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

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

Universidad de Vigo – Research Group on Sign Languages[1]
Lagoas-Marcosende (36310) Vigo
patri.alvarez@gmail.com, cbaez@uvigo.es, anafe@uvigo.es

Abstract

The creation of a Spanish Sign Language corpus has been, since 1995 until 2000, one of the main aims of our Sign Languages Research Group at the University of Vigo. This research has the aim of helping us in the description of LSE and developing tools for research: labeling, transcription, etc. We obtained language samples from 85 informants whose analysis raised several difficulties, both technical and sociolinguistic.

At this stage, with renewed energy, we have taken up again our initial aims, crossing the technical, linguistic and sociological obstacles that had hindered our proposal to reach its end.

In our panel we will present, apart from the difficulties that we have encountered, the new proposals for solving and overcoming them, thus, finally reaching our initial aim: to develop a public Spanish Sign Language corpus that could be consulted online.

We will go into details with the criteria of versatility and representativity which condition the technical aspects; the sociolinguistic criteria for selecting type of discourses and informants; the labels for marking the corpus and the utilities that we pretend to give the corpus, not only centered in the use of linguistic data for the quantitative and qualitative research of the LSE, but also centered in the use for teaching.

1. General Approach

The study of LSE should not be 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.

Our contribution to the congress, in the form of a panel, is divided into three sections that correspond with the three stages of the development of our corpus. Each step is marked by a general reflection.

The first stage covers our group work from 1995 until 2000 and it represents the beginning of the process. We will present subsequently, the aims set, the steps made for the actual conception of the corpus and the difficulties encountered.

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 at the first stage. We will here present the data obtained and the new goals that we set.

The third and last stage corresponds with the present time. It is the time of showing our advances and the decisions made on the linguistic, sociolinguistic and technical sides.

2. Initial Work

"*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.*" (A, Martí, 1999)

2.1. Aims

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

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: labeling, transcription, etc

2.2. Corpus features

We considered these the main features for creating 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 glossas (conventions adapted from Klima & Bellugi, 1979) and subtitled in written language.

2.3. Process stages

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

[1] http://webs.uvigo.es/lenguadesignos/sordos

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

2.4. Difficulties in the process

The difficulties that aroused throughout the research process are:
a) 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.

b) Creation of social networks in the Deaf community with the aim of avoiding the social identity of our informants to be threatened.
c) Technical restrictions. We have to select appropriate material in order to avoid problems in compatibility between the different devices (video cameras. video player, computers, software…)

3. Analysis and Reconsiderations

"*(…) the paradox exists that once a system is available for its use, its technology becomes obsolete with regard to the one that is operative at that moment and in many cases, it must be reprogrammed*" (A, Martí, 1999)
After these first steps, it was time to analyse the gathered data. For this purpose, we created a database of informants which we are going to present now.

3.1. 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:
a) *Identification*: name, address and phone (for future contacts);
b) *Origin and social environment*: place and date of birth, age of deafness occurrence, deafness degree, deaf/hearing family, job of closest family members;
c) *School:* degree and type of studies, special/ordinary school, use/absence of SL in school;
d) *Linguistic skills*: in LSE, oral Spanish, lip-reading, written Spanish;
e) *Place of residence:* in order to reflect and control linguistic variation.

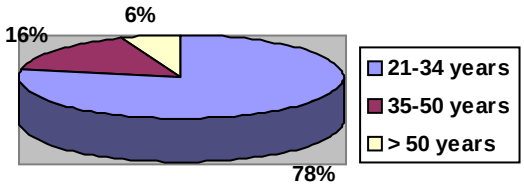

Figure 1: Distribution of informants by age group.

Distribution of informants by age group:
From 21 to 35 years: 25
From 36 to 50 years: 5
Over 50 years: 2
Total: 32 interviews.

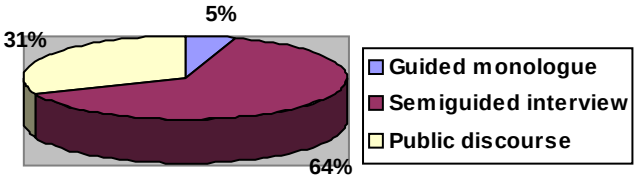

Figure 2: Distribution of language samples by gender types.

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.

3.2. Reconsiderations

After the research, we had to reconsiderate certain issues for a better development of our corpus. We will now sum these up:
a) Revision of the projects carried out in other countries.
b) 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.
Support of the LSE Standardization Center in the creation of the corpus.

4. For the time being

"*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)"
At this stage, with renewed energy, we have taken up again our initial aims, crossing the technical, linguistic and

sociological obstacles that had hindered our proposal to reach its end. In the following lines, we will present the advances achieved and the measures adopted for solving the problems already mentioned, in order to finally develop a public LSE corpus on-line.

4.1. Advances

These are the main advances that occurred in the last years:
- We are members of a network of universities for the teaching and research on Spanish and Catalan Sign Languages (Red Interuniversitaria para la investigación y la docencia de las lenguas de señas- 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"[2]
- 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[3].
- In these years, several thesis and dissertations of PhD students in topics related to sign language linguistics have been published (Fernández Soneira 2004; Iglesias Lago 2006, Álvarez Sánchez 2006). Other members of this group have published research papers on grammatical aspects in reference works (Cabeza y Fernández 2004)[4].

4.2. Current aims

We are working for creating a textual corpus of LSE as a basis for:
a) Development of LSE grammars. The grammatical analysis will focus on the determination of the relevant LSE units and the grammatical processes of relation.
b) Applied research:
- LSE interpretation
- LSE teaching
- Normalization and linguistic planning
- Transcription

c) General research:
- Language acquisition
- Linguistic universals
- Other related issues

d) 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.

4.3. 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 fulfill certain features: native signers of LSE, post lingual 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 should take into account that most of the Deaf people don't have linguistic conscience because they have never studied their language as such. Instead, they have learnt it in a natural way as a medium for communication.

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.)

4.4. Technical decisions with a view to the future

a) 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…
b) Multiple views of the signer: face, trunk, in profile…
c) Storage and backup of the recordings from the camera to the computer.
d) Editing of the recordings in chapters (monologues, semi guided interviews and free conversations) for a better handling of the images.
e) Use of the ELAN system for the notation process.
f) Corpus labeling of grammatical features and sign configuration.
g) 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.
h) Enable the search and retrieval by sign configuration, grammatical aspects and signer details.
i) Online publishing of the corpus with the aid of external financing.

---

[2] Basis for the linguistic analysis of the Spanish Sign Language (HUM2006-10870/FILO) funded by the Ministry of Education and Science. Length of the project: 2006-2008.
Spanish Sign Language: linguistic and philological aspects **(**BFF2003-05696) funded by the Ministry of Science and Technology. Length of the project: 2003-2005.
Grammatical analysis of the LSE: sociolinguistic, psycholinguistic and computational applications (PGIDT00PXIB30202PR) funded by Xunta de Galicia. Length of the project: 2000-2003.

[3] To consult LSE teaching staff profile, cfr. http://www.uvigo.es/centrolinguas/index.en.htm

[4] To consult the whole list of publications by the members of our research group, cfr. http://webs.uvigo.es/lenguadesignos/sordos/publicaciones/index.htm

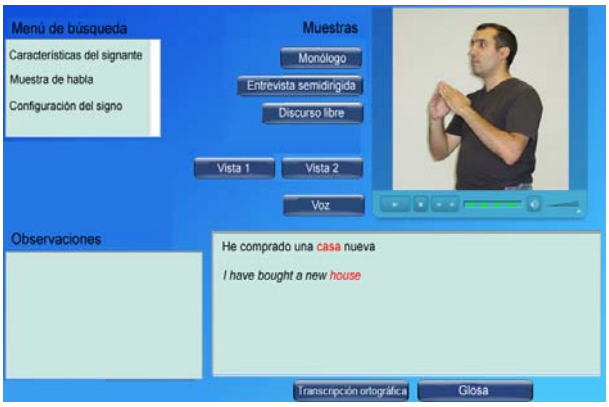

Figure 3: Sample search in the future corpus

5. 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 y 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.

6. References

Álvarez Sánchez, P. (2006). “La enseñanza de lengua extranjera a alumnos sordos”. Diploma de Estudios Avanzados. Universidad de Vigo.
Báez Montero, I. C. & Cabeza Pereiro, M. C. (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. & Cabeza Pereiro, M. C. (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. & Cabeza Pereiro, M. C. (1999). "Spanish Sign Language Project at the University of Vigo" (poster), Gesture Workshop 1999 (Gif-sur-Yvette, Francia, 17-19 de marzo de 1999).
Cabeza Pereiro, C. & Fernández Soneira, A. (2004). “The expresión of time in Spanish Sign Language”, *Sign Language and Linguistics*, vol 7/1, pp.63-82.
Iglesias Lago, S. (2006). “Uso del componente facial para la expresión de la modalidad en lengua de signos española”. Tesis doctoral inédita. Universidad de Vigo,
Fernández Soneira, A. (2004). *La cuantificación en la lengua de signos española*, Tesis doctoral. Universidad de Vigo.
López Morales, H. (1994). Métodos de Investigación Lingüística. Salamanca, Ediciones Colegio de España.
Martí Antonín, Mª A. (1999). “Panorama de la lingüística computacional en Europa”. *Revista Española de lingüística Aplicada*, pp. 11-24.