Multi-level Annotations of Nonverbal Behaviors in French Spontaneous Conversation
Ning Tan, Gaëlle Ferré, Marion Tellier, Edlira Cela, Mary-Annick Morel, Jean-Claude Martin, Philippe Blache

To cite this version:
Ning Tan, Gaëlle Ferré, Marion Tellier, Edlira Cela, Mary-Annick Morel, et al.. Multi-level Annotations of Nonverbal Behaviors in French Spontaneous Conversation. International Conference for Language Resources and Evaluation, May 2010, Malta. pp.74-79. hal-00488832

HAL Id: hal-00488832
https://hal.science/hal-00488832
Submitted on 3 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Multi-level Annotations of Nonverbal Behaviors in French Spontaneous Conversations

Ning TAN\textsuperscript{1}, Gaele FERRE\textsuperscript{2}, Marion TELLIER\textsuperscript{3}, Edlira CELA\textsuperscript{4}, Mary-Annick MOREL\textsuperscript{4}, Jean-Claude MARTIN\textsuperscript{1}, Philippe BLACHE\textsuperscript{5}

\textsuperscript{1} LIMSI-CNRS BP 133
91403 Orsay Cedex France
\textsuperscript{2} Centre International de Langues - Département d'Études Anglaises - Chemin de la Censive du Tertre BP 81227
44312 Nantes Cedex 3 France
\textsuperscript{3} Université de Provence UFR LACS - Département de FLE 29 avenue Robert Schuman
13621 Aix-en-Provence Cedex 1 France
\textsuperscript{4} Université Paris 3 UFR de Littérature et Linguistique Françaises et Latines, CLF, 13 rue Santeuil
75005 Paris France

E-mail: \{ntan, martin\}@limsi.fr, gaele.ferre@univ-nantes.fr, marion.tellier@univ-provence.fr, edlira.cela@yahoo.fr, marym@edilang.com, blache@jpl-ait.fr

Abstract

During spontaneous conversations, multiple modalities such as gesture, posture and gaze are combined in sophisticated ways for different functions such as spatial references. In the French research community, there is a lack of spontaneous multimodal corpora for the French language. This paper describes the multi-level scheme that we have defined for the annotation of gesture, posture and gaze. We explain how we applied it for the annotation of a corpus of spontaneous French conversations. Two types of analyses were made on the resulting annotations. Firstly, intercoder agreement was computed on the annotations of gesture space. Secondly, we clustered chunks of postures into categories. Such research will enable the joint study of multiple nonverbal modalities such as the relations between gesture semantics and posture.

1 Introduction

During spontaneous conversations, multiple modalities such as speech, gesture, posture and gaze are combined in sophisticated ways. Several studies considered how speech and gestures co-occur. Yet, studies on the way in which the different nonverbal modalities interact are deficient, partly due to the lack of accessible and relevant resources. For example, the multimodal corpora that are currently available for the French research community are limited in terms of accessibility, spontaneity and levels of annotation. In addition, studying relations between those modalities requires the definition of reliable coding schemes, describing multiple levels from the phonological to the gestural and postural levels.

The present work lies within the framework of the OTIM project (Blache et al., 2009), which aims at building such French digital corpus of multimodal behaviors occurring during spontaneous conversations. The project makes use of the Corpus of Interactional Data (CID) which is an audio-video database of spontaneous spoken French (Bertrand et al., 2008). Eight pairs of native French speakers take part separately in a one-hour dyadic spontaneous spoken conversation, in which each speaker tells unexpected personal experiences (Figure 1). The CID corpus is one of the first multimodal corpora in French.

In this paper we focus on a subset of nonverbal modalities which are relevant for studying spatial references occurring during conversations. This function is shared by several nonverbal modalities such as gesture, gaze and posture. The speech and linguistic aspects are beyond the scope of this paper (Bertrand et al., 2008). After specifying a coding scheme for gesture, postures and gaze (sections 2-4), we measure the inter-coder agreement between three independent coders on a subpart of the annotations (section 5.1). Finally, we study the relations between posture and gesture (section 5.2).

Figure 1: The CID French spontaneous conversation corpus.

2 Gesture

Different typologies have been adopted for the classification of gestures, based on the work by Kendon (1980) and McNeill (1992, 2005). The formal model we use for the annotation of hand gestures is adapted from the specification files created by Kipp (2003) and from the MUMIN coding scheme (Allwood et al., 2005). Both models consider McNeill’s research on gestures (1992,
produced away from the speaker in the extreme periphery, while has very small amplitude if the hand was already in this part of the gesture space.

75 minutes of the CID involving 6 speakers have been coded for hand gestures. The annotation yielded a total number of 1477 gestures. The numbers of hand gestures per semiotic type are listed in the table below.

<table>
<thead>
<tr>
<th>Gesture Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptors</td>
<td>334</td>
</tr>
<tr>
<td>Beats</td>
<td>166</td>
</tr>
<tr>
<td>Butterworth</td>
<td>36</td>
</tr>
<tr>
<td>Deictics</td>
<td>137</td>
</tr>
<tr>
<td>Emblems</td>
<td>147</td>
</tr>
<tr>
<td>Iconics</td>
<td>286</td>
</tr>
<tr>
<td>Metaphorics</td>
<td>371</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1477</strong></td>
</tr>
</tbody>
</table>

Table 1: Number of hand gestures annotated in the CID corpus.

3 Gaze

Previous studies observed that the speaker gazes away from the listener before the beginning of his speech turn (Tabensky, 1997; Bouvet & Morel, 2002), and that he returns his gaze towards the listener slightly before the end of the turn (Bouvet, 1997; Tabensky, 1997; Cuxac, 2000). There are different explanations for gaze shift: not only is it a way for the speaker to show the conversational partner that he is entitled to proceed to his turn at speech, but also that avoiding eye contact facilitates lexical retrieval, and lastly that gaze direction allows the speaker to better visualize the various referents in space (Bouvet, 1995; Cuxac, 2000; Bouvet & Morel, 2002). When the speaker gazes back at the listener, he hereby acknowledges that the listener has been addressed (Cuxac, 2000), while softening the expression of a personal standpoint. It may also be interpreted by the listener as a possible turn taking, although he may decline the offer of a turn by simply adding a verbal or gestural backchannel.

Gaze mobility reveals the fluidity of the dialogue: when the speaker gazes too long at the listener, he may become menacing and that would underline the uncooperative aspect of the interaction (Tabensky, 1997). Conversely, when the speaker gazes away from the listener too long, his gaze avoidance may be interpreted as disinterest as regards the listener’s feedback.

At last, the shift between mutual eye contact between participants and either gaze towards particular reference points in the gesture space gives it a deictic value (Cuxac, 2000), or towards the speaker’s own hands in which case the shift reveals to the listener the strong personal involvement of the speaker (Bouvet, 1997).

Gaze refers to three separate productive organs: brows, eyelids and eyes. Poggi (2001) established a set of
formational parameters to analyze gaze, roughly including eyebrows (inner part, medial part and outer part), eyelids (upper or lower), wrinkles, and eyes (humidity, reddening, pupil dilation, eye position and eye direction). Other existing gaze coding scheme focus on the direction of the speakers’ gaze as mutual gaze or non-mutual (Cerrato, 2005), eye squinting, eyebrow raising or lowering (Foster et al., 2007), coarse gaze checking whether the person is looking at the whiteboard or at another person (Carletta, 2007).

Based on these studies, we defined a gaze annotation scheme enclosing ‘gaze direction (side)’ and ‘global gaze’. The ‘gaze direction (side)’ describes eye movement orientation (up, down, sideways, left, right, around) and its deictic target (interlocutor/object). The ‘global gaze’ helps to capture a global view of gaze directions and the most common gaze poses (Wallhoff et al., 2006).

4 Posture
Posture shift may be linked to interpersonal attitudes (Argyle, 1988), emotions (Sherer, 2001), communicative styles (Richmond, 1995), and personalities (Ibister & Nass, 2000). One can reduce human postures to three categories: standing, sitting (including squatting and kneeling) and lying (Argyle, 1988).

Methods of describing postures are relatively domain-related. They vary from grid-based observational studies to technical measures. Ergonomists (Corlett et al., 1986) define postures as conditions of the body. Due to the three-dimensional character within the three-dimensional space, they use three methods to measure and describe postures by appropriate definable geometrical parameters (three coordinates of individual joints, or adjustment levels of the long axes of separate parts of the body relative to the surrounding absolute space, or those to the axis of the preceding body part). Although these interval-scale-oriented parameters can cover all possible rotational movements in any joint, only the third type takes anatomical facts into consideration (e.g. limited range of movements). For practical applications, the ordinal-scale-oriented methods are more commonly used than those interval-scale-oriented methods described above. It is due to the advantage of their being independent (or semi-independent) from technical aids for posture recording while keeping accuracy or reproducibility (e.g. using a body diagram with limb displacement segments for recording postures). Potential displacement can be assessed by the segments within the range of movement. The nominal-scale methods place the description of postures according to posture typologies, which are generally profession-related. Human anatomists (White & Folkens, 1991; Platzer & Kaltes, 2004) classify body motion as flexion vs. extension (the act of bending or straightening), abduction vs. adduction (movement away from or toward the median plan), internal rotation vs. external rotation (movement around an axis), and elevation vs. depression (movement adjusting the height). The terms describe the act of performing body movements as well as the body position after being moved. Psychologists (Scherer & Ekman, 1983) distinguish posture behavior from action behavior. The posture behavior refers to overall postures (sitting, standing, lying), frontal orientation of trunk (facing, turned away), trunk lean (forward, straight, backward, sideways), arm and leg position (folded arms, uncrossed legs) and feet (flat on floor, under chair, on other knee). In the Posture Scoring system (Bull, 1987), any movement which is taken up and maintained for at least one second is annotated as a posture. This system covers head, arms, trunk and legs. Bull also proposed a second system using a dynamic approach, which describes the posture in terms of a series of movements rather than static positions.

The previous annotation scheme for the CID corpus (Bertrand et al., 2008) only considered chest movements at trunk level. Aiming at extending the postural sphere, we added a set of tracks and attributes relevant to sitting positions met in the CID corpus. It is based on the Posture Scoring System (Bull, 1987) and the Annotation Scheme for Conversational Gestures (Kipp et al., 2007). Our scheme covers four body parts: arms, shoulders, trunk and legs. With respect to arm, Bull’s system mainly distinguishes whether the hand touches or not; while Kipp’s scheme covers four spatial dimensions to capture it. We made a trade-off decision between the two systems: we kept the four dimensions from Kipp’s coding scheme with respect to the height, distance radial orientation and swivel degree of the arm, and created a new track describing the hand touching objects to get back to the ideas of Bull’s system (Table 2). Also, we added two dimensions to describe respectively the arm posture in the sagittal plane and the palm orientation of the forearm and the hand. With respect to the leg posture (Table 2), we added three dimensions: the height of the feet, the orientation of the thigh and the way in which the legs are crossed (only suited to sitting positions).

<table>
<thead>
<tr>
<th>Arms</th>
<th>Legs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>Height</td>
</tr>
<tr>
<td>Distance</td>
<td>Distance</td>
</tr>
<tr>
<td>Radial Orientation</td>
<td>Radial Orientation</td>
</tr>
<tr>
<td>Sagittal Orientation</td>
<td>Sagittal Orientation</td>
</tr>
<tr>
<td>Arms Swivel</td>
<td>Leg-to-leg Distance</td>
</tr>
<tr>
<td>Forearm and Hand Orientation</td>
<td>Crossed Leg</td>
</tr>
<tr>
<td>Touch</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Attributes to encode the spatial configurations of arms and legs.

We annotated postures on 15 minutes of the corpus involving one pair of speakers.

The proposed coding scheme leads to annotating separately the positions of the different body parts. However, we prefer to highlight the postures, which the speakers commonly take up. To this end, we proceed in two steps. First, we export data from Anvil and retrieve them into a set of significant postures according to the
following criteria:
- exclude the frames in which there is no annotation in any track;
- exclude the repetitive and successive frames.

These criteria are relevant to those of the Posture Scoring System (Bull, 1987), in which the author emphasized that if the speaker moves from the current posture without establishing a different one and then returns to the original posture, the time spent moving should be excluded from the total duration of that posture. For this reason, we exclude the frames in which there is movement in any body parts. Second, we apply a simple Hierarchic Clustering (Euclidean distance) on the extracted data. We analyzed all parts of the video data to initialize the categorization process. This method enables us to make a global-view analysis of the posture annotations and cluster different posture combinations into similar categories, which is consistent with the idea of establishing a set of posture lexicons linking body spatial locations to posture communicative functions.

Three whole body posture types have been found for one of the speakers during the first 15 minutes of the recording. In Figure 2 we selected three representative frames to illustrate the most common postures. Over 89 extracted postures, type #1 represents 44% of the postures, type #2 represents 32.5%, and type #3 represents 15.7%. Posture type #2 occurs mostly at the opening of a turn by the speaker, while posture #1 occurs at the closing sequence; posture #3 occurs mostly in the continuing sequence for the listener.

![Figure 2: Illustrations of three typical postures occurring during the first 15 minutes of conversation for one speaker.](image)

5 Multimodal annotations

5.1 Gesture Space

The Gesture Space is a “shallow disk in front of the speaker” (McNeill, 1992: 86) where most gestures are performed. It is divided into four regions (center-center, center, periphery and extreme periphery) and eight coordinates (no coordinate, right, left, left-right, upper right, upper left, lower right, lower left, upper, lower, upper left-right, lower left-right). We used McNeill’s (1992: 89) diagram for the coding. The left-right coordinate was useful whenever a gesture was produced with both hands.

We performed a measure of inter-reliability for three independent coders for Gesture Space. The measure is based on Cohen's corrected kappa coefficient for the validation of coding schemes (Kita & Ozyurek, 2003).

Three coders have annotated three minutes for GestureSpace including GestureRegion and GestureCoordinates. Annotations of these two dimensions are based on the point of maximum extension of the gesture. When annotating GestureRegion, the gesture that occurred in the axis of the armrests of the chair is judged in periphery. When it is outside, we annotate extreme periphery. The landmark is the position of the wrist.

<table>
<thead>
<tr>
<th>Kappa (k)</th>
<th>Gesture Region</th>
<th>Gesture Coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>RightHand.GestureSpace</td>
<td>0.669</td>
<td>0.257</td>
</tr>
<tr>
<td>LeftHand.GestureSpace</td>
<td>0.674</td>
<td>0.592</td>
</tr>
</tbody>
</table>

Table 3: Average kappa values of intercoder agreement measures for each categorical item of GestureSpace for three coders

Strong agreement has been achieved for GestureRegion, which attained 64.9% agreement for the right hand; and 67.4% for the left hand. The GestureCoordinates indicated 25.7% agreement for the right hand, and 59.2% for the left hand. Such low agreement of GestureCoordinates might be due to several factors. First, the number of categorical values is high. Second, three minutes might be limited in terms of data to run a kappa measure. Third, GestureRegion affects GestureCoordinates: if the coders disagree about GestureRegion, they are likely to also annotate GestureCoordinates in a different way. For instance, it was decided that no coordinate would be selected for a gesture in the center-center region, whereas there is a coordinate value for gestures occurring in other parts of the GestureRegion. This means that whenever coders disagree between the center-center or center region, the annotation of the coordinates cannot be congruent.

Yet, strong agreement has been attained between one pair of coders for both GestureRegion and GestureCoordinates (Table 4).

<table>
<thead>
<tr>
<th>Kappa (k)</th>
<th>Gesture Region</th>
<th>Gesture Coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>RightHand.GestureSpace</td>
<td>0.748</td>
<td>0.609</td>
</tr>
<tr>
<td>LeftHand.GestureSpace</td>
<td>0.755</td>
<td>0.681</td>
</tr>
</tbody>
</table>

Table 4: Average kappa values of intercoder agreement measures for GestureSpace for two coders

5.2 Posture and gesture

We also investigated how posture overlaps with significant gestures. We chose to perform an association analysis between gesture phrases and arm postures (Kipp & Martin, 2009), and made use of all the annotated data without emphasizing individual characteristics and handedness. 18 different values derived from five spatial
location dimensions, are associated with six values of gesture phrases. We observed that, most of the metaphorical gestures co-occurred with the following postures (only at arm level):
- place hands in the height of abdomen;
- locate arm at some distance away from the body;
- let upper arm slightly hang on the side with semi-flexed forearm;
- keep arm in front of the body.

This effect that has been observed only with metaphorical gestures may be due to its high production rate compared to other semiotic gesture types, as described in Table 1.

6 Conclusions
In this study, we proposed a coding scheme that describes the nonverbal signals produced by two sitting speakers in spontaneous conversations in French. Establishing such a multi-level nonverbal coding scheme makes it possible to understand how different nonverbal modalities co-work in multimodal information production.

We see two approaches to continue the validation work. The first is to apply agreement measures with respect to gaze direction, hand shape, orientation, trajectory and velocity, with which we can run a number of association investigations about relationships between gaze and gesture. Another approach consists in validating each individual scheme using similar annotation features. For example, both GestureSpace in the gesture scheme and ArmDistance in the posture scheme, describe the arm/hand position in the body median plane. We intend to measure agreement based on these similar annotations to validate the related coding schemes.

These annotations of nonverbal behaviors will be jointly studied with previous speech and linguistic annotations of the same data as well as future annotations planned about communicative functions.

7 Acknowledgements
The work described in this paper is supported by the French Agence Nationale de la Recherche (ANR) under the project OTIM (ANR BLAN08-2 349062). The authors would like to thank Roxane BERTRAND from the University of Provence, and Michael KIPP for the Anvil tool.

8 References


