A Corpus for Analyzing Linguistic and Paralinguistic Features in 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: - This study 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 affective face-to-face conversations. The main motivation for this study is to be able to create natural co-verbal resources for automatic synthesis of highly natural co-verbal behavior on general un-annotated text and expressed through embodied conversational agents. The presented study utilizes a highly multimodal approach that investigate several linguistic levels, such as: paragraphs, sentences, sentence types, words, POS tags, and prosodic features such as phrase breaks, prominence, durations, and F0, as well as functional and formal annotation of co-verbal behavior, such as: collocutor's role (speaker, listener), semiotic classification of behavior, emotions, facial expressions, head movement, gaze, and hand gestures. The EVA corpus in this way represents a valuable resource for the algorithm for synthesizing co-verbal behavior primarily focused on gestures and semiotic intent. The EVA corpus in the presented form represent a rich empirical resource for performing several studies of complex conversational phenomena that are present in highly spontaneous face-to-face conversations, especially those related to multimodal expression of information, emotions, and communicative and non-communicative role of co-verbal expressions. In the paper also, the proposed annotation scheme and annotation procedure are presented. Preliminary studies of phenomena regarding emotions within conversations on the EVA corpus have been conducted and presented in the paper.

Key-Words: - multiparty dialog, informal conversation, multimodal corpora, linguistic and paralinguistic features, verbal and non-verbal interaction

1 Introduction

Co-verbalism has been a research paradigm in linguistic and non-linguistic research areas. In this paradigm co-verbal behavior is observed as an equivalent component of natural interaction. In terms of information presentation and understanding and active contributing to discourse it is equally relevant as speech. Allwood [1], McNeill [19,20], Duncan [8], Bozkurt[4] and Poggi[24], among others, have already made a significant effort in order to re-define the theory of communication and pushing it well beyond the realm of pure linguistics. Nowadays the co-verbal behavior is regarded as an orchestrator of communication. For instance, the gestures may clarify or re-enforce the information provided by the speech. Thus, the co-verbal behavior effectively retains semantics of the information [10], and gives a certain degree of clarity in the discourse [6,7]. Along with the theoretical research, the multimodal corpora have been developed and analyzed capturing various levels and descriptive features [2,3,29]. These multimodal corpora and annotation schemes represent a key resource for studying the complex human verbal and non-verbal behavior, for its modeling, and for the realization of the behavior on conversational agents [12,26].

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]. EVA corpus and annotation scheme is in this respect oriented specifically towards the generation of form of those co-verbal gestures and facial expressions, observed during face-to-face spontaneous multi-speaker interaction [22], and to study how they interplay with verbal part (beyond semantics). In the paper an annotation scheme for EVA corpus is presented, and preliminary observations on annotated material of extended EVA corpus are presented.

2 Background

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 [15,17]. However, most of them target narration and/or dialogues with only two participants. Furthermore, such material is often too formal and typically does not fully reflect complex natural spontaneous responses. It may also incorporate a lot of information that may be regarded as noise, and thus may obscure the effort in investigating a particular goal. To minimize the noise, another approach is the generation of multimodal corpora in a laboratory conditions [5,16,23,30]. Nevertheless, such corpora are generated with some specific purpose and usually incorporates individuals, who are instructed to implement various techniques and aspects of the spoken dialogue. Undoubtable these corpora can provide a unique opportunity for researchers to study natural multimodal phenomena. However, the conditions are still controlled, and the implications of broader context may be obscured due to the controlled and regulated set-up [11]. However, everyday natural human-human interactions are not completely ordered and synchronous. The also contain a lot of noise. This noise, if properly analyzed and incorporated, may unravel a lot of features and contexts that actually modeling the multimodal conversational expressions. Thus, the informal corpus arguably represents the most spontaneous face-to-face interaction. Namely, casual conversation is much more spontaneous than interviews and/or laboratory settings [11].

In this paper we represent a new Slovenian multimodal corpus of annotated causal multiparty conversations involving minimal three participants during interaction, which adds another set of social features to the situational contexts, such as: social imbalance, overlapping and disordered turn-management etc. Namely, in just two-participant conversation situations, the conversational context (including engagement and turn-taking) is commonly grounded between two interlocutors. On the other hand, in multiparty conversation, engagement and turn-taking cannot always be identified among the participants. Further, in multiparty conversations, participants may be left behind and may even abruptly interrupt when trying to re-incorporate the information flow. Due to constant overlaps in ideas, topics and themes, it is sometimes impossible to clearly outline the conversational context, including the scope of emotional context of responses [9,18]. Namely, the main attributes of informal multiparty settings, such as one observed, are: a) disorder in the information exchange, and b) social features incorporated in interaction. Thus, informal conversations may be especially hard to process However, the results may provide a more viable insight into various aspects of the verbal and non-verbal behavior generated during face-to-face interactions.

The proposed multimodal corpus, named EVA, is based also on studying the synchrony between verbal and co-verbal elements established during conversations, by following the concept of semiotic intent [26]. Semiotic intent incorporates grammar, which correlates communicative intent (defined through POS and prosodic features) with gestures; a concept of analysis beyond pure semantics, as presented in [26]. In this paper the concept of semiotics is extended with additional linguistic, paralinguistic features and emotion. By exploiting the EVA corpus, we can take into consideration the interplay of several conversation phenomena, such as: attitude, dialog, prosody, structuring of information and the structure of its representation, communicative intents, facial expressions and gestures, head movement, etc. In essence we try to interlink features and channels that are exploited by collocutors with linguistic and paralinguistic features that can be extracted from unannotated texts. Such links then provide us with the basis for the synthesis of more natural and more situation adaptive co-verbal behavior facilitating concepts generated especially in spontaneous and highly casual multiparty settings

3 Material for EVA Corpus

The audio/video material used for EVA corpus originate from GoS corpus [28], a corpus of spoken Slovenian that already included video and audio recordings with corresponding orthographic transcriptions of approximately 120 hours of speech. This data contains those conversations that we are exposed to on a daily basis 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, which is presented in this paper, currently 4 video recordings were selected from GoS corpus. Each selected video contains about 50 minutes of transcribed highly informal and affective multiparty conversation, with 3 – 4 collocutors exchanging information in a highly unordered and dynamical manor. Further, conversational setting is built around a talk-show with two TV presenters (natural born actors) present in all four recordings, one main guest (natural born actor or not) with two additional guests that have some personal relationship with the main guest

(close friends). Further, the topics discussed are highly changeable, informal, and casual and full of humor, resulting in several emotional responses and facial expressions. Language used by the collocutors is also quite colloquial and the dialogue contains many irregularities in turn management, and phenomena, such as overtaking and overlapping. Speech signal has been transcribed in two versions – in its original colloquial form incorporating dialects (pronunciation-based transcriptions), and in its standardized form (transcriptions based on standardized spelling). The conversation was split into 5 sessions, each session maintaining information for individual speaker. Within each session, individual's speech was segmented into statements (paragraphs), sentences and words. Currently, we have completely annotated one video file (57 min 30s of material). Thus, statistics for the annotated video given in Table 1.

.

Statements	Overall: 1516 AVG per speaker:303 (STD = 260)
Duration of statements	Overall: 93min 29s Max duration:23.22s, Min duration: 0.19s AVG per statement: 3.57s (STD = 0.54)
Sentences	Overall: 2014 AVG per collocutor:402 (STD = 364) AVG per statement: 1.32
Duration of sentences	Max duration:18.4s, Min duration: 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 shows used for EVA corpus annotation.

The goal of this research is to create a multimodal resource of natural and casual conversation to be used for automatic recreation of co-verbal behavior in TTS system PLATTOS [25]. The basic structure and nature of the material must, therefore, expose a general nature of casual interaction. As outlined in Table 1, the material consists of 1516 statements distributed among 5 speakers. The distribution per speaker further shows that some of the collocutors were participated significantly less time. As already mentioned, two of the collocutors were invited guests, as personal friends, and were present only short period of time, therefore, contributing to around 4% of the content.

However, by observing the video data, we can see that the overall duration of spoken content is almost twice longer than the overall length of the recording. This indicates that for almost half of the time statements overlapped, which is one of the phenomena in highly spontaneous multiparty dialogues. Next, if we observe the duration of sentences or statements, we can see that on average a sentence lasted 2.66s (18.4s at most) and the overall statement/paragraph, formulated with 1.3 sentences, 3.57s on average. This means that most of the time exchange of information was highly dynamic and involved shorter statements and ideas rather than monologues and narratives. This is another feature of spontaneous and casual conversation.

4 A novel annotation scheme

In order to create the EVA corpus, the video data was annotated by following the novel annotation scheme that incorporates annotation of 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 scheme is outlined in Fig. 1. This model comes as an evolution of the EVA annotation scheme [22,25]. Namely, the EVA annotation scheme has the capacity to annotate form of movement in high resolution. It also incorporates the correlation of verbal and non-verbal elements through semiotics. Further, this extension of the annotation scheme integrates linguistic, paralinguistic and non-verbal features of multimodal conversations and multiparty dialogs. The extended version of the annotation scheme allows us to analyze and incorporate these features into existing and new conversational relationships. Through obtained knowledge and co-verbal resources, we can update existing conversational models or even create new.

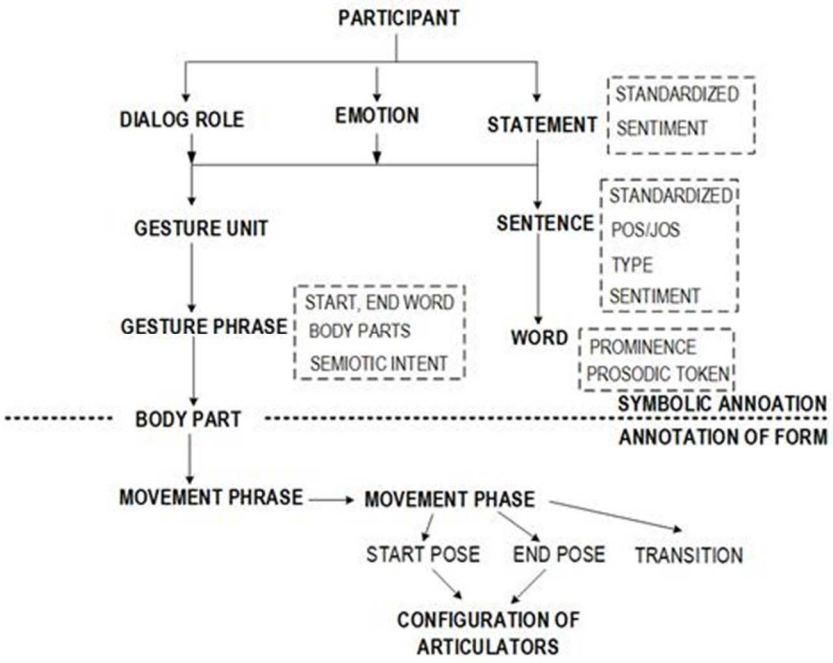

Fig. 1: The novel model for the EVA annotation scheme

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. Session for each speaker, as proposed [25], is first separated into annotation of function, and into annotation of the form. In the novel EVA annotation scheme, we have enriched the functional annotation part. We have extended verbal features of EVA Corpus by integrating additional linguistic and para-linguistic features of the spoken content.

We have also established connections linguistic and para-linguistic features and the co-verbal behavior. As outlined in the Fig. 1 the idea behind these new connections is to be able to relate gesture units (co-verbal expressions) with other low- and high-level co-verbal features in face-to-face communication, such as: emotions, dialog role, and with linguistic and paralinguistic features of verbal part, such as lemma (sentence, word), POS tags, sentence type, phrase breaks and prominence, sentiment, and semiotic intent. Trough EVA annotation scheme the annotated features 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, e.g. are there any linguistic and semiotic features related to feedback; or is there some semiotic intent that can help to indicate what kind of emotion to synthesize etc. This extends the model and the scheme well beyond natural machine responses (co-verbal behavior) and into natural language processing and understanding and even other parts of pure linguistics.

The second part of the novel EVA annotation scheme is dedicated to the annotation of form of the observed co-verbal behavior in high resolution. 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. Firstly, the annotation scheme separates between hands, arms, head, and face. 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 phrase contains a mandatory stroke and optional preparation, hold, and retraction phases. Movement phrase, therefore, joins sequential movement phases into continuous movement.

The proposed topology is outlined in Fig. 2 further joins sequential movement phases into lexical groups that can be used for building a functionally independent dictionary of co-verbal movement. Each movement phase is further segmented into start pose, end pose, and the transition trajectory that hands perform during the propagation from the start to the end pose.

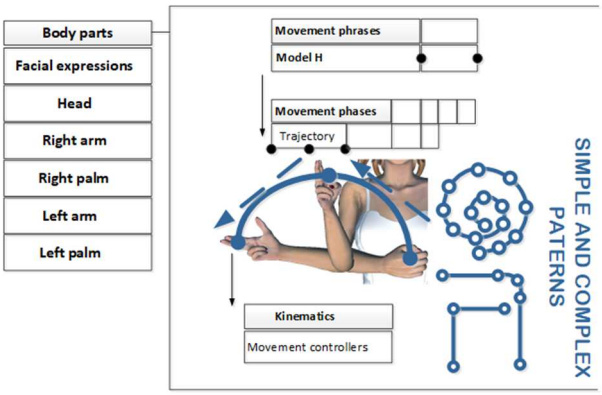

Fig. 2: Topology of the annotation of the form of co-verbal behavior and movement trajectories

One of the key features of the annotation of from is the parametric description of trajectories. As outlined in Fig. 2, It includes breaking down the trajectory in movement primes (simple patterns) such as: linear, arc. The primes can the form more complex patterns such as: chair, roof, and spiral. Each prime is further segmented into 2 (or 3 for arc) key points, each identifying a transitional hand/arm pose. At each key point the configuration of movement controllers (e.g. pose) is described only at the abstract level, e.g. in the form of the hand and arm position in 3D space relative to body and the hand-shape [25].

4.1 The EVA Corpus

The EVA multimodal corpus was generated by applying the novel annotation scheme 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) [27]. It is developed in JAVA language and can be used on Windows, Macintosh, and Linux. The tool integrates a data model that includes time information and is stored in XML format. Fig. 3 outlines the user interface and the annotation topology designed in ELAN.

We have applied the EVA Scheme and fully annotated the selected video recordings. Currently, the EVA corpus contains around 93minutes of video material. Parts of the EVA corpus have already been used (or is a part of) in studies regarding the analysis and classification of the co-verbal behavior. Nevertheless, the multimodal analysis and annotation of emotion units, as generated during multiparty casual conversations and reported in this paper and is an ongoing research in which we try to establish linguistic and paralinguistic relations between emotions, verbal content, and the identified semiotic intent. In this effort we are trying to identify a wider context of communicative features, which relate spoken content and co-verbal emotions and facial expressions. The established relations will be used for generation of spontaneous conversational emotion on our synthetic agent EVA.

5 Multimodal annotation of emotion

The main motivation regarding multimodality of emotion in spontaneous face-to-face multi-speaker conversations, is to tackle the problem of synthesizing facial expressions and emotions from unannotated texts. Namely, when virtual collocutors are capable incorporating emotions and affect in their interactions and responses, 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 [13]. 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.

In order to capture these emotions as conversational stimuli, we have applied the EVA annotation scheme to the selected video material. Annotation task by using the presented topology and performed in ELAN is outlined in Fig. 3. The annotators were asked to classify the observed emotions as expressed during the conversation. Emotion was identified in a dedicated track and regardless of the collocutors dialog role, or presence of the verbal content. The annotators were instructed to classify emotion as feeling that goes beyond listener/speaker segments, verbal content parts, or even paragraphs/sentences. E.g. emotion unit for ‘anticipation’ can span over three sentences, and is kept also during the time, where the observed collocutor acts primarily as a listener. Furthermore, emotion units could also be driven by the communicative intent as a stimulus; for instance, ‘anticipation’, ‘trust’ and ‘disagreement’ as feedback signals, while ‘anticipation’ as conversational regulator in turn-assignment.

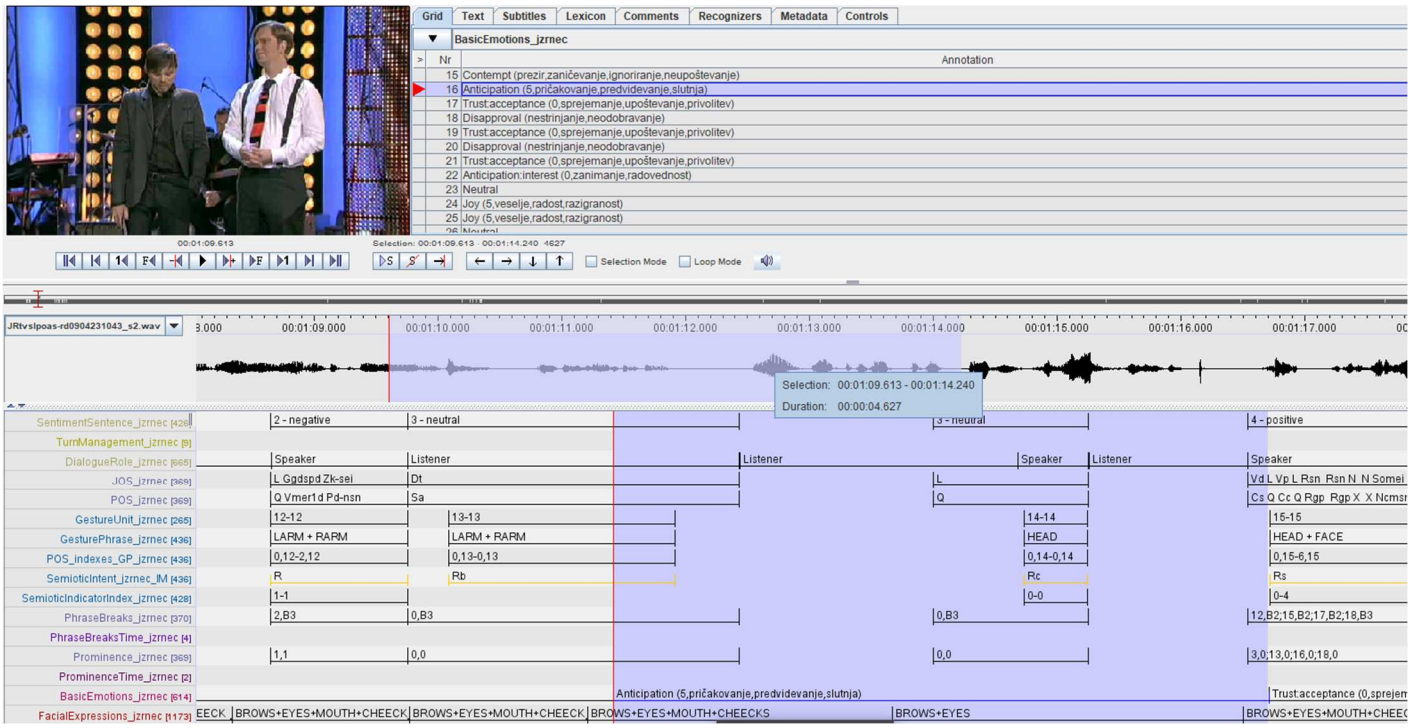

Fig. 3: The ELAN interface: multimodal annotation topology used for studying conversational emotion

For the classification of emotion units, we have used an extension of Plutchik's table with primary emotions [14]. The annotators were, therefore, asked to select between 50 emotions and 2 non-emotional states, e.g. 'rest' and 'undefined'.

6 Annotation Results

EVA corpus presented in this paper contains roughly 3000 instances of emotions. Emotion unit 'Anticipation:Interest' was observed as being the most dominant one. It was followed by 'Trust:Acceptance', 'Joy' and 'Anger:Annoyance'. The basic relation between text and emotion units can be established through sentiment units, e.g. the positive/negative connotation of content. In order to capture 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. 4 and Fig 5 present the observed distribution of sentiment units in case of emotion unit "Anticipation:Interest" and "Trust:Acceptance". Namely, through sentiment we are able to relate unannotated texts with emotions and facial expressions.

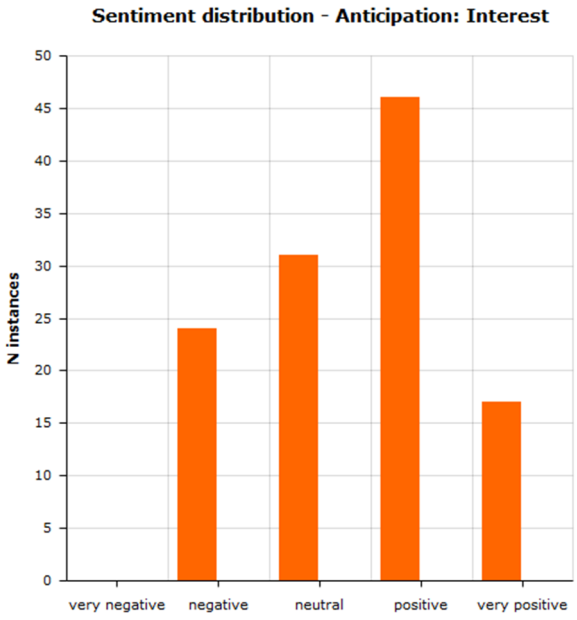

Fig. 4: Correlation of sentiment and Anticipation: Interest

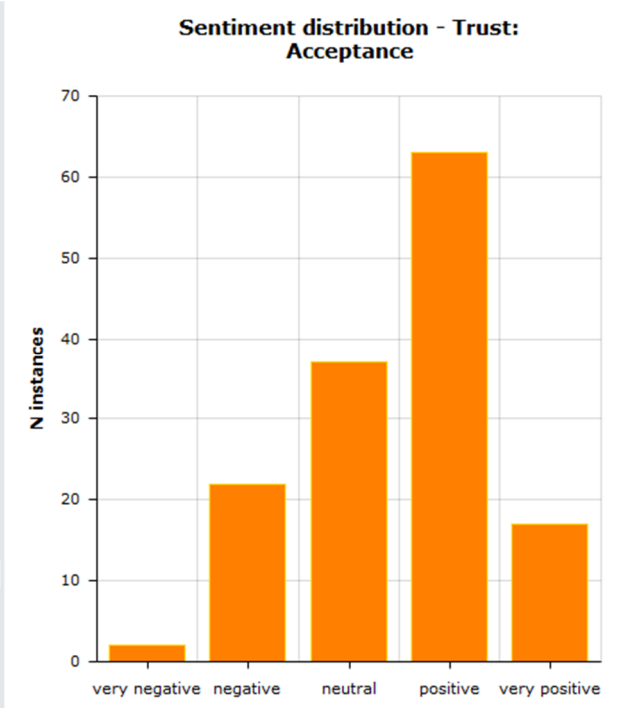

Fig. 5: Correlation of sentiment and Trust:Acceptance

Fig. 4 and 5 point out e.g. that the predominant emotions" Anticipation:Interest" and "Trust:Acceptance" are largely positive emotions. Further, they are primarily used in context, where verbal content represents positive signal. This finding is also in line with well-established emotional theory [31, 32], where 'Interest' and 'Acceptance' are defined as positive emotions, and where 'Interest' is regarded as a heightened state that calls for one's attention to something new that inspires fascination, and curiosity, while 'Acceptance' is interpreted as a mild form of 'Trust', as a willingness to see things as they are. However, neither of the emotions is obviously limited to a single sentiment value. E.g. for the emotion 'Acceptance', the more negative connotations are obvious especially 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, e.g. while collocutors are listeners. As shown in Fig. 6 and Fig. 7, this statement can be observed by relating dialog role and emotion tracks in the EVA corpus.

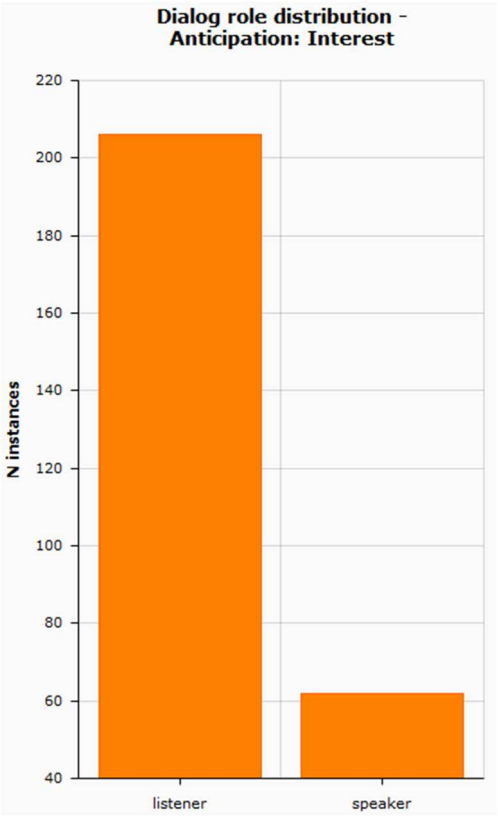

Fig. 6: Relating 'Dialog role' and emotion unit 'Anticipation: Interest' within EVA corpus.

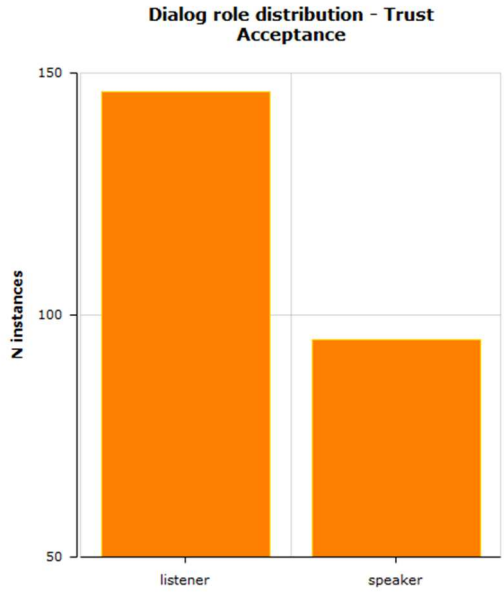

Fig. 7: Relating 'Dialog role' and emotion units 'Trust:Acceptance' within EVA corpus.

As outlined in Fig. 6, 'Interest' may be regarded as predominately part of the listener behaviour, while 'Acceptance' on the other hand is used as part of speaker and listener behaviour. Namely, it is used to signal agreement with a statement. And it can also provide emotional attitude towards a topic that collocutor is presenting, especially when the 'revelation' has negative connotation. Finally, since both emotions '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. 8, 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.

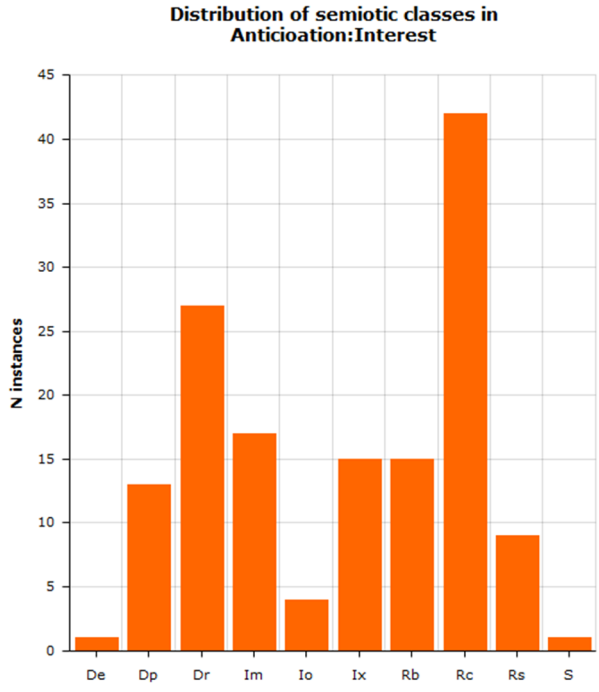

Fig. 8: Relating communicative semiotic intents and emotion units for 'Anticipation: Interest'

Therefore, in similar way we can observe and establish other statistic-based 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 gives the material a trully multimodal and multi-context attribute. Finally, the 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.

7 Conclusion

In this paper we have presented a novel EVA multimodal corpus generated by using novel EVA annotation scheme. The presented scheme is a product of proprietary research in recreation of spontaneous co-verbal behavior. It also incorporates knowledge obtained through the implementation of the first version on a machined learned model [25, 26]. The EVA annotation scheme in this paper 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, was also outlined in detail. At the end, one specific analysis regarding emotion unit in the material demonstrates, how the EVA corpus can be used for detecting and investigating several conversational phenomena and relations. The analysis 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 proposed 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 improvements will include deeper prosodic and linguistic analysis as well as detailed analysis of dialog well beyond the collocutors role.

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

References:

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

[2] Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. J. of Language Resources and Evaluation 41(3), 273–287.

[3] 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).

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] El-Assady, M., Hautli-Janisz, A., Gold, V., Butt, M., Holzinger, K., & Keim, D. (2017). Interactive visual analysis of transcribed multi-party discourse. Proceedings of ACL 2017, System Demonstrations, 49-54.

[10] 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.

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

[12] 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.

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

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

[15] 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.

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

[17] 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.

[18] Matsuyama, Y., Akiba, I., Fujie, S., & Kobayashi, T. (2015). Four-participant group conversation: A facilitation robot controlling engagement density as the

fourth participant. Computer Speech & Language, 33(1), 1-24.
[19] McNeill, D., 2005. Gesture and Thought, University of Chicago Press.
[20] McNeill, D. (2015). Why we gesture: The surprising role of hand movements in communication. Cambridge University Press.
[21] Mlakar, I., & Rojc, M. (2012). Capturing form of non-verbal conversational behavior for recreation on synthetic conversational agent EVA. WSEAS Trans. Comput. [Print ed.], 11(7), 218-226.
[22] Mlakar, I., Kačič, Z., & Rojc, M. (2012). Form-oriented annotation for building a functionally independent dictionary of synthetic movement. Cognitive Behavioural Systems, 251-265.
[23] Paggio, P., & Navarretta, C. (2016). The Danish NOMCO corpus: multimodal interaction in first acquaintance conversations. Language Resources and Evaluation, 1-32.
[24] Poggi, I. (2007). Hands, mind, face and body: A goal and belief view of multimodal communication. Berlin: Weidler.
[25] 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.
[26] 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.
[27] 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).
[28] Verdonik, D., Kosem, I., Vitez, A. Z., Krek, S., & Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. Language resources and evaluation, 47(4), 1031-1048.
[29] Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. Speech Communication, 57, 209-232.
[30] 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).
[31] Haidt, J. (2000). The Positive emotion of elevation.
[32] Seligman, M. E., & Csikszentmihalyi, M. (2014). Positive psychology: An introduction. In Flow and the foundations of positive psychology (pp. 279-298). Springer Netherlands.