Building a corpus for Italian Sign Language: Methodological issues and some preliminary results

Carlo Geraci[1], Robert Bayley[2], Chiara Branchini[1], Anna Cardinaletti[3], Carlo Cecchetto[1], Caterina Donati[4], Serena Giudice[1] Emiliano Mereghetti[1], Fabio Poletti[3], Mirko Santoro[3], Sandro Zucchi[5]

[1]University of Milan-Bicocca, [2]University of California-Davis,
[3]University Ca'Foscari at Venice, [4]Sapienza University of Rome, [5]University of Milan.
Address: Carlo Geraci,
Department of Psychology
Università di Milano-Bicocca
Piazza Dell'Ateneo Nuovo, 1
20126 – Milano, Italy
E-mail: carlo.geraci@unimib.it

Abstract

The aim of this paper is to discuss some methodological issues that emerged during the creation of a corpus of data for Italian Sign Language, LIS. Data were collected from 10 cities spread across the country. 18 signers from each city have been recruited. They are native speakers of LIS or later-exposed to LIS and are divided into 3 age groups (19-38, 39-58, 59-78) of 6 signers each (3 males and 3 females). The methodology of data collection and transcription is similar to that used in previous studies of variation in American Sign Language (Lucas, Bayley & Valli 2001) and Australian Sign Language (Johnston & Schembri 2006), with some differences that we discuss. The corpus consists of various kinds of texts collected with different strategies: free conversation (45 minutes), elicited dialogues (about 5-10 minutes), narration (10 minutes) and a picture-naming task (42 items). For the transcription we adopted the ELAN software (Johnston & Crasborn 2006). Finally, a brief report on some preliminary results is presented.

1. Introduction

Since the earliest studies (Volterra, 1987), it clearly emerged that Italian Sign Language (LIS) has an impressive degree of variation. A few studies on lexical variation pointed out some phonological processes related to historical changes (Radutzky 2009, Geraci & Toffali, 2008), and a good number of geographical variants are reported in the most important LIS dictionaries (Radutzky, 1992 and DIZLIS, www.dizlis.it), while Bertone (2007) illustrates some register variations in the use of pronominal forms. However, systematic studies of this variation at various linguistic levels have not been carried out yet. The aim of this paper is to discuss some of the methodological issues that emerged during the creation of a corpus for LIS. Data collection is close to completion at the time of writing. A large-scale corpus has been constructed as part of a national research project on sociolinguistic variation in LIS (PRIN-2007). The core part of the project involves three universities: Sapienza University of Rome, University of Milan-Bicocca and University Ca' Foscari at Venice. As part of the project, the following studies are conducted (see also section 3): variation in the distribution of wh-signs, variation in the use of the 1/G handshape, variation in sign-order, lexical variation, variation in the use of the sign DEAF.

2. Issues in data collection

A first important issue concerns the selection of the cities where data were collected. On the one hand, our choice reflected the distribution of the urban population across the country; on the other hand, it reflected other aspects of the culture and the language of the Italian Deaf community (for instance the presence in the past of important residential Deaf schools). Ten cities were selected, equally distributed across the country: four from the north (Bologna, Brescia, Milan and Turin), two from the centre (Florence, Rome), two from the south (Bari, Salerno) and two from major islands (Ragusa in Sicily, while data collection in Sardinia is imminent). The presence of two cities that are geographically close, namely Brescia and Milano, requires explanation. Despite their proximity, people from the two Deaf communities report clear differences in the use of LIS, possibly related to the existence in the past of an important residential school in Brescia.

For each city, we recruited a local contact person (usually with an active role in the deaf club) who was responsible for participant selection. A total of 180 signers from three age groups (18-30, 31-54, over 55) took part in the data collection. Both the local contacts and the participants were paid for taking part in the project, and participants also agreed to being recorded. For each city, data collection was completed in one day and a half (half a day for each age group).

The age grouping reflects the specific situation of Deaf education in Italy. Indeed, in 1977, a law of the Italian parliament stated that Deaf children could have access to mainstream education in ordinary schools. This law enabled parents to choose their children's education. Many parents (especially hearing parents) sent their deaf children to ordinary non-residential schools. Enrollment in non-residential schools undermined the only natural access to sign language for these children, and in a few years, almost all residential schools and special schools

for Deaf children closed. Hence, the older group (over 55) includes signers who attended residential Deaf schools, the middle group (31-54) includes signers who were at school age during the transition period, and the younger group (18-30) includes signers who had access to mainstream education. The protocol of data collection follows the main lines of those used for the creation of other SL corpora, in particular, the American Sign Language (Lucas, Bayley, & Valli, 2001) and Australian Sign Language (Johnston & Schembri, 2006) corpora. Data collection began with a 45-minute session of free conversation among three signers from the same age group. Then a session of question and answer elicitation followed, performed by pairs belonging to the same age group. The third task was an individual narration lasting approximately 10 minutes. Finally, each signer carried out a picture-naming task of 42 items. In contrast to Lucas, Bayley, & Valli, (2001), we opted for a smaller number of participants for the free conversation task, and we used three video cameras to record the session, one for each signer. One innovation of our study was a semi-structured question and answer task specifically designed to elicit wh-questions, a syntactic construction where variation was expected to occur (see section 3.1 and section 4). We introduced this session because it is unlikely that a number of wh-signs sufficient for a quantitative analysis would show up in free conversation signing. All participants performed the task in pairs: a scene was presented on a picture to one member of the pair. The other member could not see the picture but had to fill a form and recover the information needed by asking the partner. To illustrate, figure 1 depicts a car accident scene, while figure 2 shows the form to be filled out, which is very similar to the one Italian drivers fill out in case of small car accidents. By selecting a type of material that is mostly visual and a form that is familiar to signers, we strove to maintain as natural a situation as possible, even during a semi-structured elicitation procedure.

Figure 1: Car accident scene

In the individual narration session, signers were asked to tell some stories about their lives for about 10 minutes. In order to avoid the unpleasant feeling of signing right in front of a camera, and to reduce to a minimum the potential effects of recording, the local contact was asked to play the addressee in this part of the data collection.

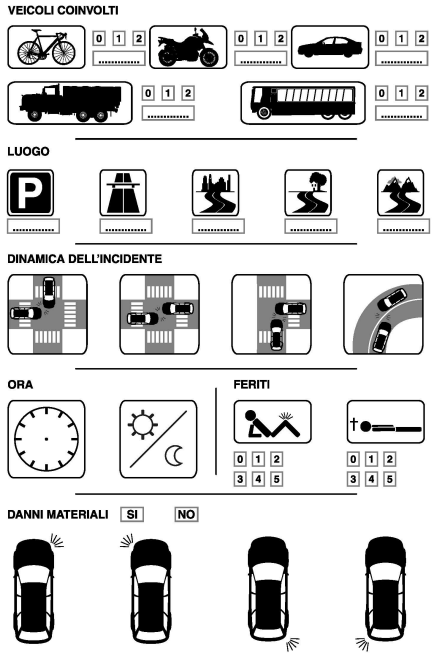

Figure 2: Insurance form

Finally, for the picture-naming task, 42 items from different lexical fields were selected in order to investigate variation in the lexicon of LIS. The list of the lexical fields includes: classifiers, compounds, color names, family names, fingerspelled words, initialized forms, month names, some specific signs known to be eligible for diachronic variation and new formations. Signers were shown an illustrated cardboard for each of the 42 items (see an example in figure 3) in a random order and were asked to name the represented object .
During data collection no hearing researcher was present. One Deaf member of the research team was present at the very beginning of the free conversation session but he left the room when the exchange took off.

Figure 3: Picture-naming cardboard

3. Issues in data coding

Depending on the linguistic variable and on the part of the corpus under analysis (free conversation, elicitation session, etc.), different procedures have been adopted to investigate sociolinguistic variation in LIS. We report here those adopted in the study of the distribution of wh-signs and in the study of the 1/G handshape variation.

For both studies, two Deaf native signers of LIS (each working on data from a different city) searched the tokens and did the first annotation of the variable by using the ELAN software (Johnston & Crasborn, 2006).

3.1 Distribution of wh-signs

Cecchetto, Geraci, and Zucchi (2009) conducted a qualitative in-depth study on wh-question formation in LIS and argued that wh-signs mostly appear in clause final position. To a lesser extent, wh-signs are reported to appear either in their argumental position, or reduplicated in situ and in clause final position. The aim of the study of the distribution of wh-signs is precisely to point out which factors are relevant in determining this variation. We analyzed the part of the corpus specifically designed to elicit questions. The first step in the annotation has been the identification of the utterances[1]. In the first tier of the ELAN file, the coders simply had to delimit the utterances for that part of the corpus. This procedure has a double function: first, it facilitates the access to the database for further studies, and second it gives a rough measure of the productivity for each signer. The second step was to identify the utterances in which a wh-sign occurred and annotate the signs included in that utterance. The third step was to annotate the signs included in the utterance preceding the one containing the wh-sign, and the answer (if present) provided by the other signer (figure 4 illustrates the timetable of an annotation file). At this level, annotations were done in Italian and every wh-sign was specifically tagged with the ID "wh-" (e.g. "what" = wh-COSA, "who" = wh-CHI). This tag allows an easy identification of wh-signs via a simple search in the ELAN files. Although not immediately relevant for this phase of the study, further tags have been added in order to keep track of lexical variants for the wh-signs. In particular, a progressive number indicates alternative variants (e.g. wh-COSA, wh-COSA1, etc.), and a "0" right after the wh-tag indicates that the wh-sign is not the appropriate one (e.g. wh-0COSA means that the wh-sign for "what" is used instead of another wh-sign which is supposed to be more appropriate in that environment). These three steps were carried out by two Deaf native signers of LIS. In the fourth step, carried out by a CODA member of the research group, all the information coded with ELAN has been extracted in a worksheet file and further coding has been done. In particular, for each token, both linguistic and non-linguistic information has been added. As for linguistic information, we coded for the position of the wh-sign in the clause (reduplicated, before or after the predicate), utterance type (direct question, indirect question, echo question, alternative question, non-interrogative clause, pseudocleft), grammatical function of the wh-sign (subject, object, adjunct, etc), wh-type (who, what, when, etc.). As for social information, we coded for geographical origin, gender, presence of Deaf people in the family (parents, relatives or none), education (kindergarten, primary school, middle school or higher education) and work experience (blue collar, white collar, professional or student).

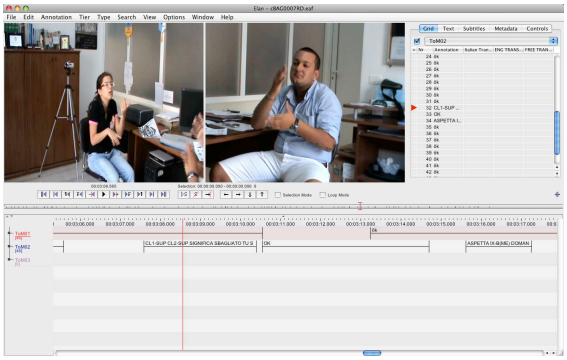

Figure 4: View of the ELAN workspace

3.2 1/G handshape variation

Our aim in the study of phonological handshape variation is to replicate a similar study conducted on ASL by Lucas, Bayley, & Valli (2001). The two crucial methodological differences between our study and that of Lucas et al. are the use of a dedicated camera for each signer instead of a single camera for all the signers involved in the conversation, and the use of ELAN for the coding. Differently from the study of the distribution of wh-signs, where the coding was done in two separate steps, in this case all the coding is done within the ELAN file. This has been made possible by using multiple tiers organized hierarchically (see figure 5). The organization in figure 5 may look complicated but, coding was in fact quite simple since most of the tiers adopt a controlled vocabulary, resulting in a pull-down menu. This choice allows the coder to control for the effects both of single features (such as number of selected fingers, or their hooked vs. straight status) and of combinations of features (i.e. groups of hanshapes).
In figure 5, the first two tiers, namely the main tier (fo1, i.e. Firenze Old signer number 1) and the GLOSS tier are devoted to highlight the sign with the 1/G handshape, the preceding sign and the following sign. The rest of the relevant tiers depends on the GLOSS tier and can be grouped in three main sets, 1-Dhand, 1-Ante Pause, 1-Post Pause, which provide information about the dominant hand, the preceding and following sign, respectively. The main characteristic of these tiers is that each of them is made up with a controlled vocabulary. For sake of exposition we illustrate here the case of the set of 1-Dhand tiers. The 1-Dhand tier specifies the number of selected fingers (other than the index finger and thumb) for the variable token (0, 1, 2, 3). The 1-Dindex tier specifies whether the index is extended, closed (as in the S handshape) or hooked. The 1-Dthumb specifies whether the thumb is extended or not, while the 1-Dhooked specifies whether the selected fingers are extended or hooked. Finally, the 1-Class tier specifies

[1] We are aware that the definition of utterance is controversial both for sign and spoken languages, and that native users of a language have different intuitions about where an utterance ends (see Barrett, 2008 for a recent discussion of this issue in spoken languages).

the grammatical class of the token (pronoun, noun, verb, adjective, adverb, functional sign). The advantage of this coding is immediate once the data are extracted for statistical analyses. Indeed, each tier is converted into a factor group already in columns.

Figure 5: 1/G handshape study tier dependencies

Furthermore, each factor group (including the dependent variable) is already fully specified, since its values come from the close array determined by the controlled vocabulary.

4. Preliminary results: the case of wh-signs

Although the coding for the cities has not yet been completed, some preliminary results about the distribution of wh-signs in LIS are worth mentioning. In particular, the data reported in table 1 are from three cities (Bari, Bologna, and Turin), and illustrate the percentages of the distribution of wh-signs occurring reduplicated (in situ and in clause final position), before and after the predicate.

The general observation made by Cecchetto, Geraci and Zucchi (2009) that the most natural position for wh-signs is the right periphery of the clause is confirmed for all age groups. Furthermore, the data nicely show a diachronic pattern of development in that the proportion of wh-signs occurring in preverbal position decreases across the three age groups from 35% to 17% and then further to 10%. This reduction is compensated by a neat increment in the postverbal positioning of wh-signs and in a moderate increment of reduplicated forms.

Age	After	Before	Reduplicated
Old (over 55)	49%	35%	16%
Middle (31-54)	63%	17%	20%
Young (18-30)	68%	10%	22%

Table 1: Distribution of wh-signs by age groups

5. Conclusions

In this paper, we addressed some of the major issues related to the collection of a corpus for LIS and one preliminary result emerging from the analysis of such corpus. Although the basic structure of our project is similar to that used in other projects that have collected sign language corpora, we introduced some innovations such as the use of a camera to record each individual signer's production, more structured elicitation sessions to elicit particular syntactic constructions and specific coding steps motivated by the use of the ELAN software as a main tool for data coding.

6. References

Barrett, R. (2008). Linguistic differentiation and Mayan language revitalization in Guatemala. *Journal of Sociolinguistics* 12(3), pp. 275--305.

Bertone, L. (2007). *La struttura del sintagma determinante nella Lingua dei Segni Italiana (LIS)*. PhD. Dissertation, University Ca'Foscari at Venice.

Cecchetto, C., Geraci, C., & Zucchi, S. (2009). Another way to mark syntactic dependencies: The case for right peripheral specifiers in sign languages. *Language*, *85*(2), pp. 1--43.

Geraci, C., Toffali, L. (2008). Tendenze conservatrici e innovative nell'uso delle lingue: la variabile dell'età nella Lingua dei Segni Italiana. In G. Bella, D. *Diamantini (Eds.), La qualità della vita nella società dell'informazione*. Milano: Guerini e associati, pp. 97--115.

Johnston, T., Crasborn, O. (2006). The use of ELAN annotation software in the creation of signed language corpora. In *Proceedings of the EMELD'06 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art*. Lansing, MI.

Johnston, T., Schembri, A. (2006). Issues in the creation of a digital archive of a signed language. In L. Barwick, & N. Thieberger (Eds.), *Sustainable data from digital fieldwork*. Sydney: University of Sydney Press, pp. 7--16.

Lucas, C., Bayley, R., & Valli, C. (2001). *Sociolinguistic variation in American Sign Language*. Washington, D.C.: Gallaudet University Press.

Radutzky, E. (Ed.). (1992). *Dizionario bilingue elementare della Lingua dei Segni Italiana LIS*. Roma: Edizioni Kappa.

Radutzky, E. (2009). Il cambiamento fonologico storico della lingua dei segni italiana. In C. Bertone & A. Cardinaletti (Eds.) *Alcuni capitoli della grammatica della LIS. Atti dell'Incontro di studio "La grammatica della Lingua dei segni italiana"*. Venezia: Cafoscarina, pp. 17--42.

Volterra, V. (Ed.). (1987). *La lingua dei segni italiana. La comunicazione visivo-gestuale dei sordi*. Bologna: Il Mulino.

Acknowledgements

The work reported in this paper has been funded by PRIN 2007 project "Dimensions of variation in Italian Sign Language". The Italian team warmly thanks Ceil Lucas and Adam Schembri for sharing their expertise in various phases of this project.