The representation Issue and its Multifaceted Aspects in Constructing Sign Language Corpora: Questions, Answers, Further Problems

Elena Antinoro Pizzuto[1], Isabella Chiari[2], Paolo Rossini[1,3]

[1]Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche, Via Nomentana, 56 – 00161, Roma

[2]Dipartimento di Studi Filologici, Linguistici e Letterari, Università di Roma "La Sapienza", P.le A. Moro, 5 – 00185 Roma

[3]Istituto Statale Sordi di Roma, Via Nomentana, 56 – 00161, Roma

E-mail: elena.pizzuto@istc.cnr.it, isabella.chiari@uniroma1.it, paolo.rossini@istc.cnr.it

Abstract

This paper aims to address and clarify one issue we believe is crucial in constructing Sign Languages (SL) corpora: identifying appropriate tools for representing in written form SL productions of any sort, i.e. lexical items, utterances, discourse at large. Towards this end, building on research done within our group on multimedia corpora of both SL and spoken or verbal languages (vl), we first outline some of the major requirements and guidelines followed in current work with vl corpora (e.g. regarding transcription, representation [mark-up], coding [or annotation] Chiari, 2007; Edwards & Lampert; 1993; Leech & al, 1995; Ochs, 1979; Powers, 2005, among others). We highlight that a basic requirement of vl corpora is an easily readable transcription that, aside from specialist linguistic annotations, allows anyone who knows the object language to reconstruct its forms, and its form-meaning correspondences. Second, we point out how this basic requirement is not met in most current work on SL, where the 'transcription' of SL productions consists *primarily* of word-labels taken from vl, inappropriately called 'glosses'. As argued by different authors (e.g. Cuxac, 2000; Pizzuto & al, 2006; Leeson & al, 2006), the use of such word-labels as a *primary representation tool* grossly misrepresents SL, even when supported by specialist linguistic annotations. Drawing on a relevant work on SL lexicon and discourse (e.g. Cuxac, 2000; Brennan, 2001; Cuxac & Sallandre, 2007; Russo, 2004; Pizzuto & al, 2008), we illustrate how the 'transcriptions' most widely used for SL are especially inadequate for representing complex sign units that are very frequent in SL discourse, and exhibit highly iconic, multidimensional/multilinear features that have no parallel in vl. Third, we discuss findings from ongoing research on Italian Sign Language (LIS) in which experienced deaf signers explore the use of SignWriting (SW: Sutton, 1999) as a tool for both composing texts conceived in written form -- thereby creating a corpus of written LIS -- and for transcribing corpora of face-to-face LIS discourse (Di Renzo & al, 2006; Di Renzo, in press; Lamano & al, in press). The results show that deaf signers can easily represent the form-meaning patterns of their language with an accuracy never experienced with other representation or annotation systems. We illustrate examples of SW-encoded vs. 'gloss'-based transcripts which suggest that SW can be a valuable tool for addressing the representation issue in constructing SL corpora. However, the present computerized form of SW poses problems that constrain its use. We conclude specifying some of the problems that need to be faced on the route towards identifying more appropriate written representations of SL.

1. Premises

The observations reported in this paper are based on one assumption we wish to make explicit. We assume that, in order to qualify as a 'corpus', any collection of linguistic and communicative productions must include not only the 'raw data' themselves (as recorded, stored and more generally accessible via and on, for example, audiovisual tools such as digital video, videotapes, CD, DVD, multimedia tools of various sort) but also, as a necessary requirement, an easily readable transcription that aside from specialist linguistic annotation, and in the absence of the raw data, allows anyone who knows the object language to reconstruct its forms, and its form-meaning correspondences.

We also believe that such a requirement remains (and most likely will remain) a substantial tool for the linguistic analysis and description of any language, in spite of substantial advancements in computer-assisted treatments of language data. In the last decade or so, several researchers have pointed out how the vast majority of Signed Languages (SL) corpora currently available do not meet the requirement specified above. The 'transcriptions' of SL data (even within multimedia tools) consist primarily of word-labels taken form verbal languages (vl), inappropriately defined 'glosses'. It has been argued from different perspectives that the use of such word-labels as a *primary representation tool* grossly misrepresents SL, and renders extremely difficult to capture and analyze distinctive properties of SL lexicon and grammar, most notably with respect to complex, highly iconic structures, and multidimensional / multilinear features that have no parallel in vl (see among others Cuxac, 2000; Brennan, 2001; Leeson & al, 2006; Russo, 2004; 2005; Di Renzo & al, 2006; Pizzuto, Rossini & Russo, 2006; Vermeerbergen, 2006; Pizzuto & al, 2008, Slobin, 2008 – and references cited therein).

In this paper, we first provide an overview of general transcription requirements in linguistic corpora, and of available tools created for the most for spoken language (multimedia) corpora. We then focus on issues concerning the selection of relevant data, and of linguistic models in the construction of SL corpora, highlighting the representation problems posed by features that are unique of SL. We illustrate how we are currently addressing these problems in ongoing work using SignWriting (SW: Sutton, 1999) for representing Italian Sign Language (LIS) texts. We conclude specifying technical and methodological problems that need to be faced on the route towards identifying more appropriate written representations of SL.

2. Transcription Requirements in Linguistic Corpora

SL corpora share with vl corpora the need of different layers of representation of linguistic and communicative aspects of discourse. The design and construction of language corpora make the issue of representation of linguistic data a crucial element both from a theoretical and a practical point of view, determining the need of explicit linguistic and logical models of representation and of formal standards employed in linguistic annotation. Some of the problems involved in transcription, mark-up and annotation of vl are relatively comparable with similar issues in SL corpora planning, even though the levels of representation of signed discourse are still open to debate and far from evident and shared in the scientific community.

In spoken and multimedia corpora the issue of transcription has been faced from a theoretical point of view (e.g. Ochs, 1979; Edwards & Lampert, 1993; Powers, 2005), from a descriptive point of view (e.g. Edwards & Lampert, 1993; Chafe, 1995; Cook, 1995; Leech, Myers & Thomas, 1995; Derville, 1997; Lapadat, 2000; Pallaud, 2003), and from a psycholinguistic perspective (e.g. Lindsay & O'Connell, 1995; Pallaud, 2003; Chiari, 2007), noting how it is inherently infused with linguistic theory and interpretation.

Recently the debate has been focused on different aspects of annotation of multimodal corpora, involving not only speech as the main linguistic signal, but also communicative information conveyed by posture, gesture, visual elements of context, and their interaction with verbal communication. *Linguistic annotation* is defined most generally as "any descriptive or analytic notations applied to raw language data […]. In the simplest and commonest case, 'linguistic annotation' is an orthographic transcription of speech, time-aligned to an audio or video recording." (Bird & Liberman, 2001). As Leech (1993: 275) states, corpus annotation is a procedure "for adding interpretative information to a text corpus", centering on the *interpretative nature* of the process as defined by specific choices made by the annotator. The term annotation thus is generally used to cover both transcription practices and activities of addition of information of other nature (mainly metalinguistic glosses) such as part of speech, rhetorical, semantic description. The nature and typology of information deriving from transcription and description is nevertheless radically different even when coded within a similar format.

Transcription always involves a set of representational choices (Ochs, 1979), even when the aim is the reproduction of the spoken words, and not the more complex metalinguistic task of linguistic annotation. Transcribers' errors are common and, to a certain extent, unavoidable, following regular patterns of substitution, deletion, insertion and inversion typically semantically-driven (Chiari, 2007). The mere act of converting spoken language into written language often involves practices of *naturalization*, such as conventions used to make speech conform to written standards. The process of approximating to speech thus involves the production of a text, the transcript, which becomes less readable in a conventional way: "the more a text reflects the oralness of speech, the less transparent it becomes for readers unaccustomed to encountering oral features in written discourse" (Bucholtz, 2000: 1461).

Annotation tools are thus the direct product of a specific linguistic model that declares the typology of layers required and the possible interconnection of the layers that the system can manage. Under this point of view many papers have focused on the issue of requirements for multimodal annotating tools, mostly devoted to vl (e.g. among others Ide & Brew, 2000; Bird & Liberman, 2001; Dipper, Götze & Stede, 2004; Garg et al., 2004). Among the properties of annotation that need to be fulfilled are reusability, flexibility and extensibility. Moreover the need of levels of annotation

that cover new aspects and meet different purposes generally poses the question of developing specific tools that live and die within one project, or to use common frameworks that share an exchange format and API (Application Program Interface) (Ide & Brew, 2000).

3. Available tools and requirements for multimodal corpora

A number of different tools have been specifically created to perform the task of annotating multimodal audiovisual corpora. Among these: ATLAS [1], MediaTagger (Senghas, 2001), TASX [2], Anvil [3], SyncWRITER (Hanke, 2001), NITE [4], MMAX [5] - EXMARaLDA - EXtensible MARkup Language for Discourse Annotation (Schmidt, 2004), ELAN [6] and ANNEX, the web-based ELAN upgrade. Among tools specifically designed for SL corpora (but adequate for multimodal VL corpora too) are SignStream (Neidle, Sclaroff & Athitsos, 2001) and partly ELAN and ANNEX.

The suggestion of keeping transcription and mark-up separated is a capital issue both for verbal and SL corpora, determining the need of what has been called *stand-off annotation*. Most tools nowadays share the preference for the XML format, still controversy is open on what information to encode in this framework and how to encode it for general purposes of scientifically accurate corpus-based research and for possible commercial uses. Annotation schemes should be implemented using XML coding, or at least should envisage the possibility of exporting annotation in XML format as suggested in current guidelines for linguistic corpora following the TEI, Text Encoding Initiative (Barnard & Ide, 1997), Eagles and CES, Corpus Encoding Standard (Ide & Brew, 2000). Stand-off annotation is a need for it provides separate storage for data (audio, video signal) and description (transcription, annotation at different levels).

Common to most of these tools is an architecture that involves multiple layers of annotation, seen as tracks, that are filled with time-anchored elements. Layers are generally separate elements of annotation that represent different aspects of the communicative acts in a linear way, as for Anvil (Figure 1):

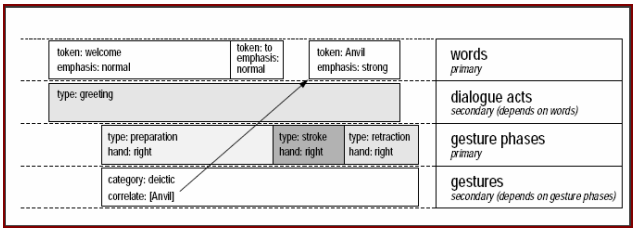

Figure 1: Anvil tracks (Kipp, 2001)

Anvil, a Java-based XML package for audiovisual annotation, enables multiple tracks that can be dependent on each other, and also links to mark co-reference. The direction in tool development runs towards the use of an object oriented system (like Java) and XML mark-up. A similar solution is that of EXMARaLDA and TASX, with the "single timeline, multiple tiers model", where elements are connected directly to a basic transcription tier, which is connected to audio by absolute time-values (Figure 2).

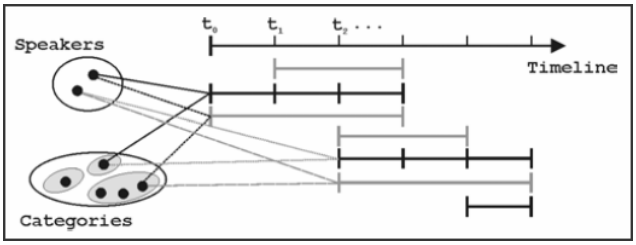

Figure 2: EXMARaLDA timeline model from (Schmidt, 2004: 2)

TASXS adds to the single timeline a multiple tiers model, with the capability of including hierarchical annotation structures, enabling the possibility of linking any annotation element belonging to one layer to the timeline or elements from other layers. Annotation should support labelling of time-aligned annotation layers and it should provide a supporting annotation scheme that takes into account the spatial and tri-dimensional space of the signed interaction. The annotation should be directly related to the video signal files.

While in spoken corpora specific instructions and conventions and annotation layers involve the representation of non-verbal data, such as contextual information, paralinguistic features, pauses, overlaps and other vocal phenomena, in SL corpora is presumable that processing, representation and annotation of these elements are involved in an often radically different manner, and interact in the discourse process in new ways that need to be taken into account specifically.

From the reading of the annotation only it should be possible to predict the exact communicative gesture that was performed without looking at the video, taking into account the inevitable variability of the linguistic act itself. This task cannot be fulfilled by a representation system that contains only glosses. Some tools make a distinction between a basic, canonical or primary layer of annotation, which mainly consists of a word level transcription, and secondary or dependent levels of annotation.

The annotation format should be as neutral as possible from a theoretical point of view, simple, and it should be based on choices generally shared by the scientific community (Barnard & Ide, 1997). This means that transcription of linguistic data should be more descriptive than interpretative when dealing with basic or primary levels of annotation (that corresponding to spoken language transcription for verbal corpora). Müller & Strube (2003: 2) in the development of the MMAX tool for multilevel annotation argue, for example, that the annotation of what they call *base* (transcription of words) for spoken corpora "can be performed on a mere formal (i.e. surface-based) level, we believe these elements to be sufficiently objective to

[1] ATLAS: http://jatlas.sourceforge.net/
[2] TASX: http://tasxforce.lili.uni-bielefeld.de/
[3] Anvil: http://www.dfki.uni-sb.de/kipp/anvil/
[4] NITE: http://nite.nis.sdu.dk
[5] MMAX: http://www.eml-research.de/english/research/nlp/download/mmax.php
[6] ELAN: http://www.lat-mpi.eu/tools/elan/

serve as the structure for what we call annotation *base data*". Even if the notion of base data is not an undisputable and definite concept, it is clear that a formal description of linguistic data is an inevitable task to be performed on any kind of corpus data, while description of further levels of interpretation, from semantic to morpho-syntactic levels, is only a further step in linguistic data description.

In vl corpora tools, the base transcription (the verbal-tier) can be also used to represent a sort of temporal point of reference for all other entities at other annotation layers, like in the EXMARaLDA tool, if time-alignment is not directly linked to source video, as happens in ELAN. Using the v-tier as temporal reference would not be the best solution for SL corpora, since the lack of superficial linearity of transcription for the formal properties of signs can cause problems in relating other tiers to the base, which is by definition multidimensional. To overcome this problem an appropriate transcription system for SL should include the possibility of linking other annotation layers to single portions of the complex gestural sign (e.g. the formal elements representing eye direction or body positioning used to express reference should be individually linked to the co-reference annotation layer, even if they are part of a larger compositional and simultaneous whole sign). A general requisite is the possibility of using a common timeline independent from transcription and annotation layers, and directly connected with primary linguistic data source, namely digital video recordings.

The question to be posed would then be: what can constitute base data for SL corpora? Is there an annotation scheme that can be used? Are there appropriate tools to perform this annotation?

A separate question involves the different users on an annotation scheme, as claimed by Dipper et al. (2004), among which the annotator himself with different skills and training needs, and the corpus explorer who needs readable, clearly understandable transcripts and annotations. Under this perspective while for vl corpora we have a number of usual standards for transcript presentation, SL corpora need to develop representation systems which are specific for the formal properties of the language, and can be easily recognized as such by signers, given the absence of a tradition in a shared written SL.

4. Base data and linguistic models for SL corpora: some key issues

The question of what constitute base data for SL corpora cannot be appropriately addressed without taking in due account broader, strictly intertwined issues concerning the theoretical models adopted, and how they influence the data (and inevitably also the base representations) we select. Within the limits of this paper we can only touch these questions. With respect to theoretical models, we would like to recall here two major theoretical perspectives in past and current work on SL. One can be defined 'assimiliationist' , and exhibits a strong tendency to focus on SL data and features that demonstrate how, beyond 'surface differences', SL are deeply similar to vl. The other is a 'non-assimilationist' view highlighting several structural properties that sharply differentiate SL from vl. This latter view has been articulated most explicitly in extensive research conducted on French Sign Language (LSF) discourse and grammar, but also in work independently developed, in similar and compatible directions, on LIS and other SL (e.g. American, British, Irish, Flemish SL -- see for overviews, among others, Cuxac, 2000; Brennan, 2001; Cuxac & Sallandre, 2007; Pizzuto, 2007; Russo, 2004; Slobin, 2008; Vermeerbergen, 2006, Vermeerbergen, Leeson & Crasborn, 2007).

These two different perspectives provide equally different descriptions of two major kinds of units that can be identified in SL discourse: 'standard' signs, more or less easily translatable with spoken language 'words' (the so-called 'frozen lexicon'), and complex, highly iconic constructions, consisting of manual and nonmanual elements arranged in a multidimensional and multilinear fashion that appears to be unique of SL (the so called 'productive' lexicon and morphology, e.g. Brennan, 2001). Following Cuxac (2000), we describe these complex sign units as Highly Iconic Structures (HIS).

A crucial point to be noted is that, in the theoretical framework proposed by Cuxac, the key feature for distinguishing HIS from standard signs (and the different, metalinguistic communicative intentions they express) is a SL-specific use of eye-gaze, which in fact renders eye-gaze a constituent parameter of sign units, and more generally of signed discourse, at several different structural levels (Cuxac & Antinoro Pizzuto, 2007). Standard signs are preceded or accompanied by eye gaze directed towards the interlocutor, whereas HIS are marked by gaze patterns directed towards the hands (in the production of two major subtypes of HIS characterized as Transfer of Form [TF] and of Situation [TS]), or via a gaze which mirrors the gaze of the referent(s) represented, in producing a third major type of HIS characterized as Transfer of Person (TP).

Focusing on different formal and functional aspects of sign production, what we define here HIS have been characterized in most SL literature with different terms. For example, most researchers have focused on the manual components of TF and TS structures, describing them for the most as "classifiers" or, more recently, "property markers" (see Emmorey, 2003; Slobin, 2008, among others). TP structures are most often characterized as 'role taking / shifting', or 'impersonation' devices (but see also Dudis, 2004; Liddell, 2003; Slobin, 2008, for different proposals).

Figures 3a-3c below illustrates, with examples taken from LIS texts, the differences between the standard signs for 'dog' (3a) and 'child' (3b), and a TP construction (3c) used in the frame of a story to represent a 'dog' and 'cat' referents and the actions they performed. The structure in (3c) actually corresponds to what is characterized as a double TP, in which several different manual and nonmanual constituents are simultaneously arranged in time and space to represent two distinct referents and their actions, encoding the meaning "the child holds the dog in his arm, while the dog licks him on his cheek".

3a: ‘dog’ 3b: ‘child’ 3c: TP

Research on LSF but also LIS and, more recently, crosslinguistic work on LSF, LIS and ASL, has documented that HIS are very frequent in SL discourse, with important variation related to discourse genre (e.g. Cuxac, 2000; Sallandre, 2003; 2007; Russo, 2004; Pizzuto, 2007). For example, Sallandre’s detailed analyses of a LSF corpus of narrative and prescriptive (cooking recipes) texts produced by 19 signers show that HIS constitute on average as much as 70% of the referential expressions identified in narrative texts, and 30% of those found in prescriptive texts. Similar patterns have been reported (using a different terminology), in Russo’s (2004) analyses of LIS poetic and non-poetic texts. A recent crosslinguistic study of narrative ASL, LSF and LIS texts shows that across these three SL HIS are by far the primary means for carrying out anaphoric reference in SL discourse: they constitute between 80% and 95% of the referential expressions used for this purpose (Pizzuto & al, 2008).

In summary, it is clear in our view that these data, framed within the non-assimiliationist theoretical approach sketchily recalled above, challenge any description of SL that tends to underestimate (and underdescribe) the extent to which SL differ from vl. It is equally clear, however, that appropriate investigations of the structural features that distinguish SL from vl, and the creation of adequate corpora that can also be used for probing alternative theoretical models, require representation and annotation tools that are still to be developed. Towards this end, we believe that it is crucial to develop a more profound reflection on *both* ‘annotation’ and ‘data editing’ problems, and how these are faced in vl as compared to SL research (see for example Blanche- Benveniste, 2007), *and* the tools that are used as primary means for creating the ‘basic transcripts’of SL corpora. In our view, any basic transcript must satisfy at the very least the requirement of allowing the transcriber, and all researchers who use the transcript, to reconstruct the forms of the language (and not just the interpretation and/or analyses performed). It is useful to consider in some detail how different representational means succeed or do not succeed in meeting this requirement, and how they influence the analyses that can be done.

5. Experimenting with SignWriting: results achieved and perspectives

As reported elsewhere (Pizzuto & al, 2006; Di Renzo & al, 2006), since 2005 we have begun experimenting the use of SignWriting (SW: Sutton, 1999), in our research on LIS, as a tool for both composing LIS texts conceived directly in written form, and transcribing videorecorded corpora of narrative texts originally produced in a signed, face-to-face form. This research activity has been developed with the direct involvement of experienced Italian deaf signers who have produced all written texts, and transcriptions of LIS (signed) texts, realized thus far. Compared with any other written annotation tool previously experimented for LIS, SW has proven to be a tool much more effective, and easy to use, for representing the form-meaning patterns of the language (Di Renzo, in press; Di Renzo & al, 2006; Lamano, Lucioli & Gianfreda, in press).

Drawing on research conducted by Di Renzo (2006; in press), transcripts (1) and (2) below, adaptated from handwritten and computerized transcriptions made by Di Renzo, illustrate some of the insights that can be gained comparing the information encoded in two different transcriptions of the same, 9” fragment of LIS text. Transcript (1), was made with the graphic symbols of SW (hereafter ‘glyphs’, after Di Renzo et al, 2006); transcript (2) is realized with spoken-word labels (i.e. what are inappropriately defined ‘glosses’)

.

N	TIME		SIGN	TP
21	0.02.11		1	
22	0.02.11		1	
23	0.02.12		1	1
24	0.02.13		1	
25	0.02.14		1	1
26	0.02.16		1	

N	TIME		SIGN	TP
27	0.02.17		1	1
28	0.02.17		1	1
29	0.02.18		1	1
30	0.02.19		1	1
31	0.02.20		1	

Transcript (1)

N	TIME	GLOSS	SIGN	TP
21	0:02:11	GO-TO-BED (center to left)	1	
22	0:02:12	(Conf B) SURFACE + (Conf 2) FEET slip inside	1	1
23	0:02:13	DOG	1	
24	0:02:14	IMPERS (PAWS) REACH-FORWARD (centre to front)	1	1
25	0:02:15	IMPERS (PAWS under chin) FALL-ASLEEP	1	1
26	0:02:16	(Conf C C) JAR	1	
27	0:02:17	FROG (NMC: sense-opportunity)	1	1
28	0:02:18	IMPERS (LEGS 3 3) CLIMB (alternated movem)	1	1
29	0:02:19	IMPERS (LEGS 3 3) JUMP-FORWARD (centre to right)	1	1
30	0:02:20	ESCAPE-AWAY (Conf L)	1	

Transcript (2)

The text represented is taken from a classical ‘Frog story’ narrative produced by a 15 yrs old deaf girl in a study on the development of narrative skills in 15 deaf

signing children and adolescents. The fragment can be translated in English as follows: "(the child) goes to bed, slips under the blanket. The dog searches for a place where to lie and falls asleep with his head on his paws. In the jar, the frog senses the opportunity, climbs up the jar, jumps out, escapes away".

One of the aims of Di Renzo's study was to ascertain the incidence of HIS compared to standard signs, and to assess at the same time the extent to which different transcription methodologies influenced the analysis and coding of signed texts. Di Renzo's transcripts include numbers for each sign unit progressively identified in the text (e.g. N 21-31 / 21-30 in the leftmost columns of transcripts [1] and [2], respectively, where the number sequences reflect the temporal sequencing of the signs in the original signed text), annotation of the point in time marking the beginning of each unit (in minutes and seconds, column 2 in the transcripts), coding for HIS. For the present purposes, we focus only on the coding for TP constructions (marked by a '1' under the TP column, while no marks in the same column indicate that the sign produced was a standard one). The '1' under the 'SIGN' column are to be ignored: they were used for calculations performed on excel sheets linked to the transcriptions. All coding for HIS vs. standard signs was performed on the basis of the transcripts.

Both transcripts must be read from top to bottom. For anyone unfamiliar with the SW glyphs (and the way they have been adapted for realizing transcripts of LIS), the graphemic units in transcript (1) would require more explanations than it is possible within the limits of this paper. It will be hopefully sufficient to note the following. Each graphemic unit within a box aligned with the unit number (N) and time code identifies a major sign unit parsed by the transcriber. Within each such unit, the glyphs on the top (i.e. circle-like shapes, and the symbols within them) stand for head postures and facial expressions and their constituent elements (e.g. significant eye-gaze pattern, head movements). Especially important are the glyphs for eye-gaze patterns used in this transcript, partially modified with respect to the set provided within the SW system. Di Renzo uses a 'ii' graphic symbol for representing eye gaze directed to the interlocutor (e.g. see sign N. 22, 24, 27). The glyphs below the 'circle-shapes' represent hand configurations, location, movements and orientation patterns proper of the manual constituents of the signs identified.

In transcript (2) the signs are represented via 'glosses' in CAPS which render their basic meaning (e.g. N. 21: GO-TO-BED). Specific annotations before or after the 'gloss', in CAPS or in smalls letters, are annotations on structural/functional features the transcriber felt useful to mark (e.g. the annotation for sign N. 22 indicates that the sign unit produced was not a standard sign but a complex unit produced with a 'B' configuration representing a SURFACE and a '2' or 'V' configuration representing somebody's FEET, conveying the meaning 'slip inside').

An appropriate contrastive analysis of these two transcripts would require far more space than it is available here. We point out only some of the major differences. A first, striking difference is the following. Transcript (1) actually allows any LIS signer familiar with the SW system to reconstruct ("read aloud", if you wish) easily and accurately the signs, in the absence of the original videorecorded data. We found this could be done not only by the transcriber of the text (and even after one year since he made the transcript), but also by other signers in our group who had never seen the original data. The same was not true for transcript (2): no one (not even the transcriber himself!) was able to produce an accurate reconstruction of the forms of the signs 'represent' via 'glosses' and 'ad hoc' annotation. In our view, and in agreement with Di Renzo & al (2006), Di Renzo (in press); Lamano & al (in press), this shows an unquestionable advantage of SW-based representations that deserves to be highly valued and which, as far as know, has no parallel in any other form of written representation that has been proposed for SL. The possibility of reconstructing the original forms of the signs, hence also the relevant sign-meaning correspondences, is obviously crucial for any metalinguistic reflection on the text represented which, in turn, is a prerequisite for segmenting, analyzing and coding it.

A second, very important difference concerns the prominence and high visibility/readability of nonmanual elements, in transcript (1), and more generally of the multidimensional and multilinear features of the signs. In the SW-encoded representation, all sign units include glyphs not just for the hand-articulators, but also for head/face/gaze/mouth patterns (indeed these proved to be essential for both reconstructing the signs and analyzing/coding them). The multilinear arrangements of the different manual and nonmanual articulators, and the distinct functions they play is immediately apparent, for example, in sign N. 23: the marked gaze (represented by upwards arrows) and facial expression easily identify the non manual component of the TP, and how this co-occurred with a complex manual construction where the two hands, and their simultaneously arrangement in space, concur to encode the meaning of a human referent 'slipping in under a flat surface'. In contrast, none of this information is easily readable or reconstructable from the representation of the same sign unit as annotated in transcript (2) under N. 22: the forms of the elements implicated simply are not there, but must be immagined/inferred (of course not necessarily correctly) from linearly arranged words which, in turn, provide a mix of translation, interpretation and annotation. This rendition of the sign unit also ignores nonmanual elements. Yet, it is interesting to note that in both transcripts (1) and (2) the sign construction under discussion is coded as a TP. The relevant difference is that transcript (1) provides us information on the form of the TP structure, expressed by a specific gaze, and represented via the SW glyphs, while transcript (2) does not. We have thus just to trust the transcriber on the accuracy of his coding. Similar remarks could be made for all the sign units represented in the two transcripts.

There are also relevant differences in the ways the text is parsed in the two transcripts, and these affect not simply the total number of meaningful units identified (10 in transcript [1] vs. 11 in transcript [2]), but especially, and more interestingly, what for lack of better terms we can define more qualitative aspects of the segmentation process. For example, unit 21 in transcript (2) is represented as one unit (following the usual convention that hyphenated glosses correspond to a

single sign requiring more than one word to be translated with spoken language words). On this ground, one could make the generalization that the sign transcribed is a single lexical unit encoding a meaning comparable to the Italian (or English) complex verbal locution 'go to bed' (made of a verb, a preposition and a noun),. However, transcript (1) provides a different representation, distinguishing two units (N. 21 and 22) for, respectively, 'go' and 'bed'. The different segmentation certainly does not tell us whether or not, in LIS, the meaning expressed is categorizable as a 'verbal locution' (as in Italian or in English), or as a simple verb + simple noun combination which has not the properties of a verbal locution. This remains to be decided on the basis of further analyses. Transcript (1), however, is more accurate because it allows us to reconstruct the forms that were produced, indicating that two (not one) lexical elements were implicated, whereas transcript (2) provides a misleading parsing, which, upon reflection, the transcriber felt was influenced by the use of word-labels.

Similar segmentation differences emerge comparing units 27-30 of transcript (1) with units 27-29 of transcript (2). Space constraints do not allow us to describe these differences as it would be necessary, but it is important to note them because they highlight the crucial relevance that the representation system adopted has in describing and analyzing the elements that constitute a corpus.

Di Renzo (2006; in press) has provided clear evidence that, in the corpus of 15 narrative texts he analyzed, the use of two different transcription systems led to different results. These concerned both the global number of sign units identified in the corpus, and their structural properties. For example, there were significant differences between the number of TP identified via SW-based transcripts (markedly higher), and that of TP identified via 'gloss'-based transcripts (markedly lower). In our view, these data and results cannot be ignored if we wish to construct appropriate SL corpora, and adequate tools for analyzing and describing such corpora.

For example, suppose that we could implement as needed the computerized version of SW (e.g. to allow an easy transposition of handwritten SL texts in a computerized format), and that we could incorporate SW in multimedia tools currently available for handling multimodal language data, creating appropriate retrieval and searching devices whereby we could search the texts transcribed in a similar fashion as we normally do with spoken language multimedia data, i.e. by written representations of the forms of the language (linked as needed, depending upon the specific project, to both the original 'raw data', and the related annotation and coding that may have been introduced). The appropriate implementation we have in mind would of course have to be flexible enough to permit easy retrieval of *both* the constituent elements of sign units (an utility already available for lexical databases -- see the 'SignPuddle' tool created by Sutton and collaborators: http://www.signbank.org/signpuddle/), *and* the 'global gestalt' of simple signs (i.e. standard signs) and complex constructions such as HIS (e.g. N. 23 in transcript [1]).

If we could do this, we could obtain basic information on type-token frequencies (an information that is still so difficult if not impossible to obtain with the tools currently available), and begin to search for a wealth of morphological, lexical, morphosyntactic elements and structural regularities that are undoubtedly present in SL discourse, as it is evident from signers' actual processing and segmentation of signed discourse, and the metalinguistic analyses they are able to provide. Yet these regularities are extremely difficult to detect and describe, in a reliable manner, with the tools currently at our disposal. Most importantly, we could do our searches bypassing the strong limitations that any annotation / interpretation / coding *not accompanied by a representation of the forms of the language* inevitably imposes on any analysis.

6. Problems to be faced

At present, to the extent that we have been able to ascertain, the computerized implementation of SW still poses problems that constrain its use, especially for composing and analyzing *texts*. It is not by chance that our example of SW-encoded transcript was given above in its original, handwritten form. Transposing this text into a computerized format would have taken us more time than we could afford within the time constraints we had. More importantly, the tools available at present appear to be designed primarily for storing and retrieving primarily individual lexical items, rather than texts.

This bias towards creating lexical (rather than text) corpora, a bias that has a long history in SL research, may be particularly undesirable, especially if the individual lexical items that come to constitute 'databanks' are identified 'out of context', instead of being drawn from actual usage, i.e. from actual SL discourse. The danger is that of involuntary introducing significant distortions in the corpus, and in the sign representation / annotation process. In research on LIS, we have found that if the signs to be included in a databank are identified 'out of context', there is a strong tendency to: a) exclude a wealth of complex sign constructions that are very frequent in discourse (most notably HIS), and thus provide description limited to so-called 'standard' signs; b) disregard relevant nonmanual components which appear to be crucial for an appropriate description of the signs, as it is evident when the same signs (and even 'standard' signs) are drawn from discourse (Di Renzo & al, 2006; Di Renzo, in press – see also Leeson & al, 2006). These differences linked to the 'sources' from which signs are drawn, along with more general considerations linked to the 'face-to-face' status of SL (see Pizzuto & al, 2006; Di Renzo & al, 2006; Vermeerbergen, 2006), support Russo's (2005) views on the need of constructing usage-based corpora of SL, i.e. of relying much more systematically than it is often done on actual SL discourse as a *primary* source for corpus construction.

With respect to storing and retrieving SW-encoded signs, we wish to note the following. We have not conducted a sufficient amount of research to evaluate whether the difficulties we have encountered are linked to relatively 'trivial' technical problems, or to more serious 'design-features' problems implicated in encoding and handling the complex graphic symbols of SW.

We would like to conclude posing a question that is

admittedly naïve: is it possible that, in order to face (if not solve) the problems we all encounter and recognize in developing appropriate written representations for SL, we could all benefit from: a) knowledge stemming from the history and evolution of non-alphabetical writing systems (e.g. written Chinese), and b) the large repository of experience and 'know-how' implicit in the tools that have been developed for adapting non-alphabetic and/or non-roman alphabet writing systems (e.g. including Hangul) to a computerized format?

7. References

Barnard, D. T. & Ide, N. (1997). The Text Encoding Initiative: Flexible and Extensible Document Encoding. *Journal of the American Society of Information Science*, 48(7), 622-628.

Bird, S. & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, 33, 23-60.

Blanche Benveniste, C. (2007). Linguistic units in signed and verbal languages. In Pizzuto, E., Pietrandrea, P. & Simone, R. (eds.), *Verbal and Signed Languages - Comparing structures, constructs and methodologies.* Berlin / New York, Mouton De Gruyter, pp. 81-106.

Brennan, M. (2001). Encoding and capturing productive morphology. *Sign Language and Linguistics*, 4: 1/2, 47-62.

Bucholtz, M. (2000). The politics of transcription. *Journal of Pragmatics*, 32, 1439-1465.

Chafe, W. (1995). Adequacy, user-friendliness, and practicality in transcribing. In G. Leech, G. Myers & J. Thomas (eds.), *Spoken English on computer: Transcription, mark-up, and application.* Harlow, England: Longman.

Chiari, I. (2007). Transcribing speech: errors in corpora and experimental settings. *Proceedings of the 2007 International Corpus Linguistics Conference*, Birmingham http://ucrel.lancs.ac.uk/publications/CL2007

Cook, G. (1995). Theoretical issues: transcribing the untranscribable. In G. Leech, G. Myers & J. Thomas (Eds.), *Spoken English on Computer: Transcription, Markup and Applications.* Harlow: Longman.

Cuxac, C. 2000. *La Langue des Signes Française (LSF). Les voies de l'iconicité*. Faits de Langues, n. 15-16. Paris: Ophrys.

Cuxac, C. & Sallandre, M.-A. 2007. Iconicity and arbitrariness in French Sign language: Highly iconic structures, degenerated iconicity and diagrammatic iconicity. In Pizzuto, E., Pietrandrea, P. & Simone, R. (eds.), *Verbal and Signed Languages - Comparing structures, constructs and methodologies*. Berlin / New York, Mouton De Gruyter, pp. 14-33.

Cuxac, C. & Antinoro Pizzuto, E. (2007), The role of eye gaze direction in the construction of reference and the semanticization of the space of enunciation in signed languages. Paper presented at the *10th International Cognitive Linguistics Conference (ICLC)*, Kracow, Polland, July 15-20, 2007.

Derville, B. (1997). Expression orale: du bon usage des transcriptions. *Le Francais dans le Monde*, 287 (Feb-Mar), 29-32.

Dipper, S., Götze, M. & Stede, M. (2004). Simple Annotation Tools for Complex Annotation Tasks: an Evaluation. *Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora, Lisbon*, 54-62

Di Renzo, A. (2006). *Le produzioni narrative in LIS di bambini e ragazzi sordi*. Tesi di laurea, Università di Roma "La Sapienza", Facoltà di Psicologia 2 (undergraduate thesis).

Di Renzo, A. (in press). Esperienze e riflessioni su metodi di trascrizione della LIS. *Atti del 3° Convegno Nazionale sulla Lingua dei Segni*, Verona, March 2007.

Di Renzo, A., Lamano, L., Lucioli, T., Pennacchi, B., Ponzo, L., (2006), Italian Sign Language: Can we write it and transcribe it with Sign Writing ? In C. Vettori (ed.), *Proceedings of the Second Workshop on the Representation and Processing of Sign Languages, LREC-2006*, Pisa: ILC-CNR, pp. 11-16.

Dudis, P. (2004). Body partitioning and real-space blends. *Cognitive Linguistics*, 15: 223-238.

Emmorey, K. (2003). *Perspectives on Classifiers Constructions in Sign Languages*. Mahwah, NJ: Erlbaum.

Edwards, J. A. & Lampert, M. D. (1993). *Talking data : transcription and coding in discourse research*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Garg, S., Martinovski, B., Stephan, S. R. J., Tetreault, J. & Traum, D. R. (2004). Evaluation of Transcription and Annotation Tools for a Multi-modal, Multi-Party Dialogue Corpus. *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, 1707-1710.

Ide, N. & Brew, C. (2000). Requirements, Tools, and Architectures for Annotated Corpora. *Proceedings of Data Architectures and Software Support for Large Corpora*, Paris: European Language Resources Association, 1-5.

Hanke, T. (2001). Sign language transcription with syncWRITER. *Sign Language & Linguistics*, 4, 275-283.

Kipp, M. (2001). Anvil - A Generic Annotation Tool for Multimodal Dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, 1367-1370.

Lamano, L., Lucioli, T. , Gianfreda (in press) I segni sulla carta : analisi e riflessioni sui primi testi di LIS scritta. *Atti del 3° Convegno Nazionale sulla Lingua dei Segni* - Verona, March 2007.

Lapadat, J. C. (2000). Problematizing transcription: purpose, paradigm and quality. *International Journal of Social Research Methodology*, 3, 203-219.

Leech, G. (1993). Corpus Annotation Schemes. *Lit Linguist Computing*, 8(4), 275-281.

Leech, G. N., Myers, G. & Thomas, J. (1995). *Spoken English on computer : transcription, mark-up, and application.* Harlow, essex, England ; New York: Longman.

Leeson, L., Saeed, J, Byrne-Dunne, D., Macduff, A. & Leonard, C. (2006). Moving heads and moving hands: Developing a digital corpus of Irish Sign Language. The 'Signs of Ireland' Corpus Development Project. *Information Technology and Telecommunication Conference 2006*, Carlow, Ireland, 25-26 October 2006.

Liddel, S.K. *Grammar, Gesture and Meaning in*

American Sign Language. Cambridge: Cambridge University Press.

Lindsay, J. & O'Connell, D. (1995). How do transcribers deal with audio recordings of spoken discourse? *Journal of Psycholinguistic Research*, 24(2), 101-115.

Muller, C. & Strobe, M. (2003). Multi-level annotation in MMAX. *SigDial,2003, Proceedings of the 4th SigDial Workshop on Discourse and Dialogue*, Sapporo, Japan.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. Schieffelin (Eds.), *Developmental pragmatics*. New York: Academic Press.

Neidle, C., Sclaroff, S. & Athitsos, V. (2001). SignStream: A tool for linguistic and computer vision research on visual-gestural language data. *Behavior Research Methods, Instruments, & Computers*, 33, 311-320.

Pallaud, B. (2003). Splendeurs et misères de la transcription. *Cliniques méditerranéennes*, 68(2), 59-73.

Pizzuto, E. (2007), Deixis, anaphora and person reference in signed languages. In E. Pizzuto, P. Pietrandrea & R. Simone (eds.), *Verbal and Signed Languages - Comparing structures, constructs and methodologies*. Berlin / New York: Mouton De Gruyter, 275-308

Pizzuto, E., Rossini, P., & Russo, T. (2006). Representing signed languages in written form: questions that need to be posed. In C. Vettori (ed.). *Proceedings of the Second Workshop on the Representation and Processing of Sign Languages, LREC-2006*, Pisa: ILC-CNR, 1-6.

Pizzuto, E., Rossini, P., Sallandre, M-A, Wilkinson, E. (2008). Deixis, anaphora and Highly Iconic Structures: Cross-linguistic evidence on American (ASL), French (LSF) and Italian (LIS) Signed Languages. In R. M. de Quadros (ed.), *Sign Languages: Spinning and Unravelling the Past, Present and Future - TISLR 9.*, http://www.editora-arara-azul

Powers, W. R. (2005). *Transcription techniques for the spoken word*. Lanham, MD: AltaMira Press.

Russo, T. (2004). *La mappa poggiata sull'isola – Iconicità e metafora nelle lingue dei segni e nelle lingue vocali.* Rende: Università della Calabria, Centro Universitario e Librario.

Russo, T. (2005). Un lessico di frequenza della LIS. In De Mauro, T. & Chiari, I. (eds.), *Parole e Numeri – Analisi quantitative dei fatti della lingua*. Roma: Aracne.

Sallandre, M-A. (2003). *Les unités du discourse en Langue des Signes Française (LSF) – Tentative de de categorization dans le cadre d'une grammaire de l'iconicité.* Thèse de Doctorat en Sciences du Langage, Paris, Université Paris 8.

Sallandre, M-A. (2007). Simultaneity in French Sign Language Discourse, In Vermeerbergen, M., Leeson, L. & Crasborn, O. (eds.), *Simultaneity in Signed Languages. Form and function*. Amsterdam: John Benjamins. 103-125.

Schmidt, T. (2001). The transcription system EXMARaLDA: An application of the annotation graph formalism as the Basis of a Database of Multilingual Spoken Discourse. *Proceedings of the IRCS Workshop On Linguistic Databases, 11-13 December 2001*, 219-227.

Senghas, A. (2001). Spatial and temporal coding of Nicaraguan Sign Language in MediaTagger: Documenting three dimensions with a two-dimensional tool. *Sign Language & Linguistics*, 4, 229-240.

Slobin, D.I. (2008). Breaking the Molds: Signed Languages and the Nature of Human Language, *Sign Language Studies*, 8:2, 114-130.

Sutton, V. (1999). *Lessons in SignWriting. Textbook & workbook*. La Jolla, CA: Deaf Action Commitee for Sign Writing (2nd edition, 1st edition 1995).

Vermeerbergen, M. (2006). Past and current trends in sign language research. *Language and Communication*, 26, 168-192.

Vermeerbergen, M., Leeson, L. & Crasborn, O. (eds.), (2007). *Simultaneity in Signed Languages. Form and Function*. Amsterdam: John Benjamins.

8. Acknowledgements.

We gratefully acknowledge partial financial support from the Istituto Statale Sordi di Roma (ISSR) and the Association "Progetti Felicità" (Joint ISTC-CNR/ISSR project "Writing LIS & SignWriting"), and from CNR and CNRS (French- Italian Project "Language, its formal properties and cognition: what can be learned from signed languages" - ISTC-CNR, Rome & UMR 7023 CNRS, Paris, 2006-2007).