

Technical, linguistic and sociological difficulties in the development of a LSE corpus

Patricia Álvarez Sánchez, Inmaculada C. Báez Montero
y Ana M^a Fernández Soneira

Universidade de Vigo
SPAIN



1 > Introduction

2 > Data

3 > Analysis

4 > Future projects

5 > Acknowledgments

6 > Bibliography

Let's see a sample view of our web page...



[Grupo de Investigación
sobre Lenguas Signadas](#)

[Quienes somos](#)
[Líneas de trabajo](#)

[Publicaciones](#)
[Proyectos](#)
[Cursos](#)
[Corpus](#)

[Colaboraciones](#)
[Bibliografía](#)
[Enlaces](#)

¿QUIÉNES SOMOS?

Empezamos a trabajar sobre la LSE en el año 1995. Desde entonces, el grupo de investigación ha asumido como propio el reto de superar las barreras de comunicación que afronta la comunidad sorda. El desarrollo de la investigación ha estado siempre vinculado a las personas sordas. En las primeras etapas nos acercamos a las asociaciones de sordos. Después apostamos por la presencia de las personas sordas en la Universidad, formando parte del grupo de investigación y con una responsabilidad plena en las tareas.

Ahora estamos en un momento de consolidación y de expansión.

Coordinación y dirección:

Báez Montero, Inmaculada: cbaez@uvigo.es [curriculum](#)

Cabeza Pereiro, Carmen: cabeza@uvigo.es [curriculum](#)

Investigadores:

Álvarez Sánchez, Patricia: patri.alvarez@gmail.com

Corvo Sánchez, M^a José: mcorvo@uvigo.es

Fernández Soneira, Ana: anafe@uvigo.es [curriculum](#)

Iglesias Lago, Silvia: siglesias@uvigo.es

Lorenzo García, Lourdes: llorenzo@uvigo.es

Mateos, Berta: bertamateos@yahoo.es

Pereira Rodríguez, Ana M^a: pereira@uvigo.es [curriculum](#)

Pérez Rodríguez, M^a Rosa: rosa@uvigo.es

Valverde, Rocío: rociovm@uvigo.es

Colaboradores externos y expertos en LSE:

Cabeza Pereiro, Elena: elena.cabeza@gowww.usc.es [curriculum](#)

Cardama Barrientos, Jose C.: jcardama@correo.cop.es [curriculum](#)

Eijo Santos, Francisco: Frank_eijo@hotmail.com

Valiño Freire, Juan Ramón: juan_deaf@yahoo.es [curriculum](#)

Our long journey...

- ❖ The first stage covers our group work from 1995 until 2000 and it represents the beginning of the process.
- ❖ The second phase goes from 2000 till 2007. It was stressed by an analysis process of the work done, and reconsiderations on our basis due to the problems arisen in the first stage.
- ❖ The third and last stage corresponds to present time. It is the time of showing our advances and the decisions made on the linguistic, sociolinguistic and technical aspects.

Our corpus...

What is it?

A group of real linguistic data that is stored in an electronic medium (computer, video, etc.) gathered with the aim of studying a language or a variety of it.

How is it?

It is representative of different social groups and linguistic styles. It has also been computerized.

What do we need it for?

It makes possible the grammatical description of a language based on real data. It makes also easy to compare the different sign languages.

1 > Introduction

2 > Data

3 > Analysis

4 > Current state

5 > Acknowledgments

6 > Bibliography

Informants' data

We have developed an interview filing card with the purpose of ascertaining the social and linguistic profile of the Galician deaf people that were later registered in videotapes.

The questions in the cards are easily understandable and the identity of the informant is safeguarded in every case.

This is the data gathered from our 85 informants:

1. **Identification:** name, address and phone (for future contacts);
2. **Origin and social environment:** place and date of birth, age of deafness occurrence, deafness degree, deaf/hearing family, job of closest family members;
3. **School:** degree and type of studies, special/ordinary school, use/absence of SL in school;
4. **Linguistic skills:** in LSE, in oral Spanish, lip-reading, written Spanish;
5. **Place of residence:** in order to reflect and control linguistic variation.

Let's see a sample view of our informants' database...

1 Identificación	Nombre	Alfonso	Santos Feijoo
3 Formación	Hasta que edad estudio	a los 21 años	
	Profesión:	Delineante (Paro)	
	Otros estudios	Delineante Proyecto y Construcción	
	Escuelas a las que asistió:		
	Fue algun colegio de sordos	<input checked="" type="radio"/> si <input type="radio"/> no	¿Cual?: Santiago y Madrid
	¿A que edad?:	6 al 16	Los profesores le hablaban L:S <input checked="" type="radio"/> SI <input type="radio"/> NO <input type="radio"/> MUY POCO
	¿Fue a algun colegio de oyentes?:	<input checked="" type="radio"/> si <input type="radio"/> no	¿A cual?: Orense y Vigo
	¿A que edad?:	16 al 21	¿Alguien le hablaba en L:S? <input type="radio"/> SI <input checked="" type="radio"/> NO <input type="radio"/> MUY POCO
	¿Quien?:		
4 Lenguas	L:S:E NIVEL DE DESTREZAS:	Usa la LSE en cualquier situación	
	ESPAÑOL ORAL NIVEL	Entiende por lectura labial a cualquier	
	ESPAÑOL ESCRITO NIVEL:	Escribe algo.	
5.Residencia	Tiempo de residencia aqui:	Vigo, 26 años	
	Lugares de residencia y tiempo:	Orense , 9 años. Madrid , 5 años. Madrid 5 años (alumno interno)	
Menu	Nuevo registro	Buscar	Ordenar
Identificación y procedencia	Formación y lenguas	SALIR	

Language samples

❖ **Guided monologue**

We ask the deaf informant to describe his family and his house, to tell a short story, etc. Thus, we obtain both description and narration, but the results are not completely natural.

❖ **Semiguided interview**

Signers are interviewed (deaf people, most of all) on diverse topics, depending on their age, preferences, etc. The objective is to obtain natural signing, as spontaneous as possible.

❖ **Public discourse (conferences, round tables, etc.)**

These give us programmed discourse in a more formal style.

Guided monologue sample



Semiguידed interview sample



Public discourse sample



Our aims...

Our work has focused on obtaining a LSE textual corpus of Galician signers from which to start the research on LSE. These were our initial researching aims:

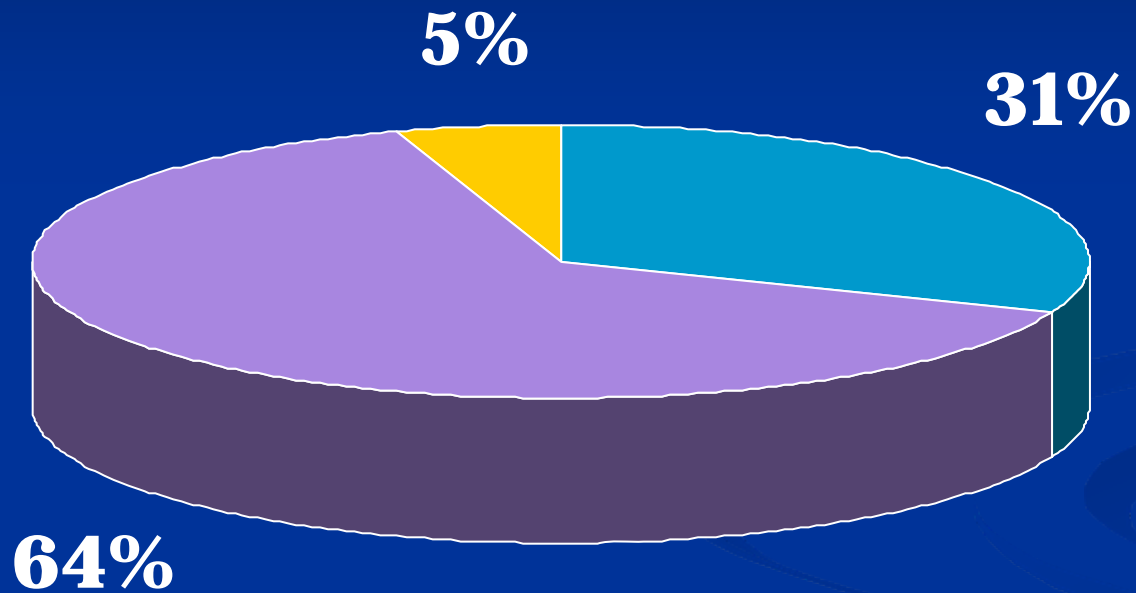
- ❖ Starting the description of LSE
- ❖ Determining which are the relevant linguistic units in SL
- ❖ Understanding the grammatical relational processes
- ❖ Developing tools for research: labeling, transcription, etc.

Corpus features

We considered these the main features of a corpus:

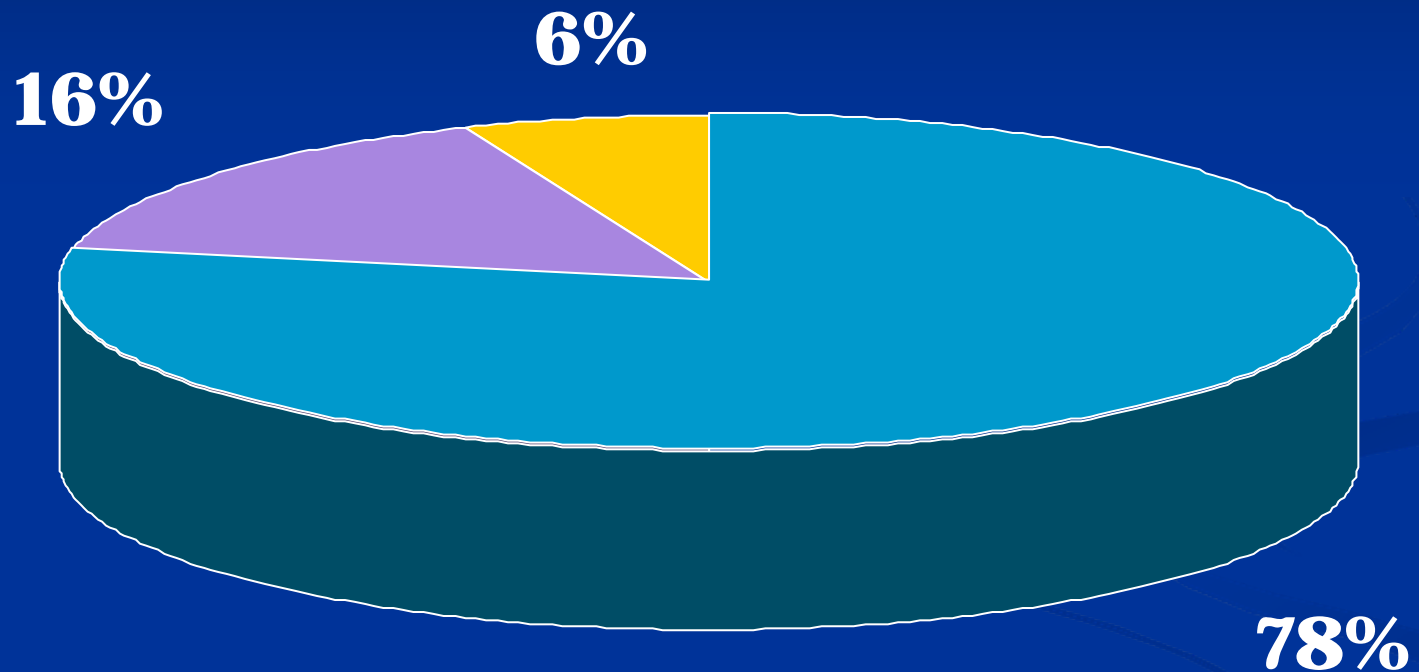
- ❖ It must contain real data.
- ❖ It must constitute an irreplaceable basis for linguistic description.
- ❖ It must be completed with computing support in order to make easy its use.
- ❖ It must gather:
 - a) Informants data
 - b) Different types of discourse samples
 - c) Wide range of topics depending on the type of discourse we want to obtain, etc.
- ❖ It must be transcribed in Spanish glosses (conventions adapted from Klima & Bellugi, 1979) and subtitled in written language.

Type of discourse distribution



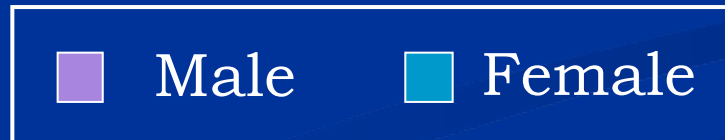
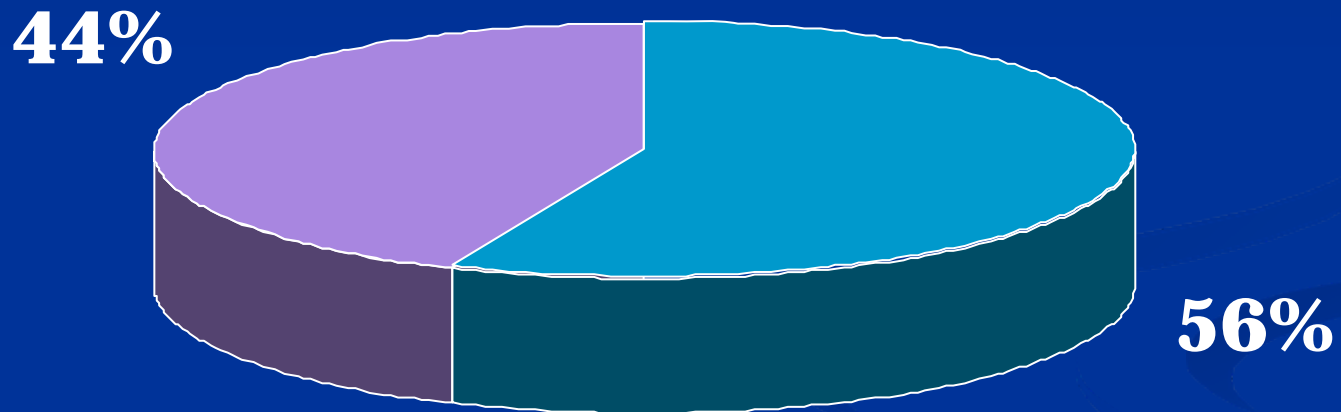
■ Guided Monologue ■ Semiguided interview ■ Public discourse

Age distribution



■ 21-34 years ■ 35-50 years ■ > 50 years

Gender distribution



Process stages

We have divided into seven stages the process of creating our corpus:

- a. Tool design for the creation of a corpus
- b. Criteria for the selection of informants
- c. Creation of a database of informants' details
- d. Collection of language samples
- e. Data storage
- f. Data labeling and marking
- g. Transcription and notation systems

Difficulties in the process

- ❖ The lack of a research tradition on Sign Languages in Corpus Linguistics forces us to solve problems from the beginning:
 - How to delimit units in sign languages.
 - How to label the different formations for their later analysis.
 - Other related issues.
- ❖ Creation of social networks in the Deaf community with the aim of avoiding the social identity of our informants to be threatened.
- ❖ Technical restrictions. We have to select appropriate material in order to avoid problems in compatibility between the different devices (video cameras, computers, software...)

1 > Introduction

2 > Data

3 > Analysis

4 > Future projects

5 > Acknowledgments

6 > Bibliography

Analysis of the interviews

We have put different strategies into practice for the recording of the interviews. These are the results of our experience:

1st strategy: hearing interviewer and deaf informant

- ❖ The results are not too natural
- ❖ “Foreign speaking” hazard
- ❖ Limited success

2nd strategy: contribution of an interpreter

- ❖ The participation of an interpreter in the interaction between the researcher and the informant is of great help. They contribute to creating a comfortable atmosphere.

3rd strategy: deaf interviewer and deaf informant

- ❖ The performance of deaf people as interviewers contributes to more natural dialogues and it involves the deaf community in the project.

Reconsiderations

After the research, we had to reconsider certain issues for a better development of our corpus.

- ❖ Revision of the projects carried out in other countries.
- ❖ Creation of social networks:

Inside the Deaf community:

Preparation of the members of the community for the carrying out of the interviews

In the institutions:

Participation in national networks in order to contact with the Deaf community all over Spain.

Support of the LSE Standardization Center in the creation of the corpus.

Linguistic and Sociolinguistic Decisions

LSE is not an standardized language and there are very few descriptive studies on this language. This forces us to propose what kind of recordings do we want, how many people do we need in order for the corpus to be representative and real, and finally, what conclusive analysis could we obtain from it.

Taking into account these determining factors, we have decided:

- Ask for the collaboration of signers of different regions of Spain for obtaining a good representation of the different geographic registers.
- Select signers that fulfill certain features: native signers of LSE, post lingual users of LSE and interpreters.
- Choose deaf interviewers. Their dialogues are more natural and they obtain a higher degree of involvement of the Deaf community in this project.
- Recordings should be adapted to the personality of the informants. We have prepared several models of the interview, with questions that may arouse interest in the informants (on deafness, family, friends, tobacco, etc.)

Technical decisions with a view to the future

- 1. Standardization of the recording format.** Use of a recording set: similar wall background, light conditions, signers clothes, position and framing.
- 2. Multiple views of the signer:** face, trunk, in profile, etc.
- 3. Storage and backup** of the recordings from the camera to the computer.
- 4. Editing** of the recordings in chapters for a better handling of the images.
- 5. Use of the ELAN system** for the notation process.
- 6. Corpus labeling** of grammatical features and sign configuration.
- 7. Use of P2P tools** for making easy the cooperation between universities or research groups.
- 8. Enable the search and retrieval** by sign configuration, grammatical aspects and signer details.
- 9. Online publishing** of the corpus with the aid of external financing.

1 > Introduction

2 > Data

3 > Analysis

4 > Future projects

5 > Acknowledgments

6 > Bibliography

Advances

- ❖ We are members of a network of universities for the teaching and research on Spanish and Catalan Sign Languages (RIID-LLSS).
- ❖ We collaborate in the creation of a LSE Standardization Center (whose creation will be possible thanks to the pass of the Law 27/2007, 23 October, on the Use and Recognition of the Sign and the Support Media for Oral Communication).
- ❖ Our group has obtained state financing for its research project “Basis for the linguistic analysis of the Spanish Sign Language”
- ❖ We count with three deaf teachers and four interpreters for the research and teaching tasks.
- ❖ We also count with specialized researchers in subtitling that will deal with the subtitling and marking tasks in the corpus.
- ❖ In these years, several thesis and dissertations of PhD students on topics related to sign language linguistics have been published (Fernández Soneira 2004; Iglesias Lago 2006, Álvarez Sánchez 2006; Cabeza y Fernández 2004).

Current aims

We are working for creating a textual corpus of LSE as a basis for:

- ❖ **Development of LSE grammars.** The grammatical analysis will focus on the determination of the relevant LSE units and the grammatical processes of relation.
- ❖ **Applied research:** LSE interpretation, LSE teaching, Normalization and linguistic planning, Transcription, etc.
- ❖ **General research:** Language acquisition, Linguistic universals and other related issues.
- ❖ **Use of the corpus in the teaching platforms** as a didactic element in order to provide the pupils with real language samples.

Sample search in the future corpus

Menú de búsqueda


- Características del signante
- Muestra de habla
- Configuración del signo

Muestras

- Monólogo
- Entrevista semidirigida
- Discurso libre

Vista 1 Vista 2

Voz



Observaciones

He comprado una **casa** nueva
*I have bought a new **house***

Transcripción ortográfica Glosa

1 > Introduction

2 > Data

3 > Analysis

4 > Future projects

5 > Acknowledgments

6 > Bibliography

Acknowledgements

- ❖ This work is part of a larger research project carried out by the Research Group on Sign Languages at the University of Vigo. Its final outcome, in the form of this paper, would not have been possible without the collaboration of Francisco Eijo Santos and Juan Ramón Valiño Freire (two of our deaf collaborators).
- ❖ This research was funded by the Ministry of Education and Science, grant number HUM2006-10870/FILO. This grant is hereby gratefully acknowledged.

1 > Introduction

2 > Data

3 > Analysis

4 > Future projects

5 > Acknowledgments

6 > Bibliography

Bibliographic References

- ❖ Alvarez Sanchez, P. (2006) “La enseñanza de lengua extranjera a alumnos sordos”. Diploma de Estudios Avanzados (Universidad de Vigo).
- ❖ Báez Montero, I. C. & M. C. Cabeza Pereiro (1995) "Diseño de un corpus de lengua de señas española", XXV Simposium de la Sociedad Española de Lingüística (Zaragoza).
- ❖ Báez Montero, I. C. & M. C. Cabeza Pereiro (1999) "Elaboración del corpus de lengua de signos española de la Universidad de Vigo", Taller de Lingüística y Psicolingüística de las lenguas de signos (A Coruña).
- ❖ Báez Montero, I. C. & M. C. Cabeza Pereiro (1999) "Spanish Sign Language Project at the University of Vigo" (poster), Gesture Workshop (Gif-sur-Yvette, Francia).
- ❖ Cabeza Pereiro, C. & A. Fernández Soneira (2004) “The expression of time in Spanish Sign Language”, *Sign Language and Linguistics*, 7 (1) pp63-82.
- ❖ Fernández Soneira, A. (2004) *La cuantificación en la lengua de signos española*, Universidad de Vigo, Tesis doctoral.
- ❖ Iglesias Lago, S.(2006): “Uso del componente facial para la expresión de la modalidad en lengua de signos española”, Universidad de Vigo, Tesis doctoral inédita.
- ❖ López Morales, H. (1994) *Métodos de Investigación Lingüística*. Ediciones Colegio de España: Salamanca.
- ❖ Martí Antonín, M^a A. (1999) “Panorama de la lingüística computacional en Europa”, *Revista Española de lingüística Aplicada*, 11-24.

Other corpora

❖ **UNIVERSIDAD DE CALIFORNIA BERKELEY**

Base de datos y transcripciones de la Lengua de Signos de Holanda.

Coordinadores: Dan Slobin & Jennie E. Pyers.

❖ **UNIVERSIDAD DE NJMEGEN**

Creación y notación de un corpus (± 120 horas)

Coordinador: Onno Crasborn.

❖ **UNIVERSIDAD DE BÉLGICA**

Grabación en video sin digitalizar (± 10 horas)

Coordinadora: Myriam Vermeerbergen.

❖ **UNIVERSIDAD DE OREBRO (SUECIA)**

Base de datos en vídeo que recoge situaciones naturales en el aula.

Coordinadora: Sangeeta Bagga Gupta.

❖ UNIVERSIDAD DE ESTOCOLMO

“Digital version of Swedish Sign Language Dictionary” (6000 clips)

Coordinador: Brita Bergman

❖ CENTRO NACIONAL DE INVESTIGACIÓN EDUCATIVA

Diccionario de lengua de signos noruega. Creación de un corpus en lengua de signos noruega digitalizado.

Coordinador: Svein Arne Peterson

❖ UNIVERSIDAD DE AMSTERDAM

Digitalización de un corpus (± 500 horas).

Coordinadores: Anne Baker y Ybonne Jobse

❖ UNIVERSIDAD DE MAGDEBURG

Proyecto Berlín: 200 horas de grabaciones de signantes sordos utilizados para la elaboración de materiales de enseñanza.

Proyecto Hamburgo: 20 horas de grabaciones de signantes sordos.

Proyecto Magdeburg: 4 años.

Coordinadores: Andreas Goeze y Jens Hebman

❖ **UNIVERSIDAD DE PARÍS 8**

30 horas de grabación convertidas en Quick-Time clips in Cd and DVD for LS-COLIN Project.

Coordinadores: Christian Cuxac and Marie-Anne Sallandre

❖ **INSTITUTE OF COGNITIVE SCIENCES AND TECHNOLOGIES (ISTC), National Research Council (CNR)**

200 horas de grabación: adult and child LIS corpora.

Corodinator: Elena Pizzuto

❖ **UNIVERSIDAD CENTRAL DE PRESTON (LANCASHIRE)**

Análisis de la lengua de signos británica de los intérpretes al traducir para sordos de la escuela secundaria (40 horas)

Coordinador: Frank Harrington

❖ **UNIVERSIDAD DE KAGENFURT (AUSTRIA)**

Análisis electrónico de léxico en lengua de signos.

Coordinador: Franz Dotter

❖ **UNIVERSIDAD NACIONAL DE AUSTRALIA (CAMBERRA)**
[Centro de Investigaciones de Tipología Lingüística]

Corpus de discurso libre en lengua de signos de India y Pakistan.
Analizado con el CHILDES. Proyecto de análisis tipológico de
interrogativas y negativas en lenguas de signos.

Coordinador: Ulrike Zeshan

❖ **UNIVERSIDAD VICTORIA DE WELLINGTON**
(NUEVA ZELANDA)

Analizadas y transcritas cerca de 100.000 palabras signadas.
Una glosa para cada palabra analizada con el WordSmith.

Coordinadores: Graeme D. Kennedy & David McKee

❖ **UNIVERSIDAD DE BOSTON**

Proyecto que incluye una herramienta para transcribir y analizar los
datos (SIGN STREAM)

Coordinador: Robert G. Lee

❖ UNIVERSIDAD DE HAMBURGO

Proyecto internacional sobre publicaciones en lengua de signos:
“SIGNING BOOKS”

Coordinador: Liesbeth Pyfers

❖ UNIVERSIDAD DE NUEVO MÉXICO

Corpus de LSA en vídeo. Base de datos con información sobre sexo, edad, diferencias sociales, etc.

Coordinador: James MacFarlane

Contact us

Our webpage:

www.webs.uvigo.es/lenguadesignos/sordos/home

Patricia Álvarez: patri.alvarez@gmail.com

Concha Báez: cbaez@uvigo.es

Ana Fernández: anafe@uvigo.es