Multimodality Issues in Conversation Analysis of Greek TV Interviews

Maria Koutsombogera[1,2] and Harris Papageorgiou[1]

[1] Institute for Language & Speech Processing, Artemidos 6&Epidavrou, 15125 Athens, Greece
[2] University of Athens, Department of Linguistics, University Campus, 15784 Athens, Greece
{mkouts,xaris}@ilsp.gr

Abstract. This paper presents a study on multimodal conversation analysis of Greek TV interviews. Specifically, we examine the type of facial, hand and body gestures and their respective communicative functions in terms of feedback and turn management. Taking into account previous work on the analysis of non-verbal interaction, we describe the tools and the coding scheme employed, we discuss the distribution of the features of interest and we investigate the effect of the situational and conversational interview setting on the interactional behavior of the participants. Finally, we conclude with comments on future work and exploitation of the resulting resource.

Keywords: non-verbal/multimodal expressions, feedback, turn management, conversation analysis, interview corpus.

1 Introduction

Relations between distinct modalities in natural interaction have been thoroughly studied [1,2,3] in order to deeper understand research questions related to multimodal communication areas such as emotional behavior [4], human-avatar interaction [5,6] etc. In this paper we present a cross-disciplinary research on the communicative role of multimodal expressions in TV face-to-face interviews occurring in various settings.

In general, TV discussions present a mixture of characteristics oscillating between institutional discourse, semi-institutional discourse and casual conversation [7,8]. This media content spans a variety of discourse types such as information exchange, entertainment and casual talk. The setting in which the interview takes place, as well as the social and discursive roles of the speakers are features that formulate the discourse structure and further influence the interactants' conversational behavior in all its expressive dimensions (speech, hand and facial gestures etc.).

Our motivation is to identify and interpret gestural features that critically contribute to the conversational interaction. Specifically, we describe their interrelations as well as their distribution across different types of TV discussions in an attempt to find evidence about their potential systematic role. In this context, we

take a first step towards the description and annotation of a multimodal corpus of Greek TV interviews, available for further development and exploitation.

2 Corpus Description

The corpus comprises 66 minutes of interviews extracted from 3 different Greek TV shows. Their structure consists of question-answer sequences performed by an interviewer (the *host*) to an interviewee (the *guest*). Apart from the commonly shared features, each interview is uniquely outlined by its setting, its topic, the roles and personalities of its speakers, their interests and their commitments.

The first interview (I1) takes place in a TV studio between the host and a politician and provides information concerning current political issues. It can be regarded as an institutionalized interaction, as it appears to be more standardized in role distribution and turn management predictability.

The second interview (I2) is a pre-filmed informal discussion in a classroom, where the guest (an entertainer) gives an account of social and personal issues and is subsequently confronted with the host's reactions and suggestions. Due to the spontaneous and intimate character of the interaction as well as a relatively lower degree of predictability, the interview is oriented towards casual conversation.

The third one (I3) is a discussion of intellectual bias between the host and a writer taking place in an office. It displays a semi-institutional character, because it is not strictly information-focused; it also promotes the personal and emotional involvement of both speakers, allowing spontaneous and unpredictable behavior to be expressed.

3 Coding Scheme

Multimodality is annotated by discriminating between the description of the form of non-verbal expressions and their respective communicative functions. The descriptive level comprises features related to hand gestures, facial expressions and body posture for each speaker involved, as well as their semiotic type.

Table 1. Descriptive and functional attributes.

Descriptive attributes				**Functional attributes**		
Facial display	Hand gesture	Body posture	Semiotic type	Feedback	Turn management	Multimodal relations
Gaze Eyes Eyebrows Mouth Lips Head	Handedness Trajectory	Torso	Deictic Non-deictic Iconic Symbolic	Give Elicit	Turn gain Turn end	Repetition Addition Substitution Contradiction Neutral

At the functional level there are features encompassing the annotation of multimodal feedback and turn management, as well as multimodal relations between speech and a respective non-verbal expression [9]. The labeling of the elements of interest was based on the MUMIN coding scheme [10].

4 Annotation Process

The annotators' tasks involved the gradual annotation of the material in line with the coding scheme. Initially, the audio signal was extracted from the relevant video, orthographically transcribed and further enriched with information about pauses, non-speech sounds, etc. using Transcriber[1]. The output was imported to ELAN[2], the tool that was used for the entire video annotation, and it was ensured that the speech transcript was kept synchronized with the video.

The coders identified facial, hand and body gestures of interest marking their start and end points and assigned the respective tags according to their surface characteristics (e.g. smile, single hand up etc.) and their semiotic type at distinct annotation tracks. Next, they labeled the implied communicative functions based on the interconnection between the gestures and the corresponding utterance.

Finally, the data were revised in order to correct possible errors and to assess the consistency of the annotations throughout the corpus.

5 Annotation Output

The study of the annotated corpus reveals the multiple functions of non-verbal (NV) communication. Its significance lies in that it provides information and sheds light on the interplay between verbal and non-verbal signals. Speakers communicate numerous messages by the way they make use of NV expressions, which may repeat, complement, accent, substitute or contradict the verbal message. Moreover, NV means of communication outline the conversational profile of the speakers as they are a powerful means for self expression.

In order to interpret the NV behavior in an interview setting we have to clarify the subtle dynamics of the situations the speakers find themselves in; the content of the communication is framed by the perception of its context (environmental conditions where the interview takes place, identity/role of the participants, and their behaviors during interaction).

NV cues may also explicitly regulate the interaction, as they convey information about when speakers should take or give the turn. In that sense, they give signs of the effective function of the conversation and the degree of success of the interaction. Finally, through NV communication the speakers project their attitudes and feelings towards either their own statements (e.g. confidence about what they say) or to their interlocutor's statements.

[1] http://trans.sourceforge.net/
[2] http://www.lat-mpi.eu/tools/elan/

In the interviews examined, we attested 1670 NV expressions. The speakers frequently employ simple or more complex multimodal expressions using their facial characteristics (gaze, eyebrows, nods etc.), hand gestures (single or both hands, fingers, shoulders etc.) and upper part of the body (leaning forward and backward) in order to reinforce their speech and express their emotions towards the uttered messages. However, the types and functions of the expressions may vary, as they depend on the role that the speakers assume, the constraints imposed by the host and the discursive nature of the show from which the interview is taken.

Table 2. Number of non-verbal expressions attested in each interview and distribution between host and guest.

Interview	Non-verbal expressions	Host	Guest
I1	613	40.5%	59.5%
I2	665	22.5%	77.5%
I3	392	35.4%	64.6%

5.1 Descriptive Features

A closer look on the data shows that there are repeated patterns that are independent of the message content. For example, there are standard gestures in opening and closing utterances as well as in the hosts' prefatory statements (the gaze direction and hand orientation are towards the interlocutor). This sort of standardized multimodal behavior is aligned to the succession of turns across time and may lead to the formulation of possible conversational multimodal scenarios.

For example, a common behavior by the interviewees in our data was that, when asked a question, they take some time to contemplate on their answer, remember or find an appropriate word. In this time span, they usually have an unfocused gaze (eyes fixed up/down, sometimes head turn away from the interviewer). When they finally take the turn, they gaze towards the host and move the hands or torso in order to denote that they are ready to answer. If they are willing to hold the turn they reinforce their speech with repeated hand gestures or eyebrows raising, while they are using mostly the eyes to elicit an acknowledgement or ensure that the host keeps up. Finally, they complete their turn either by gazing down, or by staring at the host. This kind of scenario describes a regular flow that is subject to modification in case of interruptions, overlapping talk, or strong objections and reactions that declare the speakers' emotional involvement.

Moreover, the turn extent of each interview has an effect in the production of NV expressions. In case of I1 the shifts are largely monitored, so the guest has to express his views in a restricted time interval, he cannot deviate from the agenda or be spontaneous, instead he is more reserved in his expressivity and makes a closed set of simple, accompanying gestures. Conversely, in I2 and I3 where the shifts and topics are more negotiated, the guests are entitled to elaborate on their views; they feel less restricted and thus more prone to produce a variety of facial and gestural expressions.

Semiotic types of NV expressions seem to be independent of the setting; they are however related to the content and the discursive features of the interview. For

example, in I1 the politician builds his argumentation based on concrete facts and supports it mainly by non-deictic expressions. At the same time, in more casual discussions such as I2 and I3 a large part of the guests' talk involves narration, a discourse type that is complemented with iconic multimodal expressions. Deictic and symbolic types are quite equally distributed in the three interviews (cf. Table 3).

5.2 Communicative Functions

The correlations between the distinct modalities are captured through the multimodal relations and the annotation of the NV expressions with regards to feedback exchange and the coordination of turns as well.

Multimodal Relations. The annotation of the multimodal relations gives evidence of the contribution of NV expressions to the speakers' message. There is a relatively high percentage of NV expressions that complement the message by providing additional information that is not overtly expressed by speech only (addition) or by replacing unsaid words (substitution). The majority of substitution labels is related to acknowledgements and is represented mostly by facial displays that usually correspond to head nods or smiles. The repetition type comes next, and it is used to denote that no extra information is provided. Few relations were characterized as neutral, where the annotators believed that the NV expression has no significant contribution to the message. Finally, the contradiction relation is rarely used, namely in cases of irony. It is important to report that the contradiction type was found mainly in I1, possibly as a feature of argumentative speech, while it is rarely used in less institutionalized interviews such as I2 and I3.

Feedback. Multimodal feedback in terms of perception and acceptance[3] of the uttered message is usually expressed through gaze and nods rather than hand gestures. NV expressions of feedback-giving evolve in the course of time, as the speaker denotes that he has perceived the message, he is interested or willing to contribute and, as the turn is elaborated, he shows signs of certainty about the content of his talk, possibly by forming an opinion.

The topic of discussion and the role that the participants assume is related to the expression of their emotional and attitudinal feedback towards the message. Emotions and attitudes (e/a) such as joy and satisfaction are non-verbally expressed more overtly in I2 and I3 where the speakers feel more spontaneous. On the contrary, non-verbal declaration of e/a such as disappointment, anger etc. is manifested in I1 in the context of a strictly framed political discussion. E/a of surprise, certainty/uncertainty and interest are attributed to NV expressions of all participants and they seem to be independent of the interview type; they are rather related to the speaker's attitude towards a specific message content.

[3] The annotation values for Feedback pertain to 3 groups: *perception*, *acceptance* and *emotions/attitudes*.

Turn Management. I1 is more host-controlled and therefore shows certain predictability in turn management, whereas I2 and I3 are more participant-shaped and present a relatively lower degree of predictability and weaker talk control.

Gestures and expressions that are evoked in turn management are different in type and frequency when they take place in a normal flow rather than overlapping talk. During overlapping speech there is a high density of annotated multimodal expressions; apart from prosodic features (e.g. higher intonation) the speakers engage all their expressive potentials (gestures, facial displays) to maintain their turn. In our annotated data, overlapping speech is usually triggered by a pause, a hold or a repair, which are often accompanied by an unfocused gaze or a head side turn. A common scenario is that the guest wants to hold the turn but sometimes he hesitates to answer or takes his time to focus, think or remember what to say. The host then takes advantage and he takes the turn. In most of the cases, the speaker who manages to gain the turn is the one who is most expressive.

Moreover, In I1 we rarely see expressions related to the unbiased completion of the turn and its subsequent yielding to the interlocutor. In most of the times, the speakers take the turn without explicitly being asked to do so by their interlocutors.

Table 3. Distribution of semiotic and communicative features over interview types.

Attributes	**Values**	**I1**	**I2**	**I3**
Semiotic types	Deictic	5.2%	6.9%	**7.4%**
	Non Deictic	**88.1%**	73.1%	75.9%
	Iconic	4.6%	**15.9%**	13.8%
	Symbolic	2.1%	**4.1%**	2.9%
Multimodal relations	Repetition	15.6%	**17.8%**	6.8%
	Addition	51.7%	54.6%	**71.4%**
	Contradiction	**1.7%**	0.4%	0.3%
	Substitution	19.8%	19.9%	**20.2%**
	Neutral	**11.2%**	7.3%	1.3%
Turn management	Turn Take	**44.8%**	12.1%	16.4%
	Turn Yield	1.9%	8.2%	**9.2%**
	Turn Accept	2.7%	**8.8%**	7.4%
	Turn Hold	42.1%	**45.2%**	40.3%
	Turn Elicit	5.4%	**15.3%**	14.2%
	Turn Complete	3.1%	10.4%	**12.5%**

6 Further work and exploitation

In order to provide more accurate and systematic descriptions of the multimodal features contributing to this kind of interaction we are planning to enrich the corpus with more interviews. The communicative function and role of speech features (disfluencies, prosody elements like pitch, loudness) should also be further explored.

Furthermore, we plan to exploit this multimodal corpus in our multimedia processing framework which allows the fusion of different metadata [11] in accordance with the typology and semantic characteristics of the original audiovisual material. In this respect, conversation analysis metadata can be further exploited in order to accommodate multimedia retrieval & summarization applications [12].

This kind of study may also provide descriptive cues for building believable interactive embodied agents for a variety of applications or friendly interfaces.

References

1. Ekman, P.: Emotional and Conversational Nonverbal Signals. In: Messing, L.S., Campbell, R. (eds.) Gesture, Speech and Sign, pp. 45--55. Oxford University Press, London (1999)
2. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press (2004)
3. MacNeill, D.: Hand and Mind: What Gestures Reveal About Thought. University of Chicago Press, Chicago (1992)
4. Douglas-Cowie, E. Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J.-C., Devillers, L. Abrilian, S., Batliner, A.: The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In: Paiva, A., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 488--500. Springer, Heidelberg (2007)
5. Poggi, I., Pelachaud, C., Magno Caldognetto, E.: Gestural Mind Markers in ECAs. In: Camurri, A., Volpe, G. (eds.) Gesture-Based Communication in Human-Computer Interaction. LNAI, vol. 2915, pp. 338--349. Springer, Berlin (2004)
6. Vilhjalmsson, H.: Augmenting Online Conversation through Automated Discourse Tagging. In: 6th annual minitrack on Persistent Conversation at the 38th HICSS. Hawaii (2005)
7. Heritage, J. Conversation Analysis and Institutional Talk. In: Sanders, R., Fitch, K. (eds), Handbook of Language and Social Interaction, pp. 103--146. Lawrence Erlbaum, New Jersey (2005)
8. Ilie, C.: Semi-institutional Discourse: The Case of Talk Shows. Journal of Pragmatics 33, 209--254 (2001)
9. Poggi, I., Magno Caldognetto, E.: A Score for the Analysis of Gestures in Multimodal Communication. In: Messing, L.S. (ed.) Proceedings of the Workshop on the Integration of Gesture and Language in Speech, Applied Science and Engineering Laboratories, pp. 235--244. Newark and Wilmington, Del, (1996)
10. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P.: The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena. Multimodal Corpora for Modeling Human Multimodal Behaviour. Journal on Language Resources and Evaluation, vol. 41(3-4), pp. 273--287. Springer, Netherlands (2007)
11. Papageorgiou, H., Prokopidis, P., Protopapas, A., Carayannis, G.: Multimedia Indexing and Retrieval Using Natural Language, Speech and Image Processing Methods. In: Stamou, G., Kollias, S. (eds.) Multimedia Content and the Semantic Web: Methods, Standards and Tools, pp. 279--297. Wiley, (2005)
12. Georgantopoulos, B., Goedemé, T., Lounis, S., Papageorgiou, H., Tuytelaars, T., Van Gool, L.: Cross-media summarization in a retrieval setting. In: LREC 2006 Workshop on Crossing media for improved information access, pp. 41--49 (2006)