Multitrack annotation of child language and gestures

Jean-Marc Colletta, Aurélie Venouil, Ramona Kunene, Virginie Kaufmann and Jean-Pascal Simon

Lidilem – IUFM and Université Stendhal

BP25 – 38040 Grenoble Cedex 9, France

E-mail: jean-marc.colletta@u-grenoble3.fr, a.venouil@free.fr, kuneneramona@yahoo.com, virginie.kaufmann@gmail.com, jean-pascal.simon@grenoble.iufm.fr

Abstract

This paper presents the method and tools applied to the annotation of a corpus of children's oral and multimodal discourse. The multimodal reality of speech has been long established and is now studied extensively. Linguists and psycholinguists who focus on language acquisition also begin to study child language with a multimodal perspective. In both cases, the annotation of multimodal corpora remains a crucial issue as the preparation of the analysis tools has to be in line with the objectives and goals of the research. In this paper we present an annotation scheme aimed at studying linguistic and gesture development between childhood and adulthood, with emphasis to the relationship between speech and gesture and the way it develops. We also present a validation method for gesture annotation.

1. Introduction

This paper deals with an interlinguistic and intercultural perspective of child's speech development in its multimodal and semiotic aspects. It is grounded on the multimodal reality of speech as established by gesture researchers, as well as on the evidence of the complementary semiotic nature of speech signs and gesture signs. Research on gesture as well as cognitive science has shown data which reveal that the listener, or speaker, integrates auditory and visual information from linguistic, prosodic and gesture sources into a single message (McNeill, 1992, 2005; Beattie, 2003; Goldin-Meadow, 2006).

In relation to child language development, several researchers have revealed evidence that a gesture-speech system begins to operate from 16–18 months of age (Capirci, Iverson, Pizzuto & Volterra, 1996; Butcher & Goldin-Meadow, 2000; Volterra, Caselli, Capirci & Pizzuto, 2005; Ozcaliskan & Goldin-Meadow, 2006). Furthermore, there is additional evidence that coverbal gesture - hand or head gestures as well as facial expressions linked to speech - develop as well as vary as the child grows older (Colletta, 2004; Colletta & Pellenq, 2007; Sekine, 2007). However, how does this speech-gesture system develop in children older than 5 years? Does the relationship between gesture and speech become modified under the influence of new linguistic acquisitions and new communicative behaviour? Do new coverbal gestures appear through late speech development? When and how does culture influence this co-development of gesture and speech?

Four research teams from France, Italy and the United States, involving linguists and psychologists and previous experience in multimodal and discourse development, joined forces in order to tackle these questions (French ANR Multimodality Project NT05-1_42974, 2005-2008). Our aim for this workshop is to present the methodological procedures and annotation scheme used in our study in the collaboration as stated above .

Currently, several researchers are interested in the multimodal complexity processes of oral communication. This issue has brought about increased interest to researchers aiming to transcribe and annotate different kinds of multimodal corpora, for instance, researchers in computer sciences take into account the multimodal clues in order to improve the Embodied Conversational Agents (cf. Ech Chafai, Pelachaud & Pelé, 2006; Kipp, Neff & Albrecht, 2006; Kopp et al., 2006; Vilhjalmsson et al., 2007). Other researchers, as Abrilian (2005), work on the annotation of emotional corpora in order to examine the relationship between multimodal behaviour and natural emotions. Other researchers working in the field of autism (*inter alia* Grynszpan, Martin & Oudin, 2003) or language development (Colletta, 2004) also take into consideration these multimodal clues in their studies. It is without doubt that these methods and tools of annotation have paved the way for more exploratory means to study multimodal corpora in detail.

Our data collection is based on a protocol aimed at collecting spoken narratives from American, French, Italian and Zulu children and adults under controlled

experimental conditions. A 2 minute extract of an animated "Tom & Jerry" cartoon is shown to each subject. He/she is then asked to recount the story he/she has just seen to the experimentator. Each interaction is filmed with a camcorder. From each language group, 60 spoken narratives (performed by 20 children aged 5 years, 20 children aged 10 years, and 20 adults) were collected.

The collected data are analysed using the software ELAN (EUDICO Linguistic Annotator)[1]. Two main levels of transcription are selected for annotation: a verbal level and a gesture level (see table 1: annexures). We will briefly present the first level and we will elaborate on the second level as well as on the validation process.

2. Annotation of the verbal level

The main aim of our work is the narrative abilities and the way they develop in children. As children grow older, their linguistic performance in narratives changes; as they include longer, more complex sentences as well as changes in the use of tense, determiners and connectors (Fayol, 2000; Hickmann, 2003; Jisa, 2004). The pragmatic and discourse performance also changes, as the children include changes on the processing of ground information: background *versus* foreground, more freedom in the processing of the event frame (Fayol, 1997), and various speech acts such as narrating, explaining, commenting on the narrative or on the narration (McNeill, 1992 ; Colletta, 2004). The verbal level of our annotation scheme thus includes not only an orthographical transcription, but also a syntactic analysis and a discourse analysis (see figure 1: annexures).

2.1 Speech transcription and syntactic analysis

The transcription of the speakers' words appears on two tracks: one track for the experimentator and one for the child or the adult. The transcription is orthographical and presents the entirety of the remarks of the speakers.

In order to study age related changes in the subject's narrative performance, we first segment the speech into speech turns. To annotate and cut down the speech turns is important to see from what age the child is able to achieve a monologic narrative task in one-go, without assistance from the adult (on the development of monologic discourse, see Jisa & Kern, 1998; Hickmann, 2003; Colletta, 2004; Jisa, 2004). We then segment the speech into clauses and words. The number of clauses or the number of words contained in an account provides a good indication of its informational quantity, which is likely to grow with age. We also classify the clauses of the corpus in order to see whether there is or there isn't a change towards complex syntax in the course of age development. We annotate the words to identify clues of subordination such as conjunctions, relative pronouns or prepositions. Our coding scheme relies on Berman & Slobin's work, (1994), and on Diessel's analysis of the children's syntactic units, in Diessel, (2004). The annotation of words also serves to identify connectives and anaphoric expressions (pronouns, nouns, determiners, etc.) which play an important role in discourse cohesion (de Weck, 1991; Hickmann, 2003).

2.2 Discourse analysis

Before the annotation grid was completed, the extract of the Tom & Jerry cartoon was segmented into macro and micro-episodes. During the annotation process, each clause with narrative content is categorised as processing one of these macro and micro-episodes in order to have an estimate of the degree of accuracy of the retelling of the story by each subject as well as to study his/ her processing of the event frame (Fayol, 1997). Each clause is also categorised as expressing the part or whole of a speech act (narrating, explaining, interpreting or commenting) and as expressing foreground *versus* background of the story. It is a question of studying how age and culture affect pragmatic and discourse dimensions of the narrative activity, as also seen in Hickmann (2003).

The mean duration for the annotation of the verbal level, which includes the transcription of the words of the speakers, syntactic analysis, discourse analysis and validation of all annotations, is 6 hours per file.

3. Annotation of the gesture level

In general, the annotation schemes developed by researchers in computer sciences mainly focus on the description of corporal movements and the form of gestures. It is a question of capturing, as finely as possible, the corporal movements, or to allow for an automatic synthesis. (Kipp, Neff & Albrecht, 2006; Le Chenadec, Maffiolo, Château & Colletta, 2006; Kipp, Neff, Kipp & Albrecht, 2007; Le Chenadec., Maffiolo & Chateau, 2007). Our objective is very different as the annotation has to allow us to study the relationship between gesture and speech. As a consequence, only the corporal movements maintaining a relation to speech - coverbal gesture - interest us. This relationship as well as the function filled by the gesture has a lot of significance for us.

The gesture annotation is carried out in parallel by two independent coders 1 and 2, who annotate on five stages (see figure 2: annexures). Why five stages? In our developmental perspective, the five following parameters

[1] Available from http://www.mpi.nl/tools/ . Also see Brugman and Russel (2004).

prove to be interesting; To begin with; the number of coverbal gestures, which as one would expect, increases with age as we see longer, more detailed and more complex narratives (cf. Colletta, 2004). Another key parameter is the function of gesture. If the hypothesis of a gesture-word system is valid, then we ought to observe age related changes in gesture, with more gestures of the abstract and gestures marking discourse cohesion in the older children's and the adults' performance. The third important parameter is the gesture-speech relationship, which should evolve in parallel with linguistic acquisition and provide evidence of the evolution of language performance towards a more elaborated pragmatic and discursive use (McNeill, 1992; Colletta and Pellenq, 2007). The fourth parameter which is likely to vary with the age of the subjects is the manner which gestures and speech occur on the temporal level (synchrony and anticipation). The fith parameter is gesture form, which in addition to representational accuracy (for representational gestures) in the older children and the adults, should gain more precision in use. (see our three criteria below).

Other than the developmental perspective, every one of these five parameters is likely to vary with the language and culture of the subjects. The study of the interactions between age on one side, and language and culture on the other side, should lead us to a better understanding of the role played by linguistic, cognitive and social factors in multimodal language acquisition.

3.1 Identification of the gestures

In Kendon's work (1972, 1980, 2004), a pointing gesture, a representational gesture or any other hand gesture (an excursion of the body during speech) is called a "gesture phrase" and it possesses several phases including the "preparatory stage, the stroke, i.e., the meaningful part of the gesture phrase, the retraction or return and the repositioning for a new gesture phrase". Yet, some gestures are nothing else but strokes: a head gesture or a facial expression, for instance, are meaningful right from the start till the end of the movement and have no preparatory nor any retraction phases. As a consequence, our premise is that the "gesture stroke" is any coverbal gesture phrase or isolated gesture stroke that needs to be annotated.

To identify the gesture, each coder takes into account the three following criteria (based on Adam Kendon's proposals in Kendon, 2006):

(i) If the movement is easy to perceive, of good amplitude or marked well by its speed,

(ii) If location is in frontal space of locutor, for the interlocutor.

(iii) If there is a precise hand shape or a well marked trajectory.

Once a gesture has been identified, the coder annotates its phases using the following values (based on Kendon, 2004):

<Stroke> = the meaningful height of the excursion of the gesture phrase of a hand gesture, or a movement of the head, shoulders or chest, or a facial display.

<Prep> = the movement which precedes a hand gesture stroke, which takes the hand(s) from its (their) initial position (at place of rest) to where the gesture begins. Contrary to hands, the position of head, the bust or shoulders is fixed. These movements can therefore not be "prepared" as hand movements and consequently can only be annotated as "strokes".

<Hold> = the maintaining of the hand(s) in its (their) position at the end of a hand gesture stroke, before the returning phase or a chained gesture.

<Chain> = the movement which brings the hand(s) from its (their) initial position at the end of a hand gesture stroke to the place where a new stroke begins, without returning to a rest position between the two strokes.

<Return> = the movement which brings back the hand(s) from its (their) position at the end of a hand gesture stroke to a rest position, identical or not to the preceding one (called "recovery" in Kendon, 2004).

3.2 Attributing function to gesture

The coder then attributes a function to each gesture stroke. In literature about gesture function, there generally appears to be agreement amongst gesture researchers, although they do not always agree on terminology. According to several researchers, Scherer (1984), McNeill (1992), Cosnier (1993), Calbris (1997), Kendon, 2004), 4 main functions are always mentioned:

(i) gestures that help identify (pointing gestures) or represent concrete and abstract referents;

(ii) gestures that express social attitudes, mental states and emotions and that help perform speech acts and comment on own speech as well as other's;

(iii) gestures that mark speech and discourse, including cohesion gesture;

(iv) gestures that help to synchronise own-behaviour with interlocutor's in social interaction.

Our gesture annotation scheme mostly relies on Kendon's classification and covers the whole range of these functions. The coder selects between:

<Deictic> = hand or head gesture pointing to an object present in the communication setting, or to the interlocutor, or to oneself or a part of the body, or indicating the direction in which the referent is found from the actual coordinates of the physical setting. Not all pointing gestures have a deictic function as deictic pointing gesture strictly implies the presence of the referent or its location from the actual physical setting. Thus, gestures which locate a virtual character, object or action (like in sign languages of deaf communities) are to be annotated under <representational>.

<Representational> = hand or facial gesture, associated or not to other parts of the body, which represents an object or a property of this object, a place, a trajectory, an action, a character or an attitude (ex: 2 hands drawing the form of the referent; hand or head moving in some direction to represent the trajectory of the referent; 2 hands or body mimicking an action), or which symbolises, by metaphor or metonymy, an abstract idea (ex: hand or head gesture pointing to a spot that locates a virtual character or object; hand or head movement towards the left or the right to symbolise the past or the future; gesture metaphors for abstract concepts).

<Performative> = gesture which allows the gestural realisation of a non assertive speech act (ex: head nod as a "yes" answer, head shake as a "no" answer), or which reinforces or modifies the illocutionary value of a non assertive speech act (ex: vigorous head nod accompanying a "yes" answer).

<Framing> = gesture occurring during assertive speech acts (during the telling of an event, or commenting an aspect of the story, or explaining) and which expresses an emotional or mental state of the speaker (ex: face showing amusement to express the comical side of an event; shrugging or facial expression of doubt to express incertitude of what is being asserted).

<Discursive> = gesture which aids in structuring speech and discourse by the accentuation or highlighting of certain linguistic units (ex: beat gesture accompanying a certain word; repeated beats accompanying stressed syllables), or which marks discourse cohesion by linking clauses or discourse units (ex: pointing gesture with an anaphoric function, e.g. pointing to a spot to refer to a character or an object previously referred to and assigned to this spot; brief hand gesture or beat accompanying a connective).

<Interactive> = gesture accompanied by gaze towards the interlocutor to express that the speaker requires or verifies his attention, or shows that he has reached the end of his speech turn or his narrative, or towards the speaker to show his own attention (ex: nodding head while interlocutor speaks).

<Word Searching> = Hand gesture or facial expression which indicates that the speaker is searching for a word or expression (ex: frowning, staring above, tapping fingers while searching for words).

3.3 Definition of the relation of gesture to corresponding speech

The third stage consists in giving a definition of the relation of the gesture to corresponding speech.

<Reinforces> = the information brought by the gesture is identical to the linguistic information it is in relation with (ex: head nod accompanying a yes answer; face expressing ignorance while saying "I don't know"). This annotation does not concern the representational gestures, because we consider that information brought by the representational gesture, due to its imagistic properties, always says more than the linguistic information, as per McNeill (1992) or Kendon (2004). See <Integrates>.

<Complements> = the information provided by the gesture brings a necessary complement to the incomplete linguistic information provided by the verbal message: the gesture disambiguates the message, as in the case of deixis (ex: pointing gesture accompanying a location adverb like « here », « there »; pointing gesture aiming at identifying an object not explicitly named).

<Supplements> = the information brought by the gesture adds a supplementary signification to the linguistic information, like in the case of framing gestures and certain performative gestures (ex: vigorous shaking of head accompanying a no answer; face showing amusement signs to express a comical side of an event; shrugging or showing a mimic of doubt to express incertitude of what has been asserted).

<Integrates> = the information provided by the gesture does not add supplementary information to the verbal message, but makes it more precise, thanks to the imagistic properties of gesture. For instance, drawing a trajectory provides information on the location of the characters or objects we refer to, drawing the shape of an object may at the same time give information on its dimensions.

<Contradicts> = the information provided by the gesture is not only different from the linguistic information in which it is linked but contradicts it, as in the case of certain framing and performative gestures.

<Substitutes> = the information provided by the gesture replaces linguistic information, as in the case of certain performative and interactive gestures (ex: the speaker nods as a yes answer, shakes head as a no answer, shrugs to express his ignorance of the information required).

3.4 Indication of the temporal placement of the

gesture in relation to the corresponding speech

The fourth stage indicates the temporal placement of the gesture stroke in relation to the corresponding speech:

<Synchroneous> = the stroke begins at the same time as the corresponding speech segment, whether it is a syllable, a word or a group of words.

<Anticipates> = the stroke begins before the corresponding speech segment: the speaker starts his gesture while delivering linguistic information prior to the one corresponding to it.

<Follows> = the stroke begins after the corresponding speech segment: the speaker begins his gesture after having finished speaking, or while delivering a linguistic information posterior to the one corresponding to it.

3.5 Gesture form

Kipp, Neff & Albrecht (2006) mention two distinct ways to describe gesture form: "gesture form is captured by either a free-form written account or by gestural categories which describe one prototypical form of the gesture". In our work, as we focus on gesture function and gesture-speech relation, we rely on basic linguistic descriptions of the body movements.

The coder gives a brief linguistic description of each annotated gesture stroke, sticking to its most salient points:

- body part of movement: head, chest, shoulders, 2 hands, left hand, right hand, index, eyebrows, mouth, etc.

- if there is a trajectory: direction of the movement (towards the top, bottom, left, right, front, back, etc.)

- if there is a hand shape: the form of the hand (flat, cutting, closed in a punch-like form, curved, palm up, palm down, fingers pinched, fingers in a circle, etc.)

- the gesture itself: head nod, beat, circular gesture, rapid or not, repeated or not, etc.

4. Validation of the gestures' annotation

In most cases, the validation of the gestural annotation is based on the comparison of the annotations done by two independent coders, and even more rarely, on re-creating gestures by an animated agent (Kipp, Neff and Albrecht, on 2006). These methods are useful to test the validity of an annotation scheme, but they do not allow to check and to stabilise the analysis of a corpus at the end of an annotation procedure. Indeed, in our case, it is not only a question of testing a gestural annotation grid, but it is also a question of validating the annotation of a multimodal corpus (gestures+speech) before using the results of the annotation in statistical analyses.

As a consequence, the last step of the analysis covers two objectives:

- firstly, to finalise the gestural annotation from choices made by both coders and decide in case of disagreement;

- secondly, calculate the interreliability of agreement between all the coders.

The validation phase only applies to the first three parameters (identification of a gesture unit, function and relation to speech), as our goal is to check whether they vary as a function of age and culture. It does not apply to the fith parameter because gesture form is written in free form and therefore the coders can see the same gesture differently, which will be usefull in a more detailed and qualitative analysis. Nor does it apply to the fourth parameter (temporal placement), which will be useful too in such an analysis.

In order to achieve the validation task, a third coder independent of the first two, controls all the annotations. She first adds a supplementary track and annotates "agreement" when she agrees with both coders on the presence of a gesture, or when at least two coders on three agree ont the presence of a gesture, and "disagreement" on the contrary. She then adds two additional tracks to annotate using the same method of "agreement" *versus* "disagreement" for gesture function and gesture relation to speech. She furthermore proceeds to create three new tracks which have the definite annotation of gestures, gesture functions and gesture-speech relation which will help in quantitative analysis.

This last analysis step allows a measure of interreliability amongst the coders and is useful to enhance the process of validation of the annotation. We then calculate:

- Interreliability for the identification of gestures: number of agreement / number of gesture strokes per file.

- Interreliability for the identification of gesture function: number of agreement / number of gesture strokes per file.

- Interreliability for the identification of gesture-speech relation: number of agreement / number of gesture strokes per file.

The mean duration for the annotation of the gesture level, including the validation and final annotation, is 12 hours per file. The duration time varies a lot and is certainly dependant on the subject's communication behaviour, as some gesture far more than others.

5. Final remarks

The project described in this presentation requires the use of a transcription tool and the annotation of both verbal and gesture data. To fulfil our aim, we chose to use the annotation software *ELAN*, a multi-track software with the alignment of transcription of audio and video sources. A multilevel annotation makes it possible to study the gesture-word relations in a very concise manner. It makes it possible to identify, count and describe concrete *versus* abstract representational gestures, marking of connectives, syntactic subordination, the anaphoric recoveries, hesitation phenomena, etc. as well as to study narrative behaviour from a multimodal perspective.

Yet, some technical issues need to be enhanced: the gesture annotation can be more precise if one dissociates the body parts: head, face, hand(s), the whole body. This would avoid the fact that for the same complex gesture involving several body parts, several coders code different aspects of the same behaviour. Moreover, this is painstaking, particularly for adult gestures, where the same gesture can perform two, even three functions simultaneously, which means that the values given in the drop-down menus should, in the future, include this pluri-function feature.

Presently, the analysis in progress will make it possible to appreciate the use of our validation procedure of the gesture annotations… a crucial issue …

6. References

Abrilian, S. (2005). Annotation de corpus d'interviews télévisées pour la modélisation de relation entre comportements multimodaux et émotions naturelles. *6ème colloque des jeunes chercheurs en Sciences Cognitives (CJCSC'2005), Bordeaux, France.*

Beattie, G. (2003). *Visible Thought: The New Psychology Of Body Language*. Routledge, London.

Berman, R.A., Slobin, D.I. (1994). *Relating events in narrative : A crosslinguistic developmental study*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Butcher, C., Goldin-Meadow, S. (2000). Gesture and the transition from one- to two-word speech : When hand and mouth come together. In D. McNeill (Ed.), *Language and gesture.* Cambridge, Cambridge University Press, pp. 235--257 .

Brugman, H., Russel, A. (2004). Annotating Multi-media / Multi-modal resources with ELAN. In *4th International Conference on Language Resources and Language Evolution (LREC2004), Lisbon, 26-28 may*.

Calbris, G. (1997). Multicanalité de la communication et multifonctionnalité du geste. In J. Perrot, *Polyphonie pour Yvan Fonagy*, Paris, L'Harmattan.

Calbris, G. (2003). *L'expression gestuelle de la pensée d'un homme politique*. Paris, Editions du CNRS.

Capirci, O., Iverson, J.M., Pizzuto, E., Volterra, V. (1996). Gesture and words during the transition to two-word speech. *Journal of Child Language*, 23, pp. 645--673.

Colletta, J.-M. (2004). *Le développement de la parole chez l'enfant âgé de 6 à 11 ans. Corps, langage et cognition*. Hayen, Mardaga.

Colletta, J.-M., Pellenq, C. (2007). Les coverbaux de l'explication chez l'enfant âgé de 3 à 11 ans. *Actes du 2e Congrès de l'ISGS : Interacting bodies, corps en interaction, Lyon, 15-18 juin 2005, CDRom Proceedings*.

Cosnier, J. (1993). Etude de la mimogestualité. In R. Pléty, *Ethologie des communications humaines : aide-mémoire méthodologique*. Lyon, ARCI et Presses Universitaires de Lyon, pp. 103--115.

Diessel, H. (2004). *The acquisition of complex sentences*. Cambridge, Cambridge University Press.

Ech Chafai, N., Pelachaud, C., Pelé, D. (2006). Analysis of gesture expressivity modulations from cartoons animations. In *LREC 2006 Workshop on "Multimodal Corpora", Genova, Italy, 27 May*.

Fayol, M. (1997). *Des idées au texte. Psychologie cognitive de la production verbale, orale et écrite*. Paris, P.U.F.

Fayol, M. (2000). Comprendre et produire des textes écrits : l'exemple du récit. In M. Kail et M. Fayol, *L'acquisition du langage, T.2 : Le langage en développement. Au-delà de trois ans*. Paris, P.U.F., pp. 183--213.

Goldin-Meadow, S. (2006). Talking and thinking with our hands. *Current Directions in Psychological Science*, 15, pp. 34--39.

Grynszpan, O., Martin, J.C., Oudin, N. (2003). On the annotation of gestures in multimodal autistic behaviour. In *Gesture Workshop 2003*, *Genova, Italy, 15-17 April*.

Hickmann, M. (2003). *Children's discourse : person, space and time across languages*. Cambridge, Cambridge University Press.

Jisa, H. (2004). Growing into academic French. In R. Berman (ed.), *Language Development across Childhood and Adolescence, Trends in Language Acquisition Research, vol.3*. Amserdam/Philadelphia, John Benjamins Publishing Company, pp. 135--161.

Jisa, H., Kern, S. (1998). Relative clauses in French children's narrative texts. *Journal of Child Language*, 25, pp. 623--652.

Kendon, A. (1972). Some relationships between body motion and speech. In A.W. Siegman et B. Pope (eds.), *Studies in dyadic communication*. Elmsford, NY, Pergamon Press, pp. 177--210.

Kendon, A. (1980). Gesticulation and speech, two aspects of the process of utterance. In M.R. Key (ed.), *The*