

## Chapter

# Can Turn-Taking Highlight the Nature of Non-Verbal Behavior: A Case Study

*Izidor Mlakar, Matej Rojc, Darinka Verdonik  
and Simona Majhenič*

## Abstract

The present research explores non-verbal behavior that accompanies the management of turns in naturally occurring conversations. To analyze turn management, we implemented the ISO 24617-2 multidimensional dialog act annotation scheme. The classification of the communicative intent of non-verbal behavior was performed with the annotation scheme for spontaneous authentic communication called the EVA annotation scheme. Both dialog acts and non-verbal communicative intent were observed according to their underlying nature and information exchange channel. Both concepts were divided into foreground and background expressions. We hypothesize that turn management dialog acts, being a background expression, co-occur with communication regulators, a class of non-verbal communicative intent, which are also of background nature. Our case analysis confirms this hypothesis. Furthermore, it reveals that another group of non-verbal communicative intent, the deictics, also often accompany turn management dialog acts. As deictics can be both foreground and background expressions, the premise that background non-verbal communicative intent is interlinked with background dialog acts is upheld. And when deictics were perceived as part of the foreground they co-occurred with foreground dialog acts. Therefore, dialog acts and non-verbal communicative intent share the same underlying nature, which implies a duality of the two concepts.

**Keywords:** non-verbal behavior, non-verbal communicative intent, multimodal analysis, background expressions, regulators, deictics, turn-taking, dialog acts, ISO 24617-2

## 1. Introduction

Turn-taking is an indispensable part of spontaneous and authentic human communication. Despite its significance, it is not always as obvious and straightforward as one might want it to be. Rather, it is sometimes conveyed by elusive and subtle cues. These cues can be of verbal or non-verbal nature, but, in successful communication, all of them can be picked up by the human observer. To facilitate effective natural communication between machines and humans, significant effort must be put towards understanding and recognizing the inter-dynamics and intent of non-verbal communication, of which turn-taking is also a part.

The theory of dialog acts offers one possible way to gain insight into the functionality of verbal and non-verbal expressions of communication. Dialog act (hereinafter DA) theory has its origins in speech act theory [1, 2]. But despite its name, DA theory is not merely a theoretical concept. As Bunt [3] emphasizes, its goal is to provide a computational model of language in actual use. According to Searle [2], a DA represents the meaning of an utterance at the level of illocutionary force, and hence, it constitutes the basic unit of linguistic communication.

There are numerous DA annotation schemes, some of which are more purpose-specific, such as the Verbmobil scheme, which is based on business appointment-scheduling dialogs [4], the TRAINS scheme, annotating dialogs about train freight management [5], or the Coconut annotation scheme, with dialogs about buying dining or living room furniture [6], while the ISO 24617-2, the DIT++, the DAMSL and the Switchboard annotation schemes, for example, cover various topics and apply to a wider range of material. The Switchboard scheme was created for a corpus of various authentic, spontaneous telephone calls in the United States and defined 42 types of DAs [7]. The DAMSL scheme, moreover, filled the need for applying multiple tags to a single segment [8] and was the first multidimensional scheme [3]. The concept of dimensions is best described by the ISO 24617-2 annotation scheme, whereby it is defined as a “class of DAs with the same type of semantic content” ([9]: 2). In comparison to multidimensional schemes, one-dimensional schemes use several tags, which are, however, mutually exclusive. Multidimensional schemes are, therefore, more appropriate for the annotation of naturally occurring dialogs. Another example of a multidimensional scheme is the DIT++ annotation scheme, which is partly based on the DAMSL scheme. It distinguishes between general-purpose and dimension-specific functions, which together form a set of ten dimensions – the Task/Activity dimension, the Auto-Feedback, the Allo-Feedback, the Turn Management, the Time Management, the Contact Management, the Own Communication Management, the Partner Communication Management, the Discourse Structuring Management, and the Social Obligations Management dimensions [3]. Furthermore, the DIT++ is not limited to verbal communication only; it also considers non-verbal communication, such as head gestures and prosody. The ISO 24617-2 annotation scheme is partially based on the DIT++ taxonomy. As Bunt [10] elaborates, it was created as a consolidation of selected taxonomies with the aim of avoiding confusion among the several existing annotation schemes and their inconsistent terminology [9]. Moreover, in addition to its multidimensionality, the ISO scheme strives to be a domain-independent scheme. Regarding dimensions, it contains functionally the same dimensions as the DIT++ with the exemption of the Contact Management dimension, which is not included in the ISO 24617-2. Among these nine dimensions, the scheme specifies 57 different functions. Six of these functions pertain to the dimension of Turn Management, namely, the functions of accepting, taking, grabbing, assigning, releasing, and keeping a turn. The functions are relatively self-explanatory as long as we remember that the function is always carried out by the sender, i.e. the “dialogue participant who produces a dialog act” ([9]: 4). The functions of turn management are all dimension-specific, which means that they cannot be assigned to any other dimension. The scheme also acknowledges the need for subtle characteristics of utterances such as conditionality, modality, (un)certainly, stance, and sentiment, which Petukhova and Bunt [11] raised in their analysis of existing annotation schemes. As a solution, the ISO 24617-2 proposes function qualifiers that can be applied to a DA function. Following its predecessor, the DIT++, the ISO 24617-2 also considers non-verbal behavior in terms of DA annotation. After all, in its definition of DAs, the ISO 24617-2 does not discriminate between verbal and non-verbal behavior, since it defines DAs as “a semantic unit of communicative behaviour”.

Hence, the ISO 24617-2 is well-suited for the annotation of multimodal material and was implemented in research of non-verbal behavior. Yoshino et al. [12] utilized the scheme to annotate information navigation and attentive listening dialogs to improve natural conversation modeling for caretakers that communicate with the elderly. Navaretta and Paggio [13] explore non-verbal behavior occurring when providing feedback among persons who just met, i. e. in highly spontaneous settings. The one-hour recordings, annotated with the tool Anvil, specifically analyze what kind of head movement or facial expressions accompany a certain subtype of the feedback dimension. Their classification of non-verbal behavior is based on the MUMIN scheme. Petukhova and Bunt [14] utilize almost an hour-long recording from the corpus AMI, which consists of project meetings. They analyze DAs according to the DIT++ and the ISO 24617-2 schemes together with co-occurring non-verbal behavior, which is classified according to the CoGest scheme. In their previous work, Petukhova and Bunt [15] annotate recordings from the AMI corpus according to the DIT scheme. Both the annotation of DAs and the annotation of non-verbal behavior is carried out with the DIT scheme, since, as they emphasize, non-verbal behavior helps us understand the true function of a DA. The pragmatical annotation of the multimodal corpus HuComTech [16], however, is not based on the ISO scheme, yet its main annotation units are very similar. They are referred to as communicative acts, which denote the function or purpose of an utterance (e.g., agreement, turn management, information). The annotation of non-verbal behavior, including facial expressions, eyebrow movement, head movement, touch motions, posture, or emotions, was performed manually and partially automatically with the tool Qannot.

Although there seems to be strong evidence to support the multimodal and multi-signal nature of the human-human interaction, for decades, spoken language understanding has first and foremost focused on speech a priori [17]. The classification of non-verbal behavior by Mlakar et al. [18], draws upon McNeill's [19] common growth point theory, according to which speech and gestures both stem from a common growth point of a concept and mutually influence one another, Pierce's [20] semiotics, that provides analysis of non-linguistic signs and symbols as the meaning of non-verbal behavior, Ekman and Friesen's [21] categories and coding of non-verbal behavior, and Birdwhistell's [22] insights into the importance of kinesics. Moreover, the classification by [18] utilizes the communication management theory [23, 24] and, therefore, also encompasses discourse functions to some extent. Mlakar et al. [18] refer to 'gestures' as behavior generated by moving body parts (i.e., head, hands/arms, face, and posture) performing a communicative purpose, i.e., containing a discourse function, as a non-verbal communication intent (hereinafter NCI). These non-verbal expressions represent the basis of cognitive capabilities and understanding [25]. Namely, although not bound by grammar, non-verbal expressions co-align with language structures and compensate for the less articulated verbal expression models, thus providing a certain degree of clarity of discourse [26]. The non-verbal behavior retains the semantics and at the same time helps in providing suggestive influences and serves for interactive purposes, even such as content expression of one's mental state, attitude, and social functions. The classification proposed by Mlakar et al. [18] positions the role/intent of non-verbal concepts into five main NCI classes of regulators or adapters, deictics or pointers, illustrators, symbols or emblems, and batons.

Cooperrider's [27] classification of gestures, on the other hand, concerns itself with the question of whether the gesture "communicates a critical part of a message" ([27]: 179) or not. He divides gestures into foreground and background gestures. Foreground gestures are those gestures of which we are aware when we perform them, such as a thumb up, whereas background gestures occur

unconsciously, automatically, such as nodding during a telephone call. Therefore, foreground gestures are also in the foreground of the interaction. Among their characteristics, he lists co-occurrence with demonstratives, absence of speech, and a significant effort in their production, i.e., gestures that are bigger and more precise. Contrary to them are background gestures. They are both smaller in size and precision and occur while the sender is speaking. Despite this clear division, Cooperrider [27] emphasizes that the line between foreground and background gestures is anything but straightforward, as some gestures can break the foreground-background barrier. He demonstrates this with pointing gestures, which are generally in the foreground, but when pointing to oneself, they occur in the background. Furthermore, even symbolic gestures can take the background if performed automatically and if they are void of their communicative message. On the other hand, beats occur only as background gestures. One can, therefore, roughly consider illustrators, symbols, and partially deictics as NCI occurring in the foreground, while regulators, beats, and partially deictics can be considered as NCI occurring in the background, while still bearing in mind that the dividing line can always be crossed.

Hence, Cooperrider [27] differentiates between gestures with a semantic or propositional content, i.e., a message that provides some kind of information, and those that are void of it. The same distinction can be made for DAs. There are DAs that primarily convey information that is indispensable for communication, such as the task dimension, and those DAs that primarily do not contain propositional content (hence, they contain metadiscursive content) yet are vital for successful natural communication, such as the turn and management dimensions. Nevertheless, we must apply the same caveat as the one in the background-foreground distinction for gestures, as some DAs can occur either in the foreground or the background. For example, the dimension of managing social obligations can generally be considered part of the foreground, such as the concept of greeting someone upon the first encounter. Still, if a social convention is performed routinely, unconsciously, and is deprived of its semantic content, such as thanking someone for the floor, such a DA can be considered as occurring in the background. The nine DA dimensions can, therefore, roughly be divided into those occurring in the foreground, such as the task and the social obligation management dimension, and those occurring in the background, such as the feedback dimensions, the time and the turn management dimensions, the discourse structuring dimension, and the own- and the partner communication management dimensions.

For successful communication, the message must be as clear as possible. An utterance with a mismatching underlying nature is potentially confusing. For example, to take a turn, which is a typical background DA, one sometimes begins one's utterance with "look". The NCI accompanying "look" is usually a subtle hand gesture (e.g., a referential deictic), completely void of meaning and therefore a background gesture. Whereas when one uses "look" in the propositional sense, one uses a pointing gesture; both the DA and the NCI are, in this case, of foreground nature. To use a pointing (foreground) gesture with the mentioned turn-taking (background) DA in the "look" example would therefore be confusing, steering the collocutor to search for an object in sight, which does not exist. Therefore, to ensure cohesion and for the communication to be more effective, it seems plausible that a non-propositional episodes should require a background DA as well as a background NCI.

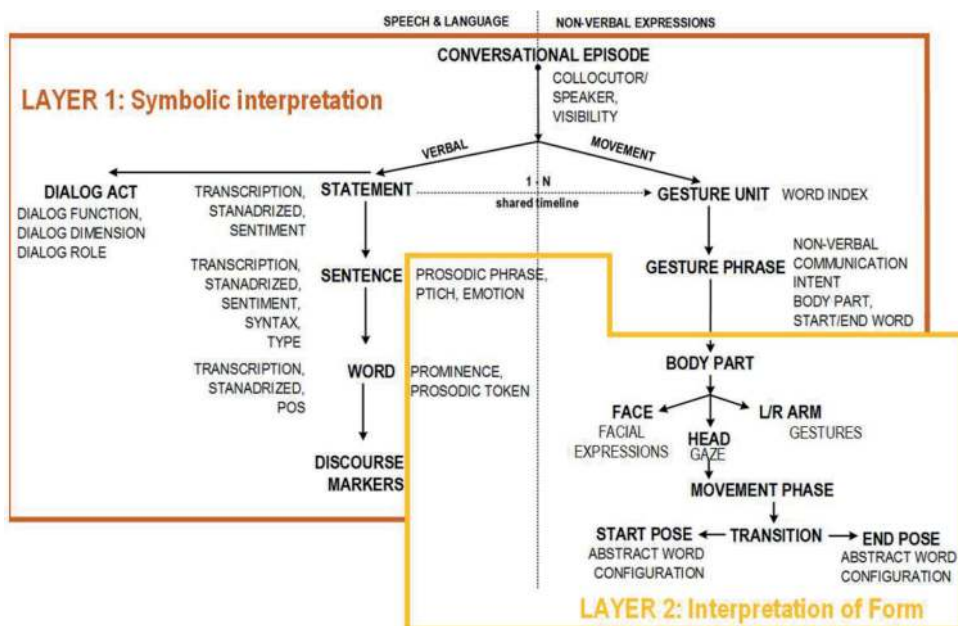
In light of this foreground-background link between DAs and the NCI of gestures, we set out to explore whether the theory of DAs can help predict the nature of the NCI of the corresponding unit. Specifically, we hypothesize that turn management DAs correlate with background gestures. Therefore, we propose the following hypothesis:

Turn management DAs, as background expressions, will tend to co-occur with NCI of background nature. In particular, turn management DAs will co-occur primarily with communication regulators.

## 2. Data and methodology

In order to perform research into authentic non-verbal behavior during turn-taking, we utilized a 57-minute long video recording from the Corpus EVA [18]. Our annotation scheme, adapted from Mlakar et al. [28], outlined in **Figure 1**, was applied in the dataset to perform conversational analysis. For this research, dialog acts were added as a linguistic branch.

The main objective of the scheme is to identify inferred meanings of co-verbal expressions as a function of linguistic, paralinguistic, and social signals (e.g., where and when to gesture) on a symbolic level, and to identify the physical nature (e.g., articulation of body language) and use of the available “imaginary forms” (e.g., how to gesture, how to vocalize), i.e., the level of the interpretation of non-verbal forms. The first layer, in **Figure 1**, the symbolic interpretation, is the focus of this research. It is used to analyze the interpretation of the interplay between various conversational signals, that is, verbal and non-verbal (i.e., DAs, gestures, syntax, discourse markers) at a symbolic level. The second layer, the interpretation of form, is concerned with how information is expressed beyond language, through prosody and embodied expressions, as an abstract concept of a non-verbal conversational expression with a specific communicative intent, i. e. how it is physically realized. For example, 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). As already mentioned, in this chapter, however, we focus on the first layer. The layer which aims to find patterns and tendencies in how people communicate through joint use of language, prosody, gaze, gesture, facial expressions, and other articulation of the



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

body, specifically focused on turn-taking and analysis of DAs and NCIs overlapping in conversational expressions (episodes).

## **2.1 The EVA Corpus**

The EVA Corpus consists of 228 minutes in total, and includes four video and audio recordings, each 57 minutes long, with corresponding orthographic transcriptions. The discourse in all four recordings is a part of the entertaining evening TV talk show *'A si ti tut not padu'*, broadcast by the Slovene commercial TV in 2010. In this research, we utilize one of the videos.

In total, five different collocutors are engaged in each episode in multiparty discourse. The conversational setting is relaxed and unrestricted. The hosts are skilled interlocutors who engage in witty, humorous, and sarcastic dialog with the guest. Therefore, the discourse is highly spontaneous, authentic, and, in this case, since all the participants know each other privately, also relaxed and full of emotional responses. Overall, the video contains 1,516 utterances, with an average of 303 utterances per speaker. The episode contains 1,999 sentences, with an average of 399.8 per participant. The average sentence duration is 2.8 seconds, whereby the longest is 18.1 seconds, and the shortest is 0.19 seconds. Overall, there are 10,471 words in the episode, and on average, a speaker uttered 2,094 of them, with a mean value of 7.9 words per sentence. While the total length of the recording is just under one hour, the total duration of all utterances without overlapping is 1 hour 33 minutes and 26.3 seconds, which suggests a substantial amount of overlapping speech. Consequently, the dialog is characterized by a vivid and rapid exchange of speaker roles, which makes it ideal for the study of non-verbal behavior that accompanies turn-taking.

## **2.2 DA annotation**

The entertainment show was segmented and transcribed with the transcription tool Transcriber 1.5.1 and annotated in the annotation tool ELAN. The annotation of DAs was performed with the web-based annotation tool Webanno. For the classification of DAs, we applied the ISO 24617-2 scheme, however, it was partially consolidated in accordance with our research's aim. In the dimension of information-providing functions, we specified the function Correction as it does not clarify whether the sender corrects themselves or the interlocutor. 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 filled the need for argumentative acts and 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. With regard to 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 InitGreeting 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 annotation of sentiment included

the qualifiers Disappointment, Disgust, Emphasis, Hurt, Negative, Positive, Satisfaction, and Surprise.

In line with Cooperrider's [27] foreground-background distinction, we divided DAs according to whether they are conveying a vital part of the message without which the encounter would be void of propositional content or not. Since task-oriented DAs include the functions of information-seeking and -providing, as well as commissive and directive functions, they are part of the foreground. Similarly, the social obligations management DAs perform functions such as greetings, introductions, apologies, thanking, and valedictions. They contain propositional content and can, therefore, be considered part of the foreground. On the other hand, the feedback DAs, turn management DAs, time management DAs, discourse structuring DAs, and own- and partner communication management DAs perform background functions as their main purpose is not to convey information but to steer the dialog or to provide active listenership. For example, when correcting oneself after misspeaking, the act of correction is not in the foreground; it is the underlying information-related DA. Similarly, when helping the interlocutor to find the correct ending to a word, the act of completion is in the background, while the interlocutor's primary utterance that is being completed by the partner is in the foreground. As emphasized in the Introduction, some functions can cross this distinction. Let us consider an example with the function of completion. When people try to demonstrate their connection by finishing each other's sentences, the partner's act of completing the interlocutor's primary utterance is in the foreground, since both interlocutor's purpose of communication was to demonstrate their connection by completing each other's sentences. Nevertheless, for the majority of cases, the proposed distinction of DAs can be applied as proposed.

In terms of the background-foreground distribution of observed DA episodes, we can conclude that the material is well balanced. It consists of 1,897 instances where the primary role of the DA was recognized as of foreground nature, and 2,020 instances where the primary role of the DA was of background nature.

### 2.3 NCI annotation

The annotation of non-verbal expressions focusing on gestures, mimics, was carried out in Elan. The annotation of each phenomenon highlighted in **Figure 1** (e.g., gesture unit, phrase, NCI) was conducted individually, but by two or three annotators at a time. In terms of annotation disagreement, diverging values were elaborated and argued until consensus was reached. Moreover, before the annotation process began, all annotators were familiarized with the nature of the signal to be annotated and notified with the possible values from which they could choose.

In terms of NCI annotation, we used the following classification:

- **Illustrators (I)** define body movement (embodiment) that illustrates what a speaker is saying. Regarded as foreground behavior, they accompany or reinforce verbal cues and are accompanied by an actual word referent in the speech. Illustrators are further classified into outlines, ideographs, and dimensional illustrators. The outlines ( $I_O$ ) subclass encompasses embodiments that reproduce a concrete aspect of the accompanying verbal content (explicit referents in speech). The ideographic/metaphoric illustrators ( $I_i$ ) subclass refers to a concretization of the abstract through a specific shape. The spatial/dimensional ( $I_d$ ) subclass refers to the spatial movements outlining or depicting dimensional relations. They are used to 'paint' characteristics of entities and actions to further highlight their physical properties.

- **Regulators/adaptors (R)** define embodiments that are primarily used to model the flow of information exchange. Adaptors are regarded as part of background behavior and can be produced even without speech. They exist without a specific speech reference and do not link with a specific speech structure. The regulators are further classified into self-adaptors ( $R_S$ ), the communication regulators ( $R_C$ ) subclass, the affect-regulators ( $R_A$ ) subclass, the manipulators ( $R_M$ ) subclass, and the social function and obligation regulators ( $R_O$ ) subclass. Self-adaptors relate to how a speaker continuously manages the planning and the execution of the speaker's own communication. The communication-regulators refer to managing interactions with other interlocutors through systems of turn-taking, feedback, and sequencing, e.g., interactive communication management (ICM). The affect-regulators are either self- or person-addressed and are used to further emphasize or express attitude or emotion regarding a topic, object, or person. Manipulators convey relief or release of emotional tension or outline states of the body or mind, such as anxiety, uncertainty, or nervousness. Finally, social function and obligation regulation primarily deals with embodied behavior used in social settings, such as greetings, goodbyes, introductions.
- **Deictics (D)** include entities that can actually be present in the real environment of the gesturer (e.g., indicating objects, persons, or places) or are ideally present in the discourse content or abstract (e.g., pointing upwards or pointing backward to indicate the past). If deictic expressions are actual word referents with a semantic interlink, they are regarded as part of the foreground. If the semantic link does not exist or is weak, deictic expressions will also be recognized as part of the background. We further distinguish between pointers ( $D_P$ ), indexes/referential pointers ( $D_R$ ), and enumerators ( $D_E$ ).
- **Symbols/emblems (S)** tend to establish a strong semantic link with verbal counterparts. They are regarded as foreground and include all symbolic gestures and symbolic grammars. Their specific meaning is often cultural-specific, as the same emblem can have different meanings in different cultures. Nevertheless, there are cross-cultural hand emblems, which are easily recognizable because, despite their arbitrary link with the speech they refer to, they have a direct verbal translation, which usually consists of one or two words or a whole sentence (often a traditional expression shared in a specific culture).
- **Batons (B)** are those staccato strikes that create emphasis and grab attention, such as a short and single baton that marks an important point in a conversation. Whereas repeated batons can “hammer” a critical concept. Batons are equivalent to beats, however, beats may appear as a more random movement (e.g., outlining rhythm). Batons, on the other hand, may also set the rhythm and signal importance but, more importantly, they also outline the structure of verbal counterparts, e.g., tag a set of words that should be processed together (e.g., to produce a summary of the meaning of an utterance).

In terms of the background-foreground distribution of observed NCIs, we can observe that the material contains predominantly non-verbal behavior “functioning” in the background. Overall, we have observed roughly 1,684 non-verbal expressions, out of which 1,274 belonged to regulators (75.65 percent) and 136 (8.08 percent) to illustrators and symbols. The rest, 275 (16.33 percent), belonged to deictic expressions. The majority of NCI is, therefore, of background nature.



A rough classification of NCIs and DAs according to their underlying nature, which can be of background and/or foreground nature is represented in **Table 1**. It must be emphasized that this classification is purely provisional, as the foreground-background barrier is vague and can, depending on the wider context, be crossed by both NCIs and DAs.

## 2.4 Annotation agreement

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. The annotation was performed in pairs, i.e., two or three annotators annotated the same signal. After the annotation, consensus was reached by observing and commenting on the values where 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 they could use. 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 2** also cover cases with a possible duality of meaning, the level of agreement is acceptable and comparable to other multimodal corpus annotation tasks [29].

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 [30] was high, namely, between 0.75 and 0.9 on the Kappa score. The signals such as, Part-of-Speech, Syntax, Word Segmentation, were annotated (semi)automatically and the two expert annotators (linguists) overviewed the process and corrected the tags manually. The agreement was measured over the agreement on the corrections made. Pitch was annotated completely automatically, 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 starts. Moreover, in a lot of cases, a retraction phase of a gesture can be recognized as stroke phase of the next gesture phrase.

As summarized in Table 3, for the more complex signals that involve multiple modalities for their comprehension (including speech, gestures, and text) the disagreements in interpretation were expectedly higher.

	Background nature	Foreground nature
NCIs	Regulators, Batons, Deictics	Illustrators, Symbols, Deictics
DAs	Turn management, Social obligations management, Time management, Discourse structuring, Feedback, Communication management	Task, Social obligations management

**Table 1.**  
*A coarse-grained classification of the underlying nature of NCI classes and DA dimensions.*

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	0.82
Gesture phrase	0.53
Modality	0.88
Prosodic phrases	0.71
Sentiment	0.67
Dialog function	0.64
Dialog dimension	0.71
Intent (semiotic class)	0.48
Emotion label	0.51
Gesture unit	0.75
Movement phase	0.66

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

### 3. Case study

Example 1: DAs as part of background and foreground conversational expressions.

*Guest: Ampak se izkaže, da ta zdravnik ne zna nič drugega delat kot vedno iste in samo iste (A) obraze in so vsi poklonirani (B) – no to je (1) to. Fajn, ne (2)?*

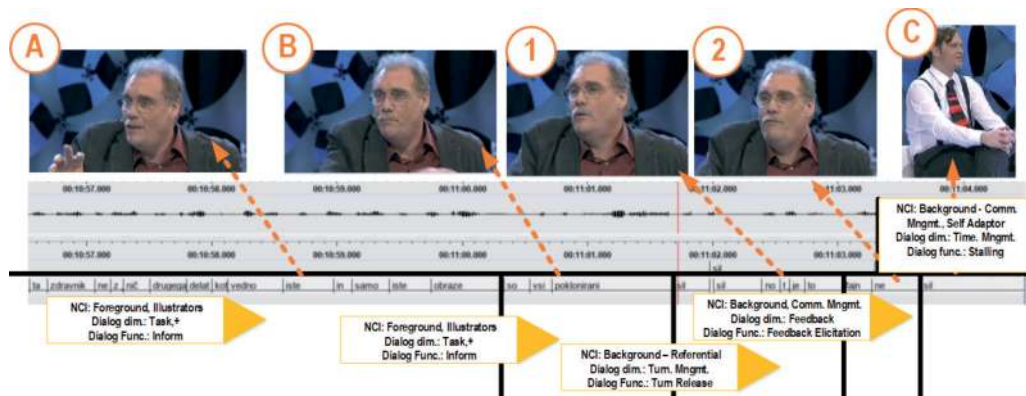
*Co-host: (C) Samo v bistvu, a veš, v bistvu sej če pri nas gledaš sj so tud pol vsi glih.*

*Guest: But it turns out that this doctor can create only one and the same (A) face and nothing else and that they are all cloned (B) – well this is it (1)<sup>1</sup>. Great, huh (2)?*

*Co-host: (C) But actually, you know, actually if you took a look at where we are then they are also all the same.*

This segment represents a case of sudden turn release by the main guest. Previously, the participants were discussing the effects of aging, during which several sarcastic comments were uttered. The show's host afterwards tries to transition to the next topic, which is the play the guest was directing, called *The Ugly One*. However, the guest is offended by the co-host's snide remark, where he compares the name of the play and the guest, suggesting that the guest might also be an ugly one. Nevertheless, after being asked to tell the audience about the play, he briefly outlined the plot, which deals with cosmetic surgery in connection with the feeling of self-worth and success. He is still mid-sentence and speaking with a rising intonation (see **Figure 2**: B, *in so vsi poklonirani* "and they are all cloned") when he suddenly takes a deep breath and decides to stop summarizing the play with the words *no to je to* "well this is it". Additionally, he emphasizes that he no longer wishes to talk about the topic, as he adds *fajn, ne?* "great, huh?". With it, he simultaneously elicits feedback, which is yet another way to assign his turn to someone else.

<sup>1</sup> The literal translation of the utterance is »this is this«.



**Figure 2.**  
 Multimodal analysis of the conversational expressions: use of DAs in background and foreground expressions.

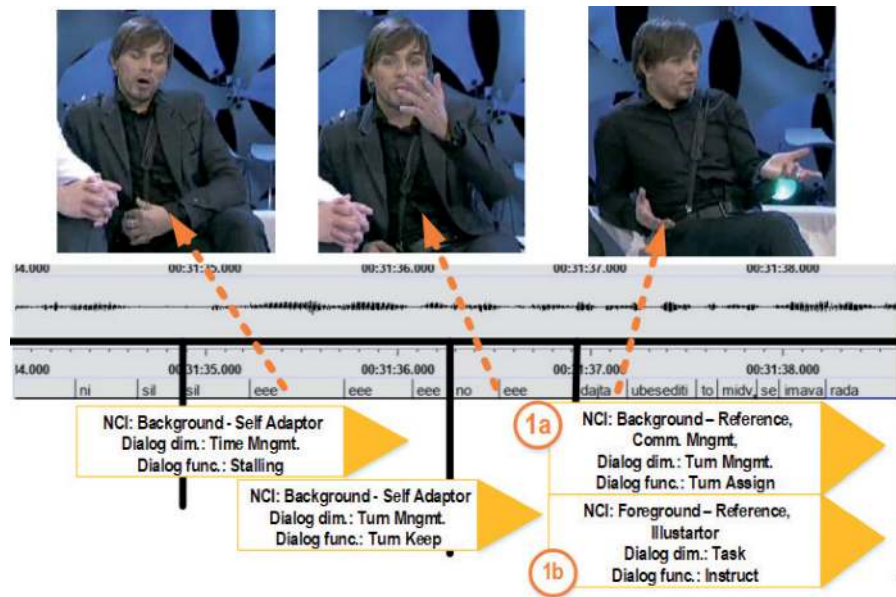
The utterance “well this is it” cannot be characterized as any other DA than turn management with the function of turn release as it serves no other communicative purpose. The phrase itself is tautological, deprived of any propositional content. The analysis of the accompanying body behavior also corroborates this fact. While outlining the plot of the play, he uses foreground NCI (A, B), namely illustrators, represented by two very prominent hand gestures. As he decides that he no longer wishes to explain the gist of the play, his NCI also changes. The body behavior is no longer prominent but very quick and even difficult to notice. The guest swiftly turns his head slightly to the right and back again (see **Figure 2: 1**) as if he was trying to point to the abstract “this” in “this is it” while still keeping eye contact with the host and co-host. The head movement was classified as a deictic NCI, specifically, a referent, since the guest is referring to the abstract “this”. He then adds the utterance *fajn, ne?* “great, huh?” which primarily acts as feedback elicitation, but secondarily also serves the function of turn management. The accompanying body behavior is, again, subtle, just a slight shrug of the right shoulder (see **Figure 2: 2**). It was classified as a communication regulator. The co-host perceives his turn release request and takes the turn by commenting on the essence of the play. However, since the release was unexpected, his response is yet to be formulated. This is highlighted through the use of metadiscourse (“But actually, you know, actually”) acting primarily as stalling within the time management dimension. However, of course, stalling functions also as a turn-take maneuver.

Example 2: Ambiguity of DAs in conversational expressions.

Host: *eee eee eee* (A) *no eee* (B) *dejta ubesedet to midva se mava rada* (1)

Host: *uh uh uh* (A) *well uh* (B) *come on, define this we like one another* (1)

The example above is a case of strong turn assigning. As a surprise for the main guest, his stepdaughter was invited to the show. The show’s host is trying to determine the correct nomenclature for the relation non-biological father/adoptive daughter, which are specific and probably less frequently heard words in Slovene. However, he is very clumsy when formulating his question, and neither of the guests understands him, but rather fill their answers with humor. The show’s host is dissatisfied and tries to change the evolution of the conversation. However, he needs time to formulate proper utterances and thus uses fillers (see **Figure 3: A and B**). After the first filler (A), which acts as stalling, the content is not completely formulated, which is why he uses the second filler (B). At the same time, however, the guests become impatient. The second filler, therefore, functions not only as a stalling element but primarily



**Figure 3.** Multimodal analysis of the conversational expressions: The duplicity of DAs when interpreted as background or as foreground conversational expressions.

as turn keep device. Once he formulated his idea, he begins with the imperative formulation *dajta* (for this purpose best translated as) “come on”. This utterance is accompanied by the host’s extended and raised left arm, both (temporarily) open hands, slightly raised shoulders, and a protruding head movement (see **Figure 3: 1**). This NCI was classified as a referential deictic (1a), as the host’s hands and head are extending towards the guests. At the same time NCI can also be perceived as visualizing the word *dajta*, thus being recognized as an illustrator. From the context of DAs, the utterance can be interpreted as having the underlying function of turn-taking or, due to the imperative formulation, the instruct function within the task dimension. As highlighted in the example, the use of DAs determines the perceived NCI.

Example 3: DAs in turn management within a multiparty conversation.

*Guest: grmičevje je zlo nerodno objemat*

*Co-host: zakaj?*

*Guest: ful pič ... ful ful te (A)*

*Host: ful ful pič ful pič*

*Co-host: ful me*

*Guest: drevo je fajn men*

*Host: eee (1)*

*Co-host: kosmulja (2)*

*Co-host: nadaljuj (3).*

*Host: ja (4)*

*Guest: it's very tricky to hug shrubs*

*Co-host: why?*

*Guest: totally pricks ... you get totally totally (A)*

*Host: totally totally pricks totally pricks*

*Co-host: I get totally*

*Guest: trees I like*

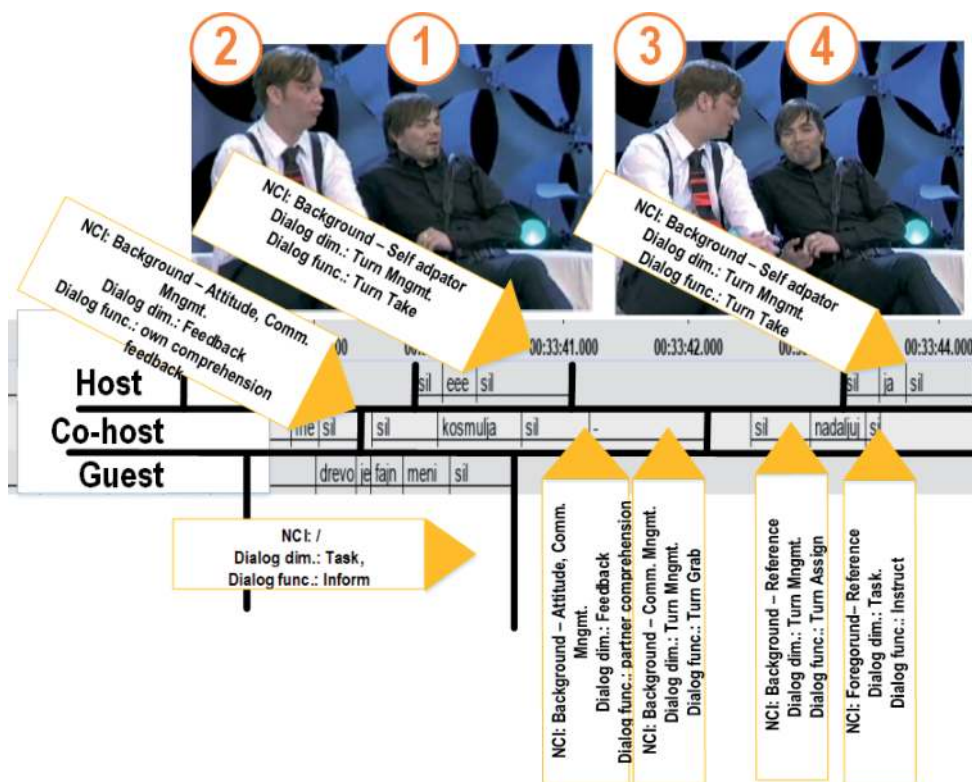
*Host: uh (1)*

*Co-host: gooseberry (2)*

Co-host: continue (3).  
 Host: yes (4)

This segment illustrates NCI in tree different turn management functions. Prior to the several turn-taking acts, the show's host is mocking the guest for his alleged morning ritual where he hula-hoops in his garden. The co-host humorously adds that he hugs surrounding trees and shrubs in the garden. First, the guest smiles at this mental image, but then his facial expression changes to serious, and he cautions that it is very hard to hug shrubbery. As the co-host asks why this is so, he turns the answer into a comical depiction of how he gets stung by thorns whereby he uses the colloquial Slovene word *ful*, which means *very* or *a lot*. The co-host is fascinated by this word choice of the guest, a theater actor, who just minutes before teased him for not enunciating correctly. He, therefore, mocks his (almost plosive) pronunciation of the word *ful* and the guest joins in in the mocking. The show's host, however, tries to join the conversation (see **Figure 4:1**), but his co-host still continues the mocking by saying "gooseberry" (see **Figure 4: A**) in a very comical manner, triggering light laughter from the guest but befuddlement from the host. The host turns to the co-host, hoping for clarification, the co-host stares back at him and finally tells him to continue with the show. After briefly gathering his thoughts, the host nods, says "yes" and changes the topic.

This excerpt, therefore, contains turn-taking, turn-assigning, and turn-accepting. Following a series of task dimension DAs, the show's host tries to take the turn by uttering the filler *eee* "uh". He fails, as his co-host drowns him out with "gooseberry". There is no NCI accompanying the host's utterance, as he barely moves (see **Figure 4: 1**). We, therefore, classified the NCI as undetermined. An indicator for his turn-take attempt is his gaze, which remains directed towards the guests (see **Figure 4: 1**) throughout the



**Figure 4.** Multimodal analysis of the conversational expressions: Use of turn-take and turn-grab to mediate the conversation.

co-host's interruption. As he turns to the co-host, he remains speechless and waits for him to elaborate. The co-host subsequent NCI, on the other hand, is a clear referential deictic (3) accompanying the DA of assigning the turn. His gaze towards the host was not enough to prompt a response, so he adds a firm head nod (see **Figure 4: 3**) towards him and verbalizes his intent of assigning the turn to him with "continue". This firm head nod is why the DA was not secondarily classified as a turn release, but as instructing within the task dimension. The host almost simultaneously responds to this NCI with a slight nod himself (see **Figure 4: 4**) and thereupon the verbal confirmation "yes". The nod was identified as a communication regulator NCI. The verbal confirmation functions as a turn accept DA. Both the DA and the NCI are of background nature.

#### 4. Discussion

An effective analysis of the non-verbal behavior that accompanies turn-taking requires material that is authentic and rather informal than formal. The episode from the entertainment show *As ti tut not padu?* offers such a resource. As elaborated in the section *The EVA Corpus*, the notion that the material is highly spontaneous is corroborated by the video's statistics. The sheer amount of overlapping speech indicates a high frequency of turn management acts. The notion that the material is spontaneous and performed in a relaxed manner is further supported in Cooperrider's [27] foreground-background distinction, as the DAs are well-balanced according to their nature. Overall, there are 1,897 foreground DAs and 2,020 background DAs. Therefore, more than half of the DAs are of background nature. This suggests that the material is not task-oriented, but instead serves the purpose of an entertainment show. Moreover, the most frequently observed NCI in the video were NCIs classified as regulators. The group of regulators represents 3/4 of all recognized NCIs in the entertainment episode. Regulators are followed by the group of deictic NCIs representing almost roughly 16 percent of the recognized NCIs. The remaining groups of illustrators, batons, symbols, and undetermined NCI each account for less than a tenth of the recognized NCIs. Again, the dominant NCI groups are of background nature, even if observing only regulators. These findings further support the notion that the material is highly spontaneous, relaxed, and entertaining. Therefore, it was a suitable choice for the analysis of natural turn-taking behavior and its accompanying non-verbal behavior.

Even though it seems relatively elementary to annotate turn management DAs, at times, the process proved to be complex. The acts of turn management are intertwined with stalling and instruction DAs. Examples 1 and 2 show how the stalling function (within the time management dimension) can also act as a turn-taking mechanism. In example 1, on the one hand, the co-host wants to take the floor since the guest suddenly released his turn. On the other hand, he does not know how to start. Hence, it is difficult to determine the primary DA, especially since the remainder of his response can be considered an information-providing DA. Example 2 illustrates how fillers, which generally pertain to the stalling function within the time dimension, act as turn-taking devices. They signal to the interlocutor that one wishes to speak but still requires additional time to properly verbalize one's thoughts. To further complicate the annotation of turn-taking management, in a conversation, each utterance by another person can secondarily be considered a turn-taking DA. Even the act of posing a question, which would primarily be annotated within the information-seeking dimension, can secondarily be annotated with a turn-assign function (since person A, who is asking person B the question, wishes person B to respond). Nevertheless, we did not annotate such secondary cases of turn management as alternative DAs. Only DAs where turn management is key for an utterance were assigned the turn dimension. Consequently, the share of turn

management DAs could be significantly greater, and the ratio between foreground and background expressions notably tilted towards the background spectrum. Again, this only highlights the nature of the material, which is by no means primarily task-oriented.

Since regulators and (partially) deictics are background expressions, we expected them to co-occur with turn management DAs. As noted, deictics are the elusive group of NCIs which can occur both in the foreground and the background. In Example 1, the explicit turn management DA is accompanied by a referential deictic. The semantic link with the word it refers to, however, is weak. The guest speaks of an abstract “this” in his utterance “well this is it”. He does not refer to anything physically present in the room; he just points to a mental image. Therefore, the deictic is part of the background. This, in turn, is in line with the assigned DA, since both concepts are background expressions.

A similar symmetry between the nature of the DA and the NCI is observed in Examples 2 and 3. In Example 2, the NCI can be considered both as a background and a foreground expression (see **Figure 3: 1**). Within the concept of DAs, it can also be considered as occurring in the background and in the foreground. In the background, it is a turn-assigning DA with which the host hopes to receive a response to his nomenclature dilemma. In this case, the non-verbal behavior is perceived as spontaneous. Rather than to visualize the referential utterance *dajta* “come on” the speaker tries to emphasize his frustration with the interlocutors, i.e., if you know better then please explain, and thereby assigns the turn someone else. The ‘open hand gesture’ is observed to signal this. In the foreground, it has an instructing function, within the task dimension, since he demands a response. The host is referencing actual people in the room and due to the imperative use of the referential utterance *dajta* “come on you two”, the observed conversational expression may be interpreted as instructing. *Dajta* is perceived as an explicit speech referent, and the non-verbal behavior seems to directly visualize it. Similarly, in Example 3, the NCI (see **Figure 4: 3**) is not a typical background referential deictic as the co-host physically leans towards the host while he nods towards him in order to prompt him to continue. The referent of the NCI is, therefore, an actual person (the host) in the room, and the NCI also occurs in the foreground. This duality is reflected in DAs as well. On the one hand, the co-host assigns the turn to the host (within the turn dimension); on the other, the co-host instructs the host to speak (within the task dimension). Again, this is an example of the difficulty in differentiating between background and foreground expressions.

Finally, as hypothesized, regulators are the group of NCIs that co-occur with unambiguous turn management DAs. In Example 1, the secondary turn management DA co-occurs with the group of regulators, specifically a communication regulator. Whereas the group of deictic NCIs crosses the foreground-background barrier, regulators are background expressions. The fact that the accompanying NCI is a regulator and not a referent from the deictic group, which would be more typical for feedback elicitation, further endorses the assignment of a turn dimension DA and highlights the turn management intent. Example 3 illustrates a similar unambiguity when regulators are used for turn-taking. There is an underlying agreement of the nature of the DAs and the NCI in the last utterance “yes” (see **Figure 4: 4**). Namely, both take place in the background. And clearly, they were well understood by the interlocutors as no one else tried to take the turn. The opposite phenomenon is observed at the beginning of the same example (see **Figure 4: 1**) with the host’s filler “uh”. It is an example of a failed turn-take attempt since the co-host interrupts the host. There was no noticeable non-verbal behavior accompanying the host’s filler, which is why no NCI could be assigned. However, one might argue that the fact that there is no NCI accompanying his turn-take attempt, contributes to the reason of why the attempt failed. Consequently, this might be considered a supporting example of Birdwhistell’s [22] findings that successful communication requires both verbal and non-verbal components.

We can therefore confirm our hypothesis that turn management DAs co-occur with regulators. The case analysis further supports the hypothesis that turn management DAs particularly co-occur with communication regulators. Moreover, we can observe that during propositional content, i.e., task-oriented DAs, use of illustrators (foreground NCIs) is more common (see **Figure 2: B** in Example 1). In accordance with Cooperrider's [27] characteristics of foreground-background gestures, we observed the spatial prominence of each type of gesture. Example 1 shows how non-verbal behavior changes in parallel with the change of DAs. As the DAs changed from task with the function of informing to turn management with the function of turn release (see **Figure 2, B and 1**), so did the NCI. It shifted from foreground behavior to background behavior. Moreover, foreground NCIs are far more prominent than the background NCIs. It seems that, as this simultaneous shift in DAs and foreground-background behavior occurs, body behavior is decelerated and minimized. Our findings, therefore, corroborate Cooperrider's [27] the special hallmarks of foreground-background gestures.

There are, however, border cases. For example, the background DA of providing feedback during active listening, such as uttering the supportive "yes" or "mm-hmm", can be accompanied by a slight nod of the head. Head nodding is generally considered a foreground gesture, if it signals a "yes" or "no" answer, since it can substitute speech altogether. However, in background use, one does not provide an answer, but merely signals to the interlocutor, that one is listening to them and wishes them to continue their turn. Hence, the act is clearly of background nature. Nevertheless, it is impossible to state that at the same time one does not also agree with what the interlocutor is saying. Agreement, however, is considered a foreground act. This is a typical case where the duality no longer applies. Hence, it is possible even for background DAs, such as feedback providing and eliciting, to co-occur with foreground NCIs. Moreover, even task-oriented DAs are often accompanied with batons, a representative background NCI, since they signal importance or set the rhythm but do not convey any propositional content. It is therefore difficult to extend the shared background-foreground nature hypothesis to other DAs. Despite this observation, the exploration of the shared nature in foreground DAs offers an interesting research question for future research.

A potential concept to further elaborate on the underlying nature is to observe whether the gesture is prominent (in its iteration or spatial dimensions) or subtle [27], as observed in Example 1. In accordance with this distinction, a subtle nod suggests background nature whereas a prominent nod suggest that the gesture is of foreground nature. Moreover, the relative timing may also, provide additional insight in the communicative intent. Although, not directly investigated in this research, it seems that when the stroke phase of the embodiment (especially a hand gesture) co-occurs with a specific speech referent (i.e. the gesture starts at the same time as the spoken articulation) the information provided is propositional, i.e., of foreground nature, whereas when the stroke phase occurs outside boundaries of the targeted referent (or without one) the information provided is of background nature. An example would be phrases "look over there!" and "what do you mean?". In general, deictics will accompany both phrases. On the one hand, the phrase "look over there" is clearly a task-oriented DA and will be accompanied by a pointer, the stroke of which will occur aligned with the verbal articulation of "there". On the other hand, the stroke phase of a similar gesture 'visualizing' the "you" in "what do you mean?" will co-occur with "mean" and will be recognized as a referential deictic in turn management (i.e., as turn offer). Thus, in our future investigations, we tend to analyze 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 shed further light on the true purpose of the shared nature.



## 5. Conclusions

In this chapter, we examined what kind of non-verbal behavior accompanies turn management DAs. For the annotation of turn management DAs, the ISO 24617-2 scheme's functions sufficed. Nevertheless, turn management DAs frequently overlap with other DAs, especially within the time management dimension. The fact that it is sometimes very difficult to decide which dimension and function is the most fitting shows the importance of multidimensional DA tagging. As a future endeavor, it would be more functional to create annotation schemes that, besides being multidimensional, denote the hierarchical order of the tags assigned, for example, the primary, secondary, tertiary, etc. dimensions and functions.

Cooperrider's [27] distinction between gestures that occur in the foreground or background proved an effective method within the concept of DAs. We hypothesized that there is an interlink between background NCI and background DAs. Since regulators, specifically, communication regulators, convey typical background NCI, we predicted their co-occurrence with turn management DAs. Indeed, the present case study confirms this hypothesis. Moreover, an interlink with deictic NCI was observed. As they can be of either background or foreground nature, the premise that background DAs co-occur with background NCI is maintained. This duality is not observed only within NCI but also within DAs. An utterance can have alternative expressions, one of background nature and one of foreground nature. However, the duality occurs simultaneously for NCI and for DAs. Hence, the fact that there is the same duality at the NCI level and at the DAs level strengthens the hypothesis of an interlink between the two concepts.

## Acknowledgements

This paper is partially funded by the Slovenian Research Agency, project HUMANIPA (research core funding No. J2-1737 (B)). This paper is partially funded by European Union's Horizon 2020 research an innovation program, project PERSIST (grant agreement No. 875406).

## Conflicts of interest


The authors declare no conflict of interest.

## Author details

Izidor Mlakar\*, Matej Rojc, Darinka Verdonik and Simona Majhenič  
Faculty of Electrical Engineering and Computer Science, University of Maribor,  
Maribor, Slovenia

\*Address all correspondence to: [izidor.mlakar@um.si](mailto:izidor.mlakar@um.si)

## IntechOpen

© 2021 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Austin J L. How to do things with words. Oxford: Clarendon Press; 1962.
- [2] Searle, J R. Speech acts: An essay in the philosophy of language. Volume 626. Cambridge university press; 1969.
- [3] Bunt, H. The DIT++ taxonomy for functional dialogue markup. In: Heylen D, Pelachaud C, Catizone R, & Traum D, editors. Towards a Standard Markup Language for Embodied Dialogue Acts. Proceedings. Budapest; 2009. p. 36-48.
- [4] Schmitz B, Quantz J J. Dialogue Acts in Automatic Dialogue. In: Interpreting Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, TMI-95. Leuven; 1996. p 33-47.
- [5] Heeman P, Allen, J. The Trains 93 Dialogues: TRAINS Technical Note 94-2. Computer Science Department. The University of Rochester; 1995.
- [6] Di Eugenio B, Jordan P W, Pylkkänen L. The COCONUT project: dialogue annotation manual. ISP Technical Report 98-1. University of Pittsburgh; 1998.
- [7] Godfrey J J, Holliman E C, McDaniel J. SWITCHBOARD: Telephone speech corpus for research and development. In: Proceedings of the ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing; 23-26 March San Francisco, CA, USA; 1992. Volume 1. p. 517-520.
- [8] Core M G, Allen J F. Coding dialogues with the DAMSL annotation scheme. In: Traum D R, editor. Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines. Boston; 1997. p. 28-35.
- [9] ISO 24617-2. Language resource management – Semantic annotation framework (SemAF): Part 2: Dialogue acts. Reference number ISO 24617-2:2012(E). Geneva; 2012.
- [10] Bunt H. The Semantics of Dialogue Acts. In: Proceedings 9th International Conference on Computational Semantics (IWCS 2011); January 12-14; Oxford, UK, p. 1-14.
- [11] Petukhova V, Bunt H. Introducing Communicative Function Qualifiers. In: Fang A, Ide N, Webster J, editors. Language Resources and Global Interoperability. Proceedings of the second International Conference on Global Interoperability for Language Resources (ICGL 2010); City University of Hing Kong; 2010. p.123-131.
- [12] Yoshino K, Tanaka H, Sugiyama K, Kondo M, Nakamura S. Japanese Dialogue Corpus of Information Navigation and Attentive Listening Annotated with Extended ISO-24617-2 Dialogue Act Tags. LREC; 2018.
- [13] Navarretta C, Paggio P. Dialogue Act Annotation in a Multimodal Corpus of First Encounter Dialogues. LREC; 2020.
- [14] Petukhova V, Bunt H. The coding and annotation of multimodal dialogue acts. In: Proceedings 8th International Conference on Language Resources and Evaluation (LREC 2012); Istanbul: 2012.
- [15] Petukhova V, Bunt H. A multidimensional approach to multimodal dialogue act annotation. In: Proceedings of the Seventh International Workshop on Computational Semantics (IWCS); 2007. p. 142-153.
- [16] Hunyadi L, Váradi T, Kovács G, Szekrényes I, Kiss H, Takács K. Human-human, human-machine

communication: on the HuComTech multimodal corpus. In: CLARIN Annual Conference 2018; Pisa, Italy; 2018.

[17] Feyaerts K, Brône G, Oben B. Multimodality in interaction. In: Dancygier B, editor. *The Cambridge Handbook of Cognitive Linguistics*. Cambridge University Press: Cambridge; 2017. p. 135-156.

[18] Mlakar I, Kačič Z, Rojc M. A Corpus for Analyzing Linguistic and Paralinguistic Features in Multi-Speaker Spontaneous Conversations–EVA Corpus. *International Journal of Computers*, 2. 2017;136-145.

[19] McNeill D, Duncan S. Growth points in thinking-for-speaking. In: McNeill D, editor. *Language and Gesture (Language Culture and Cognition)*. Cambridge: Cambridge University Press; 2000. p. 141-161.

[20] Peirce C S. *Collected papers of Charles Sanders Peirce*. Vol. 5. Harvard University Press; 1965.

[21] Ekman P, Friesen W. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*. 17 (2). 1971;124-129.

[22] Birdwhistell R L. *Kinesics and Context: Essays on Body Motion Communication*. Philadelphia: University of Pennsylvania Press; 1970.

[23] Allwood J. A framework for studying human multimodal communication. In: Rojc M, Champbell N, editors. *Coverbal Synchrony in Human-Machine Interaction*. Ed. 1. Boca Raton: CRC Press; 2014. p. 17-39.

[24] McNeill D, Levy E, Duncan S D. *Gesture in Discourse*. In: Tannen D, Hamilton H E, Schiffrin D, editors. *The Handbook of Discourse Analysis 2*. Wiley-Blackwell; 2015. p. 262-289.

[25] Church R B, Goldin-Meadow S. So how does gesture function in speaking, communication, and thinking? In: Church R B, Alibali M W, Kelly S D, editors. *Why Gesture? How the hands function in speaking, thinking and communicating*. *Gesture Studies 7*. Philadelphia: John Benjamins Publishing Company; 2017. p. 397-412.

[26] Keevallik L. What does embodied interaction tell us about grammar? In: *Research on Language and Social Interaction*, 51(1). 2018. p. 1-21.

[27] Cooperrider K. Foreground gesture, background gesture. In: *Gesture*, 16(2). 2017. p. 176-202.

[28] Mlakar I, Verdonik D, Majhenič S, Rojc M. Towards Pragmatic Understanding of Conversational Intent: A Multimodal Annotation Approach to Multiparty Informal Interaction–The EVA Corpus. In: *International Conference on Statistical Language and Speech Processing*. Springer, Cham; 2019. p. 19-30.

[29] Paggio P, Navarretta C. The Danish NOMCO corpus: multimodal interaction in first acquaintance conversations. In: *Language Resources and Evaluation*, 51(2). 2017. p. 463-494.

[30] Cohen J. A coefficient of agreement for nominal scales. In: *Educational and Psychological Measurement*, 20(1). 1960. p. 37-46.