Linguistic, Sociological and Technical Difficulties in the Development of a Spanish Sign Language (LSE) Corpus

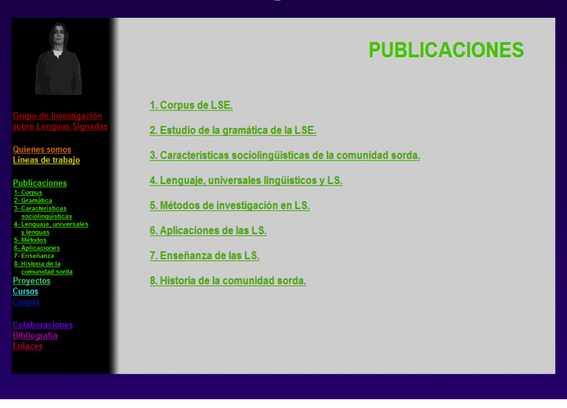
Our Research Group on Sign Languages
http://webs.uvigo.es/lenguadesignos/sordos

Reaserchers:
Ana Fernández Soneira, Patricia Álvarez Sánchez
Inmaculada C. Báez, Carmen Cabeza Pereiro
Deaf Collaborators:
Francisco Eijo Santos, Juan Ramón Valiño Freire

Contact us:
patri.alvarez@gmail.com
cbaez@uvigo.es
anafe@uvigo.es

UNIVERSIDAD DE VIGO

Patricia Álvarez Sánchez,
Inmaculada C. Báez Montero
y Ana Mª Fernández Soneira



Construction and Exploitation of Sign Language Corpora
Marrakech, 1 june 2008

Inicial Work (1995)

> *Linguistic corpora have come to fill a privileged position because they constitute a valuable source of information for the creation of dictionaries, computational lexicon and grammars. (...) As a result, a new discipline appears: CORPUS LINGUISTICS, aimed at the processing and exploitation of this type of linguistic resource.*

Aims

Obtaining a LSE textual corpus of Galician signers from which to start the research on LSE:

A) Starting the description of LSE
B) Determining which are the relevant linguistic units in SL
C) Knowing the grammatical relational processes
D) Developing tools for research: labelling, transcription, etc

How to create a corpus:

1. It must contain real data
2. It must constitute an irreplaceable basis for linguistic description
3. It must be completed with computing support in order to make easy its use.
4. 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.

Process stages

1. Tool design for the creation of a corpus
2. Criteria for the selection of informants
3. Creation of a database of informants details
4. Collection of language samples
5. Data storage
6. Data labelling and marking
7. Transcription and notation systems

Difficulties

1. The lack of a research tradition on Sign Languages in Corpus Linguistics forces us to solve problems from the very beginning:
   - How to delimit units in sign languages.
   - How to label the different formations for their later analysis.
   - Other related issues.

2. Creation of social networks in the Deaf community with the aim of avoiding the social identity of our informants to be threatened.

3. Technical restrictions. We have to select appropriate material in order to avoid problems in compatibility between the different devices (video cameras. video placer, computers, software...)

General Approach

The study of LSE should not dealt with in a different manner to that of any other oral language.
It will be mandatory to have a textual corpus.
The production of a sign language has a kinetic nature. Its reception is visual, so the conversations in sign language have to be registered in video formats.

Analysis and Reconsiderations

> *If our research manages to correct mistaken or unsuitable information, we will have made a good service to linguistics; however, this type of study usually needs for certain knowledge and experiences that do not correspond with the young researcher. (López Morales, 1994, 25)*

Where did we collect our 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.

This is the data gathered from our 85 informants:

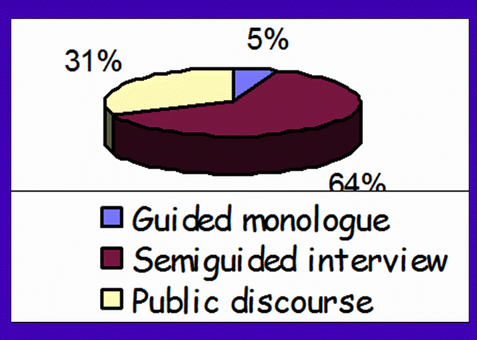
1. ***Identification***: name, address and phone (for future contacts);
2. ***Origin and social environment***: place and date of birth, age of deafness occurrence, deafness degree, deaf/hearing family, job of closest family members;
3. ***School:*** degree and type of studies, special/ordinary school, 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.

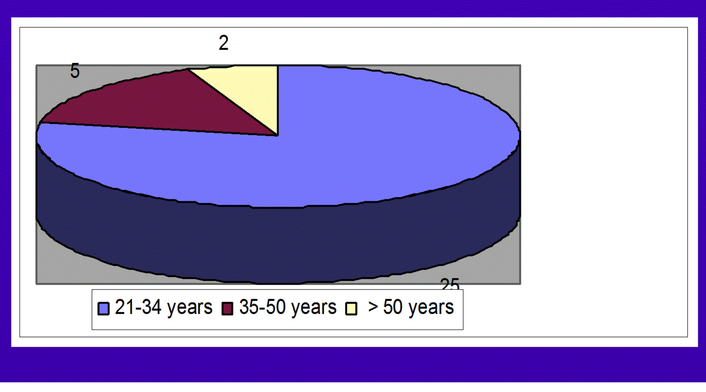
Distribution by gender types:

- *Guided monologue* - 23 minutes
The signer is asked for a description of his family, his house and a short anecdote.
- *Semiguided interviews* - 271 minutes
Signers are interviewed on several topics, depending on their age, sex, preferences, etc. Thus, the discourse is more spontaneous.
- *Public discourse* - 130 minutes.
Conferences and round tables give us a more programmed and formal style.

Distribution of informants by age groups:

- From 21 to 35 years: **25**
- From 36 to 50 years: **5**
- Over 50 years: **2**
- Total: 32 interviews.





Reconsiderations

1) Revision of the projects carried out in other countries.

2) 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 for research in order to contact with the Deaf community all over Spain.

For the time being

Avances

- We are members of a network of universities for the teaching and research on Spanish and Catalan Sign Languages.
- We collaborate in the creation of a LSE Standardization Center (Law 23/2007)
- Our group has obtained state financing for its research project "Basis for the linguistic analysis of the Spanish Sign Language"
- We count on three deaf teachers and four interpreters for the research and teaching tasks. We also count on specialised 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 in topics related to sign language linguistics have been published. Other members of this group have published research papers on grammatical aspects in reference works.

Aims

Creating a textual corpus of LSE as a basis for:

1. Development of LSE grammars. The grammatical analysis will focus on the determination of the relevant LSE units and the grammatical processes of relation.
2. Applied research:
   LSE interpretation
   LSE teaching
   Normalization and linguistic planning
   Transcription
3. General research:
   Language acquisition
   Linguistic universals
   Other related issues
4. Use of the corpus in the teaching platforms as a didactic element in order to provide the pupils with real language simples. These will complete the learning-teaching process started inside the classroom.

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 raise:

- Asking for the collaboration of signers of different regions of Spain for obtaining a good representation of the different geographic registers.
- Select signers that fulfil certain features: native signers of LSE, postlingual users of LSE and interpreters.
- Interview design:

Choice of deaf interviewers. Their dialogues are more natural and they obtain a higher degree of involvement from 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, human relationships, tobacco, etc.)

Technical decisions with a view to the future

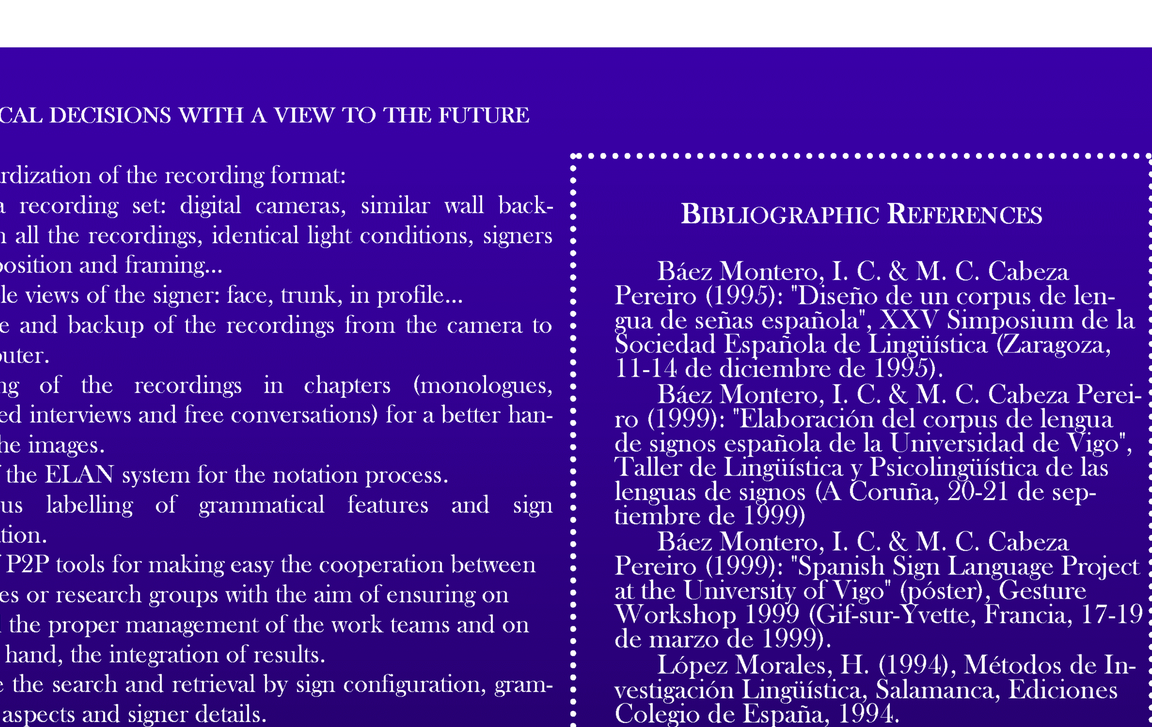
1. Standardization of the recording format:
Use of a recording set: digital cameras, similar wall background in all the recordings, identical light conditions, signers clothes, position and framing...
2. Multiple views of the signer: face, trunk, in profile...
3. Storage and backup of the recordings from the camera to the computer.
4. Editing of the recordings in chapters (monologues, semiguided interviews and free conversations) for a better handling of the images.
5. Use of the ELAN system for the notation process.
6. Corpus labelling of grammatical features and sign configuration.
7. Use of P2P tools for making easy the cooperation between universities or research groups with the aim of ensuring on one hand the proper management of the work teams and on the other hand, the integration of results.
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.

Bibliographic References

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, 11-14 de diciembre de 1995).

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, 20-21 de septiembre de 1999)

Báez Montero, I. C. & M. C. Cabeza Pereiro (1999): "Spanish Sign Language Project at the University of Vigo" (póster), Gesture Workshop 1999 (Gif-sur-Yvette, Francia, 17-19 de marzo de 1999).

López Morales, H. (1994), Métodos de Investigación Lingüística, Salamanca, Ediciones Colegio de España, 1994.

Database of the informants → Recording of language samples for the corpus in videotapes → Data processing → Corpus labelling for its linguistic analysis →