

Annotation of Sign and Gesture Cross-linguistically

Inge Zwitterlood^{1,2}, Asli Özyürek^{1,2}, Pamela Perniss^{1,3}

¹ Max Planck Institute for Psycholinguistics, Nijmegen

PO Box 310, NL-6500 AH Nijmegen, The Netherlands

² Department of Linguistics, Radboud University Nijmegen

Deafness Cognition and Language (DCAL) Research Centre, University College London

E-mail: i.zwitterlood@mpi.nl, asli.ozyurek@mpi.nl, pamela.perniss@mpi.nl

Abstract

This paper discusses the construction of a cross-linguistic, bimodal corpus containing three modes of expression: expressions from two sign languages, speech and gestural expressions in two spoken languages and pantomimic expressions by users of two spoken languages who are requested to convey information without speaking. We discuss some problems and tentative solutions for the annotation of utterances expressing spatial information about referents in these three modes, suggesting a set of comparable codes for the description of both sign and gesture. Furthermore, we discuss the processing of entered annotations in ELAN, e.g. relating descriptive annotations to analytic annotations in all three modes and performing relational searches across annotations on different tiers.

1. Introduction

In a five-year project, we compare expressions in the spatial domain, particularly related to establishing and maintaining reference, between two unrelated sign languages (German Sign language and Turkish Sign Language; henceforth DGS and TID), the co-speech gestures accompanying two spoken languages (German and Turkish), and the pantomime-like structures used by hearing non-signers in Germany and Turkey when asked to convey information without speaking. This comparison aims to discover the similarities and differences in the way in which information pertaining to the identity, location, motion, and action of is expressed between the sign languages, between the co-speech gesture expressions in the spoken languages, and between the signing, co-speech gesture and no-speech pantomime modes. To this end, a large video corpus of task-related discourse data is being constructed. The aim is to record 90 minutes of useable data per participant, with 15 participants per condition (resulting in 135 hours of useable data).

The data will be described using the IMDI metadata standards and entered in the Browsable Corpus based at the MPI for Psycholinguistics. Parts of the data will be made accessible for other researchers and for educational purposes at the end of the project. The video data are annotated using the ELAN (Eudico Linguistic Annotator) annotation tool developed at the same institute.

In this paper we report on the development of the annotation conventions we use in this project, discuss their advantages and shortcomings, and suggest further improvements. Furthermore, we

will explain how we devised our annotation templates in order to enable relational searches between tiers and across annotations after they have been entered in ELAN.

2. Data Collection

2.1 Stimuli

In order to facilitate comparison between the languages and the communication modes, the same stimulus materials were used in all three conditions. We used (and where necessary, adapted) existing materials that have been used successfully in the past, but also created new materials to target specific domains of our research questions. Future cross-linguistic research into several aspects of the sign languages and modes was also taken into account in the choice of these materials.

The stimuli include animated movies from "Die Sendung mit der Maus" (as used by Perniss 2007), the Canary Row episodes (used in various sign and gesture language research projects, e.g. McNeill 1992), and selected scenes from Charlie Chaplin movies (So *et al.*, 2005). These, as well as a subset of Zwitterlood's classifier pictures (Zwitterlood 2003) and the Balloon Story pictures (used e.g. by Kuntay 2002), were used to elicit first and subsequent mentions of referents in various types of locative and motion constructions. Furthermore, a subset of the pictures used by Volterra *et al.* (1984) as well as newly constructed movies were included to elicit verbs expressing events of *giving* and *taking* and their arguments. Finally, a large set of photographs was compiled for elicitation of quantified expressions of location of single and multiple referents. In the data collection sessions, the sets of stimuli are presented to the

speakers/signers on a laptop computer, and are worked through at the participants' own pace.

In addition to elicited data, (semi-)spontaneous data are recorded by asking participants to describe their family and their home, and to tell one or more personal narratives of their own choice. These data are very important as a control for the frequency bias of particular grammatical structures resulting from the use of elicitation materials. Targeted elicitation is necessary to ensure the presence of the phenomena under investigation in the data. The inclusion of spontaneous data is important to confirm the occurrence of these phenomena in more natural discourse.

2.2 Recording Procedure

Each recording session requires two native (or near-native) speakers or signers. One person is the main signer/speaker; the other is the addressee. First, this creates a more natural conversation. Second, in some tasks, the addressee is asked to perform a task (such as to point to a picture out of four that matches the signer's description), so we can test whether the message was understood.

The participants are seated opposite each other. Each participant is recorded individually with a Sony DV camera. Both participants are also recorded together from above by a Sony DV camera with a wide angle lens, mounted on a tall tripod. Previous research on motion and location in sign language (Zwitserslood 2003) has shown that a top view, in which the relation of the hands to the body in terms of distance and direction can be seen clearly, is indispensable when investigating the use of space by language users.

The tasks are explained by a native speaker/signer, who coordinates the recording session and helps out where necessary.

2.3 Data Processing

The data are recorded on mini-DV tapes with standard DV recorders. The content of the tapes is captured on Apple Macintosh computers and processed using Final Cut Pro version 4. The video signals from the three recordings (i.e. front view of speaker/signer, front view of addressee, top view) are synchronized on the basis of an audio peak, resulting from three claps at the start of a recording session (see also Crasborn & Zwitserslood, this volume). The segments containing relevant data are exported as separate files and compressed to MPEG1 format.

3. Annotation

The data are made accessible and searchable by providing annotations. The annotation program

used is ELAN, displaying, as needed, movies of the speaker/signer, addressee, and/or top view. In order to be able to compare annotations of data from the different languages and modes, the coding templates for each mode contain the same or similar tiers, with a comparable coding scheme for making annotations in each mode.

3.1 Existing Annotation Conventions

The annotation of sign language, gesture, and pantomime is relatively new and to date there are no clear, standard conventions. The process is extremely time consuming, especially when there are so many aspects that could be of importance. Due to time limitations and particular research targets, researchers necessarily make choices about which aspects need to be annotated and how, in order to be able to answer their specific research questions. This is unavoidable, but hampers comparison to data and annotations of other researches. We have studied annotation conventions of previous projects, in order to learn from their experiences and to use (subsets of) these annotation conventions. Regarding sign language annotation, there are various annotation systems, some even quite extensive (e.g. the Berkeley Transcription System (Slobin *et al.* 2001), conventions as used in the ECHO project (Nonhebel *et al.* 2004a,b) and the Auslan corpus project (Johnston & De Beuzeville, 2007). For gesture annotation, only one coding system is reported Kita *et al.* (1997). We combined the methods developed for gesture coding and sign coding and extended it for our purposes.

Current annotation systems and conventions for sign languages often directly analyze parts of the sign stream (e.g. by providing a sign with a gloss) or combine parts of that which is observable (e.g. the form of the sign or a particular location in space) with an interpretation or analysis (e.g. the annotation PT:PRO indicates that a sign is a pointing sign [PT] and that this pointing sign is a pronoun [PRO]). Since the structures we are interested in still need a lot of study and it is known that sign language annotation tends to involve (sometimes undesired) interpretation, at a (too) early stage (see also Leeson & Nolan, this volume), we wanted to make a clear division between annotations on a mere *descriptive* level and annotations on an *analytic* level. Annotations on descriptive level tiers describe signs/expressions in terms of their phonetic/phonological¹ form only, while annotations on the analytic level tiers provide

We use the term "phonetic/phonological", since it is still unclear in many cases if a particular sign component can be analysed as a phoneme or should be considered a particular pronunciation of a phoneme.

an interpretation and/or analysis. With both types of tiers, analytic annotations can be based on and linked to descriptive annotations, or can be independently re-analyzed if this proves to be necessary, and mismatches between descriptive and analytic annotations can easily be found and adapted. (This will be described in more detail in sections 3.3 and 3.4 below.)

However, to date, the transcription of non-oral utterances has been severely impeded by the lack of an orthographic or a phonetic/phonological notation system. Some phonetic or phonological systems have been developed for the notation of sign languages, such as the Stokoe transcription system and later developments (Stokoe *et al.* 1965 and later documentation) and HamNoSys (see Prillwitz *et al.* 2001 and later documentation). However, either such a system cannot be used in ELAN (since a system needs to have an accepted Unicode font in order to be implemented in ELAN; this is not the case for some systems); or a system is not transparent (enough), using regular fonts (letters, punctuation marks, etc.) that have no relation to what they describe. In the end, we selected and combined conventions from several systems.

3.2 Annotation of Different Modes

Some differences in annotation between the different modes are unavoidable, since in the co-speech gesture mode it is possible to annotate the spoken words, using the commonly used orthography. This is not possible in the sign/gesture mode, because of the aforementioned lack of an orthography or a clear and transparent system for phonetic/phonological annotation of gesture/sign. As a result, in the co-speech gesture there is a separate (descriptive) tier for the annotation of the German or Turkish speech. Along with this tier, there is an (analytic) tier containing the English translation, for easy access, quick reference, and comparison.

In the sign language mode, every sign/gesture is annotated by means of glosses in German or Turkish by native signers who are bilingual in DGS and German, or in TID and Turkish). Three tiers are involved: a tier for each hand and a tier for signs in which both hands are acting in unison (e.g. when both hands are clasped together). One-handed signs are annotated on the appropriate Left or Right Hand Tier, two-handed signs on both the Left and Right Hand Tier, and the special cases in which both hands act together are on the Both Hand Tier. Separate tiers (linked to the German or Turkish tiers) contain English translations of these glosses.

Annotation is a laborious, time consuming process. In view of the large amounts of data and the time allotted to the project it would be impossible to give such detailed annotations for each utterance. Therefore, besides annotating each word in the co-speech gesture mode and each sign/gesture in the sign language and no-speech pantomime modes, we focus our efforts on the utterances that are of particular interest to us. In the chosen utterances, in all modes, positions and movements of the hand(s), face, eyes, and body are annotated in a way as similar as possible.

3.3 Annotation at the Descriptive Level

3.3.1 Manual Elements

In all three modes, for the selected utterances, Position, action and shape of each hand is annotated on the appropriate Left Hand, Right Hand or Both Hands tier. Descriptions of the hand configuration (handshape and orientation combined) is described in one annotation; location or movement is described in an annotation on a separate tier.

Handshape is described using the handshape table from HamNoSys version 2, with additions as described in Van der Kooij (2002). Since the HamNoSys font is not available in ELAN and because of a lack of a generally accepted set of handshape labels, we used the Solution by Kita *et al.* (1997), who assigned letters to the rows in the handshape table and numbers to the columns. For example, a **fe** handshape, placed in the first row and in the third column of the table, is coded as “A3”.

Finger and palm orientation of each hand are coded together, in the vein of the HamNoSys codings. For palm orientation, we use a subset of the HamNoSys codings, labeled with small letters (e.g. “d” for “down”, “r” for “right”). We diverge slightly from HamNoSys in the Interpretation of the palm orientation: In handshapes where the fingers are extended we code the orientation of the *inside of the fingers* rather than the orientation of the palm of the hand. This results in the same orientation for handshapes with straight and bent fingers and is therefore, in our opinion a better description of the palm orientation. Our Interpretation of finger orientation is also slightly different from that prescribed in HamNoSys: We code the direction in which *the fingers* are actually pointing instead of the direction in which the fingers would point if they extended straight from the hand (“extended finger orientation”). This way of coding finger orientation gives a clearer indication of the orientation than the original one, because often handshapes with bent extended fingers are pronunciation variants of handshapes with

extended fingers. The finger orientations of HamNoSys are labeled with capital letters (e.g. “U” for “Up”, “LD” for “Left Down”).

An example is “E6tD”, describing the hand configuration in Figure 1.

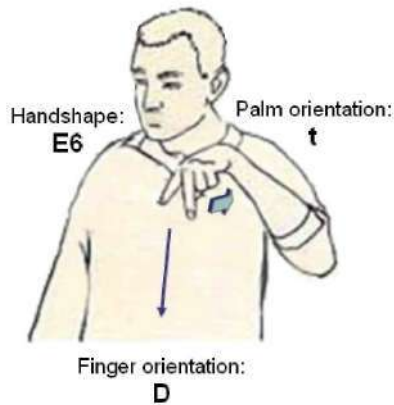


Figure 1: Hand configuration codes

In case a sign contains a handshape change and/or an orientation change, this is coded with the initial and final handshapes/orientations, separated by an arrow, e.g. “A3Ou->C10Oau” describes a handshape and orientation change.

Description of locations and movements of the hand(s) (and of non-manual elements) is, in existing systems, either not possible or too crude, e.g. “left” is not detailed enough in contexts where there may be several referents located to the left of the signer. Also, height may need to be taken into account. Therefore, we devised a 3-dimensional grid with combination codes, to which horizontal and vertical locations in signing space can be assigned. The vertical codes are shown in Figure 2, the horizontal codes in Figure 3. A combination of these codes is used within single annotations.

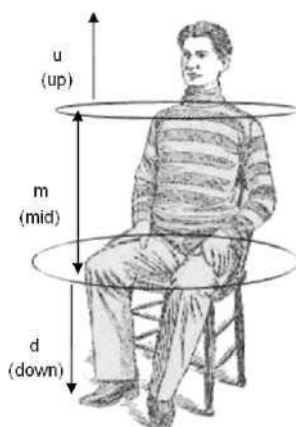


Figure 2: Vertical part of 3D location grid

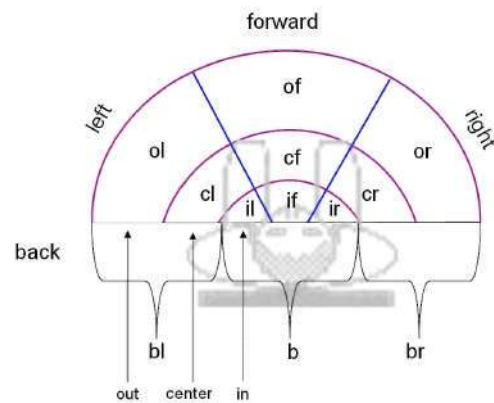


Figure 3: Horizontal part of 3D location grid

An example is “clm” (center left of signer, mid level).

In signs in which the hand moves from one location to another, this is described by the codes for the initial and the final locations, separated by an arrow, e.g. clm->clu.

3.3.2 Non-manual Elements

In communication, many non-manual elements can be used to convey information, e.g. about particular referents involved in the event that is being described. Body position, eye gaze, facial expression are well-known for this. To some extent, they are also used for referent indication in spoken languages. Therefore, we also code these elements in the utterances we select. For eye gaze, we use a separate tier, using (if possible) the codes from the 3-dimensional grid. There are also tiers for body position and head position. These are described using codes that express dynamic and static tilts, bends, and turns of the head and body, and head nods and shakes. To describe these, we selected a subset of the options described in HamNoSys (Hanke *et al.* 2001) and (as yet unpublished) in the annotation conventions used in a research project on prosody in the Sign Language of the Netherlands (Van der Kooij, p.c.). We use codes such as “sLF” and “dLL” to describe that the signer’s body shows a static turn to the left and a dynamic movement, leaning leftwards, respectively, and “tiltL” and “SNodU” to describe a head tilt to the left and a single upward nod of the signer’s head, respectively.

3.4 Annotation at the Analytic Level

Besides giving a description of the forms we see in a given discourse, we need an interpretation of the signs/gestures and other, non-verbal information. For example, we code whether a sign contains a classifier and the type of classifier. We are especially interested in coreference mechanisms in

the discourse, that is, the ways in which referents receive first and subsequent mentions. In sign languages, this can be done manually, by pointing or signing at particular locations in signing space, or by using classifier handshapes. Non-manually, it can be done by body or head shifts towards particular locations in signing space and/or by facial expression. In co-speech gesture, it is argued that similar ways of referring to referents are available. We indicate all referents that are referred to in the sign/gesture/speech signals in annotations on a separate tier, and we try to connect them to annotations on descriptive tiers. In that way, we hope to find systematicity in the expression of referents on three possible levels: language-specifically, cross-linguistically, as well as cross-modally.

4. Further Use of the Annotations

What is the next step if one has finished a set of annotations? ELAN is a powerful annotation tool with search functionality, but that functionality is, so far, restricted. It is possible to find particular annotations in one or more files and to restrict one's searches (e.g. to a particular time interval or to a subset of tiers). However, it is not possible to enter relational searches, i.e. searches where the annotations one is looking for on one tier are related to annotations on another tier. It is important to realize this before one starts to enter annotations, because the use one wants to make of the annotations influences the structure of one's ELAN templates. In our case, we wanted to be able to list annotations linked to particular annotations on other tiers, e.g. we wanted to be able to see all handshapes and locations that are used to refer to a particular referent (in all modes).

Although such relational searches cannot be done in ELAN, there are possibilities to do such searches outside the tool, in data that are exported from ELAN to another application that does have those facilities (in our case: Microsoft Excel). In order for this to work, the relations between annotations on different tiers should already be made in the ELAN template; the exported annotations then include these relations. It is possible to link annotations on parent tiers (independent annotations) with annotations on child tiers (dependent annotations). These annotations can be exported to Excel in a schematic structure, that can then easily be used for several searches.



Figure 4: Screenshot of annotation of a German narrative in ELAN

5. Concluding Remarks

In this project, we have extensively considered the possibilities and intricacies of making comparable annotations of similar types of information expressed in several modes. The first, real challenge is to find a means to describe the non-verbal expressions in such a comparable way, especially since there are no clear-cut, interpretation-neutral conventions for the annotation of non-verbal expressions. The second challenge is to find a way to relate the different types of annotations that are entered in ELAN and to be able to make easy comparisons on the basis of those annotations.

The second challenge is answered by using annotation templates, in which the relations between annotations on different tiers that we are interested in are already established, so that the relations can be viewed in another application (i.e. in Excel).

With respect to the first challenge, we use particular annotation conventions to circumvent the problems of mixing or missing information concerning the form and the interpretation of non-verbal expressions by distinguishing *descriptive level* and *analytical level* annotations, and by using non-analytic codes in annotations at the descriptive level. However, the codes we use are a combination of existing codes, adapted where these codes

appeared not to be clear (enough) and extended with extra codes, and thus, they do not form a conventional system. Furthermore, a real problem is the fact that many codes in our system are still not very transparent, as they are based on common fonts used for the description of spoken languages. We would like to encourage the linguistic community (especially that part of that community that is involved in non-verbal communication) to work on an (accepted) orthography for sign language *and* transparent phonetic and phonological annotation systems for non-verbal communication, that can and must be implemented in software applications for the annotation and processing of such communication. That way, over- and misinterpretation as often caused by mere gloss annotations and annotations that combine descriptions and analyses can be avoided in the future. Furthermore, easier and better comparison of data and analyses is facilitated.

6. Acknowledgments

The project is funded by the Netherlands Organization for Scientific Research (NWO), in the framework of the Vernieuwingsimpuls (VIDI grant (no: 276-70-009) awarded to Asli Özyürek, and also by the MPI for Psycholinguistics, The Netherlands.

7. References

- Kita, S., Van Gijn, I. & Van der Hulst, H. (1997). Movement Phase in Signs and Co-Speech Gestures, and Their Transcriptions by Human Coders. In Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction 1371, pp. 23-35.
- Hanke, T., Marshall, I., Safar, E., & Schmalin, C. (2001). "Interface Definitions", Deliverable D5-1 from the earlier ViSiCAST project - a technical description of the languages defined to interface between various components used in eSIGN work.
- Johnston, T. & De Beuzeville, L. (2007). Auslan Corpus Annotation Guidelines. Ms., Macquarie University & SOAS, University of London.
- Kuntay, A. (2002). Development of expression of indefiniteness: Presenting new referents in Turkish. In *Discourse Processes* 33(1), pp. 77-101.
- Nonhebel, A., Crasborn, O. & Van der Kooij, E. (2004). Sign language transcription conventions for the ECHO project. Version 9, 20 January 2004. Ms., Radboud University Nijmegen.
- Nonhebel, A., Crasborn, O. & Van der Kooij, E. (2004). Sign language transcription conventions for the ECHO project. BSL and NGT mouth annotations. Ms., Radboud University Nijmegen.
- Perniss, P. (2007). Space and Iconicity in German Sign Language (DGS). Nijmegen, MPI Series in Psycholinguistics 45.
- Prillwitz, S., Leven, R., Zienert, H., Hanke, T., Henning, J., (1989). Hamburg Notation System for Sign Languages: An Introductory Guide. Hamburg, International Studies on Sign Language and the Communication of the Deaf 5.
- Slobin, D. I., Hoiting, N., Anthony, M., Biederman, Y., Kuntze, M., Lindert, R., Pyers, J., Thumann, H., Weinberg, A. (2001). Sign language transcription at the level of meaning components: The Berkeley Transcription System (BTS). *Sign Language & Linguistics* 4, pp. 63-96.
- So, W.C., Coppola, M., Licciardello, V., & Goldin-Meadow, S. (2005) The seeds of spatial grammar in the manual modality. *Cognitive Science* 29, pp. 1029-1043.
- Stokoe W.C., Casterline, D.C. & Croneberg, C.G.. (1965). *A Dictionary of American Sign Language Based on Linguistic Principles*. Silver Spring, Md.: Linstok.
- Van der Kooij, E. (2002) Reducing phonological categories in Sign language of the Netherlands. Phonetic implementation and iconic motivation. Utrecht: LOT Dissertation Series 55.
- Volterra, V., Laudanna, A. Corazza, S., Radutsky, E. & Natale, F. (1984). Italian Sign Language: the order of elements in the declarative sentence. In Loncke, F. *et al.* (Eds.) *Recent Research on European Sign Languages*, pp. 19-48.
- Zwitserlood, I. (2003). *Classifying Hand Configurations in Nederlandse Gebarentaal (Sign Language of the Netherlands)*. Utrecht, LOT Dissertation Series 78.