Towards Pragmatic Understanding of Conversational Intent: A Multimodal Annotation Approach to Multiparty Informal Interaction – The EVA Corpus

Izidor Mlakar(✉), Darinka Verdonik, Simona Majhenič, and Matej Rojc

Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia
{izidor.mlakar, darinka.verdonik, simona.majhenic, matej.rojc}@um.si

Abstract. The present paper describes a corpus for research into the pragmatic nature of how information is expressed synchronously through language, speech, and gestures. The outlined research stems from the 'growth point theory' and 'integrated systems hypothesis', which proposes that co-speech gestures (including hand gestures, facial expressions, posture, and gazing) and speech originate from the same representation, but are not necessarily based solely on the speech production process; i.e. 'speech affects what people produce in gesture and that gesture, in turn, affects what people produce in speech' ([1]: 260). However, the majority of related multimodal corpuses 'ground' non-verbal behavior in linguistic concepts such as speech acts or dialog acts. In this work, we propose an integrated annotation scheme that enables us to study linguistic and paralinguistic interaction features independently and to interlink them over a shared timeline. To analyze multimodality in interaction, a high-quality multimodal corpus based on informal discourse in a multiparty setting was built.

Keywords: Corpora and language resources · Multimodal corpus · Multimodal technologies · Natural language understanding. pragmatics · Annotation · Conversational intelligence

1 Introduction

In social and spoken interaction, language is not used in isolation and does not occur in a vacuum [2]. Embodied behavior adds more than 50 percent of non-redundant information to the common ground of the conversation [3]. The sharing of information or the exchange of information in human social interactions is far more complex than a mere exchange of words. It is multilayered and includes attitude and affect, utilizes bodily resources (embodiment) as well as a physical environment in which the discourse takes place [4].

Effective communication requires the following conditions to be fulfilled: (i) the communicator must make his or her intention to communicate recognizable and (ii) the

C. Martín-Vide et al. (Eds.): SLSP 2019, LNAI 11816, pp. 19–30, 2019.
https://doi.org/10.1007/978-3-030-31372-2_2

propositional content or conceptual or ideational meaning (e.g. semantic information) that they wish the recipient to receive must be represented effectively [5].

In interpersonal discourse, verbal signals carry a symbolic or semantic interpretation of information through linguistic and paralinguistic properties, while non-verbal signals (i.e. embodiments) orchestrate speech [6]:4. Non-verbal concepts, such as prosody, embodiments, emotions, or sentiment are multi-functional and operate on the psychological, sociological, and biological level and in all time frames. These signals represent the basis of cognitive capabilities and understanding [7, 8]. Embodied behavior in particular, effectively retains the semantics of the information, helps in providing suggestive influences, and gives a certain degree of cohesion and clarity to the overall discourse [9, 10]. Non-verbal behavior, although not bound by grammar, co-aligns with language structures and compensates for the less articulated verbal expression model [2, 11]. It also serves interactive purposes, such as content representation or expression of one's mental state, attitude, and social functions [12–16].

The main motivations for the work presented in this paper is driven by the goal of enabling machine 'sensing' and more natural interaction with virtual agents. Despite the considerable interest in this topic and significant progress reported, automatically understood and machine-generated information from a set of evidence is, in general, still far from perfect or natural [11, 17]. Moreover, not only speech and language affect embodiment but embodied signals also affect what people produce through language and speech [1].

This paper presents a multimodal approach to generating 'conversational' knowledge and modeling of the complex interplay among conversational signals, based on a concept of data analytics (mining) and information fusion. Our work outlines a novel analytical methodology and a model to annotate and analyze conversational signals in spoken multi-party discourse. Moreover, the results of our annotation process (i.e. the corpus) applied to a multi-party discourse setting in Slovenian are represented. In addition to capturing language-oriented signals, naïve to modern corpus linguistics, the model also provides a very detailed description of non-verbal (and paralinguistic) signals. These disparate phenomena are interconnected through the notion of co-occurrence (e.g. timeline).

2 Background

One of the main issues in sentic computing is misinterpretation of conversational signals and non-cohesive responses. As a result, 'multimodality in interaction' became one of the fundamental concepts in corpus linguistics. Especially in interactional linguistics and conversation analysis, a significant focus was shifted to embodied behavior (an overview of such research can be found in [11, 18]). The semantic domain is particularly well-suited when investigating co-verbal alignment. Research studies show how humans 'map' semantic information onto linguistic forms [10, 19, 20]. Linguistic approaches in general tend to observe embodied behavior in discourse on a linguistic basis (i.e. language and grammar). However, as argued by Birdwhistell [21], what is conveyed through the body does not meet the linguist's definition of language. Therefore, the apparent grammatical interface between language and gestures seems to

be limited ([2]). In terms of creating conversational knowledge, such association operates in very narrow contexts and opens limited and highly focused opportunities to explore the interplay between verbal and non-verbal signals [8, 22].

In contrast, the researchers in [6, 14–16], among others, propose to involve additional modalities, such as sound and image, and investigate the functional nature of embodiments during discourse. The widely adopted approach to multimodality in interaction is Pierce's semiotic perspective (i.e. the 'pragmatics on the page'), which explores the meaning of images and the intrinsic visual features of written text. In [23], for instance, the authors correlated hand shapes (and their trajectories) with semiotic class based on a broader context of the observed phenomena. Although the approaches oriented towards semiotics (i.e. [24–26]) go beyond semantics and do not restrict embodiments to linguistic rules, they still restrict themselves functionally, that is to a specific phenomenon and a narrow discourse context.

In contrast to the aforementioned approaches inspired by linguistics, Feyaerts et al. [27] authors build on the cognitive-linguistic enterprise and equally incorporate all relevant dimensions of how events are utilized, including the trade-off between different semiotic channels. However, the discourse setting is limited to an artificial setting. Due to the challenging nature of informal, especially multiparty discourse, researchers tend to establish artificial settings [28]. These settings introduce laboratory conditions with targeted narration and discourse concepts between collocutors which focus on a specific task. Such data sources therefore clearly reveal the studied phenomena but hinder 'interference' of other, non-observed signals that would appear in less restricted settings. Furthermore, in most cases, a wider scope of conversational signals is intentionally left out of the conversational scenario [29, 30]. Following [27], we observe discourse as a multimodal phenomenon, in which each of the signals represents an action item, which must be observed in its own domain and under its own restrictions. We focus on corpus collection, structuring, and analysis. Instead of 'artificial' scenarios we utilize a rich data source based on an entertaining evening TV talk show in Slovene, which represents a good mixture of institutional discourse, semi-institutional discourse, and casual conversation.

3 Data Collection and Methodology: The EVA Corpus

3.1 Data Source

In this research, we used the EVA Corpus [31] which consists of 228 min in total, including 4 video and audio recordings, each 57 min long, with corresponding orthographic transcriptions. The discourse in all four recordings is a part of the entertaining evening TV talk show *A si ti tut not padu*, broadcast by the Slovene commercial TV in 2010. In total, 5 different collocutors are engaged in each episode. The conversational setting is relaxed and unrestricted. It is built around a general scenario, focused on day-to-day concepts. The discourse involves a lot of improvisation and is full of humor, sarcasm, and emotional responses. Moreover, although sequencing exists and general discourse structuring (e.g. role exchange, topic opening, grounding, etc.) applies, it is performed highly irregularly. Table 1 outlines the general

characteristics of the recording used to define the proposed model of conversational expression. The utterances in the EVA Corpus are mostly single short sentences, on average consisting of 8 words. The discourse contains 1,801 discourse markers (counting only those with a minimum frequency of 10). The corpus includes a lot of non-verbal interactions: 1,727 instances in which 'movement' was attributed to convey meaning (e.g., a gesture performed with an intent) were classified.

Table 1. General characteristics of discourse in the EVA Corpus.

Utterances	
Total	1,516
AVG per speaker	303
Sentences	
Total	1,999
AVG per speaker	399.8
AVG per statement	1.32
Words	
Total	10,471
AVG per speaker	2094
AVG per sentence	7.9
Metadiscourse	
discourse markers (n > 10)	1,801
AVG per speaker	599
Non-verbal behavior	
Total number of semiotic intents	1,727

The data in Table 1 clearly outline that contributors are active and that the discourse involves short statements (i.e. under 5 s) with a significant amount of overlapping speech. Individual sentence duration ranges from 0.5 s to 5 s and 2.8 s on average. Together, all participants generate roughly 93 min of spoken content in a 57-min recording. The statements are interchanging rapidly among the collocutors and with high density.

3.2 Annotation Topology

In order to realize the aforementioned 'conversational model' and observe each conversational expression in greater detail, a multimodal annotation approach typically used in conversational analysis was adopted. For this purpose, an annotation topology with various levels, as outlined in Fig. 1, was defined. The scheme applies a two-layered analysis of the conversational episode.

In the first layer (i.e. symbolics/kinesics), signals that are primarily evident in the formulation of an idea and identify the communicative intent were observed and annotated. As outlined in Fig. 1, this layer annotates linguistic and paralinguistic signals

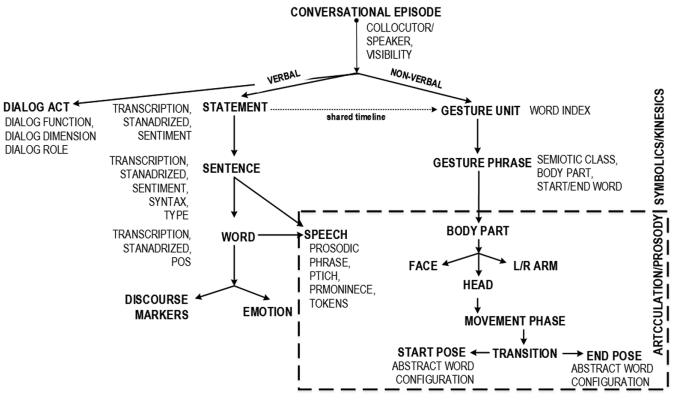

Fig. 1. The topology of annotation in the EVA Corpus: the levels of annotation describing verbal and non-verbal contexts of conversational episodes

(e.g. kinesics [21]). The second layer (i.e. articulation/prosody) is oriented towards the form and is concerned with how an abstract concept (i.e. communicative intent) is physically articulated through auditory and visual channels. It provides detailed descriptions of the structure of verbal and non-verbal components and on how the articulators (verbal and non-verbal ones) are modeled, moved, and put to use.

The material was annotated by an annotator with a background in linguistics and experience in annotation of multimodal materials. The annotations were performed in ELAN (EUDICO Linguistic Annotator) and WebAnno, converged into a single data source, specified as JSON, and visualized. The currently available annotations were performed over a nine-month period and in separate trials for each conversational concept and, in some cases, even for each signal.

3.3 Annotation Procedure and Inter-annotator Agreement

Five annotators, two with linguistic background, and three with technical background in machine interaction were involved in this phase of annotations. Annotations were performed in separate sessions, each session describing a specific signal. The annotation was performed in pairs, i.e. two or three annotators annotated the same signal. After the annotation, consensus was reached by observing and commenting on the values where the was no or little annotation agreement among multiple annotators (including those not involved in the annotation of the signal). The final corpus was generated after all disagreements were resolved. Procedures for checking inconsistencies were finally applied by an expert annotator.

Before starting with each session, the annotators were given an introductory presentation defining the nature of the signal they were observing and the exact meaning of the finite set of values they could use. An experiment measuring agreement was also performed. It included an introductory annotation session in which the preliminary inconsistencies were resolved. For the less complex signals, influenced primarily by a single modality (i.e. pitch, gesture unit, gesture phrase, body-part/modality, sentence

type, etc.), the annotators' agreement measured in terms of Cohen's *kappa* [32] was high, namely between 0.8 and 0.9. As summarized by Table 2, for the more complex signals that involve all modalities for their comprehension (including speech, gestures, and text) the interpretation was less unanimous.

Table 2. Results of the preliminary inter-coder agreement experiment.

Signal	Kappa score
Sentiment	0.67
Dialog function	0.64
Dialog dimension	0.71
Intent (semiotic class)	0.48
Emotion label	0.51
Gesture unit	0.75
Movement phase	0.66

The figures indicate that agreement was 0.63 on average. Given the complexity of the task and the fact that the values in Table 2 also cover cases with possible duality of meaning, the level of agreement is acceptable and comparable to other multimodal corpus annotation tasks [25]. For *Intent* the possible duality of interpretation was surprisingly common. The annotators in general agreed on the major class and would have a difference in opinion in the minor sub-class.

3.4 Transcription and Segmentation

The audio data was transcribed in original colloquial transcriptions (verbatim), and in their standardized transcriptions (standardized Slovenian spelling). The colloquial transcriptions also include meta information transcribed in brackets '[]' (e.g., [:laugher], [gap], [incident], [:voice]). All transcriptions are segmented into statements, sentences and words while also considering the temporal domain. The boundaries for colloquial and standardized statements match completely. The conversations are split into 5 sessions, in which each session contains information and annotation levels for each individual speaker. Additionally, each word was POS tagged following the JOS specifications.

3.5 Discourse Management and Structuring

Following the ISO 24617-2 [33] guidelines, dialogue acts (DA) in the EVA Corpus were annotated as an independent concept and some adjustments to the ISO scheme were added. The definition of the ISO functional segments as the basic unit of annotation and their several layers of information (sender, addressee, dimension, and communicative function) were retained. Some non-task dimensions were merged into a single cover dimension, the social obligation dimension was generalized into social management. The results of the annotation are listed in Table 3.

Table 3. Results of DA annotation in the EVA Corpus

DA		Dialog dimensions > 200	
Total acts	3,465	Total dimensions	3,465
With 1 dimension	2,144	Task	1,960
With 2 dimensions	1,175	Communication management	608
With 3 or more dimensions	146	Feedback	445
Dialog functions			
Total functions		3,479	
Functions with frequency > 25			
inform: 982, stalling: 291, ownComprehensionFB: 272, setQuestion: 176, answer: 163, checkQuestion: 135, retraction: 112, feedbackElicitation: 108, agreement: 104, instruct: 95, confirm: 93, positive: 78, interaction Structuring: 68, negative: 65, backchannel: 64, disagreement: 48, opening: 46, argument: 43, completion: 39, request: 38, partner ComprehensionFB: 35, turnTake: 32, suggest: 31, emphasis: 28, flattery: 26			

The most common dimension was task (e.g. information providing, agreement, confirmation, instructing) which accounted for more than half of the DAs. Communication management (stalling, retraction, etc.) was the second most frequently assigned dimension. This reflects a high level of spontaneity in dialogue. The third most frequent dimension was feedback, which can be explained with a high level of interaction and informal character of the dialogue.

3.6 Discourse Markers

The present research draws on previous work on Slovene DMs [34], which includes a vast set of expressions ranging from connective devices such as *and* and *or* to the interactional *yes* and *y'know* and to production markers such as *uhm*. Altogether 121 different expressions were tagged as DMs; however, only DMs with a minimum frequency of 10 were analyzed and classified into the following groups:

DM-s (speech formation markers): *eee* 'um' (316), *eem* 'uhm' (15), *mislim* 'I mean' (24), *v bistvu* 'actually'[1] (10)
DM-d (dialogue markers):

- **DM-d(c)** (contact): *veš* 'y'know' (14), *a veš* 'y'know' (24), *glej* 'look' (23), *daj* 'come on' (17), *ne* 'right?' (183), *a ne* 'right?' (21), *ti* 'you' (10), *ej* 'hey' (14)
- **DM-d(f)** (feedback): *aja* 'I see' (18), *mhm* 'mhm' (20), *aha* 'oh' (53), *ja* 'yes' (409), *fajn* 'nice' (14)
- **DM-d(s)** (dialogue structure): *dobro* 'alright' (39), *no* 'well' (79), *ma* 'well' (10), *zdaj* 'now' (21), *čakaj* 'wait' (22)

[1] *It is impossible to provide exact English equivalents for the Slovenian discourse markers examined in this paper as there are no one-to-one equivalents. The translations provided here are therefore only informative, giving the general meaning of each discourse marker.*

DM-c (connectives): *in* 'and' (65), *pa* 'and' (48), *ker* 'because' (13), *ampak* 'but' (16), *tako* 'so' (20), *a* 'but' (117), *pač* 'just' (16).

Altogether 1,651 DMs were annotated which accounts for 15.8% of all spoken content (i.e. 10,471 words).

3.7 Emotion

Emotional attitude in discourse primarily pertains to the way people feel about the conversational episode, the interlocutor, or the content of the ongoing conversation. For the annotation of emotions, Plutchik's three dimensional [35] model was applied. It has the capacity to describe complex emotions and how they interact and change over time and in a broader, social context. The results are listed in Table 4.

Table 4. Cross-speaker distribution of annotated emotions in the EVA Corpus

Emotion	Instances	Emotion	Instances
Anticipation: interest	1,239	Delight	19
Trust: acceptance	671	Trust: admiration	19
Joy	349	Boredom	15
Joy: serenity	221	Sadness	15
Disapproval	137	Contempt	14
Joy: ecstasy	92	Pensiveness	12
Surprise	69	Anger: annoyance	10
Amazement	49	Pride	10
Anticipation: vigilance	43	Alarm	7
Cynicism	29	Fear: apprehension	7
Disgust	23	Optimism	7
Distraction	23	Shame	7
Curiosity	22		

In the EVA corpus, 3,312 instances of emotional attitude were identified. The 'Anticipation: interest', 'Trust: acceptance' and 'Joy' category were identified as dominant emotions.

3.8 Classification of Embodied Behavior Through Semiotic Intent

This research focuses only on 'meaningful' movement defined through an extension of semiotics as the basis for symbolic interpretation of body language in human-human interaction. We applied the classification proposed in [31], which leverages between semiotics and kinesics, and also includes functions of discourse management (i.e. [15, 16]). The following classes of semiotic intent (SI) were distinguished:

- illustrators (I), with the subclasses: outlines (I_O), ideographs (I_I), dimensional illustrators (I_D), batons (I_B);

- regulators/adapters (R), with the subclasses: self-adaptors (R_S), communication regulators (R_C), affect regulators (R_A), manipulators (R_M), social function and obligation regulators (R_O);
- deictics/pointers (D), with the subclasses: pointers (D_P), referents (D_R), and enumerators (D_E), and
- symbols/emblems (S).

Table 5. The usage of embodied behavior in the EVA Corpus

SI class	SI subclass	Frequency	Total
I	I_O	20	178
	I_I	68	
	I_D	11	
	I_B	80	
R	R_A	105	1,194
	R_C	717	
	R_M	16	
	R_O	27	
	R_S	329	
D	D_P	40	275
	D_R	219	
	D_E	16	
S	S	37	37
(undetermined)	U	43	43
Total			**1,727**

As visible in Table 5, the EVA Corpus contains 1,727 instances of SIs generated during the discourse. The distribution of SIs shows that most of the observed embodied movement correlates with regulation and adaptation of discourse (SI class R). Among regulators, communication regulators (R_C) and self-adapters (R_S) were the most utilized non-verbal mechanism. Symbols (S) and illustrators (I) exhibit the most significant linguistic link and even a direct semantic link. In most cases, they are accompanied by a speech referent, although symbols do provide a clear meaning even without a referent in speech. In the EVA Corpus, they were classified as the least frequent non-verbal mechanism, which is also in line with non-prepared discourse.

3.9 Form and Structure of Non-verbal Expressions

From the perspective of kinesics, gestures and non-verbal expressions are considered body communication generated through movement, i.e. facial expressions, head movement, or posture. The approach outlined in [36] and the definition of the annotation of form, as represented in [37], were adopted for the description of non-verbal expressions (shape and motion). The distribution of non-verbal expressions based on modality (i.e. body parts) as represented in the EVA corpus is outlined in Table 6.

Table 6. Non-verbal patterns across all speakers in the EVA Corpus.

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

4 Conclusion

This paper presents the first Slovene multimodal corpus, the EVA Corpus. Its aim is to better understand how verbal and non-verbal signals correlate with each other in naturally occurring speech and to help improve natural language generation in embodied conversational agents. The various annotation levels incorporate and link linguistic and paralinguistic, verbal and non-verbal features of conversational expressions as they appear in multiparty informal conversations.

The concept proposed in this paper builds on the idea that a 'multichannel' representation of a conversational expression (i.e. an idea) is generated by fusing language ('what to say') and articulation ('how to say it'). On the cognitive level (i.e. the symbolic representation), an idea is first formulated through the symbolic fusion of language and social/situational context (i.e. the interplay between linguistic and paralinguistic signals interpreted as the communicative intent). On the representational level, one utilizes non-linguistic channels (i.e. gestures, facial expressions), verbal (i.e. speech) and non-verbal prosody (i.e. movement structure) to articulate the idea and present it to the target audience.

Acknowledgments. This work is partially funded by the European Regional Development Fund and the Ministry of Education, Science and Sport of the Republic of Slovenia; the project SAIAL (research core funding No. ESRR/MIZŠ-SAIAL), and partially by the Slovenian Research Agency (research core funding No. P2-0069).

References

1. Kelly, S.D., Özyürek, A., Maris, E.: Two sides of the same coin: speech and gesture mutually interact to enhance comprehension. Psychol. Sci. **21**(2), 260–267 (2010)
2. Couper-Kuhlen, E.: Finding a place for body movement in grammar. Res. Lang. Soc. Interact. **51**(1), 22–25 (2018)
3. Cassell, J.: Embodied conversational agents: representation and intelligence in user interfaces. AI Mag. **22**(4), 67 (2001)
4. Davitti, E., Pasquandrea, S.: Embodied participation: what multimodal analysis can tell us about interpreter-mediated encounters in pedagogical settings. J. Pragmat. **107**, 105–128 (2017)
5. Trujillo, J.P., Simanova, I., Bekkering, H., Özyürek, A.: Communicative intent modulates production and comprehension of actions and gestures: a Kinect study. Cognition **180**, 38–51 (2018)
6. McNeill, D.: Why We Gesture: The Surprising Role of Hand Movements in Communication. Cambridge University Press, Cambridge (2016)
7. Church, R.B., Goldin-Meadow, S.: So how does gesture function in speaking, communication, and thinking? Why Gesture?: How the hands function in speaking, thinking and communicating, vol. 7,p. 397 (2017)
8. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. Inf. Fusion **37**, 98–125 (2017)
9. Esposito, A., Vassallo, J., Esposito, A.M., Bourbakis, N.: On the amount of semantic information conveyed by gestures. In: 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 660–667. IEEE (2015)
10. Lin, Y.L.: Co-occurrence of speech and gestures: a multimodal corpus linguistic approach to intercultural interaction. J. Pragmat. **117**, 155–167 (2017)
11. Keevallik, L.: What does embodied interaction tell us about grammar? Res. Lang. Soc. Interact. **51**(1), 1–21 (2018)
12. Vilhjálmsson, H.H.: Representing communicative function and behavior in multimodal communication. In: Esposito, A., Hussain, A., Marinaro, M., Martone, R. (eds.) Multimodal Signals: Cognitive and Algorithmic Issues. LNCS (LNAI), vol. 5398, pp. 47–59. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00525-1_4
13. Arnold, L.: Dialogic embodied action: using gesture to organize sequence and participation in instructional interaction. Res. Lang. Soc. Interact. **45**(3), 269–296 (2012)
14. Kendon, A.: Semiotic diversity in utterance production and the concept of 'language'. Phil. Trans. R. Soc. B **369**, 20130293 (2014). https://doi.org/10.1098/rstb.2013.0293
15. McNeill, D.: Gesture in linguistics (2015)
16. Allwood, J.: A framework for studying human multimodal communication. In: Rojc, M., Campbell, N. (eds.) Coverbal Synchrony in Human-Machine Interaction. CRC Press, Boca Raton (2013)
17. Navarro-Cerdan, J.R., Llobet, R., Arlandis, J., Perez-Cortes, J.C.: Composition of constraint, hypothesis and error models to improve interaction in human-machine interfaces. Inf. Fusion **29**, 1–13 (2016)
18. Nevile, M.: The embodied turn in research on language and social interaction. Res. Lang. Soc. Interact. **48**(2), 121–151 (2015)
19. Hoek, J., Zufferey, S., Evers-Vermeul, J., Sanders, T.J.: Cognitive complexity and the linguistic marking of coherence relations: a parallel corpus study. J. Pragmat. **121**, 113–131 (2017)

20. Birdwhistell, R.L.: Introduction to Kinesics: An Annotation System for Analysis of Body Motion and Gesture. Department of State, Foreign Service Institute, Washington, DC (1952)
21. Adolphs, S., Carter, R.: Spoken Corpus Linguistics: From Monomodal to Multimodal, vol. 15. Routledge, London (2013)
22. Navarretta, C.: The automatic annotation of the semiotic type of hand gestures in Obama's humorous speeches. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 1067–1072 (2018)
23. Han, T., Hough, J., Schlangen, D.: Natural language informs the interpretation of iconic gestures: a computational approach. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Vol. 2, pp. 134–139 (2017)
24. Brône, G., Oben, B.: Insight interaction: a multimodal and multifocal dialogue corpus. Lang. Resour. Eval. **49**(1), 195–214 (2015)
25. Paggio, P., Navarretta, C.: The Danish NOMCO corpus: multimodal interaction in first acquaintance conversations. Lang. Resour. Eval. **51**(2), 463–494 (2017)
26. Lis, M., Navarretta, C.: Classifying the form of iconic hand gestures from the linguistic categorization of co-occurring verbs. In: Proceedings from the 1st European Symposium on Multimodal Communication University of Malta; Valletta, 17–18 October 2013, no. 101, pp. 41-50. Linköping University Electronic Press (2014)
27. Feyaerts, K., Brône, G., Oben, B.: Multimodality in interaction. In: Dancygier, B. (ed.) The Cambridge Handbook of Cognitive Linguistics, pp. 135–156. Cambridge University Press, Cambridge (2017)
28. Chen, L., et al.: VACE multimodal meeting corpus. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, pp. 40–51. Springer, Heidelberg (2006). https://doi.org/10.1007/11677482_4
29. Knight, D.: Multimodality and Active Listenership: A Corpus Approach: Corpus and Discourse. Bloomsbury, London (2011)
30. Bonsignori, V., Camiciottoli, B.C. (eds.): Multimodality Across Communicative Settings, Discourse Domains and Genres. Cambridge Scholars Publishing, Cambridge (2017)
31. Rojc, M., Mlakar, I., Kačič, Z.: The TTS-driven affective embodied conversational agent EVA, based on a novel conversational-behavior generation algorithm. Eng. Appl. Artif. Intell. **57**, 80–104 (2017)
32. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)
33. Mezza, S., Cervone, A., Tortoreto, G., Stepanov, E.A., Riccardi, G.: ISO-standard domain-independent dialogue act tagging for conversational agents (2018)
34. Verdonik, D.: Vpliv komunikacijskih žanrov na rabo diskurznih označevalcev. In: Vintar, Š. (ed.) Slovenske korpusne raziskave, (Zbirka Prevodoslovje in uporabno jezikoslovje). 1, 88–108. Znanstvena založba Filozofske fakultete, Ljubljana (2010)
35. Plutchik, R.: The nature of emotion. Am. Sci. **89**, 344–350 (2001)
36. Kipp, M., Neff, M., Albrecht, I.: An annotation scheme for conversational gestures: how to economically capture timing and form. Lang. Resour. Eval. **41**(3–4), 325–339 (2007)
37. Mlakar, I., Rojc, M.: Capturing form of non-verbal conversational behavior for recreation on synthetic conversational agent EVA. WSEAS Trans. Comput. **11**(7), 218–226 (2012)