A Corpus for Investigating the Multimodal Nature of Multi-Speaker Spontaneous Conversations – EVA Corpus

IZIDOR MLAKAR, ZDRAVKO KAČIČ, MATEJ ROJC
Faculty of Electrical Engineering and Computer Science, University of Maribor
SLOVENIA
izidor.mlakar@um.si, kacic@um.si, matej.rojc@um.si

Abstract: - Multimodality and multimodal communication is a rapidly evolving research field addressed by scientists working in various perspectives, from psycho-sociological fields, anthropology and linguistics, to communication and multimodal interfaces, companions, smart homes and ambient assisted living etc. Multimodality in human-machine interaction is not just an add-on or a style of information representation. It goes well beyond semantics and semiotic artefacts. It can significantly contribute to representation of the information as well as in interpersonal and textual function of communication. The study in this paper is a part of an ongoing effort in order to empirically investigate in detail relations between verbal and co-verbal behavior expressed during multi-speaker highly spontaneous and live conversations. It utilizes a highly multimodal approach for investigating into relations between the traditional linguistic (such as: paragraphs, sentences, sentence types, words, POS tags etc.) and prosodic features (such as: phrase breaks, prominence, durations, and pitch), and paralinguistic features traditionally interpreted as non-verbal communication or co-verbal behavior (such as: dialog role, semiotic classification of behavior, emotions, facial expressions, head movement, gaze, and hand gestures). The main motivation for this study is to be able to understand especially the informal nature of human-human communication, and to create co-verbal resources for automatic synthesis of highly natural co-verbal behavior from un-annotated text and expressed through embodied conversational agents. The EVA corpus designed by a novel EVA annotation scheme represents a rich empirical resource for performing such studies in conversational phenomena that manifest themselves in highly spontaneous face-to-face conversations. A preliminary analysis regarding emotions within conversations has been also conducted and presented in the paper.

Key-Words: - conversation analysis**,** informal conversation, emotions, multiparty dialog, language and social interaction, multimodality, pragmatics, verbal and non-verbal interaction, co-verbal behavior

1 Introduction

Conversation analysis has been a powerful tool for analyzing language and action in various aspects of social communication [1]. Multimodality, traditionally considered as 'non-verbal' or 'co-verbal' communication, has been recognized as a key feature in the age of new orality [2]. Namely, communication is an act of conveying information, in which humans can convey it through a variety of methods, such as: writing, speaking, body language (gestures and posture) and facial expression, and even social signals [3]. Further, interpersonal communication involves the transfer of information between two or more collocutors using verbal and non-verbal methods and channels. The verbal part carries symbolic/semantic interpretation of message through linguistic and paralinguistic features of interaction, while the co-verbal part serves as an orchestrator of communication. Thus, the co-verbal goes well beyond and add-on or a style of information representation [2]. Namely, it is equally relevant as speech, and actively contributes to the information presentation and understanding, as well as discourse itself. Further, it regulates communicative relationships and may support or even replace the verbal communication in order to clarify or re-enforce the information provided by the verbal counterparts [4]. Thus, the co-verbal behavior effectively retains semantics of the information [5], provides suggestive influences [6], and gives a certain degree of clarity in the discourse [7, 8]. Researchers such as Allwood [9], McNeill [10], Duncan [11], Bozkurt [12] and Poggi [13], among others, have made a significant effort in order to re-define the theory of communication and to push it well beyond the realm of pure linguistics. As a result, the co-verbal behavior has become one of the central research paradigms and one of important features of interaction. Namely, the multimodal nature of communication has been investigated from various perspectives e.g. from psycho-sociological fields, anthropology and linguistics, to

communication and multimodal interfaces, companions, smart homes, ambient assisted living etc. [14, 15, 16]. Along with the increasing research interest in the multimodal nature of interaction, the multimodal corpora have been designed in order to capture and to analyze various levels and descriptive features of how the informal interaction actually works [9, 17, 18]. The knowledge extracted from multimodal corpora and annotation schemes, represents nowadays a key resource for better understanding the complexity of the relations between verbal and co-verbal parts of informal human-human communication. It provides insights into understanding of various signals, and their interplay as a resource for understanding, modeling and for the realization of the co-verbal behavior on conversational agents [19, 20].

The main motivation for this study is to investigate various linguistic, paralinguistic and co-verbal features of spontaneous human conversations, in order to be used for modeling of more natural human-like affective conversational behavior realized by an embodied conversational agent EVA [21]. The represented EVA corpus and novel annotation scheme are, in this respect, oriented specifically towards the analysis of function and form of those co-verbal-expressions, observed during face-to-face spontaneous multi-speaker interaction. Thus, the proposed EVA annotation scheme is oriented towards the analysis of the multimodality in the interplay of non-verbal and verbal parts. The schema, therefore, captures form of co-verbal signals, functional and non-functional roles of co-verbal behavior, as well as the linguistic and paralinguistic features of verbal information, and the aspect of attitude expressed through integration of emotion.

2 Issues regarding Multimodal Annotation

Among video corpora, the TV interviews and theatrical plays have shown themselves to be very usable resource of spontaneous conversational behavior for the analytical observation and annotation of co-verbal behavior and emotions used during conversation [22, 23, 24]. In general, TV discussions represent a mixture of institutional discourse, semi-institutional discourse and casual conversation. However, most of the studies and set-ups target narration and/or dialogues with only two participants. Furthermore, such material is often subject to certain restrictions, such as: time restriction, agenda, and technical features (camera direction and focus, editing) that further influence especially communicative function of co-verbal behavior and its expressive dimensions (speech, gestures, facial displays). As a result, the observed material incorporates a lot of information that may be regarded as noise, and thus may obscure the effort in investigating a particular goal.

In order to minimize the noise, some approaches generate multimodal corpora under specific laboratory conditions and by following artificially constructed settings [25, 26, 27, 28]. Such corpora are usually created with some specific purpose and usually incorporate individuals, who are instructed to implement various concepts and aspects of the communication. Undoubtable these corpora can provide a unique opportunity for researchers to study several natural multimodal phenomena. However, in general the conditions and the context are controlled, and the implications of broader context may be obscured due to the controlled and regulated set-up [29]. However, everyday natural human-human interactions are not completely ordered and synchronous. They also contain a lot of noise. This noise, if properly analyzed and incorporated, may unravel a lot of features and contexts that model the natural multimodal conversational expressions. Thus, the informal corpus arguably can represent the most spontaneous face-to-face interaction. Namely, casual conversation is much more spontaneous than interviews and/or laboratory settings [29]. The ‘noise’ in this case represents ‘meta’ information, which may provide further insights into how informal communication works [2]. Finally, emotion, as expression of eyes and face or expressed as gesture or even through words and speech, is deeply integrated into informal communication [30]. It is used to render one’s relation to the situation. Moreover, emotional actions are generated by motives to alter the current state of the world so that it transfers to a more optimal state. Thus, emotions are triggered by events; directly perceived, or recollected, or even as imagined [31]. In casual conversations, especially multiparty interactions, emotions are triggered by various stimuli originating from ones-self, the environment, and the collocutors. In verbal parts, emotional contents are low and perhaps unnoticeable. However, the co-verbal parts are full of emotional expressions. These emotive concepts are expressed through voice (intonation, prosody), facial expressions, and gestures. In conversation they give the conversational models the motivational force to converse and develop [32].

In this paper we represent a novel multimodal corpus, named EVA Corpus. The synchrony/correlation between annotated verbal and co-verbal elements established during conversations were already analyzed in our previous efforts [33, 34]. Here, we go deeper into the concept of semiotic intent. Namely, the semiotic intent is a concept through which we can correlate the intent of verbal information (defined through POS, prosodic features and classification of interpretation through meaning) with gestures. Face-to-face interactions are multimodal and go well beyond pure language and semantics. Thus, the concept of the proposed correlation between verbal and co-verbal elements tries to exploit this. Therefore, the extension of semantics and communicative intents with other linguistic and paralinguistic features, dialog functions and emotion, as proposed in this paper, seems only natural. By exploiting the casual nature of the EVA corpus, we can take into consideration also the interplay of various conversation phenomena, such as: emotional attitude, dialog, prosody, communicative intents, structuring of information, and the form of its representation, e.g. through the, facial expressions and gestures, head movement, etc. Thus, giving us a true insight into how informal communication works, what stimuli triggers various phenomena, and how do these impulses interact and reflect on each other and the other phenomena. In essence, we search for relations between features and channels that are exploited by collocutors to establish a specific state of the world; e.g. to promote an idea, achieve a specific goal etc. Such links provide synthetic agents with the basis for the multimodal literacy; the capacity to construct meaning through understanding of situation and responding to situation [2, 35]. Thus, it directly enables the synthesis of more natural and more situation adaptive co-verbal behavior that facilitates concepts generated especially in spontaneous and highly casual multiparty settings.

3 The Presentation of EVA Corpus Description

The audio/video material used for EVA corpus originates from GoS corpus, a corpus of spoken Slovenian [28]. GoS includes video and audio recordings and corresponding orthographic transcriptions of approximately 120 hours of speech. The GoS corpus is focused on conversations that we are exposed to on a daily basis and in various situations e.g.: radio and TV shows, school lessons and lectures, private conversations between friends, or within the family, meetings at work, consultations, conversations in buying, and selling situations, etc. For the EVA Corpus, four video recordings were selected from the GoS corpus; this is approximately four hours of video material. These videos were selected in order to comply with multiparty condition, and to involve as much affective casual/informal conversation as possible. Each selected video contained about 57 minutes of transcribed highly informal and affective multiparty conversation, with 3 – 4 collocutors exchanging information in a highly unordered and dynamical manor. In total 5 collocutors per recording contribute relevantly. Among those, two are TV presenters and are present in all four recordings. The other collocutors represent a main guest and two other guests that have some personal relationship with the main guest (e.g. his close friends). The conversational setting is totally relaxed and free, and built around a talk-show that follows some general script/scenario, however, the topics discussed are highly changeable, informal and full of humor and emotions. Furthermore, most of the collocutors are more or less public persons, therefore, well attuned to cameras and audience. The collocutors also know each other, further enhancing the spontaneity and casualness of the conversation. Although sequencing exists, it is performed highly unorderly as are also the communicative functions. This results in a highly causal and unordered discourse, with overlapping statements and roles, vivid emotional responses and facial expressions. Language used by the collocutors is also quite colloquial incorporating dialects and a lot of grammatical irregularities. Table 1 summarizes basic statistics behind the annotated video material.

Statements	Overall: 1516 AVG per speaker:303 (STD = 260)
Statement duration	Overall: 93min 29s Max:23.22s, Min: 0.19s AVG per statement: 3.57s (STD = 0.54)
Sentences	Overall: 2014 AVG per collocutor:402 (STD = 364) AVG per statement: 1.32
Sentence duration	Max:18.4s, Min: 0.19s AVG per collocutor: 2.66s(STD = 0.26)
Words	Overall: 12067 AVG per collocutor:2414 (STD = 2300)

Table 1: Statistics for the selected TV show used in EVA corpus.

In order to study informal and free conversations, it was important to pre-select those recordings that are placed in relaxed environments and are perceived to proceed as natural as possible, with full of impulsive and emotional reactions and well beyond a script. As outlined in Table 1, the basic structure and nature of the material, exposes a general nature of casual interaction. Conversation consists of 1516 statements that are distributed among 5 speakers and split into 2014 sentences. This shows that most of the statements are single sentence, or two sentences at most. The duration of statements across speakers, AVG = 3.57s (STD = 0.54) also indicates that most of the statements are short. Thus, longer statements, or monologues, are rare and last for about 23s at most. This means that all collocutors were active contributors often overlapping each other. However, two of the collocutors were invited guests as personal friends, and were less present on the scene. Their contribution might be statistically less relevant, since they contribute to around 4% of the entire content. However, we believe it is still relevant, since it may reflect a deeper personal relationship in communicative behavior of the main guest.

When we look at overall duration of the spoken content (93min 29s) and compare it to the duration of whole video material (57min 30s), we can observe that almost half of the time spoken content had overlapped. Further, if we look at the distribution of listener/speaker dialog roles in Table 2, we can observe that the roles are evenly distributed. Thus, no collocutor is predominantly speaker or listener. However, hosts A and B are the main contributors; which is expected since they moderate the conversations.

	Listener	**Speaker**
Host A	332	326
Host B	458	455
Guest	296	292
Extra guest A	16	15
Extra guest B	32	32

Table 2: Distribution of dialog roles per collocutor

Thus, we can conclude that the exchange of information in the annotated video is casual, highly dynamic, and involve shorter statements and ideas rather than longer monologues and narratives. This are important features for studying spontaneous and casual conversations.

4 A novel EVA annotation scheme

In order to capture and analyze new phenomena in EVA corpus, the video material has been further annotated by following the novel EVA annotation scheme that additionally incorporates linguistic and paralinguistic features, as well as maintaining cultural/personal background of the speaker. The annotation process was performed separately for each speaker, where the formal model of the novel scheme is outlined in Fig. 1. This model is based on the EVA annotation scheme proposed in [33, 34]. That scheme targeted the analysis of the form of movement in high resolution and an approximation of correlation through semiotics, prosody (paralinguistic features), and other linguistic features (POS, semantic patterns etc.). On the other hand, the novel EVA annotation scheme integrates some additional linguistic, paralinguistic, and non-verbal features distinctive for multimodal conversations and multiparty dialogs, such as: communicative function, dialog role, syntax, and emotions/attitude. In this way the EVA annotation scheme now allows us to analyze the EVA corpus in even greater detail. It also allows us to incorporate various linguistic, paralinguistic and co-verbal features into existing and new ‘conversational’ relationships. Further, through newly gathered knowledge, we are going to be able to pare features into complex stimuli used for: a) triggering the generation of the conversational artefacts, and b) to improve the understanding of the situation through multimodality.

The model, outlined in Fig. 1, allows for a clear recognition of cultural background as well as language dependencies of the collocutor. Furthermore, annotations are performed independently for each speaker. The session for each speaker, as proposed [34], is separated into annotation of function, and into annotation of the form. Regarding the previous EVA annotation scheme, we have enriched the functional annotation part and have added additional tracks to capture and to describe linguistic and para-linguistic features of the spoken content in addition to co-verbal features. As outlined in the Fig. 1, the research goal is to analyze and search for various multi-dimensional relationships between conversational artefacts. Thus, to identify and establish temporal and symbolic links between verbal and co-verbal features of multi-party interaction.

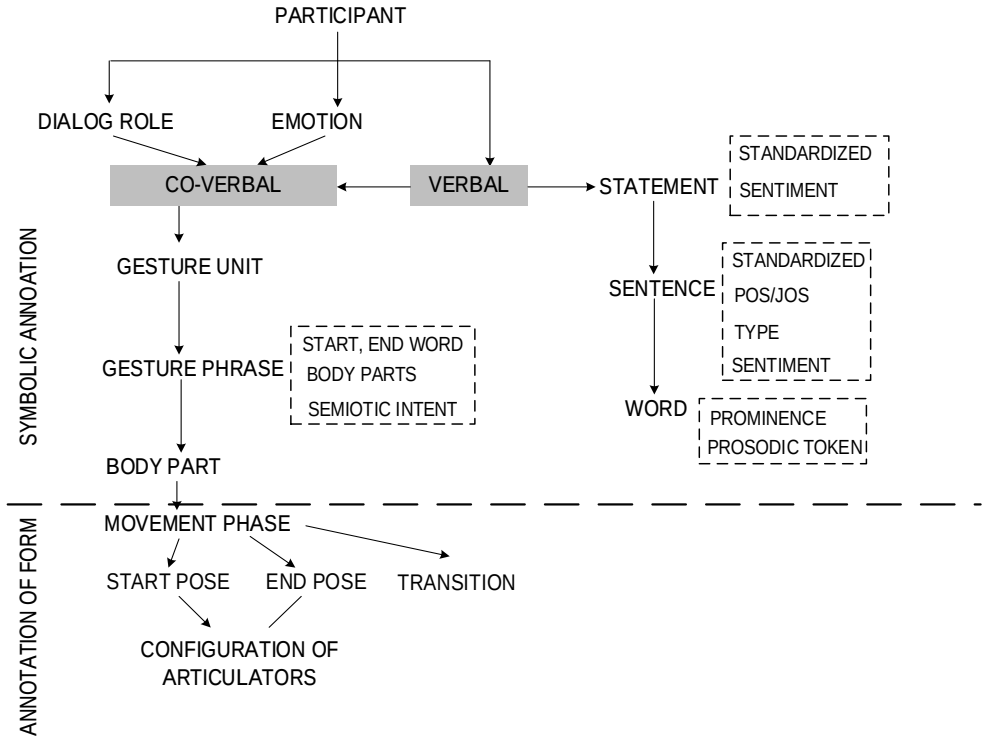

Fig. 1: The novel model for the EVA annotation scheme

Namely, to relate the form of co-verbal movement and its manifestation (e.g. gestures, gaze, facial expressions) with low- and high-level communicative artefacts, such as: emotions, dialog role, and with linguistic and paralinguistic features of verbal part, such as: lemma, POS tags, sentence type, phrase breaks, prominence, sentiment, and semiotic intent.

As outlined in Fig.1, the novel model distinguishes between symbolic part and the part that targets description of the form of co-verbal behavior, generated over symbolically defined segments. In the symbolic part the main concepts are co-verbal and verbal behavior. The stimuli for the co-verbal behavior may be verbal in nature. It may originate as a reflection of attitude/emotion or even be a supportive artefact in the implementation of the communicative function (e.g. feedback, turn taking, turn accepting, sequencing, etc.). Similarly, the verbal behavior primarily used for representation of information, may also reflect attitude/emotion or be adjusted to serve as a part of the implementation of a communicative function. All artefacts are interconnected through temporal domain, and can be related among each other in numerous ways and combinations. For instance, one can investigate the relationship between sentence, sentence type, and dialog role; or how are linguistic and semiotic features related to feedback; or, for instance, how can semiotic intent help to indicate what kind of emotion to synthesize on and ECA. This enriches the model with aspects originating from natural language processing and understanding, (such as sentiment, syntax), and even other parts mostly limited to pure linguistics. As a result, the novel EVA model can capture various relationships between, words, grammar, and inter/intra -personal communication, and can be applied to various context and research fields from behavioral sciences, psychology, anthropology and sociology.

Regarding the description of the form of co-verbal behavior, we are concerned in the shapes and movements generated during symbolically defined co-verbal intervals. Body-parts are the core objects of the observation in the annotation of the form. We adopt the idea that symbolic relations and concepts are established on the functional/symbolic level and realized via hand gestures (left, right arm and hands), facial expression, head movement, and gaze. The EVA annotation scheme separates between hands, arms, head, and face. Further, the movement of each body-part is described with movement phrase, movement phases, transitions, and the articulators propagating the observed movement. Here, the movement phrase describes the full span of movement phases (from preparation to retraction). Each movement phase contains a

mandatory stroke and optional preparation, hold, and retraction phases. In terms of physical realization of gesture on the ECA; each movement-phase identifies a pose at the beginning, and a pose at the end. Both poses are 'interconnected' with a trajectory that identifies the path over which the observed body parts propagate from the start pose to the end pose. Mathematically, each movement phase may be represented as a function of pose and trajectory (e.g. *M=f(Ps,Pe,T)*). The proposed topology of co-verbal behavior and movement phases in particular is outlined in Fig. 2. As can be seen, each movement phase is segmented into start pose *Ps*, end pose *Pe*, and the transition trajectory *T*, which hands perform during the propagation from the start to the end pose. The trajectory *T* represents a parametric description of propagation, which includes the partitioning of the trajectory *T* into movement primes (simple patterns), such as: linear and arc, each defined through the intermediate poses. Namely, a movement trajectory can reach various complexities, and outline complex forms, such as: spiral, roof, chair, etc. To properly animate it on the ECA, it has to be split into simpler forms (primes), e.g. chair is partitioned into 2 linear elements or 2 linear + 1 arcs. Further, each prime is segmented into 2 (or 3 for arc) key points, each identifying a transitional hand/arm pose.

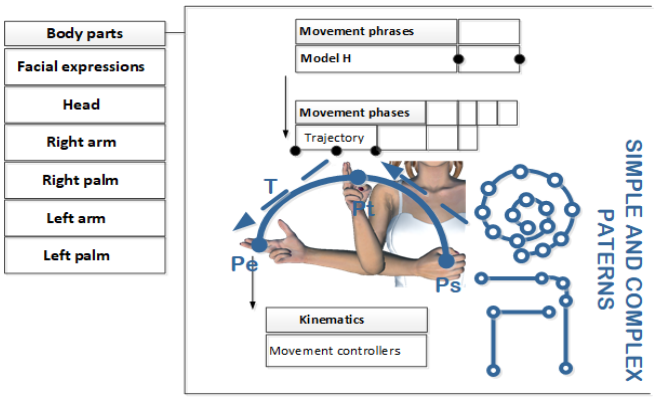

Fig. 2: Topology of the annotation of the form of co-verbal behavior and movement trajectories

As outlined in Fig. 2, a hand propagates from its starting pose to its ending pose via transition *T*. In 3D space, there are infinite transitional poses through which hand could travel in order to reach pose *Pe*. Thus, infinite hand gestures outlining infinite number of shapes. Each possible path can also have different length. And since each path must be implemented within a fixed time-slot, each resulting gesture has to be performed at a different velocity. Therefore, some gestures would appear more natural, and some gestures less. By using the center transitional key point *Pt*, we have defined the final shape of the performed gesture for given conversational context, and for given time slot.

Thus, we have ensured proper velocity of the manifestation. As a result, the ECA may during animation 'choose' from available similar gestures, or adjust the proportions of some gesture to the defined timeslot. Therefore, in order to simplify the annotation, the configuration of movement controllers (e.g. pose) is specified only at the abstract level, e.g. in the form of the hand and arm position in 3D space, and relative to body and hand-shape [34]. This gives the ECA the possibility to perform each gesture slightly different and to adjust it especially to a broader context; e.g. incorporating prominence, emotion or intent into its selection.

To sum up, through the novel EVA model we have established additional relations between symbolic features of communication generated during symbolic intervals, and multimodal expressions, as generated through hands, face and gaze. In this way we can relate e.g. smile to several linguistic and paralinguistic features; or we can correlate gesture to an emotion and modulate its 'power' (velocity) with the intensity of the emotion etc. In the same manor, a nod can represent acceptance or can facilitate thinking (word search).

5 Annotation of the EVA corpus

In order to annotate the material, three annotators were recruited with background in linguistics and experience in annotations of multimodal material. Each of the annotators was briefed on the meaning of each communication artefact observed in detail. The annotations were performed in ELAN (EUDICO Linguistic Annotator) tool, generally used for multi-level annotation of video and/or audio data that has been developed at the MPI institute (Max-Planck-Institute) [36]. Fig 3 outlines the implementation of the novel EVA annotation scheme in the ELAN environment.

The annotations were performed over 6-month period and in separate trials for each communicative concept. For instance, annotation of emotions, dialog role, sentence type, sentiment, prominence and phrase breaks, were implemented in separate trials and not at the same time.

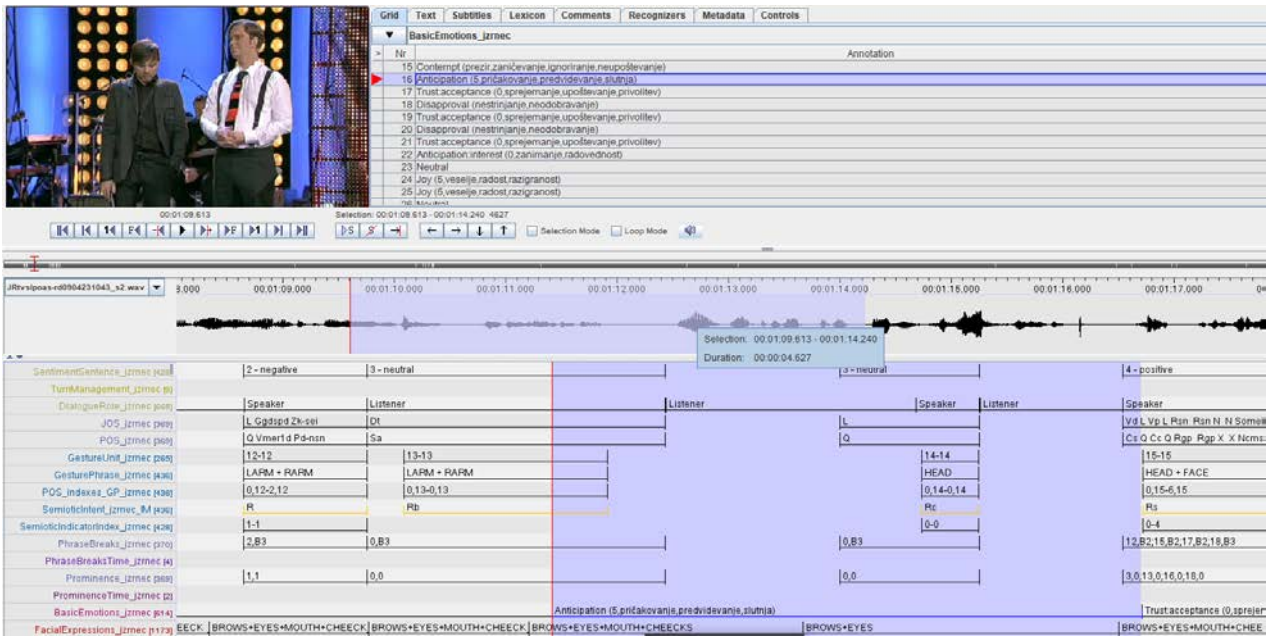

Fig. 3: The ELAN interface: multimodal annotation topology used for studying conversational emotion

5.1 Annotation of verbal part

The EVA corpus audio data are transcribed in original colloquial form (as pronounced), and in their standardized form (standardized Slovenian spelling). Each transcription was split into 5 sessions, each session maintaining information for individual speaker and possible speaker dependencies. The colloquial transcription include also meta information maintained in '[]', such as *[:laugther]*, *[gap]*, *[incident]*, *[:voice]*. In its standardized form, the 'meta' information has been replaced with '-' character. The transcriptions are also segmented into statements by considering time. The boundaries for colloquial and standardized statements completely match.

In EVA corpus presented in [33, 34], the conversations were split into 5 sessions, where each session maintains information for individual speaker. Thus, the corpus retains possible cultural, sociological and personal speaker-dependent features. The annotators were asked to first segment the standardized statements into sentences, and words. Each sentence was then also POS tagged using POS tagger provided by PLATTOS TTS engine [34] and in JOS[1] and MULTEXT-East V4[2] format. Afterwards POS tags were manually corrected, mostly linguistic fillers such as pauses, 'mmm', 'eee', 'aaa' that were assigned as interjection. In order to annotate the EVA corpus by using the novel EVA annotation scheme, we asked the annotators to assign each sentence with a sentence type (e.g. interrogative, declarative, exclamatory and imperative), and sentiment (e.g. very negative, negative, neutral, positive and very positive). The sentiment was also specified in the level of paragraphs. For the sentence type we achieve strong agreement among the three participants (93%). However, for the sentiment we achieve fair agreement with weighted kappa of 0.34. In general, the disagreement primarily originated from the degree of sentiment and the perception of neutral sentiment.

Next, the annotators were asked to assign phrase brakes (B2, B3) to each of the statements and identify the token correlated to the phrase break. Here, we have reached very good inner annotator agreement with weighted kappa of 0.87. Finally, the annotators were also asked to define the prominence on sentence level, e.g. to identify at which tokens PA occurs (are most prominent). For the annotation of prominence, we reach very good inner annotator agreement with weighted kappa of 0.92.

5.2 Annotation of the co-verbal part

The EVA corpus contains annotated gesture units, but exclusively associated with verbal behavior. Also the novel EVA annotation scheme that we have applied follows the same concept to a) gesture that do not correlate with verbal parts and may originate from attitude or communicative function,

[1] http://nl.ijs.si/jos/msd/html-en/index.html

[2] http://nl.ijs.si/ME/V4/msd/html/

and b) facial expressions and head movements (gaze). In total we have identified 4199 gesture phrases in the material. Gesture phrases are defined as units of visible and meaningful bodily actions [37], with or without overlapping verbal counterpart. Each gesture phrase can be further deconstructed into movement phases and each movement phase into multiple trajectories (Fig 2). In the annotated material 70% of gesture phases are generated over a single linear or arc trajectory, while 30% of gestures are complex and generated as a combination of multiple prime elements. In terms of modality we observed that most co-verbal behavior is generated by using face and head. A more detailed distribution of the modality of gestures across speakers is outlined in Table 3.

Body part (modality)	Total	Mean per participant
FACE	53	10,6
HEAD	704	140,8
HEAD + FACE	717	143,4
LARM	34	6,8
LARM + FACE	4	0,8
LARM + HEAD	289	57,8
LARM + HEAD + FACE	230	46
LARM + RARM	74	14,8
LARM + RARM + FACE	19	3,8
LARM + RARM + HEAD	789	157,8
ALL MODALITIES	476	95,2
RARM	57	11,4
RARM + FACE	2	0,4
RARM + HEAD	428	85,6
RARM + HEAD + FACE	323	64,6

Table 3: Distribution of modalities within identified gesture units

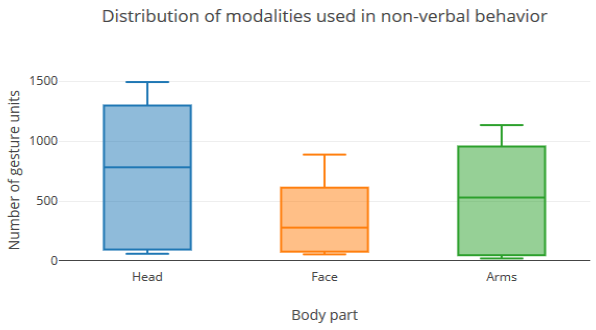

Fig. 4: The distribution of modalites used while generateing meaningfull non verbal behavior

In terms of inner annotator agreement, we have reached moderate agreement (kappa = 0.57) for the modality. In gestures, however, the participants used multiple modalities (e.g. head + face + arms). The lower kappa value is observed especially with those combinations, where face and head were used. Namely, it seems that the annotators could not strongly agree weather head movement/gaze or facial expressions were of communicative nature, or just some random/involuntary movement. Fig 4 further enhances the head as a lead modality in the overall distribution of modalities. The results are expected and in line with findings in the field. Namely, listener behavior primarily involves head (such as shake, nod, etc.), and face as modalities for the generation of non-verbal signals, especially feedback [9, 38]. In the annotated material we can observe that collocutors on average spend around 60% of their involvement as listeners and 40% as speakers.

5.3 Annotation of emotion/attitude

The main motivation regarding multimodality of emotion in spontaneous face-to-face multi-speaker conversations is to study how is the emotion related with facial expressions and gestures, also regarding time. The knowledge regarding this matter could help us to improve naturalness, also when processing unannotated texts or information in the context of dialog. Namely, when virtual collocutors are capable incorporating emotions and affect in their interactions and responses in proper ways, they are able to achieve higher degree of human like responsiveness. As a result, they could be used in a variety of applications, from true companions to sensitive and sensible tutors, and helpers. Namely, humans are social beings and affective (emotional) responses play a crucial role in such conversations. Further, emotions enable people to react to the stimuli in environment [39]. Emotion is also regarded as a multimodal feeling, which is expressed through various channels of spoken content (what is being said), the way it is spoken (vocal cues), and gestures and facial expressions (non-verbal signals) generated during emotion.

Emotion studies often deal with the basic emotions as described by Ekman [40]. However, Ekman's (and similar) notions of emotion might simply not suffice, especially in case of emotions and attitudes in conversation. In EVA corpus most perceived emotional behavior is emotional attitude, also called affective epistemic state [41]. The emotional attitude primarily considers the way people feel about the communicative situation, the interlocutor, or the content of the ongoing conversation. Plutchik's three-dimensional model [42], which describes the relations among emotions

may be helpful in understanding how complex emotions interact and change over time and in a broader, social context. Nevertheless, in order to capture emotional attitudes, and represent them as conversational stimuli in EVA Corpus, the annotators were asked to apply 50 emotional variations and 2 non-emotional states, e.g. 'rest' and 'undefined' to the selected material. The annotators have classified emotions within a dedicated track, and regardless of the collocutors dialog role, or presence of the verbal content. Thus, they classified the emotional attitudes as feelings that reach beyond listener/speaker segments, verbal content parts, or even paragraphs/sentences. As a result, emotion unit for 'anticipation' can span over three sentences, and is maintained also during the time, where the observed collocutor acts primarily as a listener as outlined in Fig. 3. Thus, emotional attitude can truly reflect emotion or even situational context, such as: regulation in turn-assignment, or anticipation in feedback signals.

In the EVA corpus we have identified 3312 instances of emotional attitude. Nevertheless, the 'Anticipation:interest', 'Trust:acceptance' and 'Joy' were identified as dominant emotions. The inner annotator agreement reached was fair. Table 4 summarizes the distribution of statistically more relevant emotions as identified in the material.

Emotion	**Number of instances**
Anticipation: Interest	1239
Trust: Acceptance	671
Joy	349
Joy: Serenity	221
Disapproval	137
Joy: Ecstasy	92
Surprise	69
Amazement	49
Anticipation: Vigilance	43
Cynicism	29
Disgust	23
Distraction	23
Curiosity	22
Delight	19
Trust: Admiration	19
Boredom	15
Sadness	15
Contempt	14
Pensiveness	12
Anger: Annoyance	10
Pride	10
Alarm	7
Fear: Apprehension	7
Optimism	7
Shame	7

Table 4: Distribution of emotions classifed across speakers

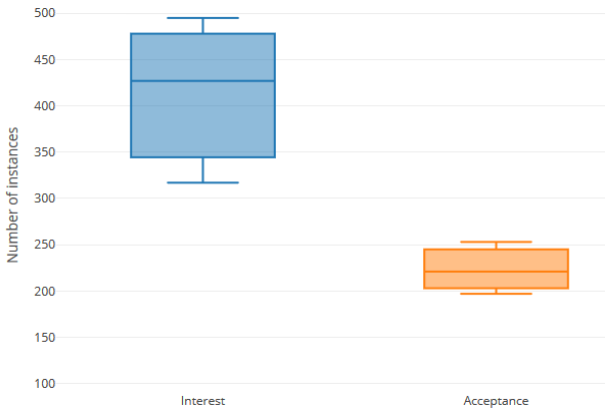

Fig. 5: The distribution of emotions used while generateing emotional attitude

Similiarly as before we can observe that most of the emotional attitude is generated as part of the feedback signals. Thus, 'Anticipation:interest' and 'Trust:acceptance' are expected to be predominant, especially since feedback is predominatly generated as part of listener behavior, which in EVA Corpus represents about 60% of material.

6 Experiments

EVA corpus currently contains roughly 3300 instances of emotions. Emotion unit 'Anticipation:Interest' is observed as being the most dominant one. It is followed by other emotions, such as 'Trust:Acceptance', 'Joy' and 'Anger:Annoyance'. The relation between text and emotion units can be studied through sentiment units, e.g. the positive/negative connotation of content. In order to study the relation between verbal content and emotion units, we have first annotated the sentiment tracks used for the verbal content on paragraph and sentence level. Each paragraph/sentence sentiment was annotated on a 5-level scale, from very positive to very negative. The analysis of dependences between sentiment and emotion units has been performed through temporal domain. Fig. 6 and Fig 7 present the observed distribution of sentiment units in case of emotion unit "Anticipation:Interest" and "Trust:Acceptance".

In general both emotions are accepted as largely positive emotions. Primarily these emotional attitudes are used, when verbal content represents/outlines positive signal or positive nature of context [43]. Namely, 'Interest' and 'Acceptance' are defined as positive emotions. 'Interest' is regarded as a heightened state that calls for one's attention to something new. It inspires fascination and curiosity. While 'Acceptance' is interpreted as a mild form of 'Trust'. It is perceived as a willingness to see things as they are. In addition to positive it can also have neutral/negative connotation. However, neither of the emotions is obviously limited strictly to a single sentiment value. E.g. for the emotion 'Acceptance', the more negative connotations are obvious. Especially in cases, when expressing sarcasm, or when one realizes a truth that is negative or has a negative connotation with one's belief. 'Interest' can also have a negative sense, especially when trying to express sarcasm or 'low quality' or 'negative attitude'. Generally, the connotation of 'interesting' is defined by the inflection used. Based on the well-established definitions of both emotions, we would expect that in terms of communicative intent, both serve as a signal explicitly targeted at the collocutor. Thus, the predominant usage would be during generating feedback, while collocutors are listeners. As shown in Fig. 8 and Fig. 9, this statement can be observed by relating dialog role and emotion tracks as annotated in the EVA corpus.

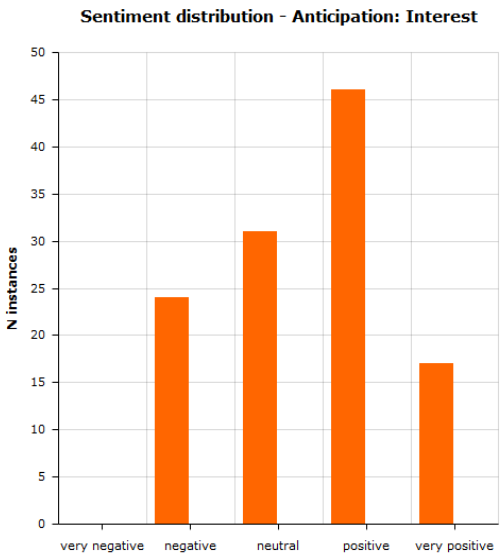

Fig. 6: Correlation of sentiment and emotion 'Anticipation: Interest'

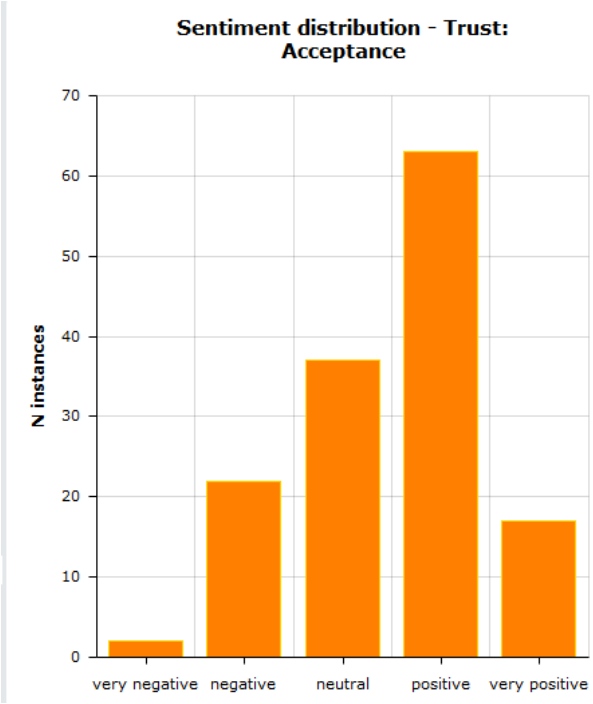

Fig. 7: Correlation of sentiment and emotion 'Trust:Acceptance'

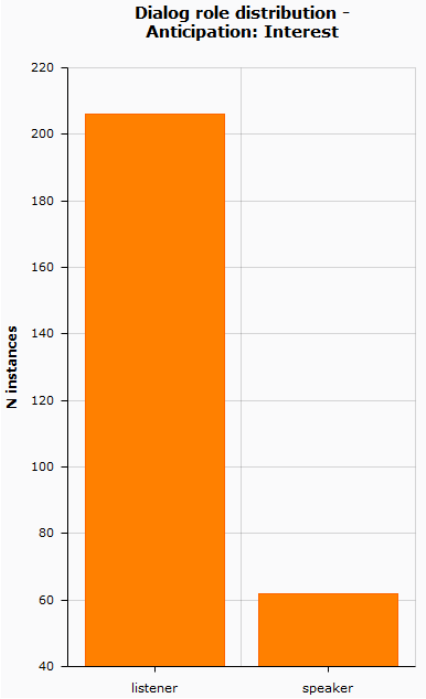

Fig. 8: Relating 'Dialog role' and emotion unit 'Anticipation: Interest' within EVA corpus.

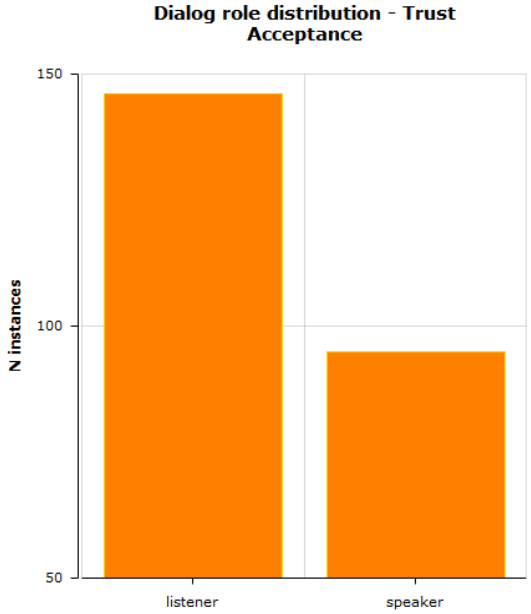

Fig. 9: Relating 'Dialog role' and emotion units 'Trust:Acceptance' within EVA corpus.

As outlined in Fig. 8, the emotion unit 'Anticipation:Interest' may be regarded as predominately part of the listener behaviour, while 'Trust: Acceptance' on the other hand is used as part of speaker and listener behaviour. Namely, it is used to signal agreement with a statement. It can also provide emotional attitude towards a topic that collocutor is presenting. Especially, when the 'revelation' has negative connotation. Finally, since both emotion units 'Interest' and 'Acceptance' are used as feedback signals for the collocutor, one would expect that a similar communicative intent would be observed alongside. As shown in Fig. 10, this generally holds true. Namely, alongside 'Interest' mostly referential and regulative intents were observed. However, in terms of 'Acceptance' we have observed predominantly metonymic nature of gestures followed by regulative intents. Therefore, in similar way we can observe and establish other relations between emotion, co-verbal, lingustic, and paralingustic features of conversations, such as: dialog role, body parts used for the generated expression, semiotic classes and subclasses along-side and emotion, prosodic features of emotion etc. This results in the annotated material with a trully multimodal and multi-context attribute. Finally, the EVA annotation schema and the EVA corpus may be used in various fields well beyond co-verbal behavior recreation. Namely, both to capture and to connect a wide variety of conversational phenomena.

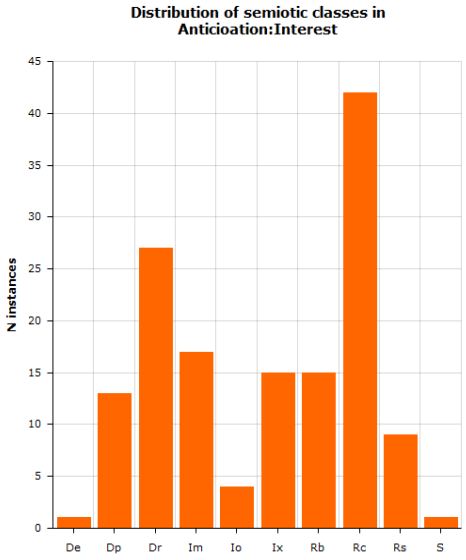

Fig.10: Relating communicative semiotic intents and emotion units for 'Anticipation: Interest'

7 Conclusion

In this paper we have presented a novel EVA multimodal corpus, as generated by novel EVA annotation scheme. The annotation scheme is a result of research regarding recreation of spontaneous co-verbal behavior. The scheme is based on findings presented in [25, 26]. The annotation scheme incorporates and correlates linguistic and paralinguistic, verbal and non-verbal features of multiparty informal conversations. The topology, the formal and functional part of the scheme were also outlined in detail. At the end, the analysis regarding emotion unit in the material is discussed in more detail, in order to demonstrate how the EVA corpus can now be used for detecting and investigating several additional conversational phenomena and relations. The annotation is based on an on-going effort in searching and investigating those features and relations that may be used as stimuli in the synthesis of co-verbal emotions as well as how emotion may influence complete co-verbal behavior.

The novel EVA annotation schema goes well beyond similar efforts and efforts of the authors in the field of co-verbal synthetic behavior and

adds a linguistic and paralinguistic dimension to the traditional verbal/co-verbal relations. It is designed in such a way that all phenomena regarding form, e.g. posture, gesture, gaze and facial expressions and higher-level phenomena regarding function, e.g. lemma & structure, POS tagging, semiotics, prosody and dialog, are described within a single session, and related via a common time-line. Thus, several relations in either track or between the tracks may be established and investigated. Additionally, the level of casualness detected in the material and the level of spontaneous detected in the intrapersonal responses among interlocutors, goes well beyond laboratory settings, plays, and interviews. Namely, it incorporates a high degree of informality with overlapping, sarcasm, disorder, and spontaneous reactions. It also contains a colorful variety of conversational emotions incorporated into highly dynamical responses.

Multimodal conversational behavior and its stimuli beyond semantics is relatively new, thus ideas, concepts and corpora are still evolving. At this point the annotation of EVA Corpus is largely a result of manual work, performed by several skilled annotators. Although the corpus incorporates various perspectives, future development will focus on deeper prosodic and linguistic analysis as well as detailed analysis of dialog well beyond the collocutors role.

The development of multimodal corpora is still ongoing process. As a result, multimodal corpora available are still rare and highly specific. The standardization of annotation methods and approaches are still developing. The annotation data in EVA corpus are generated mostly manually. Since this is largely very time-consuming process, tools and methods to at least partially automate the whole process are highly demanded. Therefore, in the near future we plan to study algorithms, which could at least partially automatize some of the annotation processes, e.g. gesture form and dynamics classification, sentiment classification, word segmentation, etc.

Acknowledgments:
This work is partially funded by the European Regional Development Fund and the Ministry of Education, Science and Sport of Slovenia.

References:

[1] Mondada, L. (2017). New Challenges for Conversation Analysis: The Situated and Systematic Organization of Social Interaction. Langage et société, (2), 181-197.

[2] Bonsignori, V., & Camiciottoli, B. C. (Eds.). (2017). Multimodality Across Communicative Settings, Discourse Domains and Genres. Cambridge Scholars Publishing.

[3] Velentzas, J. O. H. N., & Broni, D. G. (2014). Communication cycle: Definition, process, models and examples. In Proceeding of the 5th International Conference on Finance, Accounting and Law (ICFA‟ 14) (Vol. 17, pp. 117-131).

[4] Kleckova, J. A. N. A., & Mahdian, B. (2004). Nonverbal Communication in Spontaneous Speech Recognition. WSEAS Transactions on Electronics, 1(3), 531-536.

[5] Esposito, A., Vassallo, J., Esposito, A. M., & Bourbakis, N. (2015, November). On the Amount of Semantic Information Conveyed by Gestures. In Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on (pp. 660-667). IEEE.

[6] Dafinoiu, I. & Rotaru T.S. (2011). The use of suggestive influences in promoting environmental behaviours. In Proceedings of the 7th WSEAS / IASME International Conferrence on Educational Technologies (EDUTE 11). Ias.

[7] Chen, C. L., & Herbst, P. (2013). The interplay among gestures, discourse, and diagrams in students' geometrical reasoning. Educational Studies in Mathematics, 83(2), 285-307.

[8] Colletta, J. M., Guidetti, M., Capirci, O., Cristilli, C., Demir, O. E., Kunene-Nicolas, R. N., & Levine, S. (2015). Effects of age and language on co-speech gesture production: an investigation of French, American, and Italian children's narratives. Journal of child language, 42(1), 122-145.

[9] Allwood, J. (2013). A framework for studying human multimodal communication. Coverbal Synchrony in Human-Machine Interaction, 17.

[10] McNeill, D. (2015). Why we gesture: The surprising role of hand movements in communication. Cambridge University Press.

[11] Duncan, S. D., Cassell, J., & Levy, E. T. (Eds.). (2007). Gesture and the dynamic dimension of language: Essays in honor of David McNeill (Vol. 1). John Benjamins Publishing.

[12] Bozkurt, E., Yemez, Y., & Erzin, E. (2016). Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures. Speech Communication, 85, 29-42.

[13] Poggi, I. (2007). Hands, mind, face and body: A goal and belief view of multimodal communication. Berlin: Weidler.

[14] Holler, J., & Bavelas, J. (2017). Multi-modal communication of common ground. Why Gesture?: How the hands function in speaking, thinking and communicating, 7, 213.

[15] Salama, M. A. R. I. A., & Shawish, A. H. M. E. D. (2013). A Comprehensive Mobile-Based Companion for Diabetes Management. In 7th WSEAS European Computing Conference, Dubrovnik.

[16] Tsiourti, C., Moussa, M. B., Quintas, J., Loke, B., Jochem, I., Lopes, J. A., & Konstantas, D. (2016, September). A virtual assistive companion for older adults: design implications for a real-world application. In Proceedings of SAI Intelligent Systems Conference (pp. 1014-1033). Springer, Cham.

[17] Bergmann, K., Kopp, S. (2010).Systematicity and Idiosyncrasy in Iconic Gesture Use: Empirical Analysis and Computational Modeling. In: Kopp, S., Wachsmuth, I. (eds.) GW 2009. LNCS, vol. 5934, pp. 182-194. Springer, Heidelberg (2010).

[18] Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. Speech Communication, 57, 209-232.

[19] Jokinen, K., & Pelachaud, C., (2013). From Annotation to Multimodal Behavior. In Coverbal Synchrony in Human-Machine Interaction, Rojc, M. & Campbell, N., eds., Crc Press, 2013, ISBN: 978-1-4665-9825-6.

[20] Rojc, M., Mlakar, I., & Kačič, Z. (2017). The TTS-driven affective embodied conversational agent EVA, based on a novel conversational-behavior generation algorithm. Engineering Applications of Artificial Intelligence, 57, 80-104.

[21] Yumak, Z., & Magnenat-Thalmann, N. (2016). Multimodal and multi-party social interactions. In Context Aware Human-Robot and Human-Agent Interaction (pp. 275-298). Springer International Publishing.

[22] Li, Y., Tao, J., Chao, L., Bao, W., & Liu, Y. (2016). CHEAVD: a Chinese natural emotional audio-visual database. Journal of Ambient Intelligence and Humanized Computing, 1-12.

[23] Martin, J. C., Caridakis, G., Devillers, L., Karpouzis, K., & Abrilian, S. (2009). Manual annotation and automatic image processing of multimodal emotional behaviors: validating the annotation of TV interviews. Personal and Ubiquitous Computing, 13(1), 69-76.

[24] Koutsombogera, M., Touribaba, L., & Papageorgiou, H. (2008, May). Multimodality in conversation analysis: a case of Greek TV interviews. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008) Workshop on Multimodal Coorpora from Models of Natural Interaction to Systems and Applications (pp. 12-15).

[25] Caridakis, G., Wagner, J., Raouzaiou, A., Lingenfelser, F., Karpouzis, K., & Andre, E. (2013). A cross-cultural, multimodal, affective corpus for gesture expressivity analysis. Journal on Multimodal User Interfaces, 7(1-2), 121-134.

[26] Paggio, P., & Navarretta, C. (2016). The Danish NOMCO corpus: multimodal interaction in first acquaintance conversations. Language Resources and Evaluation, 1-32.

[27] Lin, Y. L. (2017). Co-occurrence of speech and gestures: A multimodal corpus linguistic approach to intercultural interaction. Journal of Pragmatics, 117, 155-167.

[28] Zhang, Z., Girard, J. M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., & Cohn, J. F. (2016). Multimodal spontaneous emotion corpus for human behavior analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3438-3446).

[29] Fitzpatrick, E. (Ed.). (2007). Corpus linguistics beyond the word: corpus research from phrase to discourse (Vol. 60).

[30] Liu, I. T., & Sun, C. S. (2007). Association between emotional reaction and visual symbols. In 3rd WSEAS/IASME international conference on EDUCATIONAL TECHNOLOGIES (EDUTE'07).

[31] Ridderinkhof, K. R. (2017). Emotion in action: A predictive processing perspective and theoretical synthesis. *Emotion Review*, 1754073916661765.

[32] Zhu, L. (2016). Language, emotion and metapragmatics: A theory based on typological evidence. International Journal of Society, Culture & Language, 4(2), 119-134.

[33] Mlakar, I., Kačič, Z., & Rojc, M. (2012). Form-oriented annotation for building a functionally independent dictionary of synthetic movement. Cognitive Behavioural Systems, 251-265.

[34] Rojc, M., Mlakar, I. (2016). An expressive conversational-behavior generation model for advanced interaction within multimodal user interfaces, (Computer Science, Technology and Applications). New York: Nova Science Publishers, Inc., cop. XIV, p. 234 str. ISBN 978-1-63482-955-7. ISBN 978-1-63484-084-2.

[35] Walsh, M. (2010). Multimodal literacy: What does it mean for classroom practice? Australian Journal of Language and Literacy, 33(3), 211.

[36] Sloetjes, H., & Wittenburg, P., (2008). Annotation by category – ELAN and ISO DCR. In: Proceedings othe 6th International Conference on Language Resources and Evaluation (LREC 2008).

[37] Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

[38] Allwood, J., Nivre, J., & Ahlsen, E. (1993). On the semantics and pragmatics of linguistic feedback. Journal of Semantics, 9(1), 1–26.

[39] Keltner, D., & Cordaro, D. T. (2017). Understanding Multimodal Emotional Expressions. The science of facial expression, 1798.

[40] Ekman, P. (1992). An argument for basic emotions. Cognition and Emotion, 6 (3/4), 169–200.

[41] Allwood, J., Lanzini, S., & Ahlse´ n , E. (2014). Contributions of different modalities to the attribution of affective-epistemic states. In P. Paggio & B. N. Wessel-Tolvig (Eds.), Proceedings from the 1st European symposium on multimodal communication University of Malta (pp. 1–6).

[42] Laycraft, K. C. (2014). Creativity As An Order Through Emotions: A Study of Creative Adolescents and Young Adults. BookBaby.

[43] Seligman, M. E., & Csikszentmihalyi, M. (2014). Positive psychology: An introduction. In Flow and the foundations of positive psychology (pp. 279-298). Springer Netherlands.