Language Resources and Evaluation (2023) 57:641–671
https://doi.org/10.1007/s10579-022-09627-y

ORIGINAL PAPER

# Understanding conversational interaction in multiparty conversations: the EVA Corpus

**Izidor Mlakar**[1] · **Darinka Verdonik**[1] · **Simona Majhenič**[1] · **Matej Rojc**[1]

Accepted: 7 November 2022 / Published online: 10 December 2022


## Abstract

This paper focuses on gaining new knowledge through observation, qualitative analytics, and cross-modal fusion of rich multi-layered conversational features expressed during multiparty discourse. The outlined research stems from the theory that speech and co-speech gestures originate from the same representation; however, the representation is not solely limited to the speech production process. Thus, the nature of how information is conveyed by synchronously fusing speech and gestures must be investigated in detail. Therefore, this paper introduces an integrated annotation scheme and methodology which opens the opportunity to study verbal (i.e., speech) and non-verbal (i.e., visual cues with a communicative intent) components independently, however, still interconnected over a common timeline. To analyse this interaction between linguistic, paralinguistic, and non-verbal components in multiparty discourse and to help improve natural language generation in embodied conversational agents, a high-quality multimodal corpus, consisting of several annotation layers spanning syntax, POS, dialogue acts, discourse markers, sentiment, emotions, non-verbal behaviour, and gesture units was built and is represented in detail. It is the first of its kind for the Slovenian language. Moreover, detailed case studies show the tendency of metadiscourse to coincide with non-verbal behaviour of non-propositional origin. The case analysis further highlights how the newly created conversational model and the corresponding information-rich consistent corpus can be exploited to deepen the understanding of multiparty discourse.

**Keywords** Corpora and language resources · Speech corpus · Multimodal corpus · Pragmatics · Conversational intelligence

✉ Izidor Mlakar
izidor.mlakar@um.si

[1] Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia

## 1 Introduction

Language in spoken interaction, i.e., speech, is influenced by other phenomena since it does not occur in a vacuum (Couper-Kuhlen, 2018). In fact, non-redundant information is added to the common ground of the conversation by roughly 50% of non-verbal behaviour (Cassell, 2001). The sharing of information in human social interactions is far more complex than solely an exchange of words. It is multi-layered and includes attitudes and affect, utilises bodily resources [depictive manual actions (Kendon, 2017)], as well as the physical environment in which the discourse occurs (Davitti & Pasquandrea, 2017). For effective communication, two main conditions need to be met: (i) the communicator must make their intention to communicate recognizable, and (ii) the propositional and non-propositional [cf. (Hyland, 2005)] intent that they wish the recipient to comprehend must be presented effectively. In line with Hyland's (2005) definition of metadiscourse, the propositional refers to the content, i.e., the conceptual, ideational meaning (e.g., "Cut the rope"), while the non-propositional refers to pragmatic or metadiscoursive content (e.g., "Well, let me see, [cut the rope]"). The latter group corresponds to McNeill's (1985) off-propositional, Kendon's (2017) pragmatic, and Cooperrider's (2017) background gestures (e.g., hand strokes for emphasis while speaking), and the first to McNeill's (1985) propositional and Cooperrider's (2017) foreground gestures (e.g., hand strokes depicting the act of cutting). In this work we investigate the non-auditory cues of non-verbal behaviour, beyond gestures, as visual cues (Riggio & Riggio, 2012). We, however, omit other visual nonverbal cues, from the definition provided by Riggio and Riggio (2012), which include hairstyle, facial hair, use of cosmetics, grooming, and dress and limit the cues to non-verbal behaviour with a communicative intent (Trujillo et al., 2018). Despite the differences, neither group is redundant since they both convey intent (cf. Cooperrider, 2017, p. 191), as the propositional part represents the content and the non-propositional signals to the recipient how to interpret it. In interpersonal discourse, the verbal signals convey a symbolic or semantic interpretation of information through linguistic and paralinguistic properties (e.g., prosody, pitch), while non-verbal signals orchestrate speech (McNeill, 2016, p. 4). Non-verbal signals, such as prosody, visual cues, emotions, or sentiment, are multifunctional and processed on psychological, sociological, and biological levels (e.g., how we perceive non-verbal components) in all time frames (Church & Goldin-Meadow, 2017; Kelly, 2017). These signals represent the basis of cognitive capabilities and understanding (Church & Goldin-Meadow, 2017). Especially visual cues effectively retain the essence of the information (e.g., learning new words is more efficiently if they are accompanied by gestures (Kelly, 2017), help in providing suggestive influences, serve interactive purposes to express mental states, attitude, and social functioning, and give a certain degree of cohesion and clarity to the overall discourse (Allwood, 2013; Arnold, 2012; Keevallik, 2018; Kendon, 2015; Lin, 2017; Mcneill et al., 2015). Visual cues (i.e., facial expressions, gestures, posture, gazing, and head movements) "communicate a critical part of a message" (Cooperrider, 2017, p. 179).

In spoken interaction, visual cues can add non-redundant information to the discourse (Cassell, 2001); and, although not bound by grammar, they can co-align with speech structures and even compensate for the less articulated verbal parts (Brône et al., 2017; Chen et al., 2015; Couper-Kuhlen, 2018; Kelly, 2017). Such is the function of foreground gestures (Cooperrider, 2017), which can, in certain cases, even substitute the verbal content (e.g., nodding your head instead of providing an affirmative answer). Typical representatives are iconic gestures, specifically those expressing spatial relations and symbols, and deferred references. These can convey meaning without any verbally expressed information (Melinger & Levelt, 2005). For example, in western culture, touching the index finger and thumb together and "writing" a wavy line in the air, as if to sign one's name, is recognized as a conventionalized gesture for requesting a receipt in a restaurant. Similarly, deliberately tapping one's wrist while moderating a session, can be understood as deferred reference, to convey a conventional meaning beyond meaning of the watch (Nunberg, 1993, 1995); i.e. to signal the speaker that the speech needs to be wrapped up. On the other hand, however, tapping one's fingers on the table while contemplating something is merely an accompanying gesture, i.e., a background gesture (Cooperrider, 2017). Background gestures are generally generated with minimal effort, i.e., 'half-heartedly' and 'half-mindedly'. Speakers rarely pay significant attention to their fine-grained details or are even unaware of them (Cooperrider, 2017). To quote Cooperrider, "They are in the background of the speaker's awareness, in the background of the listener's awareness and the background of the interaction" (Cooperrider, 2017, p. 7). Despite this, they convey conversational intent (Cooperrider, 2017, p. 191) or meaning which helps the addressee understand the message as the sender intended (e.g., a hesitant and slow nod of the head accompanying the utterance "right" suggest to the recipient that it should not be interpreted as an affirmative answer) (Mlakar et al., 2021). Background gestures, in particular, serve in providing suggestive influences and give a certain degree of clarity to the overall human–human discourse.

Although there seems to be strong evidence to support the multimodal and multi-signal nature of the human–human interaction, spoken language understanding (SLU) has, for decades, first and foremost focused a priori on speech (Vigliocco et al., 2014). Most studies in corpus linguistics primarily focused on language (e.g., a language as a structured system amenable to linguistic analysis), while non-verbal behaviour was mainly disregarded or observed over specific linguistic utterances (e.g., speech or dialog acts). However, as recently observed by many researchers, the study of human discourse must apply the concept of 'multimodality in interaction' (Feyaerts et al., 2017). And this is the main motivation of the research presented in this paper. It is driven towards the fundamental understanding of if and how signals form (i) a linguistic basis (i.e., semantics, syntax, parts of speech (POS), (ii) the paralinguistic domain (prosody, pitch), (iii) the social domain (dialog function, emotion, sentiment) and (iv) the non-verbal components (gestures, facial expressions can be utilised to interpret and convey information in discourse) (Allwood, 2017; Bozkurt et al., 2016; Chaturvedi et al., 2018; Kendon, 2014; Ma et al., 2019; Wang, 2017).

In light of the notion that the co-verbal alignment and synchronization processes are the conductors behind any affective and human-like social interaction, this paper provides novel concepts to research, fuse, and describe how complex linguistic and paralinguistic signals interact with visual cues during spontaneous and multiparty human interactions. It is built on the baselines of McNeill's (2016, p. 21) 'common growth point theory', according to which the non-verbal behaviour and the verbal content (which are synchronised) share a common source which is the initial pulse that triggered both of them, as well as the 'integrated systems hypothesis'. The motivation behind this study is to create complex conversational knowledge to be exploited when delivering a conversational model used for understanding and automatically generating natural conversational responses of embodied conversational agents (Rojc et al., 2017). To facilitate our motivation, we first apply a modular and extendable EVA (embodied virtual agent) Scheme (Mlakar et al., 2019) using the annotation tools ELAN (Wittenburg et al., 2006)[1] and WebAnno(Eckart de Castilho et al., 2016).[2] The conversational phenomena are observed on a comprehensive multimodal corpus that contains real-life, near authentic, multiparty discourse with spontaneous responses. In this way, we quantify the conversational phenomena into verbal and non-verbal cues into action items(Riggio & Riggio, 2012): (i) linguistic annotations, i.e., segmentation (token, utterance, turn), sentence type, sentiment, parts of speech tags (POS), syntax, and discourse markers (DMs); (ii) paralinguistic signals, i.e., communicative intent, emotions, prosodic phrases, and accentuation (primary accent (PA) and secondary accent (NA)); (iii) management and social signals, i.e., person/relation, dialog acts, and (iv) visual cues and non-verbal communicative intent (NCI) (Mlakar et al., 2021).

Through this framework, the 'positive' inferences can be investigated, and information resources, rules, and models automatically generated. In line with the information fusion theory (Snidaro et al., 2015), and non-verbal communication theories (Allwood, 2017; Kendon, 2017; McNeill, 2016), the action items can be analysed in exploratory research and reveal new fusion functions based on consistent knowledge regarding both the intent and complimentary use of verbal and non-verbal items; where the fusion function is defined as a process that integrates data and features from multiple sources (i.e., linguistics, paralinguistics, and kinesics) to produce more consistent, accurate, and useful information (i.e., conversational expression). Moreover, using the analytics framework, the new resources can be consumed in human–machine interaction to deliver human-like co-verbal behaviour with the embodied conversational agent (ECA) EVA (Rojc et al., 2017).

Following Feyaerts et al. (2017), 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 annotation, collection, structuring, and analysis. Instead of artificial scenarios, we utilise a rich data source based on a multiparty TV talk show in the Slovenian language, which

[1] ELAN: https://tla.mpi.nl/tools/tla-tools/elan/, last visited August, 2022.
[2] WebAnno: https://webanno.github.io/webanno/, last visited August, 2022.

represents a good mixture of institutional discourse, semi-institutional discourse, and casual conversation. Overall, the key contributions of the paper are:

- an EVA annotation scheme for observing conversational phenomena (linguistic, paralinguistic, and non-verbal) as discrete action items within their own restrictions. The cross-correlation (and fusion) is implemented over the temporal domain,
- a conversational corpus generated under close to authentic, partially spontaneous, and multiparty settings to outline the multi-layered interplay between conversational signals,
- a comprehensive visualisation and analytics framework designed for the hypothesis-driven research in the domain of multiparty discourse,
- three case studies implementing the EVA Corpus annotation topology and demonstrating the links between verbal and non-verbal signals, thus illustrating the potential of the corpus.

This paper is structured as follows; we begin by outlining related research, further highlighting the background and motivation, in Sect. 2. Section 3 describes data collection and the methodology including the annotation topology for the conversational behaviour generation model. Section 4 describes the analysis of the defined conversational signals. Results and case studies are given in Sect. 5, along with a demonstration of the knowledge utilisation on an ECA. Final thoughts and future directions are highlighted in Sect. 7.

## 2 Related works

One of the main issues in machine-based discourse understanding is the duality and misinterpretation of conversational signals, which results in non-cohesive responses. From distinguishing propositional and non-propositional content to dialogue acts (DAs) and gestures, the presence or absence of conversational intent is crucial (Cooperrider, 2017, pp. 181, 196). As a result, 'Multimodality in interaction' is becoming one of the cornerstones even in corpus linguistics (Feyaerts et al., 2017). The unimodal information, although semantically correct, may simply not be explicit enough for a machine due to its potential ambiguity (e.g., sarcastic utterances as "Well, congratulations!"). Furthermore, due to the complexity of systems in the environment, in some cases, it can even be misleading. The end of the human–human interaction cycle is an active response generated by the user, not the signals and human–human interaction itself (Opel & Rhodes, 2018).

Recent approaches in corpus linguistics have thoroughly overhauled how we understand and process spoken discourse. For instance, interactional linguistics and conversation analysis made visual cues, especially gestures, one of its focal points (Keevallik, 2018; Nevile, 2015). Still, multimodality is analysed as a relatively restricted concept of co-verbal alignment, primarily aimed at foreground non-verbal behaviour. Nevertheless, this makes it particularly well-suited for research into co-verbal alignment within a specific function (Allwood, 2013; Arnold, 2012;

Vandelanotte & Dancygier, 2017). The function of feedback and the non-verbal behaviour accompanying it, for example, was examined by Navarretta and Paggio (2020). Their study focused on the type of head movement and facial expressions that align with a subtype of the feedback dimension in encounters where persons first met. Petukhova and Bunt (2012) analyse the link between DAs and non-verbal behaviour according to the CoGest annotation scheme. The general purpose or function of an utterance, i.e., the function of disagreement, communication management, or information-providing, on the other hand, was explored based on the pragmatic multimodal corpus HuComTech (Hunyadi et al., 2018), which also contains non-verbal information, such as facial expressions, eyebrow movement, or posture. Along the lines of the present case study, Bolly and Boutet (2018) explore the use of verbal non-verbal pragmatic markers (including some discourse markers) in the elderly to determine their communication abilities. Their research shows that planning gestures co-occur with fillers and interjections, whereas gestures serving interactive functions co-occur with parentheticals and connectives (Bolly & Boutet, 2018). Similarly, Graziano and Gullberg (2018) found that when gestures co-occur with disfluencies (which can also include discourse markers), they can be both propositional and non-propositional (cf. Graziano & Gullberg., 2018), however, they are more likely to occur in non-native speakers. More generally, their research shows that disfluencies in speech are mirrored by disfluencies in gestures. In this sense, most studies tend to observe how visual cues are generated along with or close to language (Chui et al., 2018; Hoek et al., 2017). Most of them address non-verbal elements by mapping them to linguistic forms (Lin, 2017), i.e., synching their occurrence with the corresponding lexical form (e.g., making cutting gestures when uttering "cut"). Nevertheless, such approaches cannot uncover the complete interplay between verbal and non-verbal parts of discourse (Adolphs & Carter, 2013, pp. 12, 143).

The proposed research, therefore, adopts a different approach that does not ground non-verbal behaviour. It leans on Birdwhistell's (1952) understanding of body language and body motion, according to which the message transmitted through the body does not necessarily meet the linguist's definition of language. Visual cues are not conventionalized. The non-verbal behaviour generated using them is not linear as language but can overlap (e.g., one gesture melting into another) (Couper-Kuhlen, 2018, p. 23), which is why non-grounded approaches are more suitable. Therefore, the apparent semantic interface between language and gestures seems limited, as not every verbalised item can be represented with visual cues (Couper-Kuhlen, 2018). With this in mind, Peirce's (1935) semiotic perspective (i.e., the 'pragmatics on the page') should complement the linguistic comprehension of discourse (Queiroz & Aguiar, 2015). Studies that explore this theory include Brône and Oben (2015), and Navarretta (2019), where handshapes and the trajectories of gestures are examined regarding how they interlink with specific semiotic classes. A shortcoming of the studies is, however, that they still observe a specific narrow discourse context (Alahverdzhieva et al., 2018; Han et al., 2017) The field of affective computing has been one of the most prominent fields exploiting these baselines to generate targeted multimodal knowledge for content recommendation and opinion mining (Qian et al., 2019). In this sense, datasets such as HUMAINE (Douglas-Cowie et al., 2011),

**Table 1** General discourse characteristics of the implemented video in the EVA Corpus. The data pertain to five speakers

| | Total | Average per speaker | Average per utterance |
|---|---|---|---|
| Utterances | 1999 | 399.8 | |
| Tokens | 10,471 | 2094 | 7.9 |
| DMs (n > 10) | 1801 | 599 | |
| Total number of NCI | 1727 | | |

SEMAINE (McKeown et al., 2012), and AFEW-VA (Kossaifi et al., 2017) have been created. However, with the focus on non-verbal components, the relationships beyond emotional components and the verbal nature of conversational expressions are rarely established.

Inspired by cognitive linguistics, Feyaerts et al. (2017) build on the cognitive linguistic enterprise and equally incorporate all relevant dimensions of usage events, including the trade-off between different semiotic channels. A practical implementation of the methodology is the NOMCO corpus (Paggio & Navarretta, 2017), which captures various verbal and non-verbal signals over a specific discourse context, i.e., first acquaintance conversations. Another less discourse-oriented example would be EmoLite (Wegener et al., 2018), which investigates the correlation between the complex interplay of contextual features while individuals are reading texts. The multimodal corpus DUEL (Hough et al., 2016) similarly explores disfluencies, exclamations, and laughter, albeit in laboratory settings. Due to the challenging nature of informal, especially multiparty discourse, researchers, however, tend to establish an artificial setting (Chen et al., 2006). This may represent a drawback and a limitation in the context of use since artificial settings tend to reveal the studied phenomena and restrict 'interference' of other signals that are not defined in the targeted scenario. Thus, a broader scope of conversational signals, which could represent 'noise', is intentionally left out of the conversational scenario (Bonsignori & Camiciottoli, 2016; Knight, 2011). TV interviews and theatrical plays, as the TV show used in this research, have shown themselves to be a very appropriate resource of conversational expressions that appear to be significantly more suitable for research of broader concepts, such as management functions, attitudes, and emotions (Martin et al., 2009). Generally, 'public' discussions with a completely unrelated goal to the research represent a good mixture of institutional discourse, semi-institutional discourse, and casual conversation.

## 3 Data collection and methodology: the EVA Corpus

### 3.1 Data source

The EVA Corpus used in this research consists of a 57-min episode of casual multiparty interaction. Table 1 provides the general characteristics of the recording. The transcription guidelines for the episode follow the guidelines for the Slovene spoken

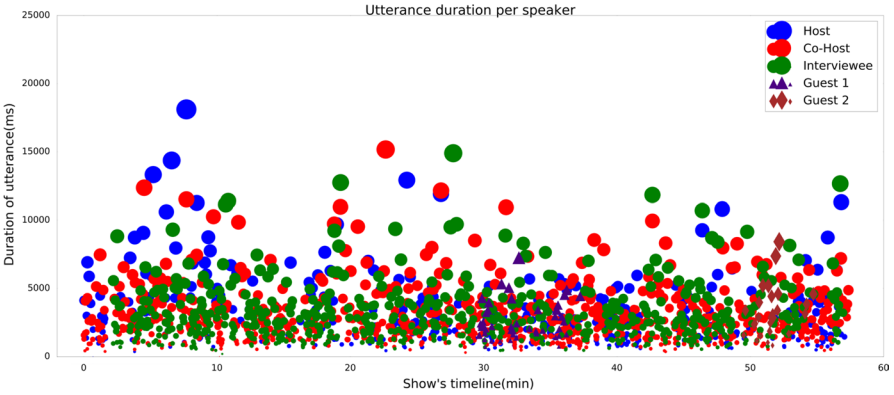


**Fig. 1** Distribution of spoken content and duration of utterances in the data resource for the EVA corpus. The diameter of the dots reflects the duration of the utterances (i.e., the longer the utterance, the wider the dot)

corpus GOS (Verdonik et al., 2013). The speech is segmented into utterances, i.e., "short semantically autonomous units of speech, marked with short pauses at the beginning and end" (Verdonik et al., 2013). Elements such as words and word fragments are considered tokens. The utterances in the EVA Corpus are predominately short, on average containing eight words. The discourse contains 1801 DMs (counting only those with a minimum frequency of ten). The corpus includes numerous non-verbal interactions, where 1727 instances of 'movement' convey meaning, e.g., a gesture performed with an intent.

In total, five different speakers are engaged in each episode. The conversational setting is relaxed and unrestricted. The hosts are skilled speakers who engage in witty, humorous, and sarcastic dialogue with the guest. Therefore, the discourse is close to authentic, and, since all the participants know each other privately, full of emotional responses. Overall, the video contains 1999 utterances, with an average of 399.8 per participant. The average utterance duration is 2.8 s, whereby the longest is 18.1 s, and the shortest is 0.19 s. Overall, there are 10,471 tokens in the episode, and on average, a speaker uttered 2094 of them, with a mean value of 7.9 tokens per utterance. While the total length of the recording is just under one hour, the total duration of all utterances without overlapping is 1 h, 33 min, and 26.3 s, which suggests a substantial amount of overlapping speech. Consequently, the multiparty dialogue is characterized by a vivid and rapid exchange of speaker roles, which makes it ideal to study non-verbal behaviour that accompanies turn-taking.

Together, all participants generate roughly 93 min of spoken content in a 57-min recording. Figure 1 outlines the distribution of spoken content between collocutors (the "Host" is the show's main host who leads the conversations, the "Co-Host" is an actor and supports the Host, the "Interviewee" is the main guest who is also an actor, "Guest 1" is the interviewee's stepdaughter, also an actress, while "Guest 2" is a childhood friend of the interviewee and a physician) and the overall distribution of the utterance duration. The data in Fig. 1 and Table 1 clearly outline that contributors are active and that the discourse involves many short utterances (i.e.,

**Table 2** Frequency of DMs in the observed EVA Corpus compared to previously analysed data of different speech genres

| | GOS-nzos-01 | Turdis-2 | BNSIint | EVA |
|---|---|---|---|---|
| ja [yes/right/ok] | 383 | 313 | 25 | 356 |
| aha [aha] | 23 | 120 | 0 | 50 |
| aja [oh] | 15 | 5 | 0 | 15 |
| mhm [um-hum] | 44 | 155 | 6 | 17 |
| (a) ne [right] | 105 | 186 | 16 | 170 |
| dobro/v redu/okej/prav [ok/right] | 6 | 69 | 14 | 36 |
| no [well] | 44 | 28 | 35 | 76 |
| (po)(g)lej(te) [look] | 3 | 24 | 15 | 22 |
| (a) veš(ste) [y'know] | 48 | 8 | 2 | 37 |
| zdaj [now] | 1 | 64 | 1 | 19 |
| eee/mmm [uh/um] | 235 | 387 | 413 | 315 |
| mislim [I mean] | 21 | 7 | 1 | 23 |
| Total | 944 | 1366 | 528 | 1136 |

It is impossible to provide exact English equivalents for the Slovenian DMs examined in this paper as there are no one-to-one equivalents. The translations provided through this paper are therefore only informative, giving the general meaning of each DM

under 5 s) with a significant amount of overlapping speech. The length of individual utterances ranges from 0.5 to 5 s and lasts 2.8 s on average. Moreover, the distribution of the dots (representing utterances by a specific collocutor at a specific time) also shows that utterances are interchanging rapidly and with a high density among the collocutors.

As outlined in Fig. 1, sequencing, i.e., the conversational organisation of speech acts, extending potentially over several turns, into meaningful parts (Allwood et al., 2007), takes place but is performed highly unorderly, as are other functions related to discourse structuring (e.g., role exchange, topic opening, etc.). This constant overlapping points to a casual and highly irregular progress of the discourse, with lots of overlapping utterances and roles, vivid emotional responses, and facial expressions, with a lot of room for improvisation without a fixed scenario. Thus, we can conclude that the exchange of information in the annotated video is casual, highly dynamic, and involves shorter utterances rather than longer monologues and narratives. The casual nature of the discourse observed is further supported by Table 2, which outlines the comparison in the number of DMs per 10,000 words among our data (the EVA Corpus) and: (1) GOS-nzos-01—based on spontaneous speech interaction in personal contact among friends or family members in everyday encounters from the Slovene reference speech corpus GOS; (2) Turdis-2—based on information-providing telephone interactions between travel agencies or hotel receptions and costumers; (3) BNSIint—based on TV-interviews in the late-night evening news on Slovene national TV (Verdonik et al., 2007). The numbers show that the frequency of DMs in the EVA Corpus is most similar to the frequency of DMs in spontaneous speech interactions in GOS-nzos-01, a corpus based on conversation among friends

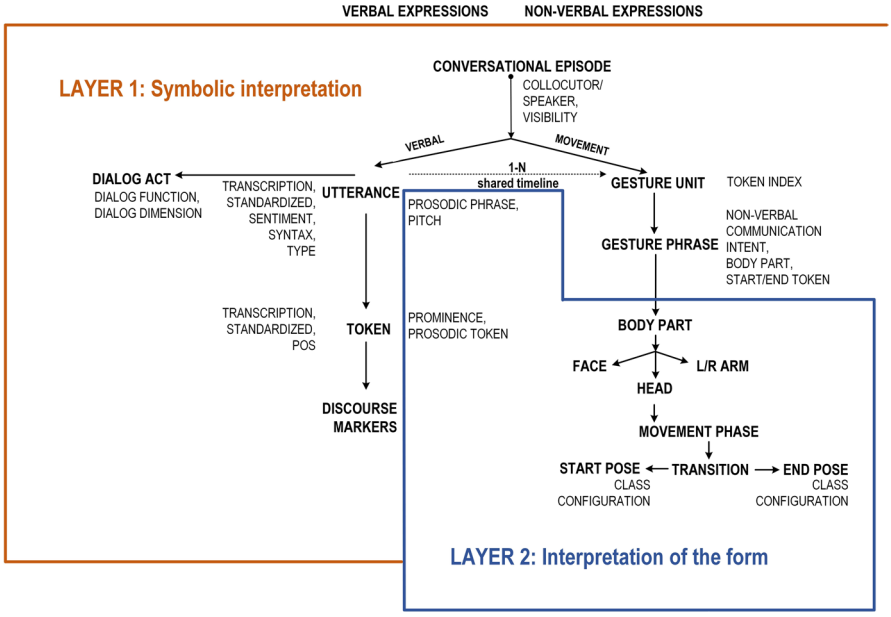


**Fig. 2** The annotation topology in the EVA Corpus: the levels of annotation describing verbal and non-verbal contexts of conversational episodes

or family members in everyday encounters. Also, the comparison among different genres shows least similarities with the BNSIint corpus which is more (Verdonik et al., 2007). This analysis strengthens the notion that the material is authentic.

### 3.2 Methodology and annotation topology

To capture the linguistic, paralinguistic, non-verbal, and social contexts and signals highlighted in Fig. 2 and to observe each conversational expression in greater detail (i.e., explore function *f* and *g*), a multimodal annotation approach was adopted for the conversational analysis. For this purpose, we defined an annotation topology with various levels, as outlined in Fig. 2. The scheme applies a two-layered analysis of the conversational episode.

The main objective of the scheme is (a) to identify inferred meanings of co-verbal expressions as a function of linguistic, paralinguistic, and social signals (e.g., *where and when to gesture*), and (b) to identify the physical nature (e.g., articulation of body language) and use the available imaginary forms (e.g., *how to gesture*). The first layer is defined to capture the 'symbolic' interlinks between the annotated signals and is the focus of this research. It is used to analyse the interpretation of the interplay between various conversational signals, i.e., verbal and non-verbal (e.g., DAs, gestures, syntax, DMs) at a symbolic level. As outlined in Fig. 2, we start with a *conversational episode,* which may be generated with verbal or non-verbal modalities or as a combination of both. Namely, we assume

that the stimuli for the co-verbal behavior are conversational in nature. It may originate as a reflection of attitude/emotion or even be a supportive artifact in the implementation of the communicative function (e.g., feedback, turn-taking, turn accepting, sequencing, etc.). Signals in both modalities are thus 'independent', however, interconnected through the temporal domain. Since the material incorporates mainly informal speech, there are a lot of colloquialisms. Thus, the standardized form is also annotated. The basic verbal element is the utterance. Each utterance is annotated according to features such as the orthographic transcription and sentiment, the standardized form, the sentence type, and syntax, and via paralinguistic signals, e.g., its prosodic phrase and pitch contour. Next, each utterance is segmented into tokens. Each token is described via the linguistic annotations, orthographic transcription, standardized form, and part-of-speech tag (POS), and its paralinguistic signals, e.g., prosodic token and prominence. Finally, tokens are used to define the boundaries of the following linguistic, paralinguistic and social signals: DMs, emotions, and DAs.

The second layer, the interpretation of the form, is concerned with how information is expressed beyond language, through prosody and visual cues, as an abstract concept of a non-verbal conversational expression with a specific communicative intent, i.e., how it is physically realized (e.g., the 'form' of a gesture or 'accentuation' of speech). Its primary goal is to provide a detailed description, the closest possible to the physical reality and the entity that will realize it (e.g., an embodied conversational agent). The basic element in the second layer is the *gesture unit*. Defined by Kendon (2015), a gesture unit denotes all hand and arm movement between rest states; e.g., from the beginning, when a hand/arm starts moving away from a rest position until it returns to the same or new rest position. Since visual cues, as non-verbal aspects of interpersonal communication, involve more than the use of hands, we adapt the definition of a gesture unit to formalize the description of a movement of any visual cue. A gesture unit consists of at least one *gesture phrase* since they can overlap and even coincide (Couper-Kuhlen, 2018, p. 23). It is defined by its purpose (i.e., NCI), the body parts used to articulate the form, and the start/end token, which is used to symbolically ground the gesture to language (if a given movement occurred partially aligned with the verbal counterparts). The internal structure of a gesture is addressed via the propagation and intensity of observed movement in the form of movement phases. Thus, each movement phrase (as a symbolical concept) is described via five consecutive movement phases (Kita et al., 1998):

- **the rest state**, a neutral/stable position from where the gesture begins,
- **the preparation phase**, during which a movement away from the resting position to the start position of the next phase occurs,
- **the stroke**, typically regarded as **obligatory**, most energetic, and with a maximum of information density, directed at manifesting the communicative intent,
- **the holds**, motionless phases potentially occurring before or after the stroke, and
- **the retraction** phase, during which the excited body parts move back to the rest position.

**Table 3** Results of the preliminary inter-coder agreement experiment

| Signal | Kappa score |
|---|---|
| Word Segmentation (semi-automatic) | 0.95 |
| Part-of-Speech (semi-automatic) | 0.81 |
| Pitch (automatic) | – |
| Syntax (semi-automatic) | 0.79 |
| Sentence type | 0.97 |
| Gesture unit | 0.75 |
| Gesture phrase | 0.53 |
| Modality | 0.88 |
| Prosodic phrases | 0.71 |
| Sentiment | 0.67 |
| Dialog function | 0.64 |
| Dialog dimension | 0.71 |
| NCI | 0.48 |
| Emotion label | 0.51 |
| Movement phase | 0.66 |

### 3.3 Annotation procedure and inter-annotator agreement

We applied the EVA Annotation scheme by first segmenting and transcribing the recordings with the transcription tool Transcriber 1.5.1 (Barras et al., 2001). The annotation of the conversational signals was performed in the annotation tool ELAN. In total, five annotators, two with a linguistic background and three with a technical background in machine interaction, were involved in this phase of annotations. Annotations were performed in separate sessions, each session describing a specific signal. This annotation choice offers the advantage of identifying the CI for verbal and non-verbal signals separately, with minimal influence of one on the other, i.e., so that the NCI of a gesture was annotated without the influence of the speech and the DA of the utterance without that of the gesture. The annotation was performed in pairs, i.e., two annotators annotated the same signal. Where there was strong disagreement, the third annotator was activated to help resolve the bias, i.e., after the annotation, consensus was reached by observing and commenting on the values where there 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[3] they could use. The use of a final set of values can be seen as a drawback to the annotation methodology, however, a too extensive set of labels

[3] For further information on the corpus details, please see https://www.clarin.si/repository/xmlui/handle/11356/1311.

where each annotator could add their own labels could lead to similar tags, which in turn would cause difficulties in analysis. An experiment measuring agreement was also performed. It included an introductory annotation session in which the preliminary inconsistencies were resolved. Overall, given the complexity of the task and the fact that the values in Table 3 also cover cases with a possible duality of meaning, the level of agreement is acceptable and comparable to other multimodal corpus annotation tasks (Paggio & Navarretta, 2017).

For the less complex signals, influenced primarily by a single modality (e.g., pitch, gesture unit, gesture phrase, body-part/modality, sentence type), the annotators' agreement measured in terms of Cohen's kappa (Cohen, 1960) was high, namely, between 0.75 and 0.9 on the Kappa score. The signals such as POS, Syntax, Word Segmentation, were annotated (semi)automatically, and the two expert linguist annotators overviewed the process and corrected the tags manually. Pitch was annotated completely automatically, therefore, no agreement was measured. The only exceptions between less complex, unimodal signals were Gesture phrase (0.53) and Prosodic phrases (0.71). The disagreements were expected since, in some cases, it is quite ambiguous to identify where a certain phrase ends and the next one stars. Moreover, in many cases, the retraction phase of a gesture can be recognized as the stroke phase of the next gesture phrase.

## 4 Comprehensive analysis of conversational signals in the EVA Corpus

### 4.1 Sentiment

Each utterance was manually assigned a sentiment, ranging from very negative, negative, neutral, positive, and very positive. The annotators were asked to consider all verbal and non-verbal signals when deciding which sentiment to assign. The results are outlined in Fig. 3.

As outlined in Fig. 3, the conversations were generally positive and rarely reached positive or negative extremes, which were nevertheless also present. This means that the conversation took place in a relaxed setting and mainly concerned topics, which generally did not incite strong attitudes.

### 4.2 Discourse management and structuring

Following the ISO 24617-2 guidelines (Bunt et al., 2012), DAs 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. In the dimension of task functions, we specified the function *Correction* as it does not clarify whether the sender corrects themselves of the

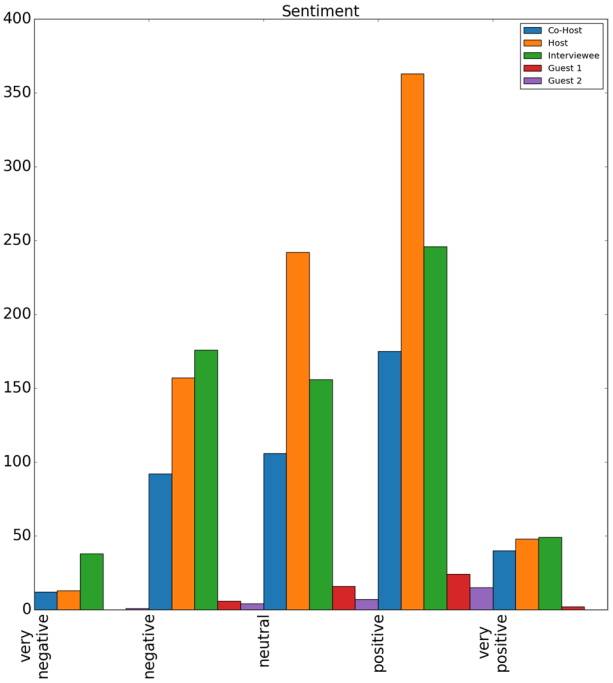


**Fig. 3** Sentiment distribution across speakers at utterance level

interlocutor, an act which in terms of non-verbal behaviour can differ. Therefore, we added the function *CorrectionPartner*, which denotes the action of the sender who is correcting the interlocutor. Among the functions *Inform* or *Agreement*, we also felt the need for argumentative acts, which are differently motivated than general inform acts (e.g., "The play is not about me" vs. "The play is not about me because I am not ugly"), and therefore added the function Argument. For occasions where the sender quotes someone, the function *ReportedSpeech* was added.

Among the directive functions, the Instruct function did not suffice for acts where the sender provides support to the interlocutor or when the sender warns the interlocutor. Therefore, the functions Encouragement and Warning were added. Regarding feedback-specific functions, we merged the *AutoPositive* and *AutoNegative* functions into the *OwnComprehensionFeedback* function. Similarly, we merged the *AlloPositive*, and the *AlloNegative* functions into the *PartnerComprehensionFeedback* function. The dimension of discourse structuring provided the function of opening but lacked the closing action, which we added. As regards the dimension that manages social obligations, we merged the *InitGreeing* and the *ReturnGreeting* functions into *Greeting*. The dimension, however, lacked the function of providing and accepting praise (or flattery), which is why the functions *Praise* and *AcceptPraise* were included.

The results of the annotation are listed in Table 4.

The most common dimension was Task (e.g., information providing, agreement, confirmation, instructing), which accounted for almost half of the DAs. Feedback

**Table 4** Frequency of DAs in the EVA Corpus

| DA dimension and function | n |
|---|---|
| DA Dimensions | 8 |
| Attitude | 160 |
| CommunicationManagement | 139 |
| Feedback | 564 |
| TurnManagement | 188 |
| TimeManagement | 236 |
| DiscourseStructuring | 102 |
| Task | 1446 |
| SocialManagement | 189 |
| Dialog functions | |
| Total functions | 3024 |
| Functions with frequency > 25 | 2501 |
| inform: 585, ownComprehensionFB: 318, stalling: 236, feedbackElicitation: 229, checkQuestion: 129, answer: 119, setQuestion: 109, turnTake: 96, agreement: 93, instruct: 93, emphasis: 78, completion: 77, confirm: 71, disagreement: 60, interactionStructuring: 50, argument: 50, retraction: 42, suggest: 34, flattery: 32 | |

was the second most frequently assigned dimension. This reflects a high level of interactivity and informal character in dialogue. The third most frequent dimension was TimeManagement, reflecting the high level of interaction in the dialogue.

### 4.3 Discourse markers

We draw on previous work on Slovene DMs (Verdonik et al., 2007), 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 analysed and classified (according to their most prominent function in the given context) into the following groups:

**DM-s** (speech formation markers): *eee* 'um' (316), *eem* 'uhm' (15), *mislim* 'I mean' (24), *v bistvu* 'actually' (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* 'um-hum' (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)

**Table 5** The usage of DMs with a minimum frequency ≥ 10 in the EVA Corpus

| | n | % |
|---|---|---|
| Speech formation DMs | 365 | 22.1 |
| Dialogue DMs, contact | 306 | 18.5 |
| Dialogue DMs, feedback | 514 | 31.1 |
| Dialogue DMs, structure | 171 | 10.4 |
| Connective DMs | 295 | 17.9 |
| Total | 1651 | |

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

As DMs are multifunctional, their functions vary from interpersonal management to discourse organization, conversation management, establishing stance towards the conversation content, etc. The same DM can perform multiple functions in different usages, and a single use of a DM can be interpreted in multiple possible functions. For instance, the DM *a veš* 'y'know' can be interpreted as a contact marker used to address the hearer and to establish a feeling of shared knowledge; or as a production marker used to gain time in the production process and fill the gap while maintaining the turn; or as an emphasis to attract the collocutor's attention. Despite the multifunctional role of DMs, they had to be categorized and summarized to perform our analysis. Altogether, 1651 DMs were annotated, accounting for 15.8% of all spoken content (i.e., 10,471 tokens).

Table 5 highlights the distribution of DM use per DM class and the percentage of all tokens that a particular DM class occupies compared to the number of all tokens (10,471) in the EVA corpus.

### 4.4 Emotions

"The expression of attitude is not, as is often claimed, simply a personal matter—the speaker "commenting" on the world—but a truly interpersonal matter, in that the basic reason for advancing an opinion is to elicit a response of solidarity from the addressee." (Martin, 2000, p. 143). Expressing emotions is much more than a personal matter, i.e., expressing one's opinion or perspective. For the annotation of emotions, Plutchik's two-dimensional model (Plutchik, 2001) was applied. The model is a wheel with eight primary emotions at its core and secondary and tertiary emotions towards its outer edges, where the intensity of the emotions lessens. In-between the primary emotions are emotions that are a mix of the primary ones. It describes the relations among emotions that may clarify how complex emotions interact and change over time and in a broader social context, which is why it is suitable for conversational settings. To capture emotional attitudes and represent them as conversational stimuli in the EVA Corpus, the 50 emotional variations and two non-emotional states, e.g., 'rest' and 'undefined', were applied. The annotators have classified emotions within a dedicated track, regardless of the collocutor's dialog

**Table 6** Cross-speaker distribution of annotated emotions in the EVA Corpus

| Emotion | n | Emotion | n |
|---|---|---|---|
| Anticipation: interest | 1239 | Delight | 19 |
| Trust: acceptance | 671 | Trust: admiration | 19 |
| Joy | 349 | Boredom | 15 |
| Serenity | 221 | Sadness | 15 |
| Disapproval | 137 | Contempt | 14 |
| Ecstasy | 92 | Pensiveness | 12 |
| Surprise | 69 | Anger: annoyance | 10 |
| Amazement | 49 | Pride | 10 |
| Anticipation: vigilance | 43 | Alarm | 10 |
| Cynicism | 29 | Fear: apprehension | 10 |
| Disgust | 23 | Optimism | 10 |
| Distraction | 23 | Shame | 10 |
| Curiosity | 22 | | |

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 turns. As a result, the emotion unit for 'anticipation' can span over three utterances and is also maintained when the observed collocutor acts primarily as a listener. Emotional attitude can, therefore, truly reflect an emotion or even situational context, such as regulation in turn-assignment or anticipation in feedback signals. The results are listed in Table 6.

In the EVA corpus, over 3200 instances of emotional expressions were identified. As the results show, in discourse, the 'secondary' emotions, i.e., those related to outer regions of Plutchik's wheel of emotions, such as love, interest, acceptance, disapproval, etc., tend to appear more often. According to Plutchik's theory, primary emotions are "idealized", and their properties must be inferred from evidence but cannot be accurately stated in full. The 'secondary' emotions are less innate, develop over time, and take longer to fade away. These are interpersonal because they are most often experienced in relation to real or imagined others. However, they have no corresponding facial expression that makes them universally recognizable. This is also quite evident by the low score of the inter-coder agreement (Table 3).

### 4.5 Syntax

To describe each utterance syntactically, the annotators were asked to first provide each word in its infinite, singular, or positive form and assign POS tags. The ontology includes lexicon features, such as frequency of occurrence, pronunciation, POS labels, syntax, sense discrimination, phraseology, etc. Moreover, the annotators were also asked to apply a form of dependency grammar to the available utterances. Namely, in informal and contemporary speech and language processing systems, the general formalisms, phrasal constituents, and phrase-structure rules do not apply directly (Jurafsky & Martin, 2018, p. 280). Instead, the syntactic structure of a

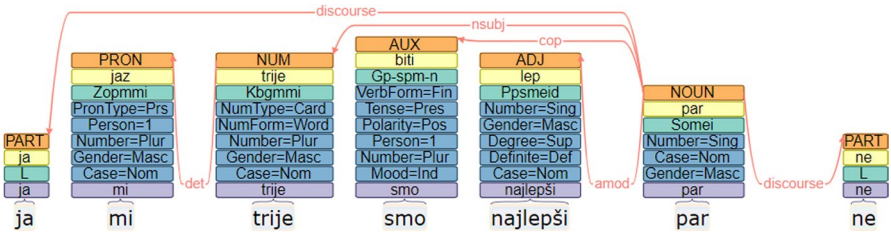


**Fig. 4** Syntax annotation

sentence is described primarily in terms of the words (or lemmas) and an associated set of directed binary grammatical relations that hold among the words (Fig. 4).

The annotation was performed in the WebAnno annotation framework and transferred into the annotation tool ELAN. Each token was extended with an array of lexicon features. The features involved with each syntax token are dynamic and depend on the word type. The static parameters, however, are POS (e.g., noun, verb, pronoun, adjective), infinitive form, and syntax structure. The syntax structure is a realization of a typed dependency structure, as defined in The Universal Dependencies Treebank for Slovenian (Dobrovoljc et al., 2017). Each word in a sentence is described as an array of directed and labelled dependencies as pair of *[j,label],* where j represents the index of the word towards which the dependency is directed, and the label indicates the type of dependencies. The origin word is the word to which the syntax token belongs.

### 4.6 Coding of non-verbal behaviour through non-verbal conversational intent (NCI)

The coding of the symbolic nature of the non-verbal behaviour focusing on gestures and mimics was carried out through the classification of NCI, according to the topology of annotation outlined in Fig. 2. We implemented the following classification:

- **Illustrators (I)** denote non-verbal movements that illustrate the speaker's message (Koutsombogera & Papageorgiou, 2012). As partially foreground behaviour (cf. Cooperrider, 2017, p. 193), they can accompany or reinforce an actual verbal referent in speech. The group is divided into outlines, ideographs, and dimensional illustrators. The outlines (Io) can reproduce a concrete aspect of the accompanying verbal content. The ideographic or metaphoric illustrators (Ii) concretualize abstract concepts with a shape. The spatial or dimensional (Id) illustrators refer to spatial movements with which outlines or dimensions are depicted. Illustrators visualise characteristics to highlight physical properties.
- **Regulators or adaptors (R)** define non-verbal messages with which we model the flow of information exchange (Esposito et al., 2001; Kendon & Birdwhistell, 1972). They are of background nature (cf. Cooperrider, 2017, p. 193) and can be produced in the absence of speech, which is why they do not link with a specific speech structure. The group is additionally divided

into self-adaptors ($R_S$), communication regulators ($R_C$), affect regulators ($R_A$), manipulators ($R_M$), and social function and obligation regulators ($R_O$). As their name suggests, self-adaptors relate to how speakers manage the execution of one's own communication. The communication regulators are used in managing interactions with interlocutors through systems of turn-taking, feedback, and sequencing, e.g., interactive communication management (ICM). Affect-regulators can be either self- or person-addressed. Their function is to further emphasize or express attitude or emotion regarding a topic, object, or person. Manipulators are a sign of the release of emotional tension, but they can also outline mental states, e.g., uncertainty, anxiety, or nervousness. Social function and obligation regulators are used in behaviour expressed during social settings, such as introductions, greetings, or goodbyes.

- **Deictics (D)** refer to real or abstract items that can actually be present in the gesturer's environment (e.g., indicating objects, persons, or places) or an abstract environment (e.g., pointing upwards or pointing backward to indicate the past) (Bühler, 2010; Krauss et al., 2010). They can be part of the foreground if they refer to an actual word and form a semantic interlink. If the semantic link does not exist or is weak, they can be part of the background. The group is divided into pointers ($D_P$), indexes, i.e., referential pointers ($D_R$) (Leavens & Hopkins, 1998) (with which one refers to abstract or real items or persons), and enumerators ($D_E$) (which serve acts such as listing items).
- **Symbols or emblems (S)** usually establish a strong semantic link with verbal referents, which is why they are generally part of the foreground. And the group includes all symbolic gestures and symbolic grammars. Predominately, they are culturally specific. Nevertheless, there are hand emblems that are understood across cultures. Despite their arbitrary link with the speech they refer to, they are recognisable, as they have a direct verbal translation, usually consisting of one or more words. Within this group we position also the deferred references, i.e. metonymic use of gestures to refer to an entity related to the conventional meaning of that expression, but not denoted by it (i.e. poiting to the keys but refereing to the car) (Nunberg, 1995).
- **Batons (B)** are staccato strikes with which we create emphasis. They also serve as "attention grabbers". Short and single batons mark important conversation points, whereas repeated batons can emphasize a critical (Leonard & Cummins, 2011). Beats are their equivalent, however, beats may appear as more random movement, with which rhythm is outlined (Bozkurt et al., 2016; McNeill, 1992). Contrary, batons also set the rhythm and signal importance, yet, they also outline the structure of the verbal counterparts, which is why they can serve as a tag for a set of words that should be processed together.

In terms of the background-foreground distribution of observed NCIs, we can observe that the material contains predominantly non-verbal behaviour functioning in the background. As visible in Table 7, we defined 1685 non-verbal expressions, out of which 1194 belonged to regulators (69.14%) (which are of background nature) and 136 (8.08%) to illustrators and symbols (which are of foreground nature). The rest, 275 (16.33%), belonged to deictic expressions

**Table 7** The distribution of use of non-verbal behaviour in the EVA Corpus

| NCI Class | NCI subclass | N | Total |
|---|---|---|---|
| I (illustrators) | $I_O$ (outlines) | 20 | 99 |
| | $I_I$ (ideographs) | 68 | |
| | $I_D$ (dimensional) | 11 | |
| R (regulators) | $R_A$ (affect) | 105 | 1194 |
| | $R_C$ (communication) | 717 | |
| | $R_M$ (manipulators) | 16 | |
| | $R_O$ (social obligation) | 27 | |
| | $R_S$ (self-adaptors) | 329 | |
| D (deictics) | $D_P$ (pointers) | 40 | 275 |
| | $D_R$ (referential) | 219 | |
| | $D_E$ (enumerators) | 16 | |
| B (batons) | – | – | 80 |
| S (symbols) | – | – | 37 |
| U (undetermined) | – | – | 43 |
| Total | 1727 | | |

(which can be of either nature). The majority of NCI is, therefore, of background nature.

Considering the background-foreground nature in which regulators and batons are regarded as background, illustrators, and symbols as foreground, and deictics as in between, the use of non-verbal behaviour is predominately of background nature. Furthermore, the behaviour seems to be generated as a supporting or coping mechanism that establishes and maintains cohesion, as most of the observed non-verbal behaviour was recognized as background gestures within the contexts of regulation and communication management, deictics, and batons. The foreground classes of illustrators and symbols, with the weakest semantic links between two-word referents and non-verbal behaviour, were observed 136 times in total.

## 5 Results: cross-modal case studies of casual discourse

### 5.1 Case analysis: speech formation DMs and NCIs in own communication management

In this case study, we investigate how speech formation markers interlink with non-verbal behaviour during the multiparty discourse. We focus on the speech formation process as part of own communication management (e.g., word search, rephrasing). As self-adaptors are generally deployed when managing (e.g., improving) our own communication, we hypothesize that speech formation DMs coincide with the background behaviour related to self-adaptors. To this end, we analyse with which NCIs the speech formation DMs coincide. The results are summarized in Table 8.

The analysis shows that the majority of speech formation DMs co-occur with self-adaptors and communication regulators, which confirms our hypothesis.

**Table 8** The distribution of co-occurrence between the most frequent speech formation DMs and NCIs.

| | n | $D_R$ | B | $R_C$ | $R_S$ | Other |
|---|---|---|---|---|---|---|
| eee 'uhm' | 250 | 18 (7.2%) | 5 (2.0%) | 40 (16.0%) | 158 (63.2%) | 29 (11.6%) |
| eem 'uhm' | 12 | 0 | 2 (16.7%) | 5 (41.7%) | 3 (25.0%) | 2 (16.7%) |
| mislim 'I mean' | 18 | 0 | 0 | 5 (27.8%) | 9 (50.0%) | 4 (22.2%) |
| v bistvu 'actually' | 10 | 2 (20.0%) | 3 (30.0%) | 1 (10.0%) | 4 (40.0%) | 0 |

The % value is calculated as the ratio between the total co-occurrence (n column) and the class-specific co-occurrence

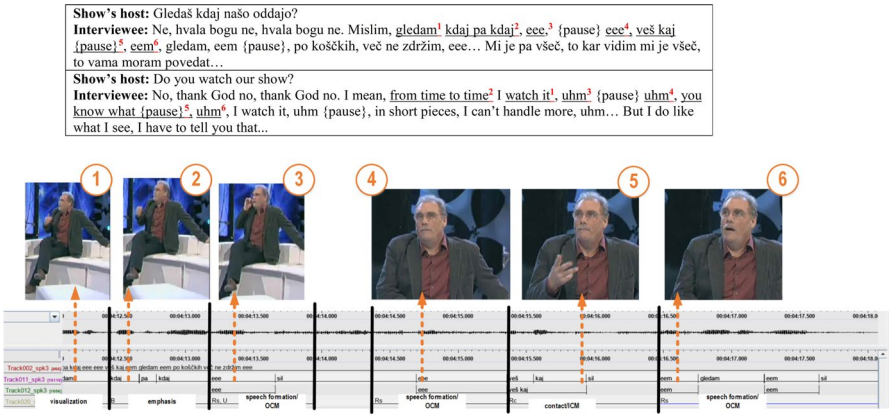


**Fig. 5** Multimodal analysis of the conversational expressions outlined in Example 1; The first track represents the orthographic transcription of spoken content, the second track represents the segmentation on the token level, the third indicates whether the utterance is a DM (single or multiword), and the final track identifies the NCI; the orange arrow connects the pose at the end of stroke phase and the utterance over which the stroke phase was carried out

Typically, the DMs are visualised using gazing combined with facial expressions of thoughtfulness. These can be accompanied by hand gestures, for instance, a raised hand, head-scratching, as highlighted in Example 1 and Fig. 5. Nevertheless, it must be noted that the results refer to the analysis of one episode of the show.

As shown in Table 1, the duality of non-verbal behaviour highlighted in Cooperrider (2017) is also well confirmed in this study. Namely, the DM *v bistvu* 'actually' also frequently co-occurs with referents. Moreover, *eee* 'uhm' can signal that the speaker wishes to keep their turn. In general, the speech formation markers had been well observed to coincide with communication regulators to interrupt the speaker's turn and signal that the participant wishes to take the turn.

***Example 1*** Speech formation DMs and self-adaptors in own communication management

**Table 9** The distribution of co-occurrence between the most frequent speech formation DMs and NCIs

| | n | $R_A$ | $R_C$ | $R_S$ | Other |
|---|---|---|---|---|---|
| aja 'I see' | 12 | 1 (8.3%) | 7 (58.3%) | 2 (16.7%) | 2 (16.7%) |
| mhm 'um-hum' | 13 | 4 (30.8%) | 6 (46.2%) | 0 | 3 (23.1%) |
| aha 'oh' | 22 | 2 (9.1%) | 13 (59.1%) | 1 (4.5%) | 6 (27.3%) |
| ja 'yes' | 240 | 18 (7.5%) | 179 (74.6%) | 18 (7.5%) | 25 (10.4%) |
| fajn 'nice' | 10 | 0 | 10 (100%) | 0 | 0 |

The % value is calculated as the ratio between the total co-occurrence (n column) and the class-specific co-occurrence

In the discourse episode in Example 1, the interviewee expresses his attitude regarding the show in which he is participating. As a form of provocation, he states he does not watch the show. Yet, it seems the provocation was too harsh and inappropriate, which is why he seems to decide to neutralise the wording. He underpins his neutralisation effort by non-verbal behaviour, illustrating the concept of him watching the show from time to time. The non-verbal behaviour that accompanied the search DM "uhm" (Fig. 5, caption (3)) visualises someone pensive while touching the microphone. During the next "uhm", the guest leans back while still seemingly thinking hard (Fig. 5, caption (4)). This changes with the following DM, "you know what", with which he lets the interlocutors know that he picked up the thread and knows how to carry on. In light of this observation, the non-verbal behaviour was assigned a communication regulator, since, as visible in Fig. 5, caption (5), the guest practically knows what he wants to say but cannot completely formulate it yet, which is why he uses non-propositional content or metadiscourse. Similarly, the communication regulator is non-propositional or of background nature. Since the content is not completely formulated, he leans on the couch, looks towards the ceiling, and seems pensive (Fig. 5, caption (6)). Finally, he neutralizes the expression and formulates a mitigating utterance ('in short parts, I can't handle more'). Again, the utterance seems offensive. To formulate the continuation, the DM *eee* 'uhm' is used, which coincides with non-verbal behaviour for searching.

### 5.2 Case analysis: feedback DMs and NCIs in interactive communication management

In this case study, we analyse how feedback DMs such as *ja* 'yes', *aha* 'I see', *mhm* 'mhm', and *fajn*' nice' and non-verbal communication regulators complement each other to support feedback functionality and the concept of active listenership in multiparty discourse. The results are summarized in Table 9.

Feedback DMs primarily tend to co-occur with communication regulators ($R_C$). The NCI class of communication regulators also encompasses the feedback behaviour when the participant is in the role of an active listener. The typical non-verbal behaviour involves head nods and shakes with gazing directed towards the source (e.g., the speaker), smiles, and hand gestures expressing confirmation, agreement, disagreement, etc. A common alternative in interpretation is the affect regulators.

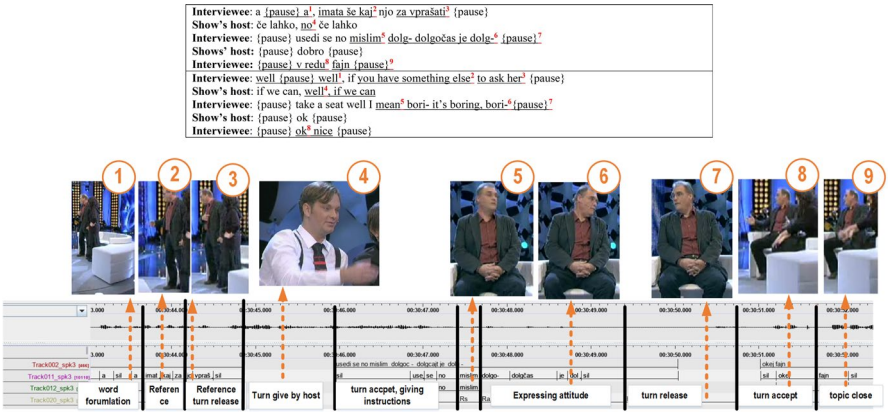


**Fig. 6** Multimodal analysis of the conversational expressions outlined in Example 2

Among the category 'other', in 2 out of 36 cases, the co-occurrence will align with other regulators (e.g., $R_O$ and $R_M$), and, in 16 cases, with referents ($D_R$) and in 18 cases with foreground behaviour (i.e., Dp, I, S). Example 2 further highlights how feedback DMs interlink with non-verbal behaviour to implement turn management.

***Example 2*** Feedback DMs and communication regulators in interactive communication management

In Example 2, the use of the feedback DM "ok" at the end of the sequence is a consequence of the guest's displeasure with the show's host. This is expressed through various referential pointing gestures towards the show's hosts (Fig. 6, caption (2)). The co-host attempts to defuse the situation by asking the guest to sit down again while using a second referential NCI, with which he specifically points towards the guests. He then releases his turn by gazing towards the show's hosts. The guest, however, ignores him and starts expressing his dissatisfaction to the new guest (Fig. 6, captions (5–7). As no comment was provided by the hosts, the guest provides feedback on his own utterance by using the DM *v redu* 'ok' and an open hand gesture extended towards the show's host (Fig. 6, caption (8)). He then further comments on himself by adding *fajn'* nice'. The utterance is accompanied by a clap of his hands, indicating that he wishes to transition to a different topic.

### 5.3 Case analysis: the interplay of DAs and NCIs in turn management

With this case, we analyse how the DAs related to turn management interlink with non-verbal behaviour to deliver multimodal ICM, specifically focused on turn management functions.

***Example 3*** DAs and communication regulators in interactive communication management

| | |
|---|---|
| **Interviewee**: | grmičevje je zlo nerodno objemat |
| **Co-host:** | zakaj? |
| **Interviewee**: | ful pič … ful ful te (A) |
| **Show's host**: | ful ful pič ful pič |
| **Co-host:** | ful me |
| **Interviewee**: | drevo je fajn men |
| **Show's host**: | eee (1) |
| **Co-host:** | kosmulja (2) |
| **Co-host:** | nadaljuj (3) |
| **Show's host**: | ja (4) |
| **Interviewee**: | it's very tricky to hug shrubs |
| **Co-host:** | why? |
| **Interviewee**: | totally pricks … you get totally totally (A) |
| **Show's host**: | totally totally pricks totally pricks |
| **Co-host:** | I get totally |
| **Interviewee**: | trees I like |
| **Show's host:** | uh (1) |
| **Co-host:** | gooseberry (2) |
| **Co-host:** | continue (3) |
| **Show's host**: | yes (4) |

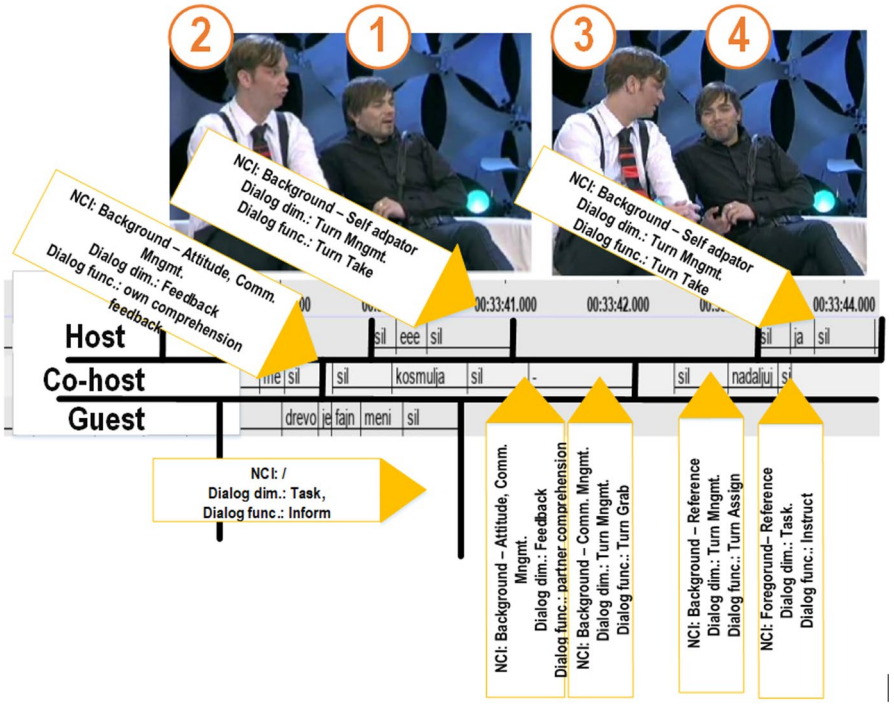


**Fig. 7** Multimodal analysis of the conversational expressions outlined in Example 3

While the show's host is seemingly trying to open a new topic with the use of the DM "uh" (Fig. 7, caption (1)), his co-host interrupts him with "gooseberry", which is a comical continuation of the topic on how shrubbery is not good for hugging acting exercises. Hence, he underpins his comical intent with a facial expression resembling swollen lips (Fig. 7, caption (2)). However, only the guests can see that, and they laugh, while the show's host is perplexed and seemingly does not understand the laughter. He remains quiet but turns to the co-host as if hoping for an explanation or continuation (Fig. 7, caption (4)), yet the co-host gazes back at him and, after a moment of silence, demands that the host carries on with the show (Fig. 7, caption (3)). The example shows that the gaze towards the host was not enough to prompt

a response, i.e., a referential deictic (Fig. 7, caption (3)), which is why he added a slight yet firm nod of the head. The verbalisation "continue" together with the nod and gazing then actually assigned the turn to the show's host, as, after a brief silent moment, the show's host also nods his head, verbally affirms that he accepts the turn with "yes" and, finally, continues with the show. These results are in line with the analysis of feedback DAs and non-verbal behaviour among first encounters by Navarretta and Paggio (2020).

## 6 Conclusions and future directions

This paper presents the first multimodal open-access corpus for the Slovenian language, the EVA Corpus.[4] It aims to better understand how verbal and non-verbal signals co-occur in naturally occurring speech and to help improve natural language generation in ECAs, i.e., to make them multimodally literate. Effective analysis of the non-verbal behaviour that accompanies natural communication requires material that is as authentic and informal as possible. The episode from the entertainment show *As ti tut not padu?* meets these requirements, as the statistics of the material speak in favour of its spontaneity. For instance, the amount of overlapping speech indicates high interactivity among the interlocutors. Moreover, the informal nature is further supported by the foreground–background distinction (Cooperrider, 2017), as the statistics show that the DAs are well-balanced according to the distinction. There are 1,446 propositional, i.e., foreground DAs, and 1,229 non-propositional or metadiscoursive, i.e., background DAs. Moreover, regulators were the most frequent NCIs in the material. They are followed by the NCI group of deictics, which account for roughly 16% of the NCIs. The remaining NCI groups of illustrators, batons, symbols, and undetermined NCI each account for less than ten percent of the observed NCIs. Since regulators are the most frequent NCI, the NCI is predominately of background nature, even if disregarding background deictics or batons. Nevertheless, the limited size of the corpus and the results based on it present a limitation, and the result must therefore be handled accordingly.

The various annotation level signals show a link between the verbal and non-verbal features of conversational expressions as they appear in multiparty informal conversations. The results outlined in this paper provide a generalization of interlinks between verbal expression and non-verbal behaviour obtained by analysing the EVA Corpus. In comparison with studies focusing a specific DA (Navarretta & Paggio, 2020) and similar to research by Hunyadi et al. (2018) or Petukhova and Bunt (2012), we look at a wide range of both propositional and non-propositional DAs and their interplay with non-verbal behaviour classified according to their communicative intent, enabling analysis, as represented by the Example 3. Moreover, we examined the occurrence of discourse markers and their interplay with non-verbal behaviour, similar to Graziano and Gullberg (2018) who focus on disfluencies,

[4] The EVA Corpus is accessible via the repository CLARIN.SI (part of CLARIN ERIH) and is open access under the license CC BY-SA 4.0. Link: http://hdl.handle.net/11356/1311, last visited August 2022.

nevertheless, the present study encompasses discourse markers in several functions (see Sect. 4.3 and Examples 2–3). Through the case study, we have shown a tendency of verbal metadiscourse to mainly coincide with non-verbal behaviour of non-propositional (background) origin (see Example 1). This is in line with results by Bolly & Boutet, 2018) in their observation of planning gestures co-occurring with DMs. Example 1 also shows that speech formation DMs are accompanied by (background) gestures even with native speakers (cf. Graziano & Gullberg, 2018). The case study results are also well-aligned with the common growth point theory (McNeill, 2013), which suggests a common 'intent' of the verbal and non-verbal counterparts.

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 (which deals with the question 'what to say') and articulation (which deals with the question '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 the social/situational context (i.e., the interplay between verbal and non-verbal signals interpreted as the communicative intent). On the representational level, one utilises non-linguistic channels (i.e., gestures, facial expressions), verbal (i.e., speech), and non-verbal prosody (i.e., movement structure) to articulate an idea and present it to the target audience. Thus, the proposed model combines the concepts of functional and descriptive annotation schemes and allows identification of the functional characteristics of verbal behaviour; identification of the intent of linguistic expressions; description of individual configurations, shapes, and poses in high resolution as abstract concepts, or movement controllers in the form of detailed 3D configuration. Moreover, the conclusions in this study corroborate the significance of non-verbal communication emphasized by Birdwhistell (2010) and Allwood, 2017). Nevertheless, as highlighted in Examples 1–3, the distinction between foreground and background gestures which remains blurry (Cooperrider, 2017) can cause a duality in the interpretation.

Still, to determine how much the proposed classification scheme discriminates between subjects and contexts, the results need to be examined through a generalizability analysis (Rubin et al., 1974). Also, as a limitation, it must be noted that a corpus based on a different genre might result in different findings. As outlined in Maricchiolo et al. (2012), almost all non-verbal interpretations are present in different social contexts, and the distribution of their use varies according to the type of the discourse. Especially in border cases, the duality in interpretation is highly context-dependent and cannot be explained in signal isolation. For instance, the DA of providing feedback, a background DA, while actively listening, verbalised as "yes" or "um hum", is usually accompanied by slight head nodding with the purpose of discourse cohesion. On the other hand, the same verbalisations used in the DAs of (dis)agreement might differ only in the frequency and strength or prominence of the execution of the non-verbal behaviour, as emphasized by Cooperrider (2017). A means of disambiguation is possible through a wider context. Thus, in future studies, we aim to analyse if the alignment of verbal structure with the prosody of non-verbal cues (i.e., the cues preceding verbal acts, cues following verbal acts, cues at the beginning or end of verbal acts) may serve as

an aid to uncover the actual purpose of the shared nature. With advances in deep learning and, in particular, image and natural language processing (i.e., autoencoders, adversarial networks, convolutional networks, and recurrent networks), there is a tangible chance to automatize the annotation of signals and complement or fuse the hypothesis-driven research with data-driven science. Still, as a limitation, it must be noted that while computer vision may partially complement manual annotations, human oversight is still needed especially in non-verbal behaviour, since it is linear and not explicitly linked to a set of conventional rules or a specific grammar and can therefore lead to low inter-coder agreement.

Overall, the development of multimodal corpora, multimodal conversational behaviour, and its stimuli are relatively new concepts. As a result, available multimodal corpora are still rare and highly focused. The annotation of data in EVA and similar corpora are generated mostly manually. Since this is a very time-consuming process, tools and methods to at least partially automate the process are highly needed. Therefore, in the near future, we plan to study algorithms, which could at least partially automatize some of the annotation processes, such as automatic gesture segmentation, gesture, and movement tracking.

**Acknowledgements** This paper is partially funded by European Union's Horizon 2020 research and innovation program, project PERSIST (Grant Agreement No. 875406). This paper is partially funded by the Slovenian Research Agency, project HUMANIPA (Research Core Funding No. J2-1737 (B)).



## References

Adolphs, S., & Carter, R. (2013). *Spoken corpus linguistics*. Routledge.

Alahverdzhieva, K., Lascarides, A., & Flickinger, D. (2018). Aligning speech and co-speech gesture in a constraint-based grammar. *Journal of Language Modelling*. https://doi.org/10.15398/jlm.v5i3.167

Allwood, J. (2013). A framework for studying human multimodal communication. In M. Rojc & N. Campbell (Eds.), *Coverbal synchrony in human-machine interaction.* CRC Press.

Allwood, J. (2017). Pragmatics: From language as a system of signs to language use. In E. Weigand (Ed.), *The Routledge handbook of language and dialogue.* Routledge.

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. *Language Resources and Evaluation, 41*(3–4), 273–287. https://doi.org/10.1007/s10579-007-9061-5

Arnold, L. (2012). Dialogic embodied action: Using gesture to organize sequence and participation in instructional interaction. *Research on Language and Social Interaction, 45*(3), 269–296. https://doi.org/10.1080/08351813.2012.699256

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication, 33*(1–2), 5–22. https://doi.org/10.1016/S0167-6393(00)00067-4

Birdwhistell, R. L. (1952). *Introduction to kinesics: An annotation system for analysis of body motion and gesture*. Department of State, Foreign Service Institute.

Birdwhistell, R. L. (2010). Essays on body motion communication. In R. L. Birdwhistell (Ed.), *Kinesics and context.* University of Pennsylvania Press.

Bolly, C. T., & Boutet, D. (2018). The multimodal CorpAGEst corpus: Keeping an eye on pragmatic competence in later life. *Cuadernos De Musica, Artes Visuales y Artes Escenicas, 13*(3), 279–317. https://doi.org/10.3366/cor.2018.0151

Bonsignori, V., & Crawford Camiciottoli, B. (Eds.). (2016). *Multimodality across communicative settings*. Cambridge Scholars Publishing.

Bozkurt, E., Yemez, Y., & Erzin, E. (2016). Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures. *Speech Communication*. https://doi.org/10.1016/j.specom.2016.10.004

Brône, G., & Oben, B. (2015). InSight interaction: A multimodal and multifocal dialogue corpus. *Language Resources and Evaluation, 49*(1), 195–214. https://doi.org/10.1007/s10579-014-9283-2

Brône, G., Oben, B., Jehoul, A., Vranjes, J., & Feyaerts, K. (2017). Eye gaze and viewpoint in multimodal interaction management. *Cognitive Linguistics*. https://doi.org/10.1515/cog-2016-0119

Bühler, K. (2010). The deictic field of language and deictic words. In *Cognitive Linguistics Bibliography (CogBib)*. Berlin, Boston: De Gruyter Mouton. Retrieved from https://www.degruyter.com/database/COGBIB/entry/cogbib.1781/html

Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., & Petukhova, V., et al. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 430–437). Istanbul, Turkey: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/530_Paper.pdf

Cassell, J. (2001). Embodied conversational agents: Representation and intelligence in user interfaces. *AI Magazine, 22*(4), 67.

Chaturvedi, I., Cambria, E., Welsch, R. E., & Herrera, F. (2018). Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*. https://doi.org/10.1016/j.inffus.2017.12.006

Chen, L., Javaid, M., di Eugenio, B., & Žefran, M. (2015). The roles and recognition of Haptic-Ostensive actions in collaborative multimodal human-human dialogues. *Computer Speech and Language, 34*(1), 201–231. https://doi.org/10.1016/j.csl.2015.03.010

Chen, L., Rose, R. T., Qiao, Y., Kimbara, I., Parrill, F., & Welji, H., et al. (2006). VACE multimodal meeting corpus. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 3869 LNCS). https://doi.org/10.1007/11677482_4

Chui, K., Lee, C. Y., Yeh, K., & Chao, P. C. (2018). Semantic processing of self-adaptors, emblems, and iconic gestures: An ERP study. *Journal of Neurolinguistics*. https://doi.org/10.1016/j.jneuroling.2018.04.004

Church, R. B., & Goldin-Meadow, S. (2017). *Chapter 18. So how does gesture function in speaking, communication, and thinking?* John Benjamins Publishing Company.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cooperrider, K. (2017). Foreground gesture, background gesture. *Gesture, 16*(2), 176–202. https://doi.org/10.1075/gest.16.2.02coo

Couper-Kuhlen, E. (2018). Finding a place for body movement in grammar. *Research on Language and Social Interaction, 51*(1), 22–25. https://doi.org/10.1080/08351813.2018.1413888

Davitti, E., & Pasquandrea, S. (2017). Embodied participation: What multimodal analysis can tell us about interpreter-mediated encounters in pedagogical settings. *Journal of Pragmatics*. https://doi.org/10.1016/j.pragma.2016.04.008

Dobrovoljc, K., Erjavec, T., & Krek, S. (2017). The Universal Dependencies Treebank for Slovenian. In *BSNLP 2017 - 6th Workshop on Balto-Slavic Natural Language Processing at the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017*. https://doi.org/10.18653/v1/w17-1406

Douglas-Cowie, E., Cox, C., Martin, J. C., Devillers, L., Cowie, R., Sneddon, I., et al. (2011). The HUMAINE database. *Cognitive Technologies*. https://doi.org/10.1007/978-3-642-15184-2_14

Eckart de Castilho, R., Mùjdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH) at COLING 2016*.

Esposito, A., McCullough, K. E., & Quek, F. (2001). Disfluencies in gesture: Gestural correlates to filled and unfilled speech pauses. In *IEEE International Workshop on Cues in Communication "Cues 2001"* .

Feyaerts, K., Brône, G., & Oben, B. (2017). Multimodality in interaction. In B. Dancygier (Ed.), *The Cambridge handbook of cognitive linguistics.* Cambridge University Press.

Graziano, M., & Gullberg, M. (2018). When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2018.00879

Han, T., Hough, J., & Schlangen, D. (2017). Natural language informs the interpretation of iconic gestures. A computational approach. In *The 8th International Joint Conference on Natural Language Processing. Proceedings of the Conference. Vol. 2: Short Papers*.

Hoek, J., Zufferey, S., Evers-Vermeul, J., & Sanders, T. J. M. (2017). Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of Pragmatics*. https://doi.org/10.1016/j.pragma.2017.10.010

Hough, J., Tian, Y., de Ruiter, L., Betz, S., Kousidis, S., Schlangen, D., & Ginzburg, J. (2016). DUEL: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*.

Hunyadi, L., Váradi, T., Kovács, G., Szekrényes, I., Kiss, H., & Takács, K. (2018). Human-human, human-machine communication: on the HuComTech multimodal corpus. In *Linköping Electronic Conference Proceedings* (Vol. 159, pp. 56–65).

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing (Continuum Discourse, 2).* Continuum.

Jurafsky, D., & Martin, J. H. (2018). Speech and language processing. In *Chapter 13: Dependency parsing, draft chapters in progress* (Vol. 3, pp. 248–273). https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf. Accessed 6 Dec 2022.

Keevallik, L. (2018). What does embodied interaction tell us about grammar? *Research on Language and Social Interaction, 51*(1), 1–21. https://doi.org/10.1080/08351813.2018.1413887

Kelly, S. D. (2017). *Chapter 11. Exploring the boundaries of gesture-speech integration during language comprehension*. John Benjamins Publishing Company.

Kendon, A. (2014). Semiotic diversity in utterance production and the concept of "language." *Philosophical Transactions of the Royal Society b: Biological Sciences*. https://doi.org/10.1098/rstb.2013.0293

Kendon, A. (2015). Gesture: Visible action as utterance. *Gesture: Visible Action as Utterance., 10*, 42–5687. https://doi.org/10.5860/choice.42-5687

Kendon, A. (2017). Pragmatic functions of gestures. *Gesture, 16*(2), 157–175. https://doi.org/10.1075/gest.16.2.01ken

Kendon, A., & Birdwhistell, R. L. (1972). Kinesics and context: Essays on body motion communication. *The American Journal of Psychology, 85*(3), 441. https://doi.org/10.2307/1420845

Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 1371). https://doi.org/10.1007/BFb0052986

Knight, D. (2011). *Multimodality and active listenership: A corpus approach*. Bloomsbury.

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., & Pantic, M. (2017). AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing*. https://doi.org/10.1016/j.imavis.2017.02.001

Koutsombogera, M., & Papageorgiou, H. (2012). Iconic gestures in face-to-face TV interviews. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7206 LNAI). https://doi.org/10.1007/978-3-642-34182-3_24

Krauss, R. M., Chen, Y., & Gottesman, R. F. (2010). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and gesture.* Cambridge University Press.

Leavens, D. A., & Hopkins, W. D. (1998). Intentional communication by chimpanzees: A cross-sectional study of the use of referential gestures. *Developmental Psychology, 34*(5), 813–822. https://doi.org/10.1037/0012-1649.34.5.813

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes, 26*(10), 1457–1471. https://doi.org/10.1080/01690965.2010.500218

Lin, Y. L. (2017). Co-occurrence of speech and gestures: A multimodal corpus linguistic approach to intercultural interaction. *Journal of Pragmatics*. https://doi.org/10.1016/j.pragma.2017.06.014

Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., & Košir, A. (2019). Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Information Fusion*. https://doi.org/10.1016/j.inffus.2018.06.003

Maricchiolo, F., Gnisci, A., & Bonaiuto, M. (2012). Coding hand gestures: A reliable taxonomy and a multi-media support. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7403 LNCS). https://doi.org/10.1007/978-3-642-34584-5_36

Martin, J. R. (2000). Beyond exchange: Appraisal systems in English. In S. Hunston & G. Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse.* Oxford University Press.

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. https://doi.org/10.1007/s00779-007-0167-y

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schröder, M. (2012). The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing, 3*(1), 5–17. https://doi.org/10.1109/T-AFFC.2011.20

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review, 92*(3), 350–371. https://doi.org/10.1037/0033-295X.92.3.350

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

McNeill, D. (2013). Gesture and thought. *Gesture and Thought*. https://doi.org/10.7208/chicago/9780226514642.001.0001

McNeill, D. (2016). *Why we gesture: The surprising role of hand movements in communication*. Cambridge University Press.

Mcneill, D., Levy, E. T., & Duncan, S. D. (2015). Gesture in discourse. In D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), *The Handbook of discourse analysis.* Wiley.

Melinger, A., & Levelt, W. J. M. (2005). Gesture and the communicative intention of the speaker. *Gesture, 4*(2), 119–141. https://doi.org/10.1075/gest.4.2.02mel

Mlakar, I., Rojc, M., Majhenič, S., & Verdonik, D. (2021). Discourse markers in relation to non-verbal behavior: How do speech and body language correlate? *Gesture, 20*(1), 103–134.

Mlakar, I., Verdonik, D., Majhenič, S., & Rojc, M. (2019). *Towards Pragmatic Understanding of Conversational Intent: A Multimodal Annotation Approach to Multiparty Informal Interaction – The EVA Corpus*. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 11816 LNAI). https://doi.org/10.1007/978-3-030-31372-2_2

Navarretta, C. (2019). The automatic annotation of the semiotic type of hand gestures in Obama's humorous speeches. In *LREC 2018 - 11th International Conference on Language Resources and Evaluation*.

Navarretta, C., & Paggio, P. (2020). Dialogue act annotation in a multimodal corpus of first encounter dialogues. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*.

Nevile, M. (2015). The embodied turn in research on language and social interaction. *Research on Language and Social Interaction, 48*(2), 121–151. https://doi.org/10.1080/08351813.2015.1025499

Nunberg, G. (1993). Indexicality and deixis. *Linguistics and Philosophy, 16*(1), 1–43. https://doi.org/10.1007/BF00984721

Nunberg, G. (1995). Transfers of meaning. *Journal of Semantics, 12*(2), 109–132. https://doi.org/10.1093/jos/12.2.109

Opel, D. S., & Rhodes, J. (2018). Beyond student as user: Rhetoric, multimodality, and user-centered design. *Computers and Composition*. https://doi.org/10.1016/j.compcom.2018.05.008

Paggio, P., & Navarretta, C. (2017). The Danish NOMCO corpus: Multimodal interaction in first acquaintance conversations. *Language Resources and Evaluation, 51*(2), 463–494. https://doi.org/10.1007/s10579-016-9371-6

Peirce, C. S. (1935). Collected papers of Charles Sanders Peirce. In C. Hartshorne & P. W. Weiss (Eds.), *Pragmatism and pragmaticism and scientific metaphysics.* Belknap Press.

Petukhova, V., & Bunt, H. (2012). The coding and annotation of multimodal dialogue acts. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist, 89*(4), 344.

Qian, Y., Zhang, Y., Ma, X., Yu, H., & Peng, L. (2019). EARS: Emotion-aware recommender system based on hybrid information fusion. *Information Fusion*. https://doi.org/10.1016/j.inffus.2018.06.004

Queiroz, J., & Aguiar, D. (2015). C. S. Peirce and intersemiotic translation. In P. P. Trifonas (Ed.), *International handbook of semiotics.* Springer.

Riggio, R. E., & Riggio, H. R. (2012). Face and body in motion: Nonverbal communication. In T. F. Cash (Ed.), *Encyclopedia of body image and human appearance.* (Vol. 1). London: Elsevier.

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*. https://doi.org/10.1016/j.engappai.2016.10.006

Rubin, D. B., Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1974). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. *Journal of the American Statistical Association, 69*(348), 1050. https://doi.org/10.2307/2286194

Snidaro, L., García, J., & Llinas, J. (2015). Context-based Information Fusion: A survey and discussion. *Information Fusion*. https://doi.org/10.1016/j.inffus.2015.01.002

Trujillo, J. P., Simanova, I., Bekkering, H., & Özyürek, A. (2018). Communicative intent modulates production and comprehension of actions and gestures: A Kinect study. *Cognition*. https://doi.org/10.1016/j.cognition.2018.04.003

Vandelanotte, L., & Dancygier, B. (2017). Multimodal artefacts and the texture of viewpoint. *Journal of Pragmatics*. https://doi.org/10.1016/j.pragma.2017.10.011

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. https://doi.org/10.1007/s10579-013-9216-5

Verdonik, D., Rojc, M., & Stabej, M. (2007). Annotating discourse markers in spontaneous speech corpora on an example for the Slovenian language. *Language Resources and Evaluation, 41*(2), 147–180. https://doi.org/10.1007/s10579-007-9035-7

Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon: Implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2013.0292

Wang, S. P. (2017). Multimodal research on tonal variations for pragmatic purposes in Mandarin. *Journal of Pragmatics*. https://doi.org/10.1016/j.pragma.2017.03.012

Wegener, R., Kohlschein, C., Jeschke, S., & Neumann, S. (2018). EmoLiTe - A database for emotion detection during literary text reading. In *2017 7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2017* (Vol. 2018-January). https://doi.org/10.1109/ACIIW.2017.8272587

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*.