annotation system also contemplates these means of expression.

The grammatical annotation of the corpus, using ELAN, is complemented by a database of predicates and arguments. It is here where predicates are semantically classified, and where the semantic role of each argument (A1, A2, …) of each predicate is defined both at a more specific level (microroles such as 'climber' and 'climbing goal') and at a more general level (macroroles such as Actor and Undergoer). To date, we have defined 250 different predicates, with 525 arguments. For comparative purposes, this database is linked to the data of the typological ValPaL database (Hartmann et al. 2013) and the ADESSE database of Spanish verbs and constructions (García-Miguel et. al 2010).

5. Repository

The whole corpus and essential tiers of primary and secondary annotations will be made freely available through an online repository. The repository includes a structured database, a graphical user interface and a number of image processing tools. The use of a database allows a simpler and faster search through different criteria. Thanks to the web-based user interface the user can gain easier access to the different materials and the administration tasks. Regarding the software repository, the database is written in MySQL language. The user interface has a client-server architecture. The client is programmed in HTML5, using JavaScript and JQuery library while the server is mainly written in PHP. This configuration allows access to the database using a web browser. There is an administrator user who is mainly in charge of updating the database information and of uploading new contents to the server. Furthermore, the image processing tools have a database component and a web component. The automatic processing is performed off-line and the results are incorporated to the database: facial parts detection (already implemented) and hands detection (work in progress). Facial parts detection consist of a face detector and an iterative 3D deformable model adaptation, composed of 68 facial points, including eyes, eyebrows, nose, mouth and face contour. The web-based user interface integrates the zooming-to-specific-parts tool. Once the search results are presented to the user, two options are available: 1) playing the excerpt of the video clip (with the detected areas of interest as an option), 2) downloading the ELAN file with the excerpt.

Figure 11

6. Uses

CORILSE expected applications are not restricted to sign language research. It has an unquestionably usage for sign language teaching (either to children or adults, to deaf community or hearing individuals). It can also provide natural and codified examples for elaborating assessment tests for specific uses (interpreters training, children development, sign language loss caused by neurological injury, etc.)

7. Conclusions and Further Work

CORILSE project was born with two purposes: firstly, to make available to users an annotated corpus of LSE, which will be extended and enriched in successive stages. Secondly, we intend to develop automatic or semiautomatic tools to facilitate the task of manual annotation.

In the preceding lines we have described our annotation system, designed on the model of Auslan corpus. Its initial aim is the description of the argument structure of LSE predicates.

CORILSE is an ongoing project with the following medium-term goals:

- To continuously increase the number of annotated recordings
- To refine the coding of secondary annotation tiers
- To incorporate other tiers, already annotated manually, into the CORILSE repository
- To use some excerpts of manual annotation to assess the automatic video recognition
- Integration of video processing tools as enabling instruments for speeding up the manual annotation

8. Acknowledgements

Other members of the research team are Juan Ramón Valiño, Francisco Eijo, Ania Pérez, Cristina Freire, Rosalía Barros, Alberto Rodriguez, Marta Martínez and Rocío Varela. We also acknowledge the collaboration of Vigo Deaf Association (Asorvigo).

CORILSE is an output of the RADIS project, funded by the Spanish Ministry of Economy and Competitivity (FFI2013-41929-P).

This research was supported by the Galician Government through the research contracts: GRC2014/024 (Modalidade: Grupos de Referencia Competitiva 2014), the "Rede Tecnoloxías e Análise dos Datos Lingüísticos (TECANDALI)" and "AtlantTIC Project' CN2012/160, and the European Regional Development Fund (ERDF).

9. Bibliographical References

Aronoff, M., Meir, I. & Sandler, W. (2005). The Paradox of Sign Language Morphology. *Language*, 81(2), 301-344.

Brugman, H., Russel, A. (2004). Annotating Multimedia/ Multi-modal resources with ELAN. *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*.

Chafe, Wallace (ed.) (1980). *The pear stories: cognitive, cultural and linguistic aspects of narrative production*. Norwood, New Jersey: Ablex

Cormier, K., Fenlon, J, Johnston, T., Rentelis, R., Schembri, A., Rowley, K., Adam, R. & Woll, B. (2012). From Corpus to Lexical Database to Online Dictionary: Issues in Annotation of the BSL Corpus and the Development of BSL SignBank. *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Language Resources and Evaluation Conference*, pp. 7-12.

Crasborn, O. & de Meijer, A. (2012). From Corpus to Lexicon: The Creation of ID-Glosses for the Corpus NGT. *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Language Resources and Evaluation Conference*, pp. 13-17.

Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. Language Resources and Evaluation Conference*, pp. 39-43.

García-Miguel, J.M., Vaamonde, G. & González Domínguez, F. (2010). ADESSE, a Database with Syntactic and Semantic Annotation of a Corpus of Spanish. *Proceedings of LREC2010. Seventh International Conference on Language Resources and Evaluation*, 1903–1910.

Hanke, T. (Ed.) (2001). ViSiCAST Deliverable D5-1: Interface Definitions. Available online at http://www.visicast.cmp.uea.ac.uk/Papers/ViSiCAST_D5-1v017rev2.pdf, (accessed on 2015-11-10)

Herrmann, A., & Steinbach, M. (Eds.) (2011). *Nonmanuals in Sign Language*. (Special issue of *Sign Language & Linguistics* 14:1). Amsterdam: John Benjamins Pub. Co.

Hodge, G. (2013). Patterns from a signed language corpus: Clause-like units in Auslan (Australian sign language). Ph.D. thesis, Sydney: Macquarie University.

Johnston, T. (2010): From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1), 106–131.

Johnston, T. (2013). Auslan Corpus Annotation Guidelines. Retrieved from http://media.auslan.org.au/attachments/AuslanCorpusAnnotationGuidelines_Johnston.pdf.

López-Ludeña,V., San-Segundo, R., González Morcillo, C., López, J.C., Pardo Muñoz J.M. (2013). Increasing adaptability of a speech into sign language translation system. Elsevier *Expert Systems with Applications*, 40, pp. 1312–1322.

Mayer, M. (1969). *Frog, where are you?* New York: Dial Press.

Meier, R. (2002). Why Different, Why the Same? Explaining Effects and Non-Effects of Modality upon Linguistic Structure in Sign and Speech. In R. P. Meier, K. Cormier, D. Quinto-Pozos (Eds.), *Modality and Structure in Signed and Spoken Languages*. Cambridge: Cambridge University Press, pp. 1-25.

Nespor, Marina & Sandler, Wendie (1999). Prosody in Israeli Sign Language. *Language and Speech*, *42*(2-3), 143–176.

Sutton-Spence, R. & Boyes Braem, P. (2001). Introduction. In R. Sutton-Spence, P. Boyes Braem (Eds.), *The Hands are the Head of the Mouth. The Mouth as Articulator in Sign Languages*. Hamburg: Signum Verlag, pp. 1–7.

10. Language Resource References

ADESSE = *Bases de datos de verbos, alternancias de diátesis y esquemas sintáctico semánticos del español.* Universidade de Vigo. http://adesse.uvigo.es

AusLan Corpus = *Australian Sign Language Corpus*. T. Johnston (dir). Macquarie University http://elar.soas.ac.uk/deposit/0001

BSL Corpus = *British Sign Language Corpus Project*. http://www.bslcorpusproject.org/

NGT Corpus= *Corpus Nederlandse Gebarentaal.* Radboud University. http://www.ru.nl/corpusngt/

ValPaL= Hartman, I., Haspelmath, M & Taylor, B. (Eds.) (2013). *Valency Patterns Leipzig*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://valpal.info (accessed on 2015-11-10)

ELAN. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands. http://tla.mpi.nl/tools/tla-tools/elan/