Universität Paderborn
Fakultät für Kulturwissenschaften

The reflection of interactional monitoring in the dynamics of verbal and nonverbal forms of explaining

Dissertation zur Erlangung des akademischen Grades Doktor der Philosophie (Dr. phil.) im Fach Psycholinguistik der Universität Paderborn

von

Stefan Teodorov Lazarov

Erstbetreuerin
Dr. Angela Grimminger

Zweitbetreuer
Prof. Dr. Geert Brône

Prüfungskommission
Prof. Dr. Katharina J. Rohlfing
Prof. Dr. Heike M. Buhl

Paderborn
2025

Eidesstattliche Erklärung

Hiermit erkläre ich gemäß §12 der Promotionsordnung:

a. dass die vorgelegte Arbeit selbstständig und ohne Benutzung anderer als der in der Arbeit angegebenen Hilfsmittel angefertigt wurde;

b. dass die Arbeit bisher weder im In- noch Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt wurde;

c. dass früher oder gleichzeitig kein Promotionsverfahren bei einer anderen Hochschule oder bei einer anderen Fakultät beantragt wurde, gegebenenfalls nebst vollständigen Angaben über dessen Ausgang.

Paderborn, 24.06.2025

S. dosarov

Contents

List of Figures

List of Tables

Acknowledgements

In the summer of 2021, I was given the opportunity to join a research project exploring how the recipients of explanations (explainees) signal understanding through multimodal behaviors. At that time, I had just finished my master's studies in linguistics. Until then, I had a great interest in phonetics and syntax, and psycholinguistics was a field that I have explored only in my introduction class as a bachelor student. However, wise people say 'you have to leave your comfort zone in order to succeed'. I remember the words of my supervisor when offering me the opportunity to work in an interdisciplinary project and do my PhD in Paderborn: "You only have to be open to the new.". And today, I am extremely grateful that I accepted this offer, which turned out to be the right depiction opening so many other doors for me.

At first glance, embracing a new opportunity may not seem difficult, but initially I made the circumstances more complicated for myself. With a purely linguistic mindset, I proposed a research topic investigating the use of deixis in explanations, arguably not the most exciting research topic in psycholinguistics. I have to admit that neither I was very enthusiastic about investigating this topic for my PhD, but I began developing my research proposal nonetheless.

Meanwhile, my investigation of explanatory interactions had begun. It was the first year after the COVID-19 pandemic, and conducting studies on human–human interaction in person requires a very complex organization. Since the main data collection for the research project that I joined had to be delayed due to the circumstances of the pandemic, I started my investigations using data from another video corpus which was also compiled for the purposes of the research projects. The student assistants in my lab had already started coding the recorded materials, including utterances and gaze behavior. A year later, several presentations took place, and I started writing my first publication together with my supervisor and another colleague from the TRR 318—Kai Biermeier. This publication was quite a game changer for me as a PhD student. My research interest in deixis gradually faded into the background—not intentionally, but because I became absorbed by the analysis of multimodal behavior and thus more interested in how multimodal behavior relates to cognitive processes. After that, another paper on the mechanism of monitoring the understanding of explainees related to the employment of co-speech deictic gestures by explainers followed. As my

research perspective had already evolved, I was able to clearly envision the research topic that I address in the present cumulative dissertation: *The reflection of interactional monitoring in the dynamics of verbal and nonverbal forms of explaining*.

While this journey might appear smooth and linear in retrospect, the people around me know very well about the challenges I faced along the way. Without the continuous support of these people, I would not have made this huge achievement (a dissertation), and I would not have learned all these valuable lessons that improved my scientific expertise.

First, I want to express my endless gratitude to Dr. Angela Grimminger for giving me the opportunity to join project the A02 and the TRR 318 "Constructing Explainability," and for supporting me throughout the past four years toward acquiring my doctoral degree. Angela, everything I have learned during this time is a product of your continuous commitment to guide me through every task and show me how to conduct quality research. I have been very fortunate to have you as my supervisor over the past four years. Thank you for being such a great scaffolding supervisor!

I would like to express my sincere gratitude to Prof. Dr. Geert Brône from the Research Group Multimodality, Interaction & Discourse at KU Leuven for seeing the potential in my research and accepting my invitation to become my second evaluator. His expertise and trust in my research gave me confidence and motivation in this very important phase of my life. Geert, I am very grateful for the constructive feedback I received from you, which contributed to improving the clarity and overall quality of my dissertation!

Another very significant person in my journey was Prof. Dr. Katharina J. Rohlfing, the Chair of the Psycholinguistics Research Group at Paderborn University. Katharina, thank you for welcoming and integrating me into your lab family and making me feel truly part of it. You taught me how important it is to help others, stay connected, and remain continuously engaged. Your support is priceless!

I also want to thank Prof. Dr. Heike M. Buhl for accepting my invitation to become a member of the examination board. Heike, our collaboration within the TRR 318 has taught me so much, especially with regard to the fine details of scientific writing, for which I am truly grateful!

My special gratitude also goes to all the participants who took part in the studies and all student assistants who annotated the materials used in the presented research. Without the work of my assistants, I would not have been able to complete any of the studies on time.

Last but not least, I must express my deep gratitude for the continuous support of my family and friends with whom I could share every exciting moment. Thank you for having faith in me!

Abstract

Explanations play a central role in everyday face-to-face interactions by helping people share knowledge, clarify ideas, and support understanding. In face-to-face explanations, explainers (i.e., the more knowledgeable part) seek to enhance explainees' understanding of an explanandum through interactional processes such as *monitoring*, *scaffolding*, and *co-constructions* (Buschmeier et al., 2023; Rohlfing et al., 2021). While co-constructions emerge from the bidirectional (non-)verbal exchange between the interlocutors, scaffolding refers to the process by which the explainers tailor an explanation by employing different forms of behavior in response to the explainees' behavior signaling their cognitive processing (Wood et al., 1976). Monitoring denotes a continuous process in which the interlocutors attend to perceptual evidence, such as (non-)verbal behaviors, to identify and interpret cues of (mis)understanding (Clark & Krych, 2004).

In the present thesis, I report five studies on dyadic human–human explanations, and based on empirical findings, I discuss the interactional dynamics underpinning some forms of the verbal and nonverbal explaining behavior. To this end, in the reported studies, I analyzed data from two video corpora on different domains of everyday explanations, such as medical and board game explanations. The corpus on medical explanations comprises eleven naturalistic interactions between physicians and caregivers about an upcoming pediatric surgery. The corpus on board game explanations consists of 87 dyadic board game explanations, from which a subsample of 24 interactions was specifically addressed in the reported studies.

To explore the verbal explaining behavior, I examined the relation between topical shifts in explanations and the explainees' multimodal behavior as monitored by the explainers in two studies addressing the domains of medical and board game explanations. The analysis of medical explanations suggests that shifts from elaborations to new topics are associated with the explainees' multimodal behavior that comprises gaze aversions, co-occurring with head nodding, and vocal backchanneling, whereas shifts into elaborations are associated with a sustained gaze direction, whether accompanied by additional cues or not (Lazarov et al., 2024). A subsequent study on board game explanations (Lazarov & Grimminger, under review) extended this analysis by incorporating the explainers' gaze behavior. The study investigated the relation

of mutual gaze and gaze withdrawals to the initiation of new topics. The findings corroborated those from the medical context: the explainees' gaze withdrawals precede topical changes more frequently than mutual gaze, which corresponds to prior research by Rossano (2013) and Rossano (2012). The analysis further investigated the relation between the initiator of gaze withdrawals and the initiator of topical changes.

To examine the nonverbal explaining behavior, I analyzed the use of co-speech gestures by different explainers in three studies on board game explanations, during which the explanandum was physically absent from the shared space. Although this absence implies a persistent need for establishing joint imagined spaces (Kang et al., 2015; Kinalzik & Heller, 2020), for example, by continuously pointing at invisible locations, the study by Lazarov and Grimminger (2025) revealed that gesture iconicity and temporal highlighting also occur variably across explanation topics about object features, action processes, and conditional rules.

Motivated by the continuous use of gesture deixis during the physical absence of the explanandum, the last two studies explored the cognitive mechanism of adapting the use of deictic gestures in relation to the monitoring of the explainees' understanding. Whether the explainers interpreted the explainees' understanding in a retrospective video recall task (Lazarov & Grimminger, 2024a), or perceived the explainees' verbal signals of understanding (Lazarov & Grimminger, 2024b), the analyses demonstrated that the frequency of gesture deixis remains stable over the course of the explanation phase in which the explanandum was absent from the shared space.

Building on the results from the studies presented in this thesis, I discuss how the continuous monitoring of the explainees' feedback behavior accounts for adaptations in both the verbal and the nonverbal modes of explaining behavior. Furthermore, the analyses I present illuminate the extent of individual variation within and across explainers, each of whom interacted with three different explainees.

Chapter 1

Introduction

Explanations are an inseparable part of daily life as they can emerge in various contexts of human–human interaction. For example, explanations may emerge in institutional settings (e.g., medical explanations), or casual daily conversations (e.g., board game explanations). Across these (and many other) interactional contexts, human–human explanations instantiate interactions in which a more knowledgeable person (the explainer) undertakes the task to enhance the understanding (and thus the knowledge) of a less knowledgeable person (the explainee) about an entity or a process (explanandum) (Buschmeier et al., 2023; Rohlfing et al., 2021). However, explanatory interactions do not reach completion in isolation; rather, they are supported by several interactional processes such as *monitoring*, *scaffolding*, and co-constructions (see Figure 1.1) (Rohlfing et al., 2021). Interactional monitoring refers to the bilateral process in which the interlocutors track perceptual evidence, such as (non-)verbal behaviors, seeking to detect and interpret (non-)verbal cues of (mis-)understanding (Clark & Krych, 2004). Scaffolding is a process by which the more knowledgeable explainer tailors (i.e., adapts) an explanation by employing different forms of behavior in response to the explainee's feedback behavior displaying their cognitive processing (Wood et al., 1976). In the context of (non-)verbal behaviors, co-constructions emerge from the bidirectional (non-)verbal interaction between the explainer and the explainee (Rohlfing et al., 2021).

Among these interactional processes that support the successful realization of explanatory interactions, interactional monitoring constitutes the central motivation to investigate the dynamics of verbal and nonverbal forms of explaining behavior in relation to the explainee's multimodal feedback behavior. In the present thesis, I report on a series of studies that examined how interactional monitoring reflects in the verbal and nonverbal co-constructive behaviors of explainers. Specifically, in the present thesis, I discuss how explainers adapt their explaining (non-)verbal behavior in relation to monitoring the explainees' feedback behavior in face-to-face explanatory interactions. To this end, it has been demonstrated that explanatory

interactions become adapted to the verbal co-constructions by both the explainers and the explainees (Fisher et al., 2023). According to Clark's 2004 theory about interactional monitoring and recent findings about the adaptive explaining behavior in the spoken discourse (Fisher et al., 2023), it is assumed that, along the verbal level, explanations also become adapted at the gestural (nonverbal) level in relation to the explainees' nonverbal feedback behavior.

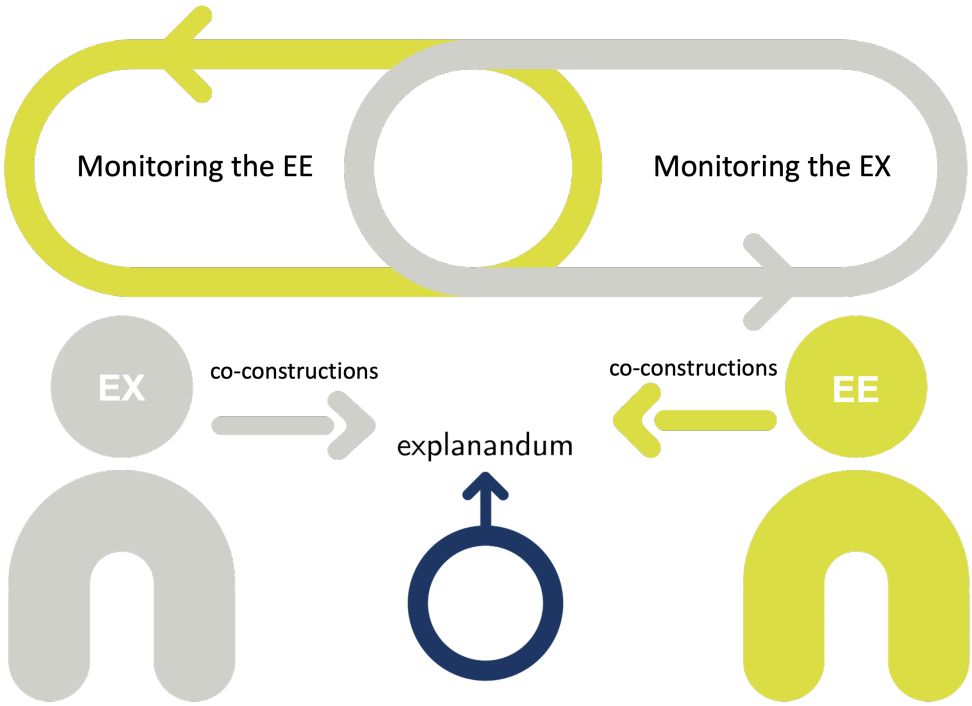

Figure 1.1: The monitoring process in dyadic explanations.

Note. EX = explainer, EE = explainee.

In the presented studies, two forms of explaining behavior were specifically addressed—(1) the topical organization of explanations (at the verbal level) and (2) the explainers' use of co-speech gestures (at the nonverbal level). The topical organization of explanations was analyzed in relation to the explainees' multimodal behavior, comprising their eye gaze behavior, head gestures, and vocal backchannels observed in two subsamples, each from a different video corpus on human–human explanations (see Chapter 5). The explainers' co-speech gestures were analyzed for the variance of frequency across different categories of explanation topics, as well as the variance of their use in relation to monitoring the explainees' (levels of) understanding during and after (in a post hoc video recall task) the interactions.

The topical organization of explanations was selected to explore the mechanisms underlying the dynamic development of multimodal interaction throughout human–human explanations. In doing so, explanations were segmented into topical units

(also, explanation episodes), which were related to different forms of the explainees' multimodal behavior. Accordingly, the dynamics of verbal explaining behavior was analyzed in the form of topical changes. In Chapter 6, two studies on the relation between the changing topical structure of explanations and the explainees' feedback behavior are presented. Study 1, by (Lazarov et al., 2024), was conducted on a subsample of ten medical explanations between physicians (explainers) and caregivers (explainees) about a child's upcoming surgery (explanandum) (see Section 6.1). In this study, topical shifts to elaborations and shifts from elaborations to new topics by different physicians were related to certain patterns of the caregivers' multimodal behavior, comprising their eye gaze, head gestures, and vocal backchannels. Study 2 (see Section 6.2) addressed a limitation from the first study by including the explainers' eye gaze behavior in the analysis, and thereby investigating the relation between two forms of interactive gaze behavior (mutual gaze and gaze withdrawals) and the topical shifts in the domain of board game explanations. As the findings of both studies complement each other, they are jointly presented and discussed in Chapter 6.

Since speech and gesture form an integrated system in which both modalities are semantically and temporally coupled (Kendon, 2004; Kita, 2009; McNeill, 2005), analyzing the explainers' use of co-speech gestures was considered alongside the explainers' verbal behavior (see Chapter 7). Concretely, the presented studies investigated the gestural dimensions (McNeill, 2006) observed in the co-speech gestures of the explainers, who explained a board game that was not visually accessible to the explainees. In particular, Study 3 investigated the use of gesture deixis, iconicity, and temporal highlighting across explanation topics addressing topical categories, such as object features, action processes, and conditional rules. In conditions when the explanandum is physically absent from the shared space, co-speech deictic and iconic gestures serve as co-constructive tools aiding the explainees to visualize the physical properties of the explanandum (and related sub-explananda), as well as to understand their spatial configuration (Clark, 2003; Congdon et al., 2017; Dargue et al., 2021; de Ruiter, 2000; Kandana-Arachchige et al., 2021; McKern et al., 2021; Stojnic et al., 2013).

Study 4 four and Study 5 focused specifically on the use of deictic gestures in relation to the monitoring of the explainees' (levels of) understanding. The main objective of these two studies was to explore a cognitive mechanism accounting for the adaptive use of co-speech gestures. This was implemented by incorporating two factorial dimensions in the analysis – (1) the explainers' interpretations about the explainees' understanding collected in a post hoc method and (2) the explainers' perception of the explainees' verbal signals of understanding during the interactions.

The main objective of the present thesis is to systematically review five recently conducted and related studies on the verbal and the nonverbal explaining behaviors

in explanatory interactions, and to demonstrate how these behaviors account for the monitoring of the explainees' (non-)verbal feedback. Prior to the main part of the thesis, a comprehensive theoretical background on interactional monitoring (see Chapter 2), relevant forms of verbal and nonverbal explaining behavior (see Chapter 3), as well as relevant forms of verbal and nonverbal feedback behavior (see Chapter 4) is provided. In the following Chapter 5, the two video corpora that were used for the analyses in the presented studies are introduced. The presentation of the studies is structured into two chapters discussing the dynamics of the topical structure of explanations (see Chapter 6) and the dynamics of the explainers' co-speech gestures (see Chapter 7). Finally, the general topic of the present thesis, i.e., how interactional monitoring reflects in the dynamics of verbal and nonverbal forms of explaining, is discussed by synthesizing the findings across the five studies (see Chapter 8).

Chapter 2

Why interactional monitoring?

An explanation is given when a less knowledgeable person (i.e., an explainee) needs to acquire understanding (and thereby knowledge) about an entity or a process (explanandum) (Buschmeier et al., 2023; Rohlfing et al., 2021). The knowledge asymmetry between an explainer and an explainee presupposes that an explanation would start at a point at which an explainee has less understanding of the explanandum (Kotthoff, 2009; Rohlfing et al., 2021). Over the course of an explanation, an explainee's level of understanding may increase or decrease in relation to the adaptive behavior of the explainer (Buschmeier et al., 2023; Rohlfing et al., 2021). Given the gradual nature of understanding as a cognitive process (Vendler, 1994) and the possibility that an explainee's understanding may gradually increase or decrease (Buschmeier et al., 2023), it is necessary for the explainer to continuously monitor and interpret the dynamically changing understanding of the explainee (Clark & Krych, 2004).

To identify understanding, both the explainer and the explainee need to reach the so-called "common ground" by sharing the same beliefs about objects or concepts (Clark, 1997; Clark & Brennan, 1991; Schober & Clark, 1989). One way in which understanding becomes identifiable to explainers is the simultaneous use of backchannel responses (Clark, 1997; Schegloff, 1982). Although understanding could be signaled in various interactive ways, it is yet little known how the monitoring of the explainees' understanding is related to the subsequent unfolding of explanatory interactions.

Understanding is not just a single product of an explanation that is unidirectionally transferred by the explainer to the explainees. Rather, understanding represents a multilevel system of cognitive processing of (abstract) concepts, and it is more narrowly defined by gradual qualities (levels) that range between *non-understanding*, *partial understanding*, and *complete understanding* (Bazzanella & Damiano, 1999; Vendler, 1994). In addition to these gradual levels, there is an additional fourth level, namely *misunderstanding*, which refers to false information processing that was earlier considered true. In the course of an interaction, a misunderstanding is communicated as complete understanding until one of the interlocutors detects a problem and initiates

a repair (Vendler, 1994). These levels of understanding are relevant for the discussion on the reflection of interactional monitoring in the dynamics of nonverbal explaining behavior. In this regard, two of the presented studies in Sections 7.2 and 7.3 relate the use of co-speech deictic gestures by explainers to monitoring the explainees' levels of understanding moment by moment.

Before discussing the specific characteristics of interactional monitoring as a process, it is important to clarify that the roles *speaker* and *addressee* are relevant to both explainers and explainees. In this relation, a study on medical explanations by Fisher et al. (2022) demonstrated that explanations are predominantly constituted by monological phases in which the explainers hold the turn, and thus become speakers, while the explainees remain addressees until they take the turn, e.g., by raising a question. In this relation, when referring to literature investigating the addressees' multimodal behavior as being monitored by the speakers, this also holds for the explainers monitoring the explainees' behavior. In addition, the cited research on monitoring in human–human interaction did not specifically address the term "explanations"; however, the examples as in Clark (1997) and Clark and Krych (2004) refer to dialogues in which a speaker instructs an addressee how to do something, which itself refers to the act of explaining *how* (Klein, 2009).

In dyadic interactions, including explanations, both explainers and explainees continuously monitor—that is, account for—each other's behavior throughout the interaction in order to track the explainees' understanding moment by moment and thereby the development of the interaction (Clark & Krych, 2004). Monitoring the explainee's behavior enables the explainer to adjust the structure of their explanations following (their interpretations of) the explainee's verbal and nonverbal feedback of understanding (Krych & Clark, 1997). In previous research on spontaneous conversations, the nonverbal behavior of the interlocutor who is listening has been shown to influence both the length and the form of turns taken by the interlocutor who is speaking (Sacks et al., 1974). Further, when addressees withdraw their gaze from the speakers—thereby indicating disattending behavior—the speakers tend to adapt their utterances, for example, by introducing repetitions (C. Goodwin, 1981).

According to Clark and Krych (2004), the speakers monitor and respond to the addressees' behavior across several levels, including their speech, facial expressions, gestures, and shared visual scenes. At the speech level, addressees may contribute with full utterances or with short feedback responses, also referred to as backchannels (Allwood et al., 1992; Betz et al., 2019; Clark, 1997; Gravano et al., 2012; Lai, 2009, 2010; Malisz et al., 2016; Neiberg et al., 2013; Schegloff, 1982; Ward, 2006; Yngve, 1970). The facial area conveys cues, such as eye gaze behavior, which signal attention to both verbal (Clark & Krych, 2004; C. Goodwin, 1981; Kendon, 1967) and nonverbal behavior (J. B. Bavelas et al., 1986), and cognitive processing (J. Bavelas & Chovil,

2018; Glenberg et al., 1998; C. Goodwin, 1981; Phelps et al., 2006). Other facial signals, including blinking (Hömke et al., 2017; Nota et al., 2021) and eyebrow movements (Nota, Trujillo, & Holler, 2023; Nota, Trujillo, Jacobs, & Holler, 2023), have been also related to (non-)understanding. Furthermore, head gestures, such as nods (Cerrato, 2005; De Stefani, 2021; Gander & Gander, 2020; Włodarczak et al., 2012; Wu & Heritage, 2025) and tilts (Ismail & Syahputri, 2022; Włodarczak et al., 2012) may signal attention, agreement, or understanding on the part of the addressee. Because many of the above-mentioned signals may lead to ambiguous interpretations by the interlocutors (see Chapter 4), monitoring the explainees' understanding could be a challenging task for the explainers due to the possibility of misinterpreting the explainees' (non-)verbal feedback.

Chapter 3

Forms of verbal and nonverbal explaining behavior

3.1 Explanation topics

Generally, everyday explanations address an overarching explanandum, which may be an object, an abstract entity, or a process (Rohlfing et al., 2021). However, in many cases, the overarching explanandum encompasses a combination of these elements—objects, abstract entities, and processes—each of which may represent distinct sub-explananda entailed by the overarching explanandum. Also, in the course of increasing the explainee's understanding of an explanandum (Buschmeier et al., 2023), explanations usually address questions such as *what*, *how*, and *why* (Klein, 2009). In this context, explanation topics can be sub-categorized into smaller units, such as new topics (what) and elaborations of previous topics (how and why).

In the presented studies, two examples of overarching explananda are investigated: (1) an upcoming surgery of a child, and (2) a board game. In the first example, the explanation about an upcoming surgery is structured around several sub-explananda, such as the diagnosis, the reasons, the medical procedure, and the post-clinical treatment. Regarding the second example, an explanation of a board game includes sub-explananda, such as the game type (e.g., collaborative or competitive), the goal, the objects constituting the game, as well as various conditional rules that govern the game actions.

One way to analyze the topical structure of explanations realized by the verbal explaining behavior of the explainers—that is, the structure of different (sub-)explananda—is to apply an episodic segmentation of explanation topics (Roscoe & Chi, 2008). By this approach, different explanation (sub-)topics can be segmented and annotated based on their semantic focus.

As dyadic explanatory interactions involve both an explainer and an explainee,

the topical organization of an explanation is shaped not only by the explainer's verbal behavior but also by the dialogical interaction between the explainer and the explainee. Providing examples from the domain of medical explanations given by physicians (i.e., the explainers) to caregivers (i.e., the explainees), Fisher et al. (2022) found that the physicians, who were verbally more active than the caregivers, introduced the majority of explanation topics. In contrast, the caregivers introduced explanation topics to a much lesser extent, for example when raising clarification questions.

While the topical organization of explanations at the verbal level is partly shaped by the verbal activity of the interlocutors, it is also influenced by nonverbal factors associated with changes in the topical structure. The distinction between new topics and elaborations of previous topics is particularly important, as elaborations may be prompted by the explainee's multimodal behavior (see Section 6.1), which is continuously monitored by the explainer (Clark & Krych, 2004).

For example, the relation between interactional monitoring and the topical organization of explanations becomes evident when explainers perceive the need to elaborate on a previous topic by providing additions, completions, or paraphrases in order to address issues of understanding (Dingemanse et al., 2015). Regarding the relation of the explainees' feedback behavior to the topical structure of explanations, the study presented in Section 6.1 demonstrates that the explainees' multimodal feedback behavior, which comprises their non-changing gaze behavior (e.g., directed toward the explainers) accompanied either by head nodding or backchanneling, is related to transitions into elaborations initiated by the explainers.

3.2 Co-speech gestures

In human–human interaction, speech is accompanied by other means of nonverbal communication, such as co-speech gestures. Together, speech and gestures form an integrated system that is characterized by a tight semantic and temporal coupling (Kendon, 2004; Kita, 2009; McNeill, 1992). Kendon (2004) defined co-speech gestures as visible actions that may appear independently as utterances or in conjunction with spoken language. Although Kendon's 2004 definition of co-speech gestures encompasses any body part (also articulator), such as the hands or the head, the present thesis focuses on the analysis of one articulator as part of the explaining behavior, namely hand gestures. Head gestures, by contrast, are examined in this work as a form of explainees' nonverbal feedback behavior (see Section 4.2 for a review).

According to McNeill (1998), co-speech gestures are defined as arm and hand movements occurring within specific spatial regions–typically in front of the speaker's torso–through which the speaker expresses a gestural reference or an emphasis on important content of speech. Co-speech gestures establish links to syntactic, semantic,

or pragmatic units of speech, for example by pointing, representing, or emphasizing key content, thereby enhancing the addressee's understanding (Clark, 2003; de Ruiter, 2000; Kendon, 2004; McNeill, 1992, 2006; Stojnic et al., 2013). For instance, a congruent semantic relationship between gestures and speech has been shown to facilitate faster reaction times and more accurate gesture interpretation while the addressees observe the speakers' gestures (Habets et al., 2011; Kelly et al., 2010; Ping et al., 2013), as well as to reduce the cognitive load in learning contexts (Li et al., 2022).

3.2.1 Gesture categorization systems

Like the signs of verbal language, co-speech gestures are defined by formal and functional characteristics, and the relationship between them varies depending on the gesture category. There are two primary approaches to categorizing co-speech gestures: a classical approach based on distinct, clear-cut categories (McNeill, 1992), and a more recent approach based on converging dimensions (McNeill, 2006).

The categorical gesture system

According to the classical categorization system, co-speech gestures are classified as either deictic, iconic, metaphoric, or beats[1], each represented by distinguishable forms (McNeill, 1992, 2006). Deictic, iconic, and metaphoric gestures typically consist of triphasic hand movements (preparation, stroke, and retraction phases), whereas beat gestures consist of biphasic movements (preparation and stroke phases)[2] (McNeill, 1992). In the presented studies on explainers' co-speech gestures (see Chapter 7), gestures were annotated at the level of gesture phrases (see Chapter 7 for details).

Each gesture category supports the establishment of a relationship between a linguistic reference and a referent to varying degrees (Kendon, 2004; McNeill, 1992). In Western cultures, deictic gestures are used to point to people, objects, or locations, often realized through index-finger or palm pointing, typically with the arm extended (Clark, 2003). In explanatory interactions, deictic gestures play a central role in the explaining behavior for increasing the explainees' understanding of an explanandum, especially when physical objects that are being explained are absent from the shared referential space between the explainer and the explainee. In order to increase the explainees' understanding about the spatial organization of physical objects, e.g., the

[1] A fifth category, emblems, refers to conventionalized (i.e., culturally specific) gestures (Kendon, 1995; McNeill, 1992). Emblems have arbitrary meanings that can be interpreted even in the absence of speech–for example, the "approval gesture" (formed by touching the tips of the thumb and index finger to create a circle, with the other fingers extended) (Kendon, 1995; McNeill, 2006; Teßendorf, 2013). Emblems were not included in the studies on explaining gestural behavior.

[2] The stroke phase, also called the apex, is the central component of a gesture phrase (i.e., a single gesture), while the preparation (raising the hand from a resting position) and retraction (returning the hand to rest) phases are not obligatory (McNeill, 1992).

pieces of a board game, the explainer may feel required to point to various locations continuously over the course of the explanation (Lazarov & Grimminger, 2024a, 2024b, 2025).

Iconic gestures involve arm and hand movements that visually depict properties of objects, actions, or events (McNeill, 1992). By creating imagery, iconic gestures have been shown to enhance memory recall and comprehension among addressees (Dargue et al., 2021; Kandana-Arachchige et al., 2021; McKern et al., 2021). A study on explanatory board game interactions between children (explainers) and adults (explainees) demonstrated that iconic gestures support the explainees' understanding of game explanations, particularly when combined with deictic gestures to create so-called "joint imaginary spaces" (Kinalzik & Heller, 2020). These findings were supported in a study on explanatory dialogues between adults that demonstrated how iconic gestures are used by explainers to increase the understanding of explanation topics, particularly dealing with the physical features of objects, e.g., the shape or the size of game pieces (Lazarov & Grimminger, 2025).

In comparison to iconic gestures, metaphoric gestures refer to abstract concepts within speech. For example, moving the left and right hands closer together can illustrate the convergence of two opposing opinions (Beattie & Shovelton, 2005; McNeill, 1992). Although metaphoric gestures may also occur in human–human explanations, this category was not investigated in the reported studies in Chapter 7. The reason to leave this category apart from the analysis was that the studies were focused particularly on the relation between the use of co-speech gestures referring to physical entities and different cognitive mechanisms that may have effect on the frequency of co-speech gestures, such as the semantic focus of explanation topics or monitoring the understanding of the explainees.

Beat gestures do not convey semantic content; rather, they are used to put emphasis on certain syntactic units within utterances. These gestures are typically aligned with the affiliated speech unit and are often marked prosodically (Beege et al., 2020; McNeill, 1992). While beat gestures can also support the addressees' understanding, they do so to a significantly lesser extent than deictic and iconic gestures (Austin & Sweller, 2014; Dimitrova et al., 2016; Rohrer, Delais-Roussarie, & Prieto, 2020). Despite being non-referential, beat gestures were considered in the studies on explanations, discussed in Chapter 7. In the course of increasing the explainees' understanding of an explanandum, explainers are shown to continuously perform a gestural emphasis related to important semantic content while using deictic and iconic gestures referring to different locations or object features (Lazarov & Grimminger, 2025).

The dimensional gesture system

In a revised version of the gesture categorization system, McNeill (2006) proposes that co-speech gestures are better understood as expressing multiple dimensions—such as deixis, iconicity, metaphoricity, and temporal highlighting—rather than being defined by distinct, clear-cut categories. This dimensional perspective allows for co-speech gestures to be realized in hybrid forms, and it has recently gained traction in research on multimodal behavior, for example being implemented in the "Multimodal MultiDimensional (M3D)" labeling system (Rohrer, Tütüncübasi, et al., 2020).

The possibilities in which the hands can use the referential space for pointing, depicting, and set prominence on important speech content are unlimited (McNeill, 2006; Müller, 2013; Rohrer, Tütüncübasi, et al., 2020). An extended index finger may be used to point to a specific location on a shared space (deixis) while performing multiple pointing strokes on the same location (gestural emphasis) to draw the addressee's attention to the corresponding spoken reference. Similarly, an extended index finger can be used to trace the shape of an invisible object in the shared space (iconicity) while simultaneously pointing at an imaginary location (deixis) and performing repeated strokes (establishing emphasis) (Müller, 2013). Human–human explanations which take place in the physical absence of an explanandum are an example of an interaction in which multidimensional co-speech gestures are used by explainers (Lazarov & Grimminger, 2024a, 2024b, 2025). The studies presented in Chapter 7 reveal the frequencies in which different combinations of dimensions, such as deixis, iconicity and temporal highlighting (used for gestural emphasis) are used by the explainers seeking to increase the understanding of explainees about a (physically absent) board game and its spatial organization on the referential space.

3.2.2 Individual differences in co-speech gestures

Although co-speech gestures are used for specific purposes–for example, to point to locations, to illustrate objects, or to highlight important units of speech (McNeill, 1992, 2006)–they are not used in the same way across individuals. Previous research on formal gesture characteristics, such as form and path, has shown that gesturing is idiosyncratic, that is, specific to individual speakers (Bergmann & Kopp, 2010; Priesters & Mittelberg, 2013). However, Bergmann and Kopp (2010) suggest that the idiosyncratic production of co-speech gestures may also vary depending on the dialogue situation and the presence of a particular addressee.

Additionally, co-speech gestures tend to be used at higher rates when there is a greater degree of expertise disparity between interlocutors (Holler & Stevens, 2007; Jacobs & Garnham, 2007; Kang et al., 2015), or when the explanandum is physically absent during an explanation (Holler & Stevens, 2007).

Based on these findings on individual gesturing behavior, one of the studies presented in this dissertation investigated the intra-individual variation in the use of gesture deixis by different explainers, specifically in relation to their interpretations of the levels of understanding of three different explainees (see Section 7.2).

Chapter 4

Forms of verbal and nonverbal feedback behavior

4.1 Interactive gaze behavior

Eye gaze behavior is one of the most prominent forms of nonverbal behavior in human–human interaction. Eye gaze facilitates the uptake of the interlocutors' visual attention and emotional expressions, and remains continuously accessible (for sighted individuals) to the interaction partners (for a review, see Hessels, 2020). Although eye gaze is continuously accessible in face-to-face interactions, interlocutors do not maintain eye contact constantly; rather, the direction of the interlocutors' gaze is related to several factors, such as the interlocutors' current role (speaking or listening) and cognitive processes, e.g., attention to other objects or thinking (Argyle & Cook, 1976; Kendon, 1967).

Concerning the visual attention of interlocutors, interactive gaze behavior can be categorized into three forms: *mutual gaze*, *gaze withdrawals*, and *shared gaze*. The term *mutual gaze* refers to moments during which the interlocutors maintain eye contact with each other (Argyle & Cook, 1976; Cook, 1977), whereas the term *gaze withdrawal* refers to moments when the eye contact with the other interlocutor is interrupted (C. Goodwin, 1981, 1985; Rossano, 2013; Rossano, 2012). In contrast to mutual gaze, the term *shared gaze* refers to the time in which interlocutors pay visual attention to the same target, i.e., an object (Argyle & Cook, 1976; Cook, 1977). Shared gaze was not analyzed along mutual gaze in Study 2 investigating the topical changes of board game explanations presented in Section 6.2 because the study focused solely on whether the interlocutors maintain eye contact prior to topic changes or not. Therefore, the remaining part of the literature review does not consider prior research on shared gaze.

The duration of mutual gaze varies across individuals and it has been related to

the interlocutors' interactive role, i.e., speaking or listening (Argyle & Cook, 1976; C. Goodwin, 1981; Kendon, 1967). For instance, listeners tend to gaze at speakers for longer periods (with brief aversions), whereas speakers tend to gaze at listeners less frequently and for shorter periods (J. B. Bavelas et al., 2002; Kendon, 1967). When speakers gaze at listeners, they often do so to elicit verbal or nonverbal feedback, such as head gestures or backchannel responses (Argyle & Cook, 1976; J. B. Bavelas et al., 2002; Brône et al., 2017; Kendon, 1967). Such feedback-seeking gaze behavior typically occurs toward the end of the speaker's utterances, and subsequent gaze withdrawals may signal the completion of a turn (Argyle & Cook, 1976; J. B. Bavelas et al., 2002; Brône et al., 2017). In this relation, gaze behavior has been demonstrated to play an important role in managing turn-taking between interlocutors (Degutyte & Astell, 2021; Jokinen, Harada, et al., 2010; Jokinen, Nishida, & Yamamoto, 2010; Kendon, 1967).

Interactive gaze behavior has also been associated with the management of conversational topics (Rossano, 2013; Rossano, 2012). In an analysis of spontaneous dyadic and triadic face-to-face interactions, using insights from Conversation Analysis, Rossano (2012) found that 84% of the observed instances of mutual gaze were linked to the expansion of current conversational topics. In contrast, 95% of the observed instances of gaze withdrawal occurred just before the closure of conversational topics. Note that the naturalistic interactions analyzed by Rossano (2012) included competing activities, such as eating or playing games while the interlocutors were interacting with each other. In the case of the studies by Lazarov et al. (2024) and Lazarov and Grimminger (under review) (see Chapter 6), the explanatory interactions took place in a way that excluded such competing activities while the explainers and the explainees were engaged with the explanation task.

Beyond turn-taking and the management of topics, addressees' gaze withdrawals (also known in the literature as gaze aversions) have been linked to cognitive processing. For example, gaze aversions are associated with increased cognitive effort among adults (Abeles & Yuval-Greenberg, 2017; Allen & Guy, 1977; Glenberg et al., 1998) and children (Phelps et al., 2006). However, previous studies have demonstrated that gaze aversions improve the individuals' task performance by enabling them to concentrate during mentally demanding tasks (Glenberg et al., 1998; Phelps et al., 2006). Moreover, gaze aversions have been identified as part of the so-called "thinking face" in language processing, often occurring alongside other modalities such as facial expressions and body posture (J. B. Bavelas & Chovil, 2018; M. H. Goodwin & Goodwin, 1986; Heller, 2021). In relation to information processing, gaze aversions have also been shown to support the mental visualization of invisible objects (Markson & Paterson, 2009). Apart from cognitive processing, while providing a dispreferred response expressing the lack of knowledge, addressees' gaze aversions from speakers are used as a face-saving strategy in spontaneous conversations (Pekarek Doehler, 2022).

4.2 Head gestures

In human–human interaction, addressees use head movements as a form of nonverbal feedback–also known as head gestures–which are typically conventionalized, that is, culturally specific (J. B. Bavelas et al., 2002; Kendon, 1970; McClave, 2000). The classification of head gestures varies across cultures. As the present thesis focuses on examples from Western cultures, only a subset of commonly used head gesture categories in these cultures are addressed here.

A very common example of head gestures in Western cultures is the head nodding, kinematically defined as down–up cyclic head movements[1] (Allwood & Cerrato, 2003; McClave, 2000; Włodarczak et al., 2012). Typically, addressees use head nodding to signal understanding, engagement, or approval (Cerrato, 2005; De Stefani, 2021; Gander & Gander, 2020; Włodarczak et al., 2012; Wu & Heritage, 2025). In contrast, head shaking—defined kinematically as side-to-side sweeps—is used to express negation, denial, or disapproval (McClave, 2000; Włodarczak et al., 2012).

These and other head gestures often co-occur with vocal backchannels (see Section 4.3), which may reinforce the communicative function of the gesture and facilitate its interpretation (Allwood & Cerrato, 2003). Although head nods and head shakes convey conventionalized polarity, such as positive or negative feedback, their interpretation can be ambiguous when assessing the addressee's level of understanding (Gander & Gander, 2020).

Additional common head gestures in Western cultures include head tilts (sideways down–up movements), jerks (inverted nods, i.e., up–down movements), and other gestures that form part of the addressee's nonverbal feedback behavior in interaction (Allwood et al., 2007; Kousidis et al., 2013; Włodarczak et al., 2012). Like head nods, head tilts are used for signaling agreement, attention, and cognitive processing (Ismail & Syahputri, 2022). In contrast to head nods and tilts, which can be interpreted ambiguously, jerks are used specifically to signal understanding (Włodarczak et al., 2012). However, Study 1 which investigated the explainees' head gestures in relation to the topical changes in medical explanation focused exclusively on head nods, as other gestures, such as head shakes, were infrequently observed in that particular dataset (see Section 6.1).

[1] A head gesture cycle refers to the complete kinematic performance of a gesture involving more than one movement, from the initial to the final position. For example, a full head nod involves both a downward and a returning upward movement. Head gestures can also be half-cyclic, such as a nod consisting only of a downward movement (Hadar, U. and Steiner, T.J. and Grant, E.C. and Clifford Rose, F., 1983; Kousidis et al., 2013; Włodarczak et al., 2012).

4.3 Vocal backchannels

Vocal backchannels are brief verbal feedback responses that may take either lexical (e.g., *alright, okay, yes*) or non-lexical forms (e.g., *mhm, uh-huh, yeah*) (Allwood et al., 1992; Gardner, 2001; Malisz et al., 2016; Park et al., 2017; Yngve, 1970). Similarly to head gestures, vocal backchannels serve various communicative functions and may be interpreted ambiguously in terms of the addressee's engagement, attention, or understanding, without being turn-taking signals (Gardner, 2001; Park et al., 2017; Yngve, 1970).

For example, backchannel responses such as *uh-huh* or *yeah* may indicate continued attention, perception, or unconditional understanding (Allwood et al., 1992; Clark & Brennan, 1991; Eshghi et al., 2015; Gardner, 2001; Schegloff, 1982). In contrast to positive backchannel responses (e.g., *yes, mhm, okay*), negative responses, such as *no*, tend to be less ambiguous in their interpretation (Allwood et al., 1992; Arnold, 2012).

Chapter 5

Corpora of naturalistic explanations

In this chapter, two video corpora of naturalistic explanations which were used for the studies in Chapters 6 and 7 are presented. The corpora have been compiled for the investigation of different interactional processes, such as the explainees' multimodal signals of understanding and the adaptive verbal and nonverbal behavior of the explainers' in relation to monitoring the explainees' understanding moment by moment over the course of the explanations.

5.1 A corpus of medical explanations

5.1.1 Participants

Seven physicians and thirteen caregivers participated in the study consisting of eleven interactions. Of the seven physicians, four participated each in two interactions. In addition, two of the explanations were attended by two caregivers (both caregivers' behavior was considered in the analysis), and ten of the explanations were attended by a child that was supposed to undergo a medical intervention. However, the behavior of the children was not included in the analysis, as most of them were at an age where they were unlikely to perceive or interpret the medical explanation.

No socio-demographic data about the physicians and the caregivers could be collected because the medical consultations were planned in short term. All participants signed a consent form, and the study was approved by the Ethics Board of the university.

5.1.2 Materials

The corpus contains eleven physician–caregiver explanatory interactions regarding the upcoming surgery of individual children. All interactions were recorded at the pediatric department of a hospital in Germany. In these sessions, the physicians

explained various aspects related to the surgery, such as the diagnosis, necessity, medical procedures, and treatment.

Figure 5.1: Data collection method of the corpus on medical explanations.

Two of the eleven interactions took place in the chief physician's office and were attended by two caregivers each, and they lasted longer–approximately 25 minutes–compared to the duration of the other nine interactions. The remaining nine interactions were conducted by different physicians, each attended by one caregiver. These interactions were recorded in the hospital's outpatient department, and their duration ranged between 04 : 54 and 14 : 23 minutes.

Ten of the eleven explanations were conducted in German, while one was conducted in English. Due to technical limitations related to the camera angles during the data collection—which hindered the observation and the analysis of the caregivers' multimodal behavior—one interaction was excluded from the analysis. Thus, ten interactions involving seven physicians and twelve caregivers were included in the final analysis presented in Section 6.1. The average duration of the ten interactions considered in the analysis was 11:06 minutes ($SD = 7 : 10$ minutes).

5.1.3 Procedure

Two cameras were used to record the interactions, each directed toward one of the participants to capture the face and torso areas of both the physicians and the caregivers. Additional techniques, such as headsets for voice recording or eye-tracking devices, were not employed as their use could have disrupted the natural interaction process and affected the participants' attention, particularly given the sensitivity and importance of the medical discussions taking place.

5.2 The MUNDEX corpus

5.2.1 Participants

The MUNDEX corpus contains data from 119 adult participants, of which 32 were explainers, and 87 were explainees. Among the explainers (age: $M = 24.41$, $SD = 3.49$), eleven were male, 20 were female, and one diverse. Of the 87 explainees, only 62 provided socio-demographic information, including age ($M = 24.16$,$SD = 6.40$), gender (27 male, 35 female), and language (59 German native speakers, 3 German non-native speakers). The lack of socio-demographic data from some explainees was not problematic, as the research questions of the presented studies did not pertain to the participants' age or gender. All participants signed an informed consent form prior to the study, which was approved by the Ethics Board of the university.

For the studies presented in Chapters 6 and 7, randomly selected data from 32 participants was analyzed, including eight explainers and 24 explainees. All eight explainers were German-speaking adults ($M = 23.6$, $SD = 3.38$); two were male and six were female. Of the 24 explainees, 18 provided socio-demographic information, including age ($M = 26.0$, $SD = 9.75$), gender (7 male, 11 female), and native language (all German).

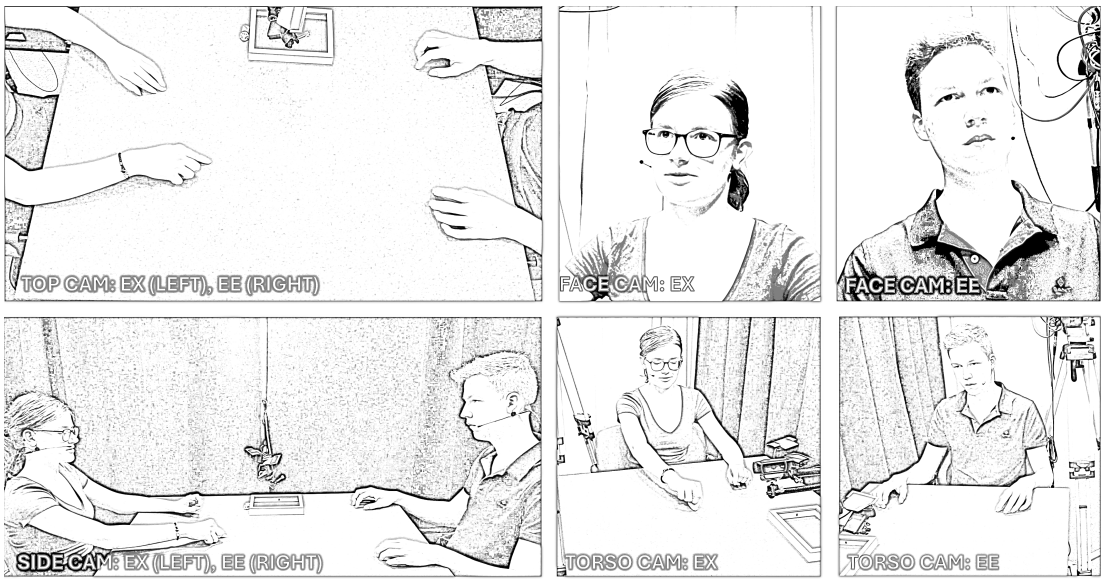

Figure 5.2: Data collection method of the MUNDEX corpus.

5.2.2 Materials

Dyadic board game explanations The video corpus MUNDEX (Multimodal UNDerstanding of EXplanations) (Türk et al., 2023) was compiled to investigate the relation between the explainees' multimodal behavior and their moment-by-moment levels

of understanding (ranging from understanding to non-understanding). MUNDEX comprises 87 dyadic explanatory interactions in which explainers explain a collaborative and competitive board game, named *Deep Sea Adventure* (Sasaki & Sasaki, 2014), to explainees in German.

In accordance with the corpus design, each explainer explained the board game to three (or two) different explainees subsequently. Each explanatory interaction included three phases that varied temporally depending on the individual interactions: (1) the board game being physically absent, (2) the board game being physically present, and (3) an interactive game play between the two interlocutors.

For the studies presented in Chapters 6 and 7, a subsample of 24 explanations from MUNDEX was randomly selected, and only the phase in which the board game was physically absent from the shared space was considered for analysis. The mean duration of the 24 explanatory interactions (including all three phases) was 26:49 minutes ($SD = 05:30$ minutes), whereas the mean duration of the phases in which the game was physically absent was 07:04 minutes ($SD = 03:44$ minutes).

The explanatory interactions were recorded using six camera perspectives: two cameras directed at the interlocutors' face area, two at their torso area, one side-positioned camera, and one top-mounted camera (Figure 5.2). In addition, the participants' speech was recorded using head-mounted microphones for subsequent speech analysis.

Video recall In accordance with the purpose of the corpus, i.e., to investigate the relation between the explainees' multimodal behavior and their levels of understanding moment-by-moment, a secondary corpus was compiled by conducting a post hoc video recall procedure[1].

The video recall corpus contains 174 audio reports about the explainees' levels of understanding recorded by the 32 explainers (each two or three times) and by 87 the explainees (each once). Each audio recorded recall was taken simultaneously while the screen on which the participants watched the explanatory interactions was also captured. Coupling the audio recordings with the screen recordings allowed the detection of the moments to which the explainers and the explainees referred during the video recall task.

[1]Video recall is an established method in psychology for self-observations and for assessing the subjective interpretation of human social behavior by subsequently presenting videotaped interactions to participants (Welsh & Dickson, 2005). The presentation aims at stimulating participants' memory about specific events from an interaction (Kuusela & Paul, 2000).

5.2.3 Procedure

Conducting the board game explanations All participants were randomly recruited and assigned to one of two roles: *explainers*, who were responsible for conveying the rules of a board game, and *explainees*, who received the explanation. The explainers were given the opportunity to learn and practice the board game independently a few days prior to the study. In contrast, the explainees were not informed in advance about which board game would be explained to them.

To ensure that the interactions remained as natural as possible, and thus resembling everyday communicative exchanges, the explainers and the explainees were required to be unfamiliar with each other prior to the study. The reason for this requirement was that the project for which the corpus was compiled aims at modeling explanations given by an artificial agent to a human explainee while the artificial agent monitors the (non-)understanding of the human explainee and adapts the explanation accordingly. To achieve this goal, the project first focused on gaining insights from studies on human–human explanations by compiling a video corpus containing rich data of human multimodal behavior that could be related to different levels of understanding. In addition, to avoid bias in the behavior of the explainers and the explainees, none of them were informed about the goal of the study.

The only specific instructions provided to the explainers were (1) to follow a prescribed sequence when to introduce the board game to the explainees and (2) to ensure that their explanations were sufficiently comprehensive to enable the explainees to play the game independently afterward.

Conducting the video recall task Following the dyadic interactions, each explainer and each explainee took part in a video recall task, in which they were asked to give short verbal reports about the explainees' understanding moment-by-moment while watching the recorded interactions captured by the side camera. Each participant did this task on their own, i.e., the experimenters did not interfere with the participants' selection of moments or the content of their reports about the explainees' understanding. The participants were instructed to reflect on any moment from the videotaped interactions which they considered relevant regarding any changes in the explainees' levels of understanding. The explainees reported on their own understanding from a first person perspective, whereas the explainers reported on the explainees' understanding from a second person (interpretative) perspective. Further, the participants were instructed to stick to the key terms *understanding*, *partial understanding*, *non-understanding*, and *misunderstanding* in their reports as much as possible.

The participants watched the videotaped explanations using a custom designed media player on which they could pause the video with a single button press. An additional microphone was connected to the computer to record the participants'

reports. This automatically created video-aligned empty segments with the length of 1 second. The empty segments were imported into ELAN, where human annotators manually transcribed the participants' reports. For the duration of the video recall task, the participants could pause and play the videos every time they wanted to comment on further understanding-related events. The explainees participated once in the video recall task, directly after their explanatory interaction with the explainers. The explainers participated in the video recall task two or three times (depending on the number of explainees they interacted with) in one go, after completing all explanations. In this way, we prevented their naturalistic explaining behavior being influenced by the video recall task in subsequent explanations. A detailed presentation of the video recall method is provided in the article by Lazarov, Schaffer, et al. (2025).

Chapter 6

Studies on the topical structure of explanations

In this chapter, two studies that specifically focused on the topical structure of explanations as a verbal form of explaining behavior will be presented. The topical structure of the explanations was examined in terms of topic changes initiated by different explainers, which were related to different forms of explainees' feedback behavior as monitored by the explainers throughout the course of the explanatory interactions.

The forms of explainees' feedback behavior related to the topical structure of explanations comprised the explainees' interactive gaze behavior, head gestures, and vocal backchannels. The first study (see Section 6.1) addressed the domain of medical explanations, while the second study addressed the domain of board game explanations (see Section 6.2).

Although both studies investigated the same overarching research topic, each analysis explored different aspects of the interactive processes underlying changes in the topical structure of explanations.

6.1 Study 1: The topical structure of medical explanations

This section presents the findings from a study by Lazarov et al. (2024) (see Appendix), in which the changes in the topical structure of medical explanations were examined in relation to different forms of multimodal feedback behavior of explainees. The analysis was based on a subsample of ten naturalistic physician–caregiver explanatory interactions. The analyzed explanations were given in preparation for the upcoming surgeries of individual children, for which the caregiver(s) consent was required.

Specifically, the study investigated the caregivers' feedback behavior that preceded two types of topical changes that were initiated by the physicians: transitions to

elaborations (T-EL) and transitions from elaborations to new topics (T-NT). Due to the limited amount of prior research on this specific topic, the study followed an exploratory approach employing two analytical methods: conditional probabilities and a statistical regression model.

6.1.1 Method

For the analysis of the present study, a subsample of 10 naturalistic medical explanations was selected (for further details about the subsample, see Section 5.1.

Data coding

The ten physician–caregiver interactions were annotated using the software ELAN (Wittenburg et al., 2006). First, the interlocutors' speech was manually transcribed, segmented, and annotated into explanation episodes. Subsequently, the caregivers' eye gaze behavior, head gestures, and vocal backchannels were also manually annotated on separate tiers. A model of the annotation procedure is provided in Figure 6.1, along with an example in Table 6.1, following the description of the coding procedure.

Explanation topics For the coding of explanation topics, the definition by Roscoe and Chi (2008), discussed in Section 3.1, was applied. For the purposes of the study, the explanation topics were annotated based on whether they constituted new topics or the elaborations of previously introduced topics. This differentiation was essential, as the study focused on transitions to elaborations and transitions from elaborations to new topics.

A segment was annotated as a new topic when the information conveyed in the utterances had not been previously introduced. A segment was annotated as an elaboration when the information was repetitive, paraphrased, or added further detail related to a previously introduced new topic. To distinguish between the two topical categories, a numerical system was applied: new topics were annotated using whole numbers (e.g., 1, 2, 3, etc.), while elaborations of previous topics were annotated using decimal extensions (e.g., 1.1, 1.2, 1.3, etc.) (see Figure 6.1, and Table 6.1). This system allowed each elaboration to be clearly associated with the new topic it expanded upon.

The point at which two segments met was defined as a transition point. Two types of transitions were extracted from the annotated data: (1) transitions to elaborations (T-EL), including both transitions from new topics to elaborations and transitions between elaborations, and (2) transitions from elaborations to new topics (T-NT). These transition points were used both to track the structural changes in the explanations and to define the temporal window for collecting the explainees' multimodal behaviors related to each transition type. It is important to note here that the segmentation of

explanation topics in this study did not include semantic labeling, as this aspect was beyond the scope of the research question.

In the first step of the annotation procedure, the coders identified all new topics in each of the ten physician–caregiver interactions. Once these were identified, they proceeded to identify the corresponding sub-categories, i.e., the elaborations which were linked to those new topics. For example, in Table 6.1, the physician introduces a new topic about the child's nutrition habits after recovery (Topic 1). In relation to this topic, two elaborations are initiated and annotated as Topic 1.1 and Topic 1.2. Following these elaborations, the physician introduces a new topic about regular check-ups following the surgery (Topic 2).

Two trained coders annotated the explanation topics across the ten interactions. Their annotations were then reviewed and compared by two additional inter-raters. Inter-rater reliability for the coding of explanation topics was assessed using the Krippendorff's α for small sample sizes (Krippendorff, 2004, 2013), resulting in a satisfactory value of $\alpha = 0.85$.

Explainees' feedback behavior For the annotation of the explainees' feedback behavior, the coders annotated the caregivers' eye gaze behavior, head gestures, and vocal backchannels. Initially, two independent coders pre-annotated the caregivers' feedback behavior. These annotations were then reviewed by two additional coders who were specifically and intensively trained for the coding procedures used in the current study. As a result, all instances of the caregivers' behavior were double-coded, each by two independent coders.

To assess inter-rater reliability, the first three minutes of three physician–caregiver interactions, which were annotated by all four coders, were analyzed using Krippendorff's 2004 reliability test. The resulting inter-rater reliability scores are reported in the corresponding subsections on the annotation of the caregivers' eye gaze direction, head gestures, and vocal backchannels.

- *Gaze behavior*. The two camera perspectives, each directed toward each of the interlocutors, allowed the manual annotation of their eye gaze behavior according to the gaze direction, i.e., toward the interlocutor, the materials, and away. The inter-rater reliability test indicated a satisfactory rate ($\alpha = 0.94$). Although the caregivers' gaze was initially coded according to the direction, the current study focused on the form of interactive gaze behavior, not the viewing target. For this purpose, the non-changing gaze directions of the caregivers were coded as *static gaze*, whereas the changes of the caregivers' gaze directions were annotated as *gaze shifts*. The two categories were mutually exclusive.
    - *Static gaze*: The caregivers' gaze was coded as static if there was no change

of the viewing direction before a topical transition.

- *Gaze shifts*: The caregivers' gaze was coded as shifting if a change of the viewing direction occurred before a topical transition. Based on previous research on gaze aversion (Glenberg et al., 1998; Phelps et al., 2006) (see Section 4.1), this category was subdivided into *gaze shifts with aversion* and *gaze shifts without aversion*. Gaze shifts with aversion were coded simply as *gaze aversions*.
  - *Gaze aversions*: Changes of the viewing direction, particularly from the interaction partner, toward the explanation-related materials or away from the shared space.
  - *Gaze shifts without aversion*: Any other changes of the gaze direction, e.g., from the explanation-related materials away or vice versa.

- *Head gestures.* The caregivers used predominantly head nods during the temporal segments of interest, i.e., around the investigated types of topical transitions. Other head gesture categories, such as head shakes were observed only fourteen times in relation to the analyzed topical transitions across the ten medical explanations. Therefore, only the caregivers' head nods were considered for the analysis. The caregivers' head nods were annotated as single or repeated up and down head movements (Allwood & Cerrato, 2003; McClave, 2000; Włodarczak et al., 2012). However, the annotations did not include details about the repeatedness of the head nods. The inter-rater reliability test for the coding of head gestures indicated a satisfactory rate ($\alpha = 0.81$)

- *Vocal backchannels.* The caregivers' short feedback signals, such as *okay, yes, mhm, alright*, etc, were annotated as backchannels (Allwood et al., 1992; Yngve, 1970). However, the backchannels were not annotated with respect to their form or function, as most of the backchannels were continuers or short affirmations.

Data analysis

Forms of the explainees' multimodal behavior Because the physician–caregiver explanatory interactions took place face-to-face, the caregivers' eye gaze behavior–categorized as static gaze, gaze shifts without aversion, and gaze shifts with aversion–was continuously observable. Given the study's focus on multimodality, the caregivers' behavior was annotated based on the type of gaze behavior and any co-occurring modalities (i.e., head nodding and / or backchanneling).

For instance, if a caregiver's static gaze occurred without any co-occurring modality, it was annotated as unimodal static gaze. If the static gaze co-occurred with either head nodding or backchanneling, it was annotated as bimodal static gaze with head

Table 6.1: Study 1: Example of an explanation about the post-treatment care.

Topic Nr.	Utterance	Annotation
1	*Later, as said, there will be no restriction regarding eating.*	a new sub-topic
1.1	*She can eat completely normally, if you also want fast food, gummy bears or similar, right?" – an elaboration of the sub-topic*	an elaboration of the sub-topic [1]
1.2	*You are not doing this. She can eat healthily. She has a normal gastrointestinal tract.*	a second elaboration of the sub-topic [1]
2	*And what we should do, of course, are check-ups with ultrasound.*	an introduction of a new sub-topic

nodding or bimodal static gaze with backchanneling, respectively[1]. When static gaze co-occurred with both head nodding and backchanneling simultaneously, it was annotated as multimodal static gaze[2].

Since the study distinguished three types of gaze behavior (static gaze, gaze shifts without aversion, and gaze aversions) and four possible combinations of multimodal co-occurrence, a total of twelve distinct forms of multimodal behavior of the caregivers were identified. These annotated forms of behavior served as the basis for the subsequent statistical analysis.

Temporal threshold To analyze the caregivers' forms of multimodal behavior in relation to the two types of topical changes initiated by the physicians, their behaviors were examined with respect to their temporal occurrence within one second before a topic change was initiated (see Figure 6.1). As there is no established temporal window for analyzing the explainers' responses to the explainees' multimodal feedback—specifically in terms of changing the explanation topic or elaborating on a previous topic—a temporal window of one second was chosen as an exploratory approach.

Within this one-second window, the onset of the caregivers' behavior was considered. For example, a transition from Topic 1 to Elaboration 1.1 (see Figure 6.1) was coded as a transition from a new topic to an elaboration (T-EL), and it was preceded by a caregiver's multimodal behavior comprising static gaze, head nodding, and backchanneling. However, the subsequent transition from Elaboration 1.1 to Elaboration 1.2 was coded as being preceded by only the caregiver's static gaze, since the onset of the head nod and the backchannel occurred after the transition point.

For the analysis of the caregivers' multimodal behavior preceding the transitions to elaborations and the transitions from elaborations to new topics, all instances in

[1]The term *bimodal* refers to the use of two modalities.
[2]The term *multimodal* refers to the use of three modalities.

which either transition type was preceded by one of the twelve identified forms of multimodal behavior within the one-second threshold were extracted.

In the statistical analysis, the occurrences of transitions to elaborations (hence T-EL) and transitions from elaborations to new topics (hence T-NT) were treated as the response variable, while the caregivers' forms of multimodal behavior served as the fixed effect.

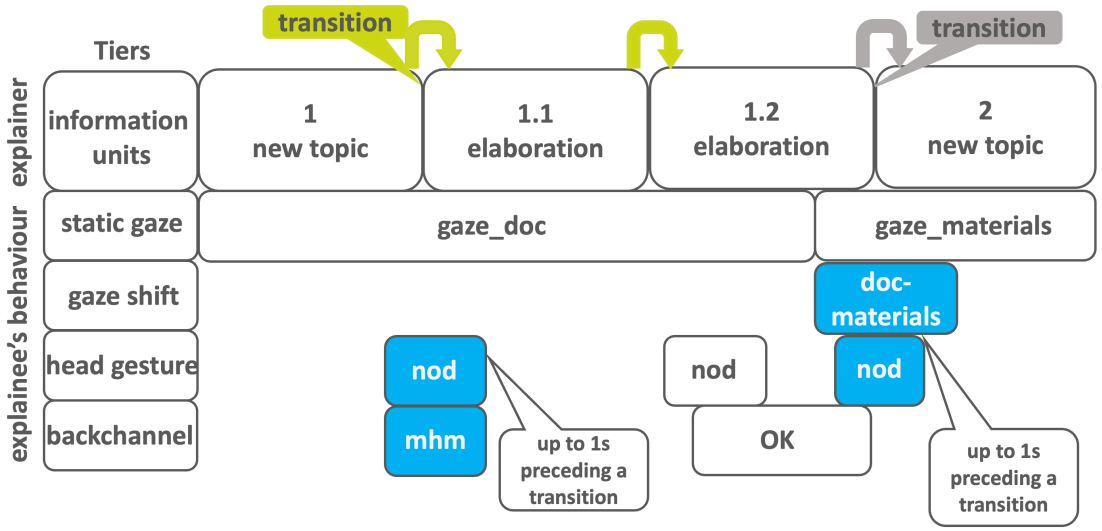

Figure 6.1: Study 1: Annotation procedure.

Note: The figure illustrates the annotation procedure in ELAN. The numbering of topics refers to the order of appearance. Full numbers represent new topics, and decimal numbers represent the category "elaborations". The boundaries between the topical segments are the actual transition points. The green colored transitions are those to elaborations, and the gray-colored transitions are those from elaborations to new topics. The light-blue colored gaze shifts, head nods and backchannels are those co-occurring in the areas of interest, i.e., up to one second before transitions into elaborations or from elaborations to new topics. Modalities in other temporal relations were not analyzed.

Conditional probabilities The first goal of the study was to examine the absolute frequencies and the conditional probabilities of the caregivers' forms of multimodal behavior occurring prior to T-EL, as well as those occurring prior to T-NT, both of which were initiated by the explaining physicians. For this analysis, the formula for conditional probability proposed by Dekking et al. (2005) was applied. This formula defines the probability of an event A, given that event B has occurred:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{\#(A \cap B)}{\#\Omega}}{\frac{\#B}{\#\Omega}} = \frac{\#(A \cap B)}{\#B}$$

For example, to calculate the conditional probability for T-EL following a caregiver's bimodal feedback behavior, the numerator and the denominator from the formula were replaced by absolute frequencies:

- $\#(A \cap B)$ – the frequency of T-EL initiated after a form of multimodal behavior, e.g., bimodal behavior with gaze aversions and head nodding (GAN)

- #B - the frequency of all transitions (T: T-EL & T-NT) initiated after caregivers' bimodal behavior with gaze aversions and head nodding (GAN)

This is an example of the application of the formula for conditional probability in the analysis:

$$P(\text{T-EL} \mid \text{GAN}) = \frac{\text{T-EL after GAN}}{\text{all T after GAN}}$$

The example is interpreted as follows: The probability that physicians initiate T-EL following the caregivers' bimodal behavior with gaze aversions and head nodding results from the division of the instances of T-EL following bimodal behavior with gaze aversions and head nodding by all transitions following bimodal behavior with gaze aversions and head nodding across the ten interactions. For all twelve forms of the caregivers' multimodal behavior, the conditional probabilities were calculated separately.

Statistical analysis The second objective of the study was to assess the effect of the different forms of the caregivers' multimodal behavior on the topical transitions initiated by the explaining physicians. A Shapiro–Wilk test indicated that the data for the response variable was not normally distributed ($W = 0.82$, $p < 0.01$). Therefore, a Generalized Linear Mixed Model (GLMM) was used for the statistical analysis.

In the model, the caregivers' forms of multimodal behavior (12 levels) and the topical transitions initiated by the physicians (2 levels) were included as two interacting fixed effects. Additionally, the model included the individual physicians, each interacting with different caregivers, as a random effect. This structure enabled the analysis of variance at the level of individual explainers across different interaction partners.

Following the model summary, additional pairwise comparisons were conducted to examine the contrasts between the two types of transitions for each of the twelve forms of multimodal behavior. To enhance interpretability, the frequencies of the transitions for each form of the caregivers' multimodal behavior were converted into proportions, calculated separately for each interaction. This allowed for normalization of the variation across the individual physician-caregiver interactions. All statistical analyses were performed using the software *RStudio* (RStudio Team, 2020).

In some of the ten interactions–and for certain transition types following specific forms of multimodal behavior–there were instances where the frequency was zero. As a result, some proportion values were 0.00 (0%) or 1.00 (100%). Such extreme proportions

often lead to errors when fitting statistical models, as they may be misinterpreted as exact data points in a binomial regression.

To address this issue, a zero-inflation procedure was applied to the response variable, following the approach described by Tang et al. (2023). Specifically, proportions of 0.00 were replaced with 0.000001, and proportions of 1.00 were replaced with 0.999999.

Consequently, the GLMM was adapted using the glmmTMB (Template Model Builder) framework, as described by Brooks et al. (2017). The statistical model was specified as follows:

```
glmmTMB(PROP_adjusted ~ BEHAVIOR * TRANSITION + (1 | EX/EE),
data = dataframe, family = beta_family())
```

In the model, `PROP_adjusted` are the adjusted proportions of the response variable, the caregivers' forms of multimodal behavior and the two types of transitions initiated by the physicians were defined as two interacting fixed effects (indicated with an asterisk), and the random effect represented the different caregivers (EE) nested within the different physicians (EX) `EX/EE`.

6.1.2 Results

Conditional probabilities

In total, 502 instances of topical changes initiated by the physicians were observed across the ten physician–caregiver interactions. The overall frequency of T-EL transitions ($n = 289$) was higher than the overall frequency of T-NT transitions ($n = 213$). It should be noted that the T-EL category included not only transitions from new topics to elaborations but also transitions occurring between elaborations.

The conditional probabilities are presented alongside the corresponding absolute frequencies (indicated in brackets) in Figure 6.3. The y-axis displays the twelve forms of the caregivers' multimodal behavior (see Note in Figure 6.3), and the x-axis represents the conditional probabilities of each form of behavior preceding either a T-EL or a T-NT transition.

Initial insights from the analysis suggest that the most frequently initiated transitions by the physicians overall occurred following the caregivers' static gaze, as well as combinations of static gaze with head nodding and backchanneling. In contrast, the lowest number of transitions was observed in relation to the caregivers' multimodal gaze shifts (without aversions). For the caregivers' unimodal behavior displayed by static gaze, T-EL were more likely to follow than T-NT (see Figure 6.3). A similar trend was observed for unimodal gaze shifts (without aversions) and for bimodal co-occurrences of gaze shifts (without aversions) and head nodding.

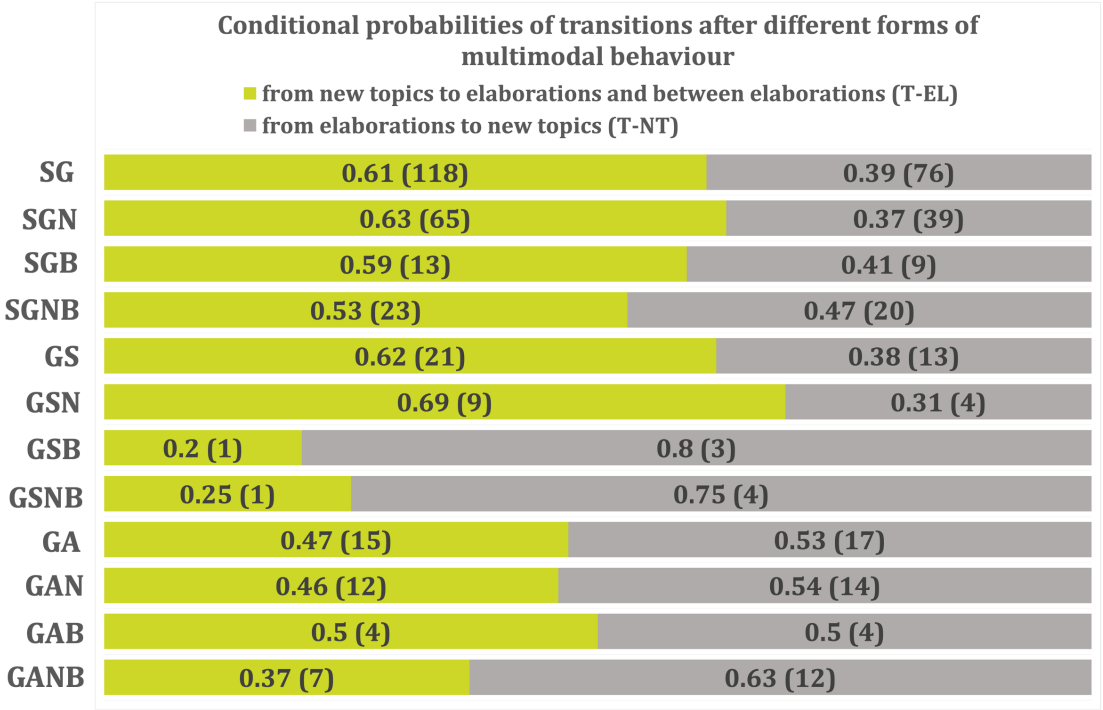

Figure 6.2: Study 1: Conditional probabilities and absolute frequencies of transitions after different forms of multimodal behavior.

Note: SG - static gaze; GS – gaze shift (without aversion); GA – gaze aversion; N – head nodding; B - backchanneling

In contrast, the caregivers' bimodal behavior involving gaze shifts (without aversions) and backchanneling, as well as multimodal behavior involving head nodding and backchanneling, were more likely to be followed by T-NT than by T-EL. For gaze aversions–whether unimodal or bimodal with either backchanneling or head nodding–the probabilities for both types of transitions were nearly equal. Only for multimodal gaze aversions, a noticeably higher probability of T-NT compared to T-EL was observed.

Statistical model

Overall, the glmmTMB model indicated a slightly better fit for the dataset ($AIC = -2292$, $BIC = -2198$) compared to a null model without the fixed effect. It also showed a high proportion of variance when considering both fixed and random effects across the ten interactions (Conditional $R^2 = 0.95$). However, minimal variability in the response variable was observed at the level of individual caregivers nested within physicians ($\sigma^2 < 0.001$, $SD < 0.001$), whereas greater variability was evident at the level of the individual physicians ($\sigma^2 = 0.03$, $SD = 0.17$). These findings suggest that the proportional variance of the topical transitions across the ten physician–caregiver interactions is more strongly associated with the individual explaining behavior of the physicians than with the variation of multimodal feedback behavior across the

different explainees.

The summary of the fixed effects (see Table 6.2) reports only those forms of the caregivers' multimodal behavior that indicated a statistically significant effect on the transitions initiated by the physicians. A significant effect on the T-EL was observed for the caregivers' bimodal behavior comprising their gaze shifts (without aversions) co-occurring either with head nodding or with backchanneling, their multimodal behavior comprising gaze shifts (without aversions), as well as their bimodal behavior comprising gaze aversions with backchanneling, and their multimodal behavior involving gaze aversions.

Table 6.2: Study 1: Forms of the caregivers' multimodal behavior with significant effect on the transitions initiated by the physicians.

Behavior	**Transition**	*M*	*SD*	β	*SE*	*z*	*p*
SG (Int.)	T-EL	0.63	0.09	0.11	0.44	0.25	$> .05$
GSN	T-EL	0.22	0.35	-1.41	0.59	-2.40	$< .05$
GSB	T-EL	0.03	0.10	-1.69	0.58	-2.92	$< .01$
GSNB	T-EL	0.05	0.16	1.71	0.58	2.94	$< .01$
GAB	T-EL	0.08	0.25	-1.68	0.58	-2.89	$< .01$
GANB	T-EL	0.15	0.27	-1.45	0.59	-2.48	$< .05$
GANB	T-NT	0.63	0.43	2.24	0.85	2.64	$< .01$

Note: SG – static gaze; GS – gaze shift (without aversion); GA – gaze aversion; N – head nodding; B – backchannelling; (Combinations of the abbreviations refer to bimodal or multimodal behavior.) T-EL – transitions to elaborations; T-NT – transitions from elaborations to new topics

Among the listed forms of the caregivers' feedback behavior in Table 6.2, only the caregivers' multimodal behavior with gaze aversions had a significant effect on the T-NT. The parameter estimates of the other forms of multimodal feedback behavior indicate that the physicians were more likely to initiate T-EL following the caregivers' multimodal behavior with gaze shifts (without aversions), and more likely to initiate T-NT following the caregivers' multimodal behavior with gaze aversions. In contrast, the likelihood that T-EL were initiated after the caregivers' bimodal behavior with gaze shifts (without aversions) co-occurring with head nodding or backchannelling, and that T-EL were initiated after the caregivers' bimodal behavior with gaze aversions and backchanneling decreased from the intercept in a negative direction. Thus, only the caregivers' multimodal behavior with gaze shifts (without aversions) predicted an increase of T-EL whereas caregivers' multimodal behavior with gaze aversions predicted an increase of T-NT.

To compare the proportions of T-EL with those of T-NT for each form of multimodal behavior, a post-hoc Tukey test for pairwise comparisons was conducted. Table 6.3 summarizes the estimated means (β) for the pairs of proportions that indicated significant differences.

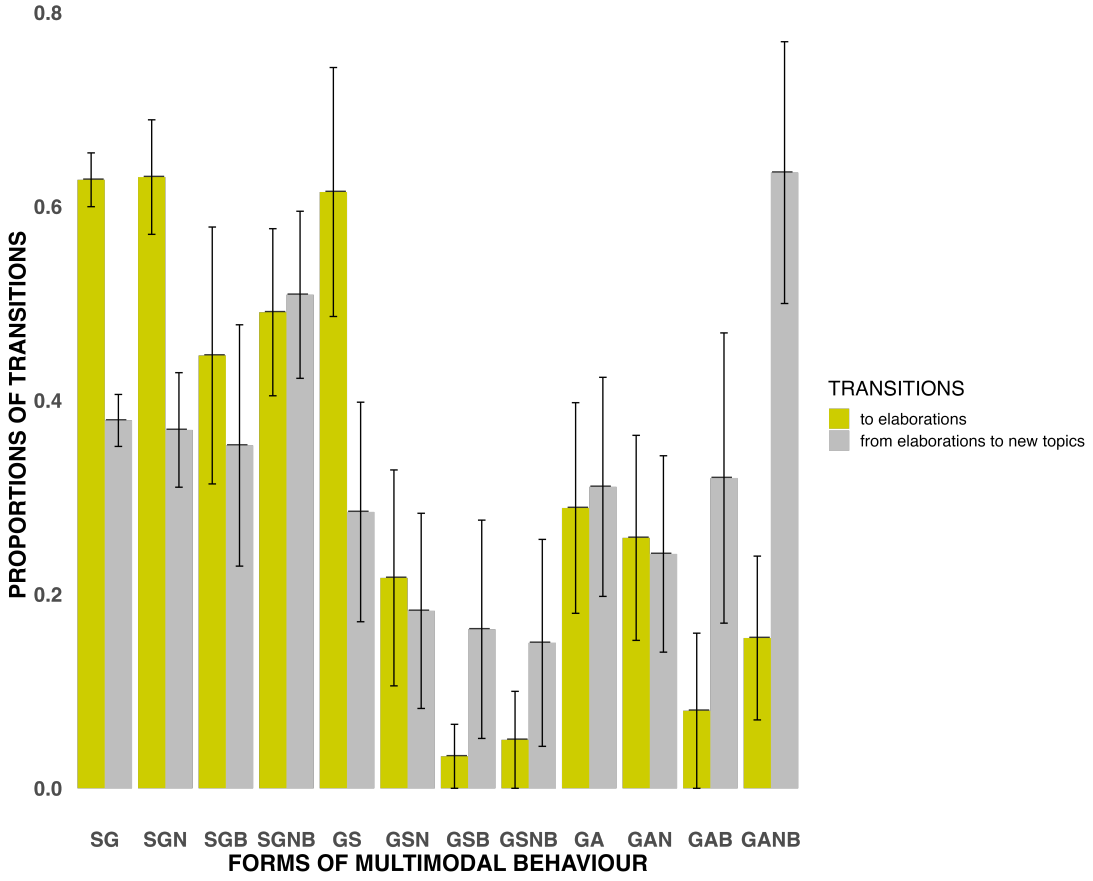

Figure 6.3: Study 1: Proportions of transitions initiated by the physicians for each form of caregivers' multimodal behavior.

Note: SG - static gaze; GS – gaze shift (without aversion); GA – gaze aversion; N – head nodding; B - backchanneling

Table 6.3: Study 1: Summary of estimated means for proportions of transitions which indicated significant differences per form of multimodal behavior.

Behavior	**Transition**	*M*	*SD*	β	*SE*	*LCL*	*UCL*
GS	T-EL	0.61	0.41	0.53	0.44	-0.33	1.39
GS	T-NT	0.28	0.36	-0.89	0.41	-1.70	-0.08
GANB	T-EL	0.15	0.27	-1.34	0.39	-2.12	-0.57
GANB	T-NT	0.63	0.43	0.70	0.42	-0.13	1.53

Note: SG – static gaze; GS – gaze shift (without aversion); GA – gaze aversion; N – head nodding; B – backchannelling; (Combinations of the abbreviations refer to bimodal or multimodal behavior.) T-EL – transitions to elaborations; T-NT – transitions from elaborations to new topics

Significant differences between the proportions of T-EL and T-NT were found for the caregivers' unimodal gaze shifts (without aversions) ($\beta = 1.42$, $SE = 0.60$, $z = 2.37$, $p < 0.05$), and for their multimodal behavior comprising gaze aversion, head nodding, and backchanneling ($\beta = -2.04$, $SE = 0.57$, $z = -3.55$, $p < 0.001$). These

findings suggest that T-EL were more strongly associated with unimodal gaze shifts (without aversions) than T-NT. Conversely, T-NT were more strongly associated with multimodal behavior involving gaze aversions, head nodding, and backchanneling, compared to T-EL.

6.1.3 Summary

The results of the exploratory analysis, examining how different forms of explainees' multimodal behavior were related to transitions to elaborations and transitions from elaborations to new topics initiated by different explainers, revealed two major findings:

1. The explainees' multimodal behavior comprising static gaze was more likely to be followed by transitions to elaborations than by transitions from elaborations to new topics. A similar pattern can be assumed for the explainees' unimodal behavior with gaze shifts (without aversions), as well as for the co-occurrence of gaze shifts (without aversions) and head nodding.

2. The explainees' multimodal behavior comprising gaze aversions from the explainers, in combination with head nodding and backchanneling, was more likely to be followed by transitions from elaborations to new topics. This relation was evident in both the conditional probability analysis and the results of the statistical model.

In sum, this study provided essential insights into the relation between the multimodal form of explainees' feedback behavior and the changes in the topical structure of explanations initiated by explainers. Further, the analyses of conditional probabilities and the statistical model demonstrated that multimodality—that is, the co-occurrence of additional modalities such as head gestures and backchanneling—has a significant effect on the type of the topical transitions initiated by the explainers. However, this study focused exclusively on the explainees' eye gaze behavior and did not take the explainers' gaze behavior into account. This limitation is addressed in the second study, presented in Section 6.2.

6.2 Study 2: The topical structure of board game explanations

In the second study, the topical structure of explanations was investigated by addressing the limitation highlighted in the summary of the previous study (see Section 6.1.3).

In this regard, the second study by Lazarov and Grimminger (under review) (see Appendix) investigated the relation between interactive gaze behavior and topic changes, addressing the domain of board game explanations. In this study, changes of the topical structure of explanations were analyzed in relation to two specific forms of interactive gaze behavior: mutual gaze and gaze withdrawals—previously investigated in spontaneous everyday conversations by Rossano (2012) (for a review, see Section 4.1).

Drawing on previous research on the relation between mutual gaze, gaze withdrawals, and the duration of conversation topics (Rossano, 2013; Rossano, 2012), as well as considering the findings of the exploratory study by Lazarov et al. (2024) presented in Section 6.1, the second study adopted a theory-driven approach. Motivated by this research, the study aimed to validate **two hypotheses**:

1. The proportion of gaze withdrawals occurring prior to topic changes is higher than the proportion of mutual gaze.

2. Gaze withdrawals are initiated more frequently by the interlocutor who initiates a topic change than by the other interlocutor who does not initiate a topic change.

As stated in the first hypothesis, the overall occurrences of mutual gaze and gaze withdrawals between the interlocutors were analyzed as proportions. A proportional analysis was chosen to ensure that the distribution of these two forms of interactive gaze behavior could be compared to the findings revealed by Rossano (2013) and Rossano (2012).

The prediction that gaze withdrawals would occur more frequently than mutual gaze prior to topic changes in explanations was motivated by findings of the study by Rossano (2012), who reported that 95% of the observed instances of gaze withdrawals were associated with the closure of conversation topics, whereas only 5% of observed instances of mutual gaze were associated with the closure of conversation topics.

Nonetheless, the potential occurrence of mutual gaze prior to topic changes in explanations was also hypothesized, based on the assumption that the speakers (here, the explainers) tend to briefly gaze at the explainees to elicit short feedback responses (Argyle & Cook, 1976; J. B. Bavelas et al., 2002; Brône et al., 2017; Kendon, 1967). Meanwhile, drawing from previous research on interactive gaze behavior demonstrating that addressees maintain prolonged gaze at speakers (Argyle & Cook, 1976; J. B. Bavelas et al., 2002; Glenberg et al., 1998; C. Goodwin, 1981; Kendon, 1967), and that in explanations, the explainees are typically the less verbally active interlocutor (predominantly addressees) (Fisher et al., 2022), it is also expected that explainees are the interlocutors who maintain prolonged gaze toward the explainers prior to topic changes. Regarding the explainees' gaze behavior prior to topic changes, Lazarov et al.

(2024) demonstrated that, despite the lack of significant relations, explainees' prolonged gaze at the explainers or the shared materials can occur prior to the introduction of new topics.

The second hypothesis focused specifically on the gaze withdrawals by the two interlocutors regarding their occurrence prior to topic changes. The hypothesis also addressed the relation between the initiator of a gaze withdrawal (e.g., the explainer, the explainee, or both) and the initiator of a topic change (e.g., the explainer or the explainee). Since the initiation of a topic change is inherently linked to the formulation of an utterance, the speaker at the moment of the topic change was considered the initiator.

Based on previous research showing that gaze withdrawals frequently precede the formulation of speech in human–human interaction (Allen & Guy, 1977; Argyle & Cook, 1976; J. B. Bavelas et al., 2002; C. Goodwin, 1981; Kendon, 1967), it was hypothesized that a similar pattern would be observed with respect to the initiation of explanation topics. Specifically, it was assumed that the explainers would briefly withdraw their gaze from the explainees before initiating a topic change and that the explainees would do the same before initiating a topic change themselves. The relation between the initiation of topic changes and the interlocutors' gaze withdrawals is associated with the formulation of spoken utterances. Spoken utterances can be split by semantic boundaries that represent interdependently analyzable segments (Haas et al., 1999). These segments in the case of the current study are the game rules (see Section 6.2.1).

In contrast to the relation between gaze withdrawals and the formulation of utterances, the study by Lazarov et al. (2024), presented in Section 6.1, demonstrated that the explainees also withdraw their gaze from the explainers even when they are in the listener role and the explainer is speaking. Hence, the explainees' gaze withdrawals occurring prior to topic changes could be related to signaling cognitive processing (Abeles & Yuval-Greenberg, 2017; Allen & Guy, 1977; Glenberg et al., 1998).

6.2.1 Method

For the analysis of the present study, a subsample of 24 board game explanations was randomly selected (for details, see Section 5.2).

Data coding

For the present study, different coders segmented the explanations into explanation episodes and annotated both the explainers' and the explainees' verbal and nonverbal behaviors. Each coding procedure was carried out in accordance with an annotation manual (for details, see Lazarov, Türk, et al., 2025).

Prior to the main annotation phase, the coders completed trial annotations (10% of the data), after which inter-rater reliability tests were conducted. All video data were annotated in ELAN (Wittenburg et al., 2006). An overview of the entire annotation procedure is illustrated in Figure 6.5.

Explanation topics The segmentation of the board game explanations into distinct explanation topics was based on previous methodological approaches outlined by Klein (2009) and Roscoe and Chi (2008) (see Section 3.1). The list of various sub-explananda (i.e., sub-topics) anticipated to emerge within the explanations was derived from the official instructions of the game *Deep Sea Adventure* (Sasaki & Sasaki, 2014), which served as the primary preparatory material provided to the explainers prior to the study. The list of explanation topics was organized according to the following structure:

- *Introduction* - Introducing the main explanandum, i.e., the name of the board game and the type of the game, e.g., a collaborative and competitive.
- *Preparation* - Explaining the overall structure and the components constituting the entire game, e.g., a submarine, an oxygen bottle, treasure chips, dices, etc., and their formal and functional features.
- *Goal* - Explaining the main goal of the game, i.e., to collect as many treasures as possible and return successfully to the submarine.
- *Turn progressions* - Explaining the players' turns in the game, including rolling the dice, collecting treasures, and the related following consequences for the players, such as the reduction of oxygen and steps.
- *End of a round* - Explaining the conditions according to which each round comes to an end.
- *End of the game* - Explaining the conditions according to which the entire game is announced to be over and which player wins or loses the game.

To represent this hierarchy more clearly and to facilitate the annotation, the different explanation topics (and sub-topics) were numbered following an approach similar to the approach suggested by Fisher et al. (2023). As shown in the schematic illustration of the annotation scheme, topic groups 2, 4, and 5 contain sub-topics. For these topic groups, the annotators were instructed to code the corresponding sub-topics (e.g., 2.1, 4.1, 5.1, etc.). Segments of explanation topics were annotated at the level of full utterances or parts of utterances, depending on the semantic structure.

Examples of coded explanation topics introduced by the explainers:

(1) Introduction: "We are going to play a little game, and I hope you enjoy diving (pause) because we are going to dive now. (pause) And, I am going to explain you how the game works (pause) without you see how the game looks. (pause)"

(2.1) Game framing structure: "And, (pause) the game consists of (pause) three rounds which are overall one."

(2.2) Submarine: "And as mentioned, we are diving. It means, we will have a (pause) little (pause) submarine (pause)"

(2.3) Oxygen supply: "in which there is a certain amount of oxygen available. (pause) And this oxygen constrains our game turns."

Two coders annotated the explanation topics initiated by both the explainers and the explainees (Fleiss' $\kappa = 0.79$).

Topic changes Following the annotation procedure in ELAN, topic changes were annotated on a separate tier by creating segments with a minimum length of 20 milliseconds. Each segment was annotated including an annotation of the interlocutor who initiated the topic change—either the explainer (EX) or the explainee (EE).

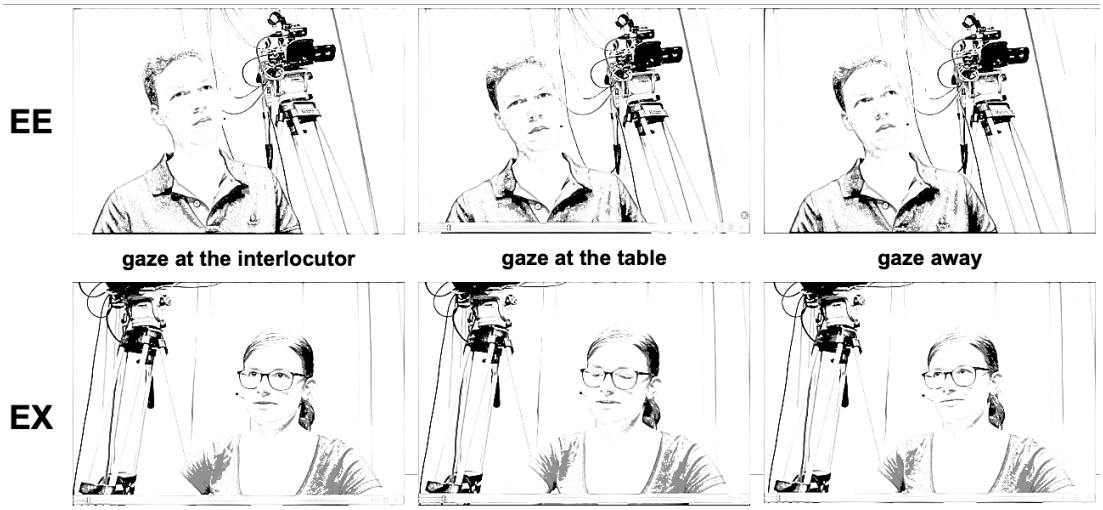

Figure 6.4: Study 2: Annotation of gaze directions.

Eye gaze behavior Eye gaze behavior was annotated for both the explainers and the explainees according to three possible viewing directions: (1) toward the explainer (for the explainees) / explainee (for the explainers), (2) toward the table (the shared referential space between the interlocutors), and (3) away (see Figure 6.4). One coder annotated the gaze behavior of both the explainers and the explainees using the video recordings from the cameras that captured the participants' facial areas. To ensure reliability, a second annotator independently coded 10% of the dataset during the training phase, resulting in a Fleiss' κ of 0.82.

Following the annotation of individual gaze behavior, mutual gaze was coded on a separate tier, based on a temporal overlap of the gaze directions of the explainer and the explainee toward each other (see Figure 6.5). Gaze withdrawals were also annotated on a separate tier and were defined as the intervals in which the participants' gaze was averted from the other person and directed either toward the table or away. Within the gaze withdrawal segments, it was further specified whether the explainer, the explainee, or both initiated a gaze withdrawal prior to a topic change.

A gaze withdrawal by either the explainer or the explainee was annotated if one participant averted their gaze without subsequently redirecting it toward the other before a topic change occurred. A gaze withdrawal by both the explainer and the explainee was annotated if both averted their gaze from each other simultaneously, without subsequently redirecting it in the direction of the other person before a topic change occurred. The duration of the segments marking the participants' averted gaze from each other prior to topic changes varied depending on the individual gaze behavior of the participants.

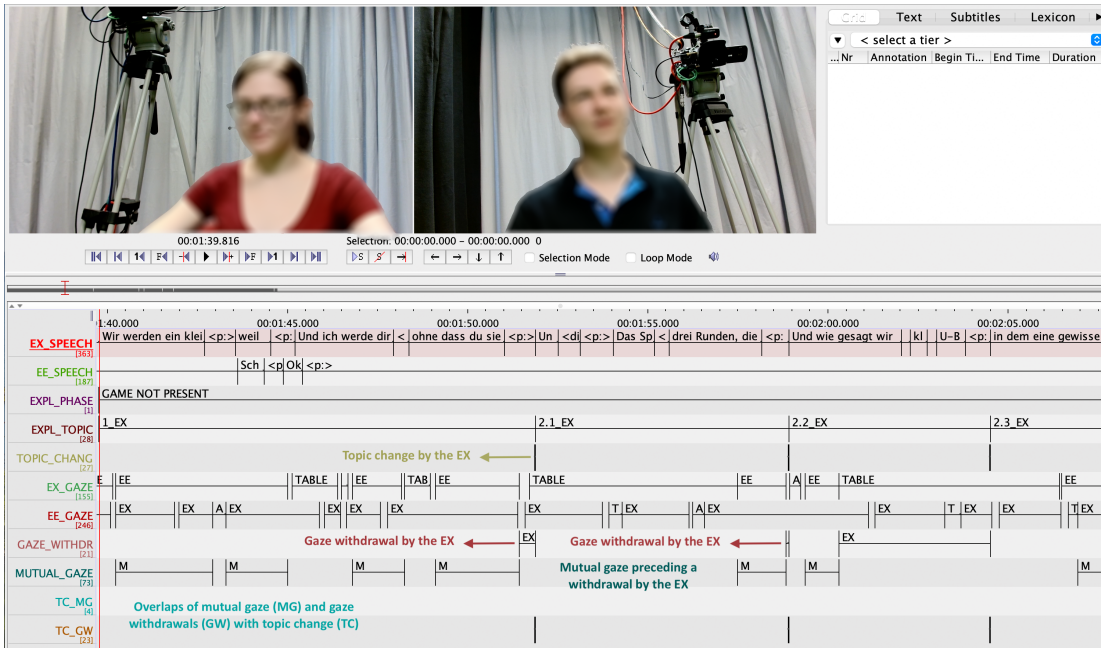

Figure 6.5: Study 2: Annotation of explanation topics, topic changes and interactive gaze behavior in ELAN.

Abbreviations: EX = explainer (left camera), EE = explainee (right camera), MG = mutual gaze, GW = gaze withdrawal, TC = topic change.

Data analysis

For the analysis of the frequencies of the participants' mutual gaze and gaze withdrawals preceding the topic changes, a detailed data sheet was created, incorporating both random and fixed effects. The random effect reflected the nested design of the

data collection, in which each of the eight explainers subsequently interacted with three different explainees (see Section 5.2.2).

Two fixed effects were included: (1) the initiator of the topic change—either the explainer or the explainee, and (2) interactive gaze behavior—either mutual gaze or gaze withdrawal, with the latter further categorized by who initiated it (the explainer, the explainee, or both simultaneously).

A Shapiro–Wilk test of normality indicated that the frequencies of mutual gaze and gaze withdrawals preceding topic changes were not normally distributed ($W = 0.68$, $p < 0.001$). Given the structure of the dataset and the non-normal distribution of the response variable, two separate Generalized Linear Models (GLMs) were conducted, each addressing the objectives of the research hypotheses. All analyses were performed using the statistical software *RStudio* (RStudio Team, 2020).

Analysis of hypothesis 1 To test the first hypothesis, which examined the overall proportional occurrences of mutual gaze and gaze withdrawals preceding topical changes (regardless which interlocutor initiated the topical change), a Generalized Linear Model (GLM) was applied. Specifically, the proportions of mutual gaze and gaze withdrawals that occurred prior to topic changes were compared across the subsample of 24 explanatory interactions.

```
glm(cbind(successes, failures) ~ GAZE_GROUPED, data = binomial_data,
family = binomial())
```

The response variable in the GLM is represented by the proportional occurrences of each form of interactive gaze behavior preceding topic changes initiated by either the explainers or the explainees. The fixed effect consisting of two levels (mutual gaze and gaze withdrawals) was represented by the variable `GAZE_GROUPED`. The model excluded the random effect as the first hypothesis addressed the overall occurrences of mutual gaze and gaze withdrawals across the dataset.

Analysis of hypothesis 2 To test the second hypothesis, which focused on a more detailed analysis of the relation between gaze withdrawals and the initiation of topic changes in relation to the participants' role in the explanatory discourse—explainer or explainee—we conducted a Generalized Linear Mixed Effects Model (GLMM). By adding the random effect to the statistical model, we provided further insights into the individual differences between the explainers, as well as the intra-individual variations for each explainer (in relation to the nested design of our dataset—one explainer interacting with three explainees)

```
glmer(FREQUENCY ~ INTERACTIVE_GAZE * TOPIC_INITIATED_BY +
(1|EX/EE), data = dataset, family = poisson())
```

In the second model, FREQUENCY of the interlocutors' gaze behavior occurring prior to topic changes was the response variable. The fixed effect was represented by the interaction between INTERACTIVE_GAZE (filtered for gaze withdrawals only: by the explainers, the explainees, or both) and TOPIC_INITIATED_BY (the initiator of topic change: the explainers or the explainees). In this model, the random effect (each explainer interacting with three different explainees) was included.

6.2.2 Results

The analysis of the first hypothesis, which focused on comparing the overall proportional occurrences of mutual gaze and gaze withdrawals prior to topic changes, indicated that the proportion of gaze withdrawals was higher ($M = 0.80$, $SD = 0.18$) than the proportion of mutual gaze ($M = 0.20$, $SD = 0.18$). Further, the results revealed a considerable variation across the analyzed subsample (see Figure 6.6). The GLM demonstrated a better model fit ($AIC = 341.69$) compared to a null-model that includes only the random effect ($AIC = 923.39$), as well as a significant effect of both forms of interactive gaze behavior: mutual gaze ($\beta = -1.55$, $S.E. = 0.10$, $z = -14.84$, $p =< 0.001$) and gaze withdrawals ($\beta = 3.11$, $S.E. = 0.15$, $z = 20.99$, $p =< 0.001$).

Figure 6.6: Study 2: Proportional distribution of mutual gaze and gaze withdrawals prior to topic changes across 24 explanatory interactions.

In addition to the statistical model, we calculated the predicted probabilities for each form of interactive gaze behavior based on the parameter estimates of each predictor:

$$p_{\text{mutual}} = \text{plogis}(-1.5572) = \frac{1}{1 + e^{1.5572}} \approx 0.174$$

$$p_{\text{withdrawal}} = \text{plogis}(1.5572) = \frac{1}{1 + e^{-1.5572}} \approx 0.826$$

According to the results, the probability of mutual gaze preceding topic changes was 0.174, whereas the probability of gaze withdrawals preceding topic changes was 0.827. With respect to the results from the statistical analysis and the post hoc probabilistic analysis, our first hypothesis could be verified.

For the second hypothesis, the relation between the initiator of gaze withdrawals and the initiator of topic changes was investigated. The results indicated that the interaction partner who initiated a topic change was the one who more often withdrew their gaze prior to the topic change compared to the frequency of gaze withdrawal of the other interlocutor who did not initiate a topic change. More specifically, when the explainers initiated a topic change, they also withdrew their gaze more often ($M = 9.79$, $SD = 7.56$) in comparison to the explainees ($M = 2.58$, $SD = 3.56$). When the explainees initiated a topic change, they withdrew their gaze more often ($M = 2.29$, $SD = 2.77$) in comparison to the explainers ($M = 0.54$, $SD = 1.18$) (see Figure 6.7). The high standard deviations of all related mean values suggest considerable intra-individual variability across the observed dataset.

In some explainer–explainee interactions, there were zero occurrences of gaze withdrawals either by the explainers, the explainees, or both observed before the topic changes. In order to avoid overdispersion in the data that may comprise the reliability of the results, the statistical model was adjusted to a negative binomial GLMM using the glmmTMB package (Brooks et al., 2017). Thus, the statistical model was modified as follows:

```
glmmTMB(FREQUENCY ~ INTERACTIVE_GAZE * TOPIC_INITIATED_BY +
(1|EX/EE), data = dataset, family = nbinom2())
```

The glmmTMB demonstrated a better model fit ($AIC = 592.03$) than the initial Poisson-GLMM ($AIC = 715.21$). Regarding the fixed effects, all factorial levels, except for the explainers' gaze withdrawals preceding the topic changes initiated by the explainees, showed a significant effect on the outcome variable (see Table 6.4). The results outlined that all three forms of gaze withdrawals, i.e., by the explainers, by the explainees, or both interlocutors simultaneously, predict topic changes initiated by the explainers.

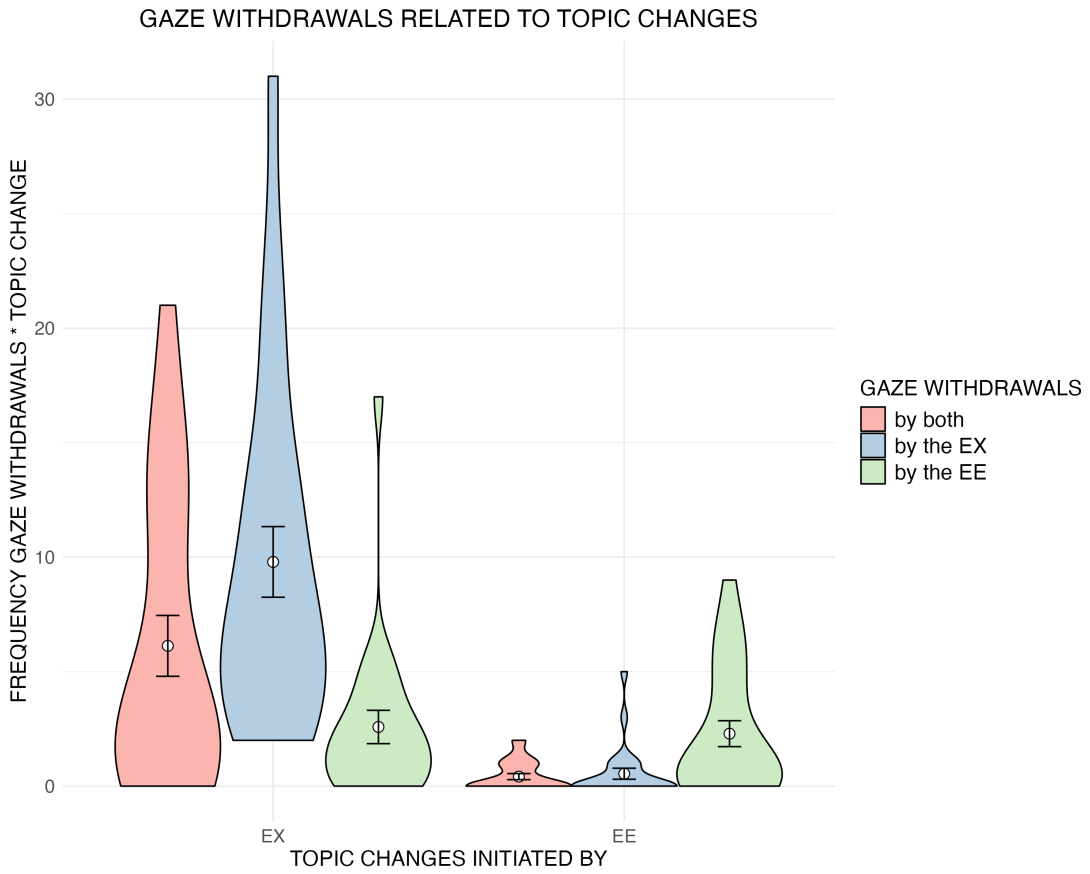

Figure 6.7: Study 2: Gaze withdrawals by the explainers, by the explainees, or by both simultaneously related to topic changes initiated by either party.

Table 6.4: Study 2: Model Summary

Gaze withdrawal by	Topic by	*Estimate*	*SE*	*z*	*p*
EX & EE	EX	1.66	0.26	6.30	< .001
EX	EX	0.56	0.27	2.07	.04
EE	EX	-0.79	0.30	-2.65	.008
EX & EE	EE	-2.72	0.41	-6.54	< .001
EX	EE	-0.31	0.56	-0.56	.575
EE	EE	2.50	0.52	4.83	< .001

Note. EX = Explainer; EE = Explainee.

To further test the second hypothesis, pairwise comparisons were conducted using simple contrasts. The results indicated significant differences between the explainers' and the explainees' gaze withdrawals prior to the topic changes initiated by the explainers ($\beta = 3.79$, $S.E. = 0.54$, $z = 9.4$, $p < .0001$). Similarly, for the topic changes initiated by the explainees, the pairwise comparisons revealed significant differences between the gaze withdrawals by the explainers and the explainees ($\beta = 0.24$, $S.E. = 0.07$, $z = -4.71$, $p < .0001$), as well as between the withdrawals by the explainees and both interlocutors together ($\beta = 0.18$, $S.E. = 0.06$, $z = -4.99$, $p < .0001$). In sum, the results from the statistical model indicate that the second hypothesis could also be

verified; that is, gaze withdrawals are initiated more frequently by the interlocutor who initiated a topic change than by the other interlocutor.

Regarding the random effect, a lower variation at the level of individual explainers interacting with three different explainees ($\sigma^2 = 0.06$, $SD = 0.24$), and a higher variation at the level of the individual explainers, regardless of the presence of different explainees ($\sigma^2 = 0.23$, $SD = 0.48$) were observed. However, the standard deviation for the random effect that included the interaction with different explainees suggested that, for some of the dyads, the presence of different explainees is also related to distribution of the interlocutors' gaze withdrawals preceding the topic changes initiated by either interlocutor. This observation was also supported by the explained proportion of variance when both the fixed and the random effects were included in the model (Conditional $R^2 = 0.72$), compared to the proportion of variance explained alone by the fixed effects (Marginal $R^2 = 0.60$).

6.2.3 Summary

The study by Lazarov and Grimminger (under review) investigated the relation between two forms of interactive gaze behavior—mutual gaze and gaze withdrawals—and the changes in the topical structure of dyadic explanations. In contrast to the first study (see Section 6.1), which examined the domain of medical explanations, the second study focused on the domain of board game explanations. In doing so, the study broadened the understanding of how the topical structure of dyadic explanations is related to the interactive gaze behavior of both interlocutors.

Based on the analysis of 24 board game explanations, **three major findings** emerged: (1) Gaze withdrawals are clearly associated with changes of explanation topics; (2) Gaze withdrawals are initiated more frequently by the explainers who are also the more frequently active part in the topical management of explanations, compared to the explainees who initiate less frequently topic changes; and (3) The variability in the revealed interactional patterns appeared to be more pronounced at the level of different explainers than at the level of different explainees interacting with the same explainer. Given the close conceptual and methodological connection between the first and second studies, the findings of both studies will be jointly discussed in the following section (see Section 6.3).

6.3 Discussion 1: The dynamics of verbal explaining behavior

In this section, the results of Study 1 (see Section 6.1) and Study 2 (see Section 6.2) are jointly discussed in relation to the first aspect of the overarching research goal, that is,

how interactional monitoring reflects in the dynamics of verbal explaining behavior. The verbal explaining behavior was analyzed in the form of explanation topics which were introduced in the spoken discourse of dyadic explanatory interactions.

Interactional monitoring refers to the bilateral process by which the explainers and the explainees continuously observe each other's verbal and nonverbal behavior throughout an explanatory interaction, in order to elicit and interpret feedback on the explainee's understanding moment by moment (Clark & Krych, 2004). In this regard, Study 1 addressed the monitoring process by investigating the relation between the topical transitions initiated by explainers and explainees' multimodal feedback behavior. The explainees' feedback behavior comprised eye gaze (static, shifting with aversion, shifting without aversion), head gestures (specifically nodding), and vocal backchannels, and it was analyzed across ten physician–caregiver interactions concerning the upcoming surgery of a child. Motivated by the findings revealed in Study 1 regarding the strong relation between the explainees' gaze aversions and the transitions toward new topics initiated by the explainers, Study 2 deepened the understanding about the relation between the interactive gaze behavior of the interlocutors and the changing topical structure of board game explanations.

6.3.1 Findings from the domain of medical explanations

The results from the study by Lazarov et al. (2024) demonstrated that the explainees' multimodal gaze aversions—when co-occurring with head nodding and backchanneling—were significantly associated with transitions from elaborations to new topics initiated by the explainers. In contrast, the explainees' static gaze behavior—particularly in the absence of head nodding and backchanneling—was more frequently associated with transitions to elaborations. What implications can be made by drawing on these contrasting relations between feedback behavior and verbal explaining behavior?

Previous research has linked addressees' gaze aversions from speakers to cognitive processing or temporary disengagement (Abeles & Yuval-Greenberg, 2017; Glenberg et al., 1998; Phelps et al., 2006), as well as to the completion of topical sequences (Rossano, 2013; Rossano, 2012). Further, the head nod gesture and vocal backchannels, such as *mhm, okay, yes, alright* expressing affirmations or agreement are signals related to ambiguous interpretations ranging between signals of ongoing attention and unconditional understanding (Allwood & Cerrato, 2003; Allwood et al., 1992; Arnold, 2012; Gander & Gander, 2020). The results of the first study suggested that when ambiguous modalities, such as gaze aversions, head nodding, and (approving) backchannels co-occur together, the degree of ambiguity becomes reduced. Although the first study did not directly assess the explainees' understanding, the completion of an elaboration and the change of the explanation topic by the explainers could

be related to (the explainers' interpretations of) the explainees' understanding of the explanation topics. However, these results are not conclusive as this research topic requires further investigation. One way to address the relation between the explainees' multimodal signals of understanding and the topical changes initiated by the explainers could be by relating the explainers' and the explainees' retrospective reports about the explainees' levels of (non-)understanding to the moments of topic changes.

The association between the explainees' static gaze behavior and the transitions to elaborations initiated by the explainers aligns with previous research showing that addressees tend to maintain prolonged gaze at speakers (J. B. Bavelas et al., 2002; Kendon, 1967). In the analyzed physician–caregiver interactions, the caregivers (the explainees) were less verbally active than the physicians (the explainers) (Fisher et al., 2022). Thus, the explainees' static gaze, also accompanied by ambiguous head nods or backchannels, may be related to sustained attention without necessarily being a request for an elaboration or a topic change.

Nonetheless, the first study did not account for the physicians' eye gaze behavior—a limitation, given that the explainers' gaze toward the explainees can function as a feedback eliciting cue (J. B. Bavelas et al., 2002; Brône et al., 2017; Kendon, 1967). Such moments tend to involve mutual gaze, which has previously been associated with the continuation of conversational topics (Rossano, 2013; Rossano, 2012), and which aligns with the definitions of elaborations in the physicians' speech (see Section 6.1.1). Thus, to this moment, the explainees' static gaze is assumed to indicate ongoing attention.

6.3.2 Findings from the domain of board game explanations

The above-mentioned limitation of Study 1 study was directly addressed in Study 2 by Lazarov and Grimminger (under review), which analyzed the interactive gaze behavior—specifically, mutual gaze and gaze withdrawals—in relation to topic changes of game explanations. Considering the eye gaze behavior of the explainers, who are the more verbally active part in explanations (Fisher et al., 2022), the second study analyzed a subsample of 24 board game explanations from the MUNDEX corpus (Türk et al., 2023). However, the analysis focused solely on the phase in which the board game was physically absent from the shared space. The corpus design allowed for conducting an analysis of variation both at an inter-individual (between explainers) and an intra-individual (within explainers) levels, as each explainer interacted sequentially with three different explainees.

Two hypotheses guided the analysis. The first—that gaze withdrawals would be more frequent than mutual gaze prior to topic changes—was supported by the results, which were in line with prior research on topical shifts in spontaneous conversations

(Rossano, 2013; Rossano, 2012). Although less frequent, mutual gaze was occasionally observed before topic changes, potentially reflecting the explainers' attempts to elicit feedback from the explainees (Argyle & Cook, 1976; J. B. Bavelas et al., 2002; Brône et al., 2017; Kendon, 1967), making the effect of interactional monitoring on the topical structure of explanations apparent (Clark & Krych, 2004). However, other co-occurring feedback signals of the explainees in response to the first study by Lazarov et al. (2024) remained a limitation in the second study.

The second hypothesis—that gaze withdrawals would be more likely initiated by the interlocutor who also initiated the topic change—was also verified by the results of the statistical analysis. This supports previous findings that the explainers take over the management of the topical structure of explanations (Fisher et al., 2022). Namely, the gaze withdrawals by explainers often preceded topic changes initiated by themselves. Furthermore, joint gaze withdrawals (i.e., by both interlocutors) also appeared to be predictors for topic changes initiated by the explainers, echoing the findings previously revealed by Rossano (2013) and Rossano (2012).

In relation to the findings of the first study by Lazarov et al. (2024), the results of the second study indicated that the explainees also withdraw their gaze from the explainers while the explainers maintain sustained gaze at the explainees and introduce a topic change. This finding was in line with the outlined findings of the first study that the explainees' multimodal gaze aversions (co-occurring with head nodding and backchanneling) from the explainers precede transitions from elaborations to new topics. Consequently, the explainees' gaze withdrawals do serve as indicators of topical completion or cognitive processing (Abeles & Yuval-Greenberg, 2017; J. B. Bavelas & Chovil, 2018; Glenberg et al., 1998; M. H. Goodwin & Goodwin, 1986; Heller, 2021). In relation to the aspect of cognitive processing, the explainees' gaze withdrawals from the explainers could be linked to the explainees' need to mentally imagine the explanandum (Markson & Paterson, 2009), which in the analyzed phases of the explanations was absent from the shared space.

It was further revealed that the explainees also withdraw their gaze from the explainers before initiating topic changes. Although the manner in which the explainees initiated topic changes was not particularly analyzed, observations of the data indicated that these topic changes were related to questions targeting the clarification of understanding-related issues. As indicated by the results, the verbal participation of the explainees during the game-absent phase was minimal compared to the verbal participation of the explainers.

In addition, the statistical model applied in Study 2 revealed that the variation of the relation between topic changes in explanations and interlocutors' gaze withdrawals becomes more prominent at the level between individual explainers than at the level of each individual explainer interacting with three different explainees subsequently.

This supports the notion that the topical organization of explanations is determined to a higher extent by the individual explaining behavior of the explainers than by the attendance of different explainees. One possible reason for this observation could be related to the asymmetry in the interlocutors' degree of knowledge about and experience with the board game (Kotthoff, 2009; Rohlfing et al., 2021). In contrast to the explainees, who were unaware of which board game would be explained, the explainers had the opportunity to prepare their explanations in advance by studying the game rules and practicing the game with others. Moreover, during the phase of the explanation in which the board game was physically absent from the shared space, it is likely that the explainees experienced difficulties imagining unfamiliar objects. As a result, they were unable to verbally co-construct the explanation in ways other than asking clarification questions related to topics previously introduced by the explainers, which had already been reported in a recent study by Fisher et al. (2023) exploring the semantic dialog patterns of board game explanations.

While discussing the results of the inter- and intra-personal variations, it is important to also reflect on the high standard deviation associated with the levels of the random effect, as well as the finding that the explainees' gaze withdrawals from the explainers can also predict the initiation of topic changes by the explainers. One way to further investigate the degree of individual variation in shaping the topical structure of explanations would be to analyze the sequence in which explanation topics were introduced by the explainers and identify which topics were subsequently reintroduced in response to the level of understanding as reported by the explainees and interpreted by the explainers. This analysis could be implemented by linking the annotations of explanation topics to the annotations of the explainees' levels of understanding from the video recall task. This aspect represents a current limitation and will be addressed in a future study.

Chapter 7

Studies on co-speech gesture dynamics

In the present chapter, the focus of the present thesis shifts toward the explainers' nonverbal explaining behavior—the use of co-speech gestures while explaining a board game to explainees. Three studies addressing this research objective are presented in the following sections. All studies analyzed a subsample from the MUNDEX corpus and focused specifically on the explanation phase in which the explanandum was physically absent from the shared space (for details, see Section 5.2).

Study 3 (Lazarov & Grimminger, 2025) expands the focus on explanation topics introduced in Chapter 6. However, here, the explanation topics were analyzed with respect to their semantic focus, allowing them to be related to different semantic dimensions being observed in explainers' co-speech gestures. The aim of the third study was to provide an overview of the presence of gesture dimensions, such as deixis, iconicity, and temporal highlighting (gestural emphasis), within different semantic categories of explanation topics.

Study 4 (Lazarov & Grimminger, 2024a) and Study 5 (Lazarov & Grimminger, 2024b) specifically addressed the dynamics of the use of gesture deixis by explainers. To deepen the understanding of the mechanisms underlying the use of gesture deixis, these two studies related the frequency of explainers' gesture deixis to explainers' interpretations of explainees' changing levels of understanding. In Study 4, the explainers' interpretations about the explainees' levels of understanding were assessed using data collected subsequently in a post hoc video recall task (for details, see Section 5.2). In Study 5, the explainees' levels of understanding were assessed based on third-person annotations of the explainees' feedback in the spoken discourse.

Note that although the explainers' co-speech gestures were annotated in the presence of speech, the spoken component—i.e., the specific words with which the gestures were temporally coupled—was not directly addressed in the research questions of the studies.

7.1 Study 3: Different explanation topics, different gestural dimensions?

Although the functional features of gesture dimensions, such as deixis, iconicity, and temporal highlighting are well established in gesture studies (McNeill, 2006; Rohrer, Tütüncübasi, et al., 2020), little is known about their proportional distribution within larger speech units, such as explanation topics, particularly in contexts in which the explanandum is physically absent from the shared space between interlocutors. Thus, the present study by Lazarov and Grimminger (2025) fills this knowledge gap, providing insights into the dynamic use of co-speech gestures within topical categories in the explanatory discourse.

The occurrence of gesture deixis, iconicity, and temporal highlighting in the explainers' behavior was analyzed across the following categories of explanation topics: *object features*, *action processes*, and *conditional rules* (see Section 7.1.1). Drawing on previous research on functional features of deictic, iconic, and beat gestures[1] (McNeill, 1992, 2006; Rohrer, Tütüncübasi, et al., 2020), the study investigated the following hypothesis: **Along the dimension of gesture deixis, gesture iconicity is expected to dominate in topics concerning object features, whereas temporal highlighting is expected to dominate in topics concerning action processes and conditional rules.**

Gesture deixis was assumed to be consistently present throughout the explanations, as the analysis focused exclusively on the phase in which the explanandum was absent from the shared space. In such contexts, the explainers may rely on gestures to construct joint imagined spaces (Kang et al., 2015; Kinalzik & Heller, 2020). The expected dominance of gesture iconicity in topics related to object features stems from the primary function of iconic gestures, which is to depict properties of objects or actions (Dargue et al., 2021; Kandana-Arachchige et al., 2021; McKern et al., 2021; McNeill, 1992). Likewise, the anticipated prominence of temporal highlighting in action processes and conditional rules is grounded in the role of beat gestures in emphasizing semantically or syntactically relevant components of speech (Beege et al., 2020; McNeill, 1992).

7.1.1 Methods

For the analysis in the present study, a randomly selected subsample of 24 board game explanations was used (for details, see Section 5.2).

[1] For details, see Section 3.2

Data coding

All data were annotated using ELAN (Wittenburg et al., 2006), following the procedures outlined in the annotation manual by Lazarov, Türk, et al. (2025).

Categories of explanation topics The board game explanations were first segmented into explanation episodes using the same coding procedure as in Study 2 (see Section 6.2.1), with an inter-rater agreement $\kappa = 0.79$. Once segmented and annotated, the explanation topics were categorized into the following three groups: *object features, action processes, and conditional rules* (see Table 7.1).

Table 7.1: Study 3: Categories of Explanation Topics

Object Features	Action Processes	Conditional Rules
Game preparation:	**Turn progressions:**	**End of a round:**
- Submarine	- Announcing directions (going down or up)	- Successful return
- Oxygen	- Reducing the oxygen	- Unsuccessful return
- Treasure chips	- Subtracting steps	- Cleaning the pathway
- Empty chips	- Action decision	
- Explorer tokens	- Skipping each other	**End of the game**
- Dices		

Explainers' co-speech gestures For annotating the explainers' co-speech gestures, recordings from the camera perspective directed toward the torso, hands, and head of the explainers were used (in the presence of speech), as this viewpoint allowed for the observation of hand shapes and movements on the shared referential space. The explainers' co-speech gestures were segmented based on the presence of single gesture strokes (for deictic and iconic gestures) or multiple strokes (for beat gestures) (McNeill, 1992).

Since initial observations of the data indicated that gesture deixis predominates in the explainers' nonverbal explaining behavior—and given that two additional studies specifically investigated this dimension—the gesture dimensions were annotated hierarchically: first for the presence of deixis, and second for the presence of iconicity and temporal highlighting. Gesture **deixis** was coded when a single pointing gesture indicated a direction or a location where an invisible object would be conceptually placed, in the co-occurrence of a relevant spoken reference. The presence of **deixis and iconicity** was coded when a categorical deictic gesture was accompanied by hand or finger shapes or movements depicting an object, its features, or a path. The explainers were frequently observed pointing at locations while simultaneously representing objects, for example, by positioning the index finger and the thumb in an object-related configuration, or by drawing shapes in the referential space using the

index finger (Streeck, 2008). The presence of **deixis and temporal highlighting** (gestural emphasis) was coded when categorical deictic gestures were paired with (repetitive) biphasic rhythmic hand or finger movements, co-occurring with prosodic emphasis. Instances in which **deixis, iconicity, and temporal highlighting** co-occurred were annotated according to the combined criteria described above. One annotator conducted the primary coding of the explainers' co-speech gestures. To ensure reliability, a second annotator independently coded 10% of the data, yielding a high inter-rater agreement for the gesture annotations ($\kappa = 0.94$)[2].

Data analysis

The annotated data were prepared for statistical analysis in accordance with the hypothesis, which addressed the distribution of the gesture dimensions—deixis, iconicity, and temporal highlighting—within each category of explanation topics individually. Accordingly, each gesture dimension was analyzed separately, irrespective of whether it co-occurred with other dimensions. For example, gesture deixis was analyzed both in its singular form and in its co-occurrence with gesture iconicity and temporal highlighting. Gesture iconicity was analyzed for its co-occurrences with gesture deixis and temporal highlighting, and likewise, temporal highlighting was analyzed for its co-occurrences with deixis and iconicity.

To account for variation in the frequency of gestural behavior between the different explainers, absolute frequencies were normalized by converting them into proportional data. Given the non-normal distribution of the proportional data across the dataset ($W = 0.92$, $p < 0.01$), a zero-inflated Generalized Linear Mixed Effects Model (GLMM) was fitted using the `glmmTMB` package (Brooks et al., 2017; Tang et al., 2023) in *RStudio* (RStudio Team, 2020):

```
glmmTMB(PROPORTION ~ 0 + TOPIC_CATEGORY * GESTURE_CATEGORY +
(1|EX/EE), ziformula = ~1, data = dataset, family = beta_family())
```

In the statistical model, `PROPORTION` is the response variable, `TOPIC_CATEGORY` and `GESTURE_CATEGORY` are the two interacting fixed effects for topic categories and gesture dimensions, followed by the random effect. The random effect represents the nested design of the dataset, in which each explainer interacted with three different explainees. The random effect was analyzed post hoc for describing the degree of variance between and within the different explainers. Additionally, a grand intercept level was included in the model in order to analyze the gesture dimension occurrences within each topical category independently.

[2]The inter-rater reliability was measured only for annotation labels, without considering the temporal length of the segments.

7.1.2 Results

The descriptive statistics indicated that deixis was the gesture dimension that occurred most frequently within all topical categories and across the 24 interactions (see Figure 7.1). Within all topical categories, the proportions of gesture iconicity were lower than the proportions of temporal highlighting (see Table 7.2 and Figure 7.1).

Table 7.2: Study 3: Proportions of gesture dimensions within categories of explanation topics.

Topical category	Gesture	M	SD	SE
object features	deixis	0.543	0.0541	0.0110
object features	iconicity	0.217	0.0477	0.00974
object features	temp. highlighting	0.240	0.0562	0.0115
action processes	deixis	0.601	0.0574	0.0117
action processes	iconicity	0.135	0.0722	0.0147
action processes	temp. highlighting	0.264	0.0670	0.0137
conditional rules	deixis	0.638	0.1350	0.0276
conditional rules	iconicity	0.129	0.0853	0.0174
conditional rules	temp. highlighting	0.245	0.1080	0.0221

The selected statistical model ($AIC = -291.3$) demonstrated substantially better performance compared to a null model without the predictor variables ($AIC = -64.4$). According to the model estimates, all factorial levels—except for gesture deixis within topics about object features—emerged as significant outcome predictors (see Table 7.3). To evaluate the hypothesis that gesture iconicity dominates in topics about object features, and that temporal highlighting dominates in topics about action processes and conditional rules, post hoc pairwise comparisons (Tukey) were conducted.

Table 7.3: Study 3: Summary of fixed effects: gesture dimensions within categories of explanation topics

Factor	Estimate	SE	*z*	*p*
object features * deixis	0.155	0.131	1.18	.238
object features * iconicity	-1.13	0.146	-7.69	***
object features * temp. highlighting	-1.03	0.144	-7.13	***
action processes * deixis	0.367	0.133	2.77	**
action processes * iconicity	-1.99	0.174	-11.40	***
action processes * temp. highlighting	-0.914	0.141	-6.47	***
conditional rules * deixis	1.02	0.144	7.09	***
conditional rules * iconicity	-2.33	0.186	-12.50	***
conditional rules * temp. highlighting	-1.47	0.157	-9.38	***
intercept (grand mean)	-32.00	613409.98	0.00	1.00

Note. p: * < .05, ** < .01, *** < .001

For topics about object features, the difference between gesture iconicity and temporal highlighting was not statistically significant ($\beta = -0.10$, $SE = 0.20$, $z = -0.49$,

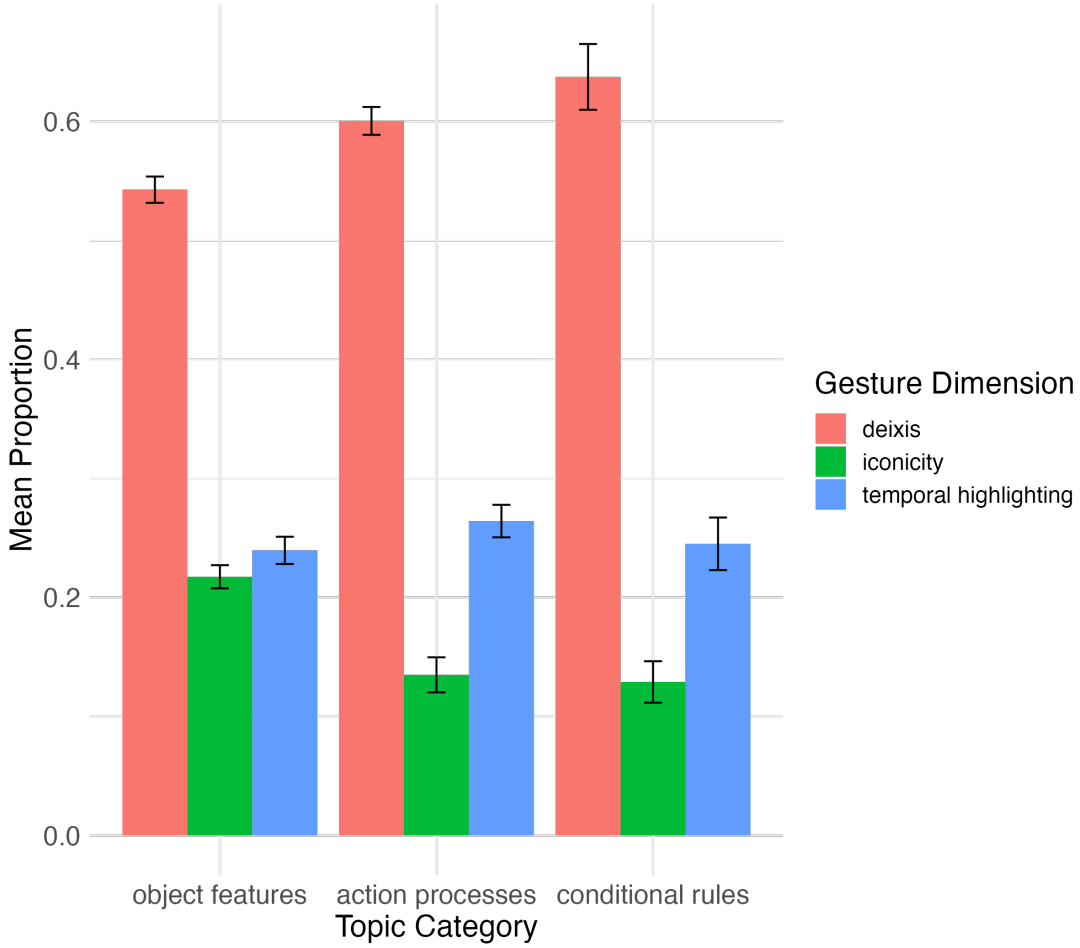

Figure 7.1: Study 3: Proportional distribution of gesture dimensions within topical categories.

$p > 0.05$), indicating that both gesture dimensions were employed to a similar extent and both contributed meaningfully to the prediction of gestural behavior. In contrast, for topics about action processes and conditional rules, significant differences were observed between gesture iconicity and temporal highlighting. Gesture iconicity occurred significantly less frequently than temporal highlighting in both topical categories: for action processes ($\beta = -1.07$, $SE = 0.22$, $z = -4.82$, $p < 0.01$), and for conditional rules ($\beta = -0.85$, $SE = 0.24$, $z = -3.58$, $p < 0.01$). Based on these results, the hypothesis could was only partly supported.

The variance within the random effect was markedly higher at the level of each explainer interacting with three different explainees ($\sigma^2 = 1.484 \times 10^{-10}$, $SD = 1.218 \times 10^{-5}$) than at the level of the eight individual explainers ($\sigma^2 = 5.768 \times 10^{-17}$, $SD = 7.594 \times 10^{-9}$). Further, the model accounted for a moderate proportion of variance across the dataset in relation to both fixed and random effects (Conditional $R^2 = 0.76$).

7.1.3 Summary

This study provided insights into the distribution of gesture deixis, iconicity, and temporal highlighting within different categories of explanation topics across 24 board game explanations, which took place in the physical absence of an explanandum from the shared space. The aim of the study was to investigate the relation between the semantic features of explainers' co-speech gestures and the semantic focus of explanation topics, which were categorized into *object features*, *action processes*, and *conditional rules*. The study investigated the hypothesis that gesture iconicity would dominate in topics about object features, whereas temporal highlighting would dominate in topics about action processes and conditional rules.

The hypothesis was only partially supported by the statistical results. Although temporal highlighting was found to occur significantly more often than gesture iconicity within explanation topics concerning action processes and conditional rules, no significant difference between the two gestural dimensions was observed for topics about object features. This suggests that, alongside gesture deixis, both iconicity and temporal highlighting were used to a similar extent in conveying information about object features.

7.2 Study 4: Gesture deixis related to interpretations about explainees' understanding

In Study 4 Lazarov and Grimminger (2024a) investigated the variation in the use of gesture deixis by eight explainers in relation to their monitoring of explainees' understanding. The study pursued two objectives: (1) Analyzing the inter-individual variation in the use of gesture deixis across different explainers in relation to their interpretations about the explainees' (levels of) understanding; and (2) Exploring the intra-individual variation in the use gesture deixis for each explainer who interacted with three different explainees.

To address the first objective, it was hypothesized that **explainers' gesture deixis would decrease following the interpretation of complete understanding in the explainees' behavior, compared to the interpretation of partial, non- or misunderstanding when the explainers' gesture deixis would decrease**. The hypothesis was motivated by prior research on the contribution of speakers' deictic gestures to addressees' comprehension (Clark, 2003; McNeill, 1992, 2006; Stojnic et al., 2013), as well as studies reporting on the benefits of observing semantically congruent co-speech gestures for cognitive processing (Habets et al., 2011; Kelly et al., 2010; Li et al., 2022; Ping et al., 2013).

To address the second objective, it was hypothesized that **the explainers would**

reveal intra-individual variations in the use of gesture deixis depending on the attendance of a different explainee. This hypothesis was motivated by prior findings on the individual differences in gesture production by different speakers (Holler & Stevens, 2007; Priesters & Mittelberg, 2013) and on the adaptation of gestures to the presence of different addressees (Bergmann & Kopp, 2010; Jacobs & Garnham, 2007). Also, this hypothesis was formulated following an exploratory approach due to the scarcity of empirical research on the variation of gesture deixis in relation to the explainers' interpretations of the explainees' (levels of) understanding. Nonetheless, drawing on the findings by Jacobs and Garnham (2007) and Bergmann and Kopp (2010), it was expected that the explainers' use of gesture deixis would vary depending on the attendance of different explainees.

7.2.1 Methods

For the present study, a subsample of 24 randomly selected board game explanations was analyzed (for details, see Section 5.2).

Data coding

The data were annotated in ELAN (Wittenburg et al., 2006), following the procedures outlined in the annotation manual by Lazarov, Türk, et al. (2025). Three coders participated in the annotation process. Coder A annotated the explainers' co-speech gestures, as well as the explainers' interpretations of the explainees' understanding collected in a post hoc video recall task. Coders B and C were involved in the inter-rater reliability testing.

Co-speech gestures For the annotation of the explainers' co-speech gestures, Coder A segmented and annotated gesture phrases (McNeill, 1992) using recordings of the camera angle directed toward the torso, hands, and head of the explainer, with audio enabled. This perspective enabled the observation of hand shapes and movements across the referential space. For reliability, 10% of the data was annotated independently by Coder B, yielding high agreement ($\kappa = 0.94$)[3].

Coders identified first the explainers' pointing behavior based on explainers' hand / finger shape, and then they annotated the relevant gesture functions according to the feature definitions provided by McNeill (1992, 2006). During the coding process, it became evident that gesture deixis did not appear solely in its singular form, but also in multidimensional forms including gesture iconicity, and / or temporal highlighting (gestural emphasis) (McNeill, 2006).

[3]The inter-rater reliability was measured only for annotation labels, without considering the temporal length of the segments.

Deictic gestures were annotated based on a single pointing towards a direction or a location where an invisible object would be placed, and co-occurring with the related spoken reference. Deictic-iconic gestures were annotated based on the criteria for categorical deictic gestures complemented by hand or finger shapes or movements depicting an object, features of an object, or a path. The explainers from our study were observed to point at locations while depicting objects by either positioning the index finger and the thumb in an object related form or drawing objects on the shared referential space, e.g., by the index finger (Streeck, 2008). Deictic-beat gestures were annotated based on the criteria for categorical deictic gestures, complemented by (repetitive) biphasic rhythmic hand / finger movements in the presence of temporal highlighting (gestural emphasis).

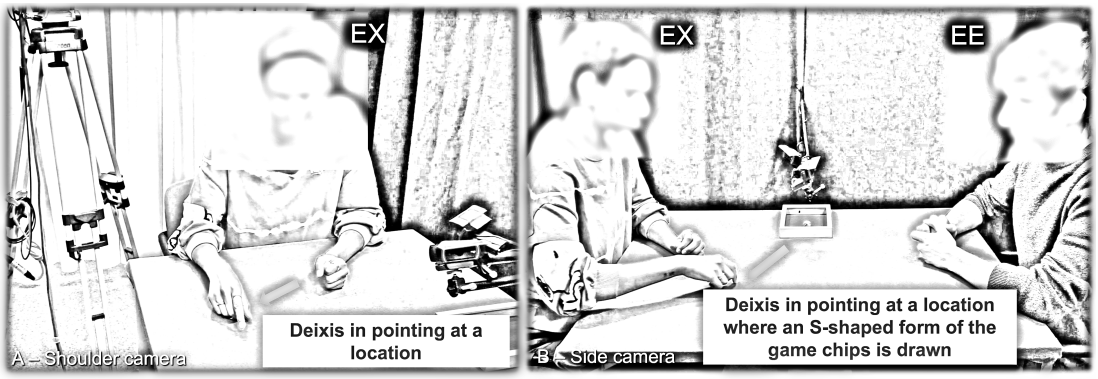

Figure 7.2: Study 4: Coding the explainers co-speech gestures.

Explainers' interpretations about the explainees' understanding Coder A annotated the explainers' interpretations of the explainees' levels of understanding collected in the post hoc video recall task. The explainees' levels of understanding were annotated according to the levels suggested by Vendler (1994): *understanding*, *partial understanding*, *non-understanding*, and *misunderstanding*. Many of the explainers' comments could be directly annotated based on the explicit use of these terms. However, some comments did not contain these key terms, but instead included synonymous or colloquial expressions—for example, "Hier hat er/sie einen Klick gemacht" (a colloquial German expression indicating understanding) or "Das konnte ich mir nicht ganz vorstellen" (indicating non-understanding). Such expressions were interpreted and coded in alignment with the implied level of understanding. Comments that were not explicitly related to the explainees' level of understanding, such as comments reflecting on the quality of the explanation, were excluded from analysis. The coders were trained to identify and interpret relevant information pertaining to the explainees' level of understanding. For reliability, Coder C annotated 10% of the data, resulting in a satisfactory inter-rater agreement ($\kappa = 0.85$).

Data analysis

For the analysis, all forms of deictic gestures (deictic, deictic-iconic, and deictic-beat) were collapsed into a single variable (gesture deixis). The number of all forms of deictic gestures produced in the gaps between the annotated levels of understanding were counted. The gaps represented the time between two levels of explainees' understanding as reported by the explainers. The number of explainers' reports on explainees' levels of understanding varied between the individual dyadic interactions (see Table 7.4).

Table 7.4: Study 4: Summary of reported levels of understanding across the 24 dyadic interactions

Reported levels of	Sum	Range	M	SD
Understanding	89	1–22	4.94	5.30
Partial understanding	58	1–9	3.41	2.53
Non-understanding	61	1–10	2.54	2.10
Misunderstanding	18	1–8	2.00	2.34

The dataset was structured to reflect the nested design of data collection, with the random effect specified hierarchically in two columns: *explainer and explainee*. Before selecting the appropriate statistical model, a Shapiro–Wilk test was conducted to assess the normality of the distribution of deictic gestures across the 24 interactions. The test indicated a significant deviation from normality ($W = 0.90$, $p < 0.05$). Given the non-normal distribution of the data, a Generalized Linear Mixed Effects Model (GLMM) was fitted using the `lme4` package (Bates et al., 2015) in *RStudio* (RStudio Team, 2020). The model was specified using the following function:

```
glmer(GEST_FREQ ~ UNDERSTAND + (1|EX/EE), data = dataset,
family = poisson())
```

The frequency of the explainers' deictic gestures (including all forms: deictic, deictic-iconic, and deictic-beat) was the response variable. The monitored levels of understanding, coded as a four-level categorical variable, were defined as the fixed effect. A simple contrast coding scheme was applied, with *understanding* set as the reference level. The remaining levels—*partial understanding*, *non-understanding*, and *misunderstanding*—were each compared against the referential level. The random effect structure reflected the nested study design, accounting for each explainer interacting with three different explainees. Accordingly, the different explainees were nested within the explainers with whom they interacted.

Results

The statistical model demonstrated a better fit (AIC = 1271.2; BIC = 1283.9) compared to a null model without the fixed effect (AIC = 1384.8; BIC = 1391.2). The model explained a modest proportion of variance regarding the fixed effect alone (marginal R^2 = 0.165), whereas a considerably higher proportion of variance was explained when accounting for both the fixed and the random effects (conditional R^2 = 0.943).

The fixed effects summary (see Table 7.5 and Figure 7.3) indicated that the levels *understanding*, *partial understanding*, and *misunderstanding* significantly predicted the frequency of gesture deixis across the 24 explanatory interactions.

Table 7.5: Study 4: Frequency of the explainers' gesture deixis related to their interpretations of explainees' understanding

Factor	M	SD	β	SE	z	p
U (int.)	46.19	32.59	3.95	0.14	27.59	***
PU	38.00	23.69	-0.33	0.06	-5.83	***
NU	61.36	44.94	0.05	0.05	0.97	ns
MU	27.28	26.48	-0.67	0.09	-7.63	***

Note. *** $p < .001$, ns $p > .05$.
U = understanding (intercept), PU = partial understanding, NU = non-understanding, MU = misunderstanding.

To test the first hypothesis—whether the explainers' use of gesture deixis decreases following the interpretation of complete understanding and increases following the interpretation of partial, non-, or misunderstanding—post hoc pairwise comparisons with simple contrasts were conducted. The results are summarized in Table 7.6.

Overall, the findings did not support the hypothesis that gesture deixis decreases following the interpretation of complete understanding. Although the use of gesture deixis increased slightly following the interpretations of non-understanding compared to complete understanding, the difference between these two conditions was not significant ($\beta = -0.05$, $SE = 0.05$, $z = 0.97$, $p > 0.05$). In contrast, gesture deixis was found to decrease significantly following the explainers' interpretations of partial understanding and misunderstanding. The statistical model indicated significant differences between (complete) understanding and partial understanding ($\beta = 0.33$, $SE = 0.06$, $z = 5.83$, $p < 0.001$), as well as between (complete) understanding and misunderstanding ($\beta = 0.67$, $SE = 0.09$, $z = 7.63$, $p < 0.001$).

Although the results indicated that monitoring the explainees' levels of understanding, partial understanding, and misunderstanding was associated with variation in the use of gesture deixis by explainers, Hypothesis 1 was not supported. Specifically, the frequency of gesture deixis following interpretations of complete understanding did not differ significantly from the frequency of gesture deixis following interpretations

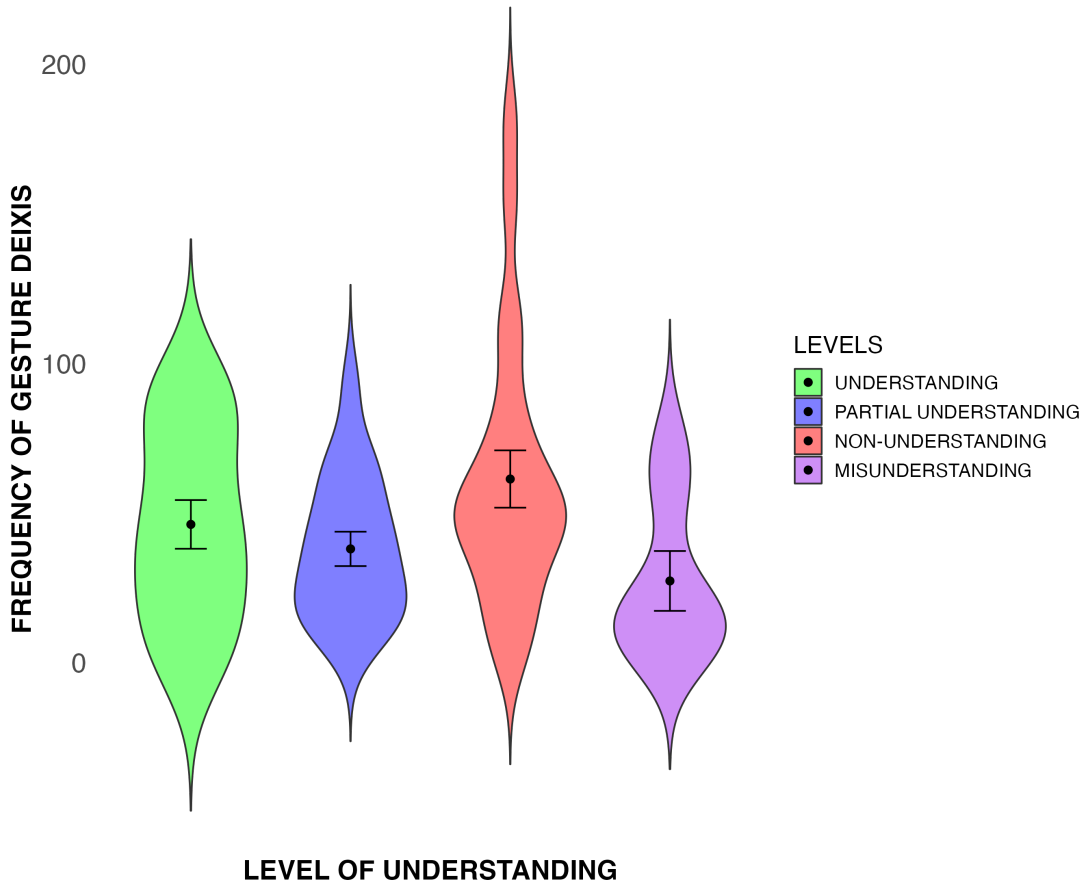

Figure 7.3: Study 4: The explainers' gesture deixis related to their interpretations of the explainees' understanding.

of non-understanding. Moreover, gesture deixis decreased significantly following interpretations of partial understanding and misunderstanding.

To test the second hypothesis, the intra-individual differences within the use of gesture deixis by the explainers were explored. To do so, the variance of the random effect revealed by the statistical model was analyzed. The nested random effect indicated greater variation in the use of gesture deixis within the individual explainers, each interacting with three different explainees ($\sigma^2 = 0.21$, $SD = 0.45$), compared to the variation between the eight different explainers, regardless of the attending explainee ($\sigma^2 = 0.08$, $SD = 0.29$). These findings suggest that the explainers adapted their use of gesture deixis more frequently in relation to the attendance of different explainees than in relation to their own explaining behavior.

Figure 7.4 illustrates the intra-individual differences for each individual interaction, by presenting normalized proportions derived from the absolute frequencies of each explainer's gesture deixis in relation to their interpretations of each explainee's level of understanding.

Table 7.6: Study 4: Gesture deixis following interpretations of understanding: EM and SE

Level	EM	SE	LCL	UCL
U	3.95	0.14	3.67	4.23
PU	3.62	0.14	3.33	3.90
NU	4.00	0.14	3.73	4.28
MU	3.28	0.16	2.97	3.59

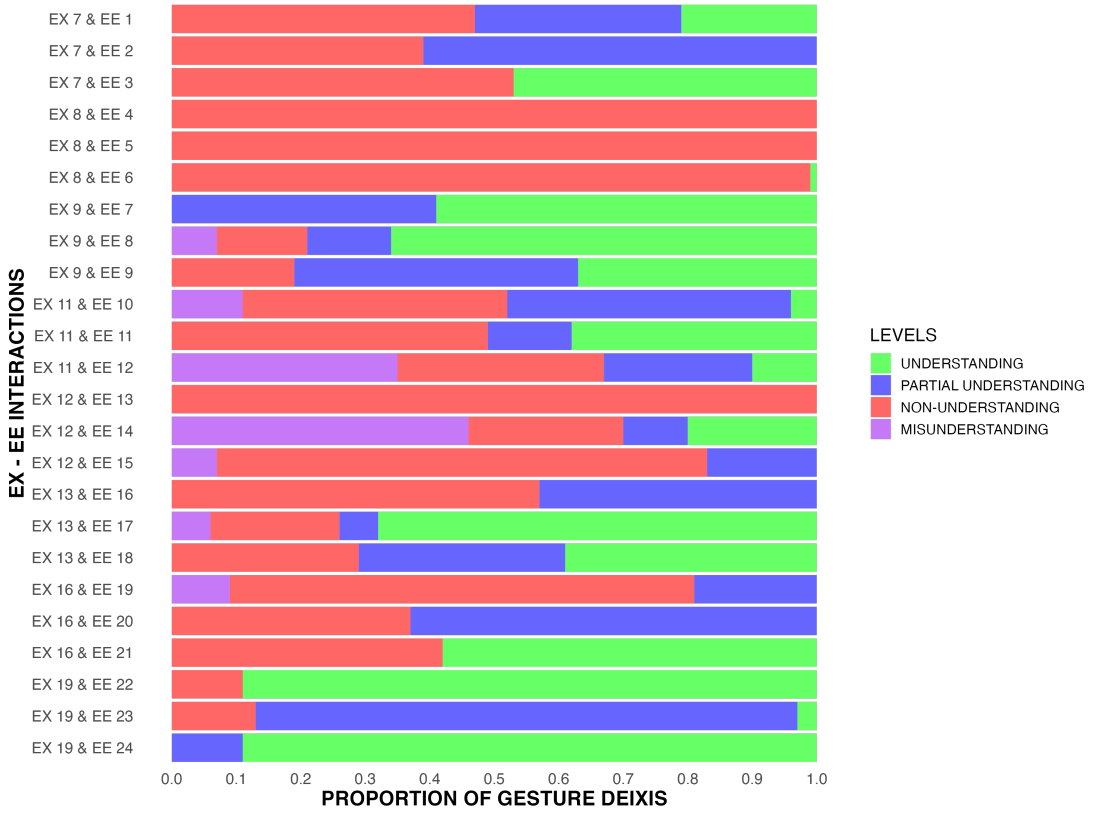

Figure 7.4: Study 4: Individual proportional variations of the explainers' gesture deixis related to interpretations of explainees' understanding

The variance in the monitored levels of understanding across the interactions between each explainer (EX) and explainee (EE) becomes immediately apparent: Not all explainers reported on all four levels of understanding while watching the video-taped interactions during the video recall task. For instance, misunderstanding was reported only for a limited number of interactions. Thus, the use of gesture deixis could be compared only for the levels *non-understanding*, *partial understanding*, and *understanding*.

In relation to the level *non-understanding*, intra-individual differences in the use of gesture deixis were observed for EX12, EX13, and EX16. The explainers who reported *partial understanding* demonstrated variation in the use of gesture deixis depending on which explainee they were interacting with. Similarly, differences in the proportions of gesture deixis related to complete understanding were observed in EX7, EX9, EX11,

EX13, and EX19.

According to these results, the use of gesture deixis is related not only to the attendance of a different explainee but also to the monitored level of explainees' understanding. In combination with the results on the random effect, the second hypothesis that the explainers exhibit intra-individual variations in gesture deixis regarding the monitored levels of understanding was supported.

Summary

Study 4 investigated the variations in the use gesture deixis by different explainers in relation to their interpretations of the explainees' evolving levels of understanding. In accordance with the nested design of the dataset, the analysis addressed this variation between and within eight different explainers.

The major finding of the study was that the frequency of gesture deixis neither decreases following the explainers' interpretations of the explainees' complete understanding, nor increases following the explainers' interpretations of explainees' non-understanding. However, this difference was not found to be statistically significant.

Focusing on the inter- and intra-individual variation in the use of gesture deixis in relation to monitoring the explainees' understanding, the results revealed that the attendance of a different explainee influences how the explainers adapt the use of gesture deixis in relation to their interpretations of the explainees' levels of understanding. In this relation, the findings demonstrated that the dynamic use of gesture deixis is shaped not only by the individual feedback behavior of the explainees, but also by the explainers' interpretations of the explainees' behavior as indicative of specific levels of understanding.

7.3 Study 5: Gesture deixis related to explainees' verbal signals of understanding

Building on the insights from Study 4, Study 5 by Lazarov and Grimminger (2024b) investigated the variation in the use of gesture deixis by the explainers in relation to the explainees' understanding as signaled in the spoken discourse during the explanatory interactions. In particular, the study investigated the question of whether verbal signals of understanding do predict a decrease in the use of gesture deixis.

To this end, the following hypothesis was tested: **The explainers' gesture deixis would decrease following the explainees' verbal signals of understanding and conversely increase following the explainees' signals of partial and non-understanding**. The hypothesis was motivated by previous research on the relation

between speakers' co-speech gestures and addressees' comprehension. Overall, co-speech gestures have been shown to facilitate addressees' understanding (Kelly et al., 2010). More specifically, in contexts where addressees provided feedback of understanding, a decrease in speakers' use of deictic and iconic gestures was observed (Holler & Wilkin, 2011). Conversely, a study conducted in classroom settings revealed that teachers increased their use of deictic and iconic gestures in response to students' feedback of non-understanding (Alibali et al., 2013).

7.3.1 Methods

For the current study, 15 explanatory interactions from the MUNDEX (see Section 5.2) corpus were analyzed. The gestural behavior of five German-speaking adult explainers (age: $M = 24.6$, $SD = 4.04$) and the verbal signals of understanding produced by 15 explainees (age: $M = 25.9$, $SD = 6.15$) were investigated. In order to compare the results from the current study with the results form Study 4 (see Section 7.2), the phase in which the board game was absent from the shared space was considered in the analysis.

Data coding

All data were annotated in ELAN (Wittenburg et al., 2006), following the procedures outlined in the annotation manual by Lazarov, Türk, et al. (2025).

Co-speech gestures The same co-speech gestures that were segmented and annotated for Study 4 were used for the analysis of the present study (for details, see Section 7.2).

Verbal signals of understanding Four coders annotated the explainees' utterances with respect to the explainees' *understanding*, *partial understanding*, and *non-understanding*, using categories adapted from established discourse annotation schemes, such as *DAMSL* (Core & Allen, 1997) and *DiT++* (Bunt, 2009, 2011). To ensure reliability, all coders independently annotated 10% of the data prior to the main annotation phase, resulting in a high inter-rater agreement ($\kappa = 0.89$).

- For annotating *understanding*, the coders identified utterances that functioned as *completions* of the explainer's speech[4], backchannels (e.g., *okay, yes, alright, good*), *repetitions* or *paraphrases* of the explainer's utterances, and *self-initiated error indications*[5].

[4] A *completion* occurs when the explainee interrupts the explainer to signal understanding by correctly completing the explainer's utterance.

[5] *Self-error* indications refer to utterances in which the explainee alone notices a mistake or misunderstanding without explicit prompting by the explainer.

- For annotating *partial understanding*, the coders considered *feedback-elicitation* utterances[6], *polar questions*, *backchannels of uncertainty* (e.g., *hm, huh*), and *holding utterances*[7].

- For annotating *non-understanding*, the coders identified o*pen-ended questions*, *direct expressions* of non-understanding (e.g., "I don't understand. / I don't get it."), *backchannels of uncertainty*, and *holding utterances*. Open questions were classified as signals of non-understanding on the basis that they typically indicate a broader knowledge gap compared to polar questions, which seek the verification of smaller than larger units of information.

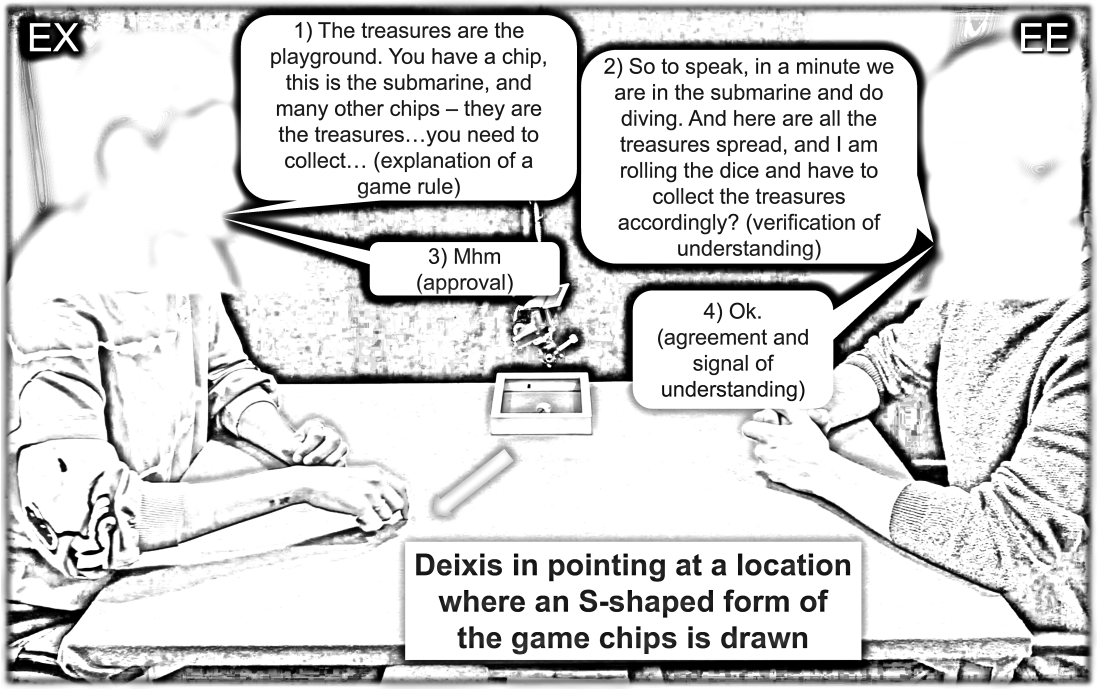

Figure 7.5: Study 5: Explaining with gesture deixis while perceiving the explainees' verbal signals of (non-)understanding.

Data Analysis

To analyze the explainers' use of gesture deixis in relation to the explainees' verbal signals of understanding, the absolute frequencies of gesture deixis were converted into proportions. A Shapiro–Wilk test indicated that the proportional distribution of gesture deixis across the dataset significantly deviated from normality ($W = 0.79$, $p < 0.01$). Given the non-normal distribution and the proportional structure of the

[6] Feedback elicitation occurs when the explainee seeks confirmation, often using declarative clauses with rising intonation, rather than typical yes/no questions.

[7] *Holding utterances* signal delayed responses or hesitation. For instance, the explainee does not immediately answer a question or postpones providing requested information, instead offering an ambiguous or incomplete reply.

response variable, a Generalized Linear Mixed Effects Model (GLMM) was fitted using the `glmmTMB` package (Brooks et al., 2017) in *RStudio* (RStudio Team, 2020):

```
glmmTMB(PROPORTION ~ SIGNALS_OF_UNDERSTANDING + (1|EX/EE),
data = dataset, family = binominal())
```

In the statistical model, `PROPORTION` was the response variable, which was analyzed in relation to the fixed effect `SIGNALS_OF_UNDERSTANDING`. The model also incorporated the nested structure of the dataset, according to which each explainer interacted with a different explainee.

Results

The GLMM indicated a better model fit ($AIC = -228.5$) compared to the null model without the fixed effect ($AIC = -181.5$). Preliminary results suggested that the majority of the explainers' gesture deixis occurred following the explainees' verbal signals of complete understanding ($M = 0.78$, $SD = 0.23$). In comparison, the proportions of gesture deixis following the explainees' verbal signals of partial understanding ($M = 0.18$, $SD = 0.22$) and non-understanding ($M = 0.04$, $SD = 0.10$) were considerably lower.

All levels of understanding emerged as significant predictors of the explainers' use of gesture deixis (see Table 7.7 and Figure 7.6). This result implied that (1) gesture deixis remains consistently present throughout explanatory phase in which the explanandum is absent from the shared space; and (2) the frequency of gesture deixis increases following verbal signals of understanding, and conversely decreases following signals of partial or non-understanding.

Table 7.7: Study 5: Model Estimates and Estimated Means for Gesture Deixis Across Levels of Understanding

Effect	M	SD	Est.	SE	z	p	EM	SE
U (int.)	0.78	0.23	1.30	0.30	4.27	***	1.30	0.31
PU	0.18	0.22	-2.75	0.47	-5.80	***	-1.45	0.31
NU	0.04	0.10	-4.21	0.51	-8.17	***	-2.91	0.34

Note. *** $p < .001$. EM = estimated mean; SE = standard error.

Post hoc pairwise comparisons using simple contrasts confirmed the pattern observed in the descriptive statistics. The proportion of gesture deixis following the explainees' verbal signals of complete understanding was significantly higher compared to the proportion of gesture deixis following the explainees' verbal signals of partial understanding ($\beta = 2.75$, $SE = 0.47$, $z = 5.80$, $p < 0.0001$) and non-understanding ($\beta = 4.21$, $SE = 0.52$, $z = 8.17$, $p < 0.0001$). Further, the proportion of gesture deixis

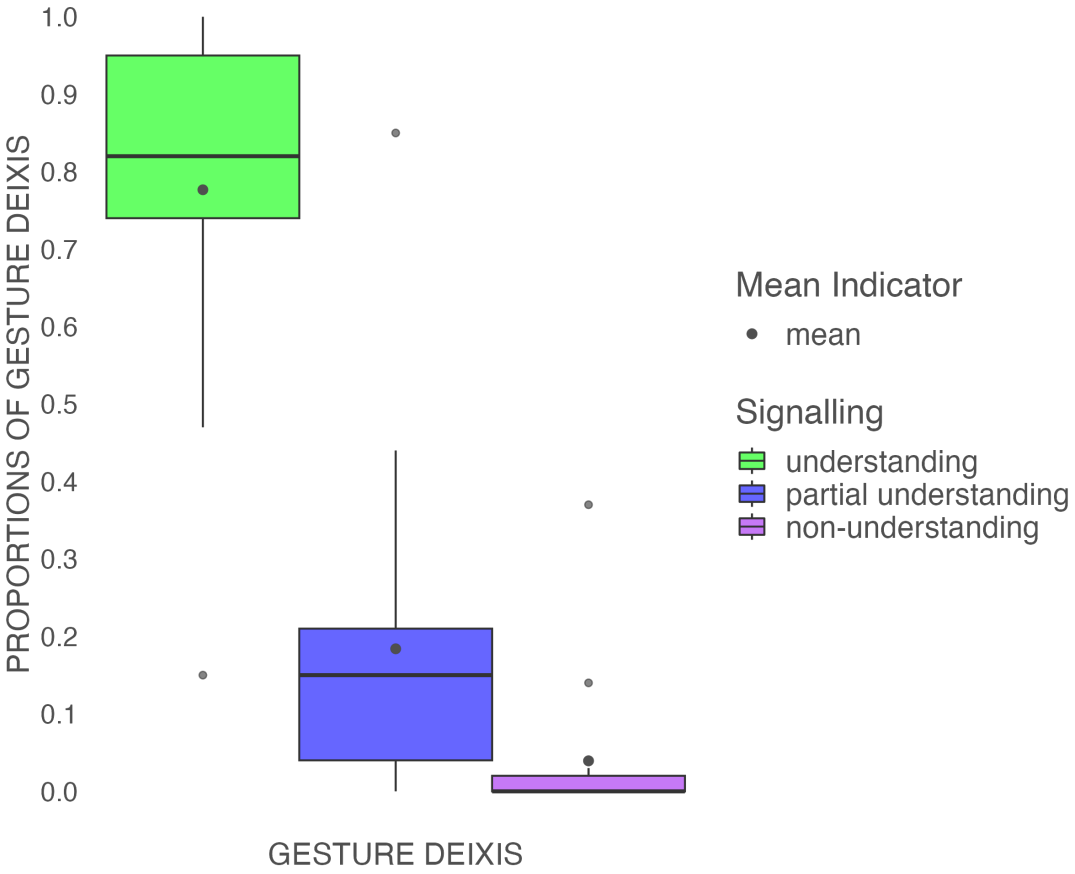

Figure 7.6: Study 5: Proportions of the EXs' gesture deixis following the EEs' verbal signals of understanding.

was significantly higher following the explainees' verbal signals of partial understanding compared to following the explainees' signals of non-understanding ($\beta = 1.46$, $SE = 0.39$, $z = 3.71$, $p < 0.001$).

Contrary to the hypothesis—which predicted that gesture deixis would decrease following the explainees' verbal signals of complete understanding and increase following the explainees' verbal signals of partial or non-understanding—the results of the statistical analysis indicated a pattern in the opposite direction. Thus, the hypothesis was not supported.

The analysis of the nested random effect indicated greater variance in gesture deixis within the explainers interacting with three different explainees ($\sigma^2 = 6.30 \times 10^{-10}$, $SD = 2.51 \times 10^{-5}$), compared to the variance between the individual gesturing behavior of the five explainers ($\sigma^2 = 8.31 \times 10^{-11}$, $SD = 9.12 \times 10^{-6}$). This means that intra-individual variation in gesture deixis was approximately 7.58 times greater than inter-individual variation.

The findings on the random effect were consistent with the findings on the random effects observed in Study 4 (see Section 7.2). Taken together, the results suggested that

the explainees' verbal signals of understanding are closely related to the dynamics in the use of gesture deixis by the explainers.

The model accounted for a moderate proportion of explained variance in the response variable when both fixed and random effects were considered (conditional $R^2 = 0.57$). This could be related to the smaller sample size compared to the sample size used for the analysis in Study 4.

Summary

The current study investigated the use of gesture deixis by different explainers in response to the explainees' verbal signals of understanding during the physical absence of an explanandum. Thus, the study built on the findings from Study 4 (see Section 7.2).

The results did not support the hypothesis that gesture deixis would decrease following the explainees' verbal signals of complete understanding and increase following the explainees' signals of partial or non-understanding. Contrary to the hypothesis, the analysis showed that gesture deixis was used most frequently following the explainees' signals of complete understanding. These findings were consistent with the findings revealed in Study 4 (see Section 7.2). To this end, it can be concluded that the frequency of explainers' gesture deixis varies in both positive and negative directions in relation to the explainees' verbal signals of (non-)understanding.

Furthermore, the analysis revealed greater variance at the explainer–explainee level than at the level between different explainers. Also this pattern mirrored the pattern reveled by the results of Study 4, concretely that the individual explaining behavior is closely related and adapting to the individual feedback behavior of explainees.

7.4 Discussion 2: The dynamics of nonverbal explaining behavior

The studies presented in Sections 7.1, 7.2, and 7.3 investigated the dynamic use of co-speech gestures—specifically, gesture dimensions—by different explainers in dyadic board game explanations while the explanandum was physically absent from the shared space. The studies were presented in a top-down manner: Study 3 examined the distribution of co-speech gesture dimensions, such as deixis, iconicity, and temporal highlighting (gestural emphasis) (McNeill, 2006) across different categories of explanation topics; Study 4 and Study 5 focused on the dimension of gesture deixis, by exploring its relation to the explainees' feedback behavior, as part of the process called "interactional monitoring" (Clark & Krych, 2004).

7.4.1 Discussion: Different topics, different gesture dimensions

Study 3 by Lazarov and Grimminger (2025) presented in Section 7.1 revealed that gesture deixis occurred not only as a standalone gesture dimension but also in combination with other dimensions, more specifically iconicity and temporal highlighting. It is important to note that the gesture annotations were performed at the level of explanation topics, covering complete utterances or parts of utterances, rather than at the level of affiliated speech units (word level) (see Section 7.1.1). As a result, the analysis captured the broader discursive function of gestures rather than their alignment with individual words.

Specifically, the study tested the hypothesis that gesture iconicity dominates over temporal highlighting in topics concerning object features, and, conversely, that temporal highlighting dominates over gesture iconicity in topics about action processes and conditional rules. This hypothesis was motivated by previous research investigating the functional features of iconic and beat gestures (Austin & Sweller, 2014; Beege et al., 2020; Dargue et al., 2021; Dimitrova et al., 2016; Kandana-Arachchige et al., 2021; McKern et al., 2021; McNeill, 1992, 2006; Rohrer, Delais-Roussarie, & Prieto, 2020).

The results of the statistical analysis provided only partial support for the hypothesis. While gesture iconicity and temporal highlighting were found to co-occur with gesture deixis in explanations about object features, their relative frequencies did not significantly differ. In contrast, temporal highlighting was observed more frequently in topics concerning action processes and conditional rules, suggesting its heightened role in marking and emphasizing structurally or semantically important content in action or rule-based contexts. Such patterns suggest that, in the absence of a physical explanandum, pointing to imagined locations via gesture deixis and emphasizing information via temporal highlighting are used more frequently than depicting object features via gesture iconicity. Thus, the results suggest a dynamic pattern in which explainers prioritize spatial referencing and prosodic emphasis over object depictions when constructing multimodal explanations in a visually constrained context.

Beyond the scope of the research hypothesis, the results also revealed that, in all three topical categories, gesture deixis was the most frequently used gesture dimension by the explainers. One plausible explanation for the constantly frequent occurrences of gesture deixis is that the explainers might have felt compelled to continuously establish ”joint imagined spaces” while the explanandum was physically absent from the shared space (Kinalzik & Heller, 2020). By employing the dimension of gesture deixis, the explainers visually conveyed information related to imagined spatial references, facilitating the explainees’ understanding of the locational organization of unknown board game objects (Lazarov & Grimminger, 2024a).

7.4.2 Discussion: Gesture deixis related to the monitoring of explainees' understanding

Both Study 4 by Lazarov and Grimminger (2024a) and Study 5 by Lazarov and Grimminger (2024b) investigated the variation in the use of gesture deixis in relation to monitoring the explainees' (non-)understanding. Specifically, the studies explored how the frequency of explainers' gesture deixis is related to the dynamically changing levels of understanding observed in the behavior of explainees.

In doing so, Study 4 (Section 7.2) investigated the relation between the frequency of explainers' gesture deixis and their interpretations of explainees' levels of understanding. These interpretations were collected in a post hoc video recall task (see Section 5.2).

The study tested the hypothesis that the explainers' use of gesture deixis would decrease following their interpretations of explainees' *complete understanding* and, conversely, increase following their interpretations of explainees' *partial understanding*, *non-understanding*, or *misunderstanding*. The hypothesis was motivated by previous research highlighting the role of co-speech gestures in facilitating addressees' comprehension (Congdon et al., 2017; Grimminger et al., 2010; Kang et al., 2015) and reducing their cognitive load (Habets et al., 2011; Kelly et al., 2010; Li et al., 2022; Ping et al., 2013). For example, the use of deictic (Clark, 2003; McNeill, 1992, 2006; Stojnic et al., 2013) and iconic gestures (Dargue et al., 2021; Kandana-Arachchige et al., 2021; McKern et al., 2021) has been reported to enhance addressees' comprehension.

Following the statistical analysis, the hypothesis was not supported. The frequency of gesture deixis did not decrease following the explainers' interpretations of the explainees' complete understanding. Instead, the use of gesture deixis remained relatively stable across the explainers' interpretations of all levels of explainees' understanding. Although it is likely that the explainers' continuous employment of gesture deixis enhanced the explainees' comprehension—as suggested in previous literature (Clark, 2003; Congdon et al., 2017; Dargue et al., 2021; Grimminger et al., 2010; Kandana-Arachchige et al., 2021; McKern et al., 2021; Stojnic et al., 2013)—the presented research focused exclusively on the explainers' interpretative perspective rather than the explainees' self-reported understanding.

As mentioned above in the discussion of Study 3, the physical absence of the explanandum from the shared space, which likely demanded the continuous establishment of "joint imagined spaces" (Kinalzik & Heller, 2020), is a reasonable explanation accounting for the consistent use of gesture deixis by the explainers. Also, this demand may have been implicitly driven by the inherent knowledge asymmetry between the explainers and the explainees (Kang et al., 2015; Kotthoff, 2009).

A follow-up study (Section 7.3) extended the topic on the variation of explainers'

gesture deixis, by relating it to explainees' verbal signals of understanding in the spoken discourse. Concretely, Study 5 tested the hypothesis that gesture deixis would decrease following the explainees' verbal signals of understanding, and increase following the explainees' verbal signals of partial or non-understanding. The hypothesis was motivated by previous research suggesting that addressees' positive feedback is associated with a a decrease in speakers' deictic and iconic gestures (Holler & Wilkin, 2011), and that the detection of trouble spots or non-understanding prompts is associated with a more frequent gesturing behavior by the speakers (Alibali et al., 2013).

As in Study 4, the results did not support the hypothesis. Gesture deixis was found to increase following the explainees' verbal signals of complete understanding, for example, when the explainees repeated or completed the explainers' utterances, provided positive feedback with backchannels, or indicated non-understanding by themselves. These findings were in line with the findings from Study 4, emphasizing the fact that the absence of the explanandum is a main factor related to the intensity and continuity of gesture deixis. This behavior likely reflects a persistent communicative strategy compensating for the absence of shared physical referents in the interaction space.

7.4.3 Discussion: Individual variations of gesturing behavior

All three studies on explainers' gestural behavior provided insights into the degree of individual variation in their use of co-speech gestures while explaining a board game to different explainees. Previous research has shown that gesturing is idiosyncratic, particularly in terms of gesture form and path (Bergmann & Kopp, 2010; Priesters & Mittelberg, 2013), and that individual variations are influenced by the dialogue situation and the attendance of a different addressee (Bergmann & Kopp, 2010). Further, the gestural behavior of has been shown to vary in relation to the degree of shared expertise between interlocutors (Holler & Stevens, 2007; Jacobs & Garnham, 2007; Kang et al., 2015), as well as the physical presence and absence of an explanandum (Holler & Stevens, 2007).

Building on this research, Study 4 (see Section 7.2) tested the hypothesis that explainers would exhibit intra-individual variation in their use of gesture deixis depending on the monitored level of understanding. The analysis took advantage of the nested design of the dataset, in which each explainer subsequently interacted with three different explainees. As intra-individual variation in explanatory interactions had not been previously examined in depth, this hypothesis was tested in an exploratory manner.

The results demonstrated that the proportions of gesture deixis varied within each

explainer, in relation to the explainee with whom they were interacting, as well as to the subjective interpretations of the explainers regarding the explainees' levels of understanding. This pattern was further supported by the results for the nested random effect, which revealed greater intra-individual than inter-individual variation in the use of gesture deixis. Thus, the hypothesis was supported.

Comparable ad hoc findings emerged from Study 3 and Study 5 although they did not specifically investigate a hypothesis related to the individual variations of the explainers' gestural behavior. In both studies, a greater variation in the gestural behavior was observed for each of the individual explainers regarding their subsequent interactions with three different explainees, rather than between the individual explainers (see Sections 7.1.2 and 7.3.1). Together, these results suggested that the attendance and the behavior of the explainees are related to the gestural behavior of the explainers. In sum, explainers appear to dynamically adapt their gesturing behavior in response to explainees' feedback behavior in the course of interactional monitoring (Clark & Krych, 2004).

Chapter 8

General discussion

In the previous Chapters 6 and 7, I presented five studies investigating the relation between the dynamics of verbal and nonverbal explaining behavior and the motioning of explainees' feedback behavior. The verbal explaining behavior was analyzed in the form of explanation topics introduced by explainers, whereas the nonverbal explaining behavior was analyzed in the form of explainers' dynamic employment of co-speech gestures during explanatory interactions with different explainees.

Monitoring is a bilateral process in which interlocutors observe the verbal and nonverbal behavior of each other to keep track of the progress of an interaction, e.g., an explanation, and assess the explainee's understanding about the explanandum moment-by-moment (Clark & Krych, 2004). Together with scaffolding[1] and co-constructions[2], an explainer attempts to successfully reach the goal of an explanation, i.e., to enhance an explainees' understanding of an entity or a process (explanandum) (Buschmeier et al., 2023; Rohlfing et al., 2021). This theoretical framework suggested that explainers' co-constructions, realized through their verbal and nonverbal explaining behavior, tend to adapt to explainees' co-constructions, realized through their multimodal feedback behavior. To what extent does interactional monitoring become reflected in explainers' adaptations of verbal and nonverbal explaining behavior to explainees' feedback behavior, and do such adaptations of multimodal explaining behavior reach any limitations?

Verbal explaining behavior The dynamics of verbal explaining behavior was explored through the perspective of the topical structure of explanations, which was analyzed as consisting of new topics and elaborations of previously introduced topics. Topical shifts initiated by the explainers were related to various forms of

[1] Scaffolding is the process by which a more knowledgeable explainer adapts the explanation based on the responses of an explainee signaling their cognitive processing (Wood et al., 1976)

[2] Co-constructions emerge from the bi-directional verbal and nonverbal interaction between an explainer and an explainee, and represent the dynamics of explaining (for the explainers) and feedback (for the explainees) forms of behavior (Rohlfing et al., 2021).

multimodal feedback by the explainees, comprising their eye gaze behavior, head gestures (e.g., nodding), and vocal backchannels. Two studies—one on the domain of medical explanations (see Section 6.1) and another one on the domain of board game explanations (see Section 6.2)—demonstrated that explainees' gaze aversions from the explainers often preceded shifts to new topics. Explainees' gaze aversions predicted such topical shifts particularly when accompanied by head nods and / or backchannels expressing agreement or acknowledgment (e.g., *okay, good, alright*). A subsequent study on the topical structure of explanations including the dynamics of interactive gaze behavior (see Section 6.2) demonstrated that explainers' gaze withdrawals from explainees can also predict topical shifts. These findings were in line with the findings of Rossano's 2013, 2012 research on spontaneous conversations.

In contrast, explainees' non-changing gaze directed at explainers was related topical shifts to elaborations of existing topics. Although Lazarov et al. (2024) did not relate the topical shifts initiated by the explainers to explainees' levels of understanding, a recent study investigating multimodal predictors of (non-)understanding by Türk et al. (2024) demonstrated that, in moments of non-understanding, explainees "freeze", i.e., their behavior (including) eye gaze remains static. Thus, it is possible that when explainees keep a consistent gaze at explainers, explainers interpret this behavior as lack of understanding and a request for elaboration.

Although Study 1 and Study 2 suggested specific patterns of multimodal behavior (e.g., gaze aversion + head nodding + backchanneling) related to the changing topical structure of explanations, the topical changes of explanations varied to a higher extent regarding the individual explaining behavior of different explainers and to a lower extent regarding the attendance of a different explainee. In other words, each explainer followed an individually structured approach toward introducing explanation topics.

There are two possible explanations for the tendency revealed in the analysis of the random effect. The first reason is related to repeated patterns of explainees' multimodal feedback behavior. Especially, concerning the explainees' gaze aversions occurring prior to topic changes initiated by the explainers, Study 1 and Study 2 demonstrated that this relation is significant and valid for two different domains of everyday explanations. The second reason concerns the anticipated knowledge asymmetry between the explainers and the explainees at the beginning of explanations (Kotthoff, 2009; Rohlfing et al., 2021). In both Study 1 and Study 2, the explainers had prepared their explanations prior to the study. In Study 1, the physicians had prepared their explanations based on their fundamental knowledge in medical science and practical experience as clinical physicians. In contrast, the attending caregivers (except one person) were non-experts in the medical domain, and their multimodal feedback behavior frequently expressed agreement and acknowledgment to the (more knowledgeable) physicians. In Study 2, the explainers had prepared their explanations

of a board game a few days prior to the study by learning the game rules and practicing the game with others. Despite the fact that the explainers were aware of the explainees' unfamiliarity with the game prior to the study, the experience gap between the explainers and the explainees in Study 2 was assumed to be smaller compared to the experience gap between the explainers and explainees in Study 1. Thus, the explainers of the board game could have adjusted the topical structure of their explanations also in relation to the attendance of different explainees. This was suggested by the high standard deviation values of the nested random effect that represented the subsequent interactions of one explainer with three different explainees.

To make stronger assumptions about the adaptations of explanation topics to monitoring the explainees' behavior, it would be beneficial to analyze the actual sequence in which the explanation topics were introduced by the explainers. This would indicate to what extent explainers follow explanation structures, as prepared by themselves, and make adjustments in relation to the monitored feedback behavior and understanding of explainees.

Nonverbal explaining behavior The dynamics of nonverbal explaining behavior was investigated regarding the use of co-speech gestures by explainers across different categories of explanation topics, and in relation to monitoring the explainees' changing levels of understanding. All studies analyzed the same materials of board game explanations, during which the explanandum was physically absent from the shared space (see Section 5.2). Assuming that the physical absence of the board game is a prerequisite for the continuous establishment of "joint imagined spaces" (Kinalzik & Heller, 2020) over the course of the explanations, the studies focused on the occurrence of gesture dimensions, concretely deixis, iconicity, and temporal highlighting (gestural emphasis) (McNeill, 2006).

In terms of interactional monitoring and adaptive explaining behavior, several findings emerged:

1. Deixis was the most frequently used gesture dimension occurring alone and together with other dimensions, such as iconicity and temporal highlighting.

2. Temporal highlighting was used more frequently than iconicity, particularly in topics related to action processes and conditional rules.

3. Gesture deixis does not decrease even when the explainers interpret the explainees' understanding as complete—based on the reports from the video recall task and on perceiving the explainees' verbal feedback during the interactions.

4. The use of gesture deixis is related to and varies depending on the attendance of different explainees and the explainers' interpretations of the explainees' levels

of understanding.

These findings suggest that explainers continuously monitor explainees' behavior and adapt the use of co-speech gestures accordingly. This was further supported by the statistical models, which revealed higher intra-individual variation (i.e., within the same explainers, each interacting with different explainees) than inter-individual variation (i.e., between the different explainers). This pattern was observed across the three gesture-related studies and points to a clear link between the attendance of different explainees, monitoring their understanding, and the dynamics of the explainers' gesturing behavior. How can this tendency of intra-personal variation related to the attendance of different explainees be explained in comparison to the inter-personal variation revealed for the adaptations of the topical structure of explanations?

In contrast to the explanation topics which constitute the external structure of an explanation, co-speech gestures are used in a tight semantic and temporal coupling within the affiliated spoken reference, which could be part of or represent an explanation topic (Kendon, 2004; Kita, 2009; McNeill, 1992). Thus, the variation in the use of co-speech gestures depends on the spoken references (words) within utterances to which the gestures are affiliated. Thus, the variation in the use of co-speech gestures by the explainers' in relation to monitoring the explainees' understanding could be observed in variations of the affiliated spoken references. The variation of co-speech gestures in relation to the affiliated spoken references remained a limitation which could be addressed in future research. Although Study 4 and Study 5 did not integrate the affiliated spoken unit of speech in the analyses, the annotation of the explainers' deictic gestures took place in the presence of speech. The common objective of both studies was to demonstrate how explainers' deictic gestures vary in relation to the interpreted levels of understanding as monitored in the explainees' behavior. In summary, all presented studies suggest that adaptations of verbal and nonverbal explaining behavior take place in two steps: (1) At the verbal level of explaining, explainers follow a previously prepared topical structure, which may vary in relation to specific patterns of feedback behavior; (2) In relation to the semantic focus of explanation topics, explainers employ specific gestural dimensions in order to increase the understanding of explainees; however, as understanding is a dynamically changing variable, the frequency of co-speech gestures also becomes adapted accordingly.

Chapter 9

Conclusion

The present thesis synthesized and discussed the results of five empirical studies that investigated the reflection of interactional monitoring (Clark & Krych, 2004) in adaptations of verbal and nonverbal explaining behavior. The studies investigated two samples from video corpora on everyday explanations, particularly from the domains of medical and board game explanations.

All studies demonstrated that explaining behavior adapts with respect to explainees' individual feedback behavior, comprising eye gaze, head gestures, and vocal backchannelling. Specifically, the first two studies on adaptations of the topical structure of explanations (verbal explaining) demonstrated that these adaptations are more pronounced rather between than within individual explainers who interacted with different explainees subsequently. In contrast, adaptations of the gestural behavior were shown to be more pronounced within different explainers and with respect to the explainers' interpretations of the explainees' levels of understanding. This variation was confirmed in the last two studies that addressed the monitoring of explainees' understanding in a post hoc video recall task and while perceiving the explainees' verbal feedback during the explanations.

In sum, interactional monitoring in explanatory interactions functions as an interactional process in which interlocutors observe, interpret, and adapt each other's behavior. Thereby, explainers are able to adapt explanations in real time by employing co-constructive verbal and nonverbal behaviors (Clark & Krych, 2004; Rohlfing et al., 2021; Wood et al., 1976). As demonstrated in this work, such co-constructive behaviors are, for example, the dynamically changing topical structure of explanations at the verbal level and the dynamic use of (multidimensional) co-speech gestures at the nonverbal level.

Bibliography

Abeles, D., & Yuval-Greenberg, S. (2017). Just look away: Gaze aversions as an overt attentional disengagement mechanism. *Cognition*, *168*, 99–109. https://doi.org/10.1016/j.cognition.2017.06.021

Alibali, M. W., Nathan, M. J., Church, R. B., Wolfgram, M. S., Kim, S., & Knuth, E. J. (2013). Teachers' gestures and speech in mathematics lessons: Forging common ground by resolving trouble spots. *ZDM Mathematics Education*, *45*, 425–440. https://doi.org/10.1007/s11858-012-0476-0

Allen, D. E., & Guy, R. F. (1977). Ocular breaks and verbal output. *Sociometry*, *40*(1), 90–96. https://doi.org/10.2307/3033550

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*, 273–287. https://doi.org/doi.org/10.1007/s10579-007-9061-5

Allwood, J., & Cerrato, L. (2003). A study of gestural feedback expressions. *Proceedings of the 1st Nordic Symposium on Multimodal Communication*, 7–22.

Allwood, J., Nivre, J., & Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, *9*, 1–26. https://doi.org/10.1093/jos/9.1.1

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge University Press.

Arnold, K. (2012). Humming along. *Contemporary Psychoanalysis*, *48*(1), 117. https://doi.org/https://doi.org/10.1080/00107530.2012.10746491

Austin, E. E., & Sweller, N. (2014). Presentation and production: The role of gesture in spatial communication. *Journal of Experimental Child Psychology*, *122*, 92–103. https://doi.org/https://doi.org/10.1016/j.jecp.2013.12.008

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bavelas, J. B., Black, A., Lemery, C. R., & Mullett, J. (1986). "i show how you feel": Motor mimicry as a communicative act. *Journal of Personality and Social Psychology*, *50*(2). https://doi.org/https://psycnet.apa.org/doi/10.1037/0022-3514.50.2.322

Bavelas, J., & Chovil, N. (2018). Some pragmatic functions of conversational facial gestures. *Gesture*, *17*(1), 98–127. https://doi.org/https://doi.org/10.1075/gest.00012.bav

Bavelas, J. B., & Chovil, N. (2018). Some pragmatic functions of conversational facial gestures. *Gesture*, *17*(1), 98–127. https://doi.org/10.1075/gest.00012.bav

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, *52*(3), 566–580. https://doi.org/10.1111/j.1460-2466.2002.tb02562.x

Bazzanella, C., & Damiano, R. (1999). The interactional handling of misunderstanding in everyday conversations. *Journal of Pragmatics*, *31*, 817–836. https://doi.org/10.1016/s0378-2166(98)00058-7

Beattie, G., & Shovelton, H. (2005). Why the spontaneous images created by the hands during talk can help make TV advertisements more effective. *British Journal of Psychology*, *96*(1), 21–37. https://doi.org/10.1348/000712605X103500

Beege, M., Ninaus, M., Schneider, S., Nebel, S., Schlemmel, J., Weidenmüller, J., Moeller, K., & Rey, G. D. (2020). Investigating the effects of beat and deictic gestures of a lecturer in educational videos. *Computers & Education*, *156*, 103955. https://doi.org/10.1016/j.compedu.2020.103955

Bergmann, K., & Kopp, S. (2010). Systematicity and idiosyncrasy in iconic gesture use: Empirical analysis and computational modeling. *Gesture in Embodied Communication and Human-Computer Interaction*, 182–194.

Betz, S., Zarrieß, S., Székely, É., & Wagner, P. (2019). The greennn tree — lengthening position influences uncertainty perception. *Proceedings of Interspeech 2019*, 3990–3994. https://doi.org/10.21437/Interspeech.2019-2572

Brône, G., Oben, B., Jehoul, A., Vranjes, J., & Feyaerts, K. (2017). Eye gaze and viewpoint in multimodal interaction management. *Cognitive Linguistics*, *28*(3), 449–483. https://doi.org/10.1515/cog-2016-0119

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Bunt, H. C. (2009). The dit++ taxonomy for functional dialogue markup. In D. Heylen, C. Pelachaud, R. Catizone, & D. Traum (Eds.), *Proceedings of edaml@aamas, workshop towards a standard markup language for embodied dialogue acts* (pp. 13–24). International Foundation for Autonomous Agents; Multi-agent Systems.

Bunt, H. C. (2011). Multifunctionality in dialogue. *Computer Speech and Language*, *25*, 222–245. https://doi.org/10.1016/j.csl.2010.04.006

Buschmeier, H., Buhl, H. M., Kern, F., Grimminger, A., Beierling, H., Fischer, J., Groß, A., Horwath, I., Klowait, N., Lazarov, S., Lenke, M., Lohmer, V., Rohlfing, K.,

Scharlau, I., Singh, A., Terfloth, L., Vollmer, A., Wang, Y., Wilmes, A., & Wrede, B. (2023). Forms of understanding of xai-explanations. https://arxiv.org/abs/2311.08760

Cerrato, L. (2005). Linguistic functions of head nods. *Proceedings of the 2nd Nordic Symposium on Multimodal Communication*. https://api.semanticscholar.org/CorpusID:60150443

Clark, H. H. (1997). Dogmas of understanding. *Discourse Processes*, *23*(3), 567–598. https://doi.org/10.1080/01638539709545003

Clark, H. H. (2003). Pointing and placing. In S. Kita (Ed.), *Pointing: Where Language, Culture, and Cognition Meet* (pp. 243–268). Psychology Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 222–233). American Psychological Association.

Clark, H. H., & Krych, M. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*, 62–81. https://doi.org/10.1016/j.jml.2003.08.004

Congdon, E. L., Novack, M. A., Brooks, N., Hemani-Lopez, N., O'Keefe, L., & Goldin-Meadow, S. (2017). Better together: Simultaneous presentation of speech and gesture in math instruction supports generalization and retention. *Learning and Instruction*, *50*, 65–74. https://doi.org/10.1016/j.learninstruc.2017.03.005

Cook, M. (1977). Gaze and mutual gaze in social encounters: How long—and when—we look others "in the eye" is one of the main signals in nonverbal communication. *American Scientist*, *65*(3), 328–333. https://www.jstor.org/stable/27847843

Core, M. G., & Allen, J. (1997). Coding dialogs with the damsl annotation scheme. *AAAI Fall Symposium on Communicative Action in Humans and Machines*, *56*, 28–35.

Dargue, N., Phillips, M., & Sweller, N. (2021). Filling in the gaps: Observing gestures conveying additional information can compensate for missing verbal content. *Instructional Science*, *49*, 637–659. https://doi.org/10.1007/s11251-021-09549-2

de Ruiter, J. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and Gesture* (pp. 284–311). Cambridge University Press. https://doi.org/10.1017/CBO9780511620850.018

De Stefani, E. (2021). Embodied responses to questions-in-progress: Silent nods as affirmative answers. *Discourse Processes*, *58*(4), 353–371. https://doi.org/10.1080/0163853X.2020.1836916

Degutyte, Z., & Astell, A. (2021). The role of eye gaze in regulating turn taking in conversations: A systematized review of methods and findings. *Frontiers in Psychology*, *12*, 616471. https://doi.org/10.3389/fpsyg.2021.616471

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2005). *A modern introduction to probability and statistics: Understanding why and how*. Springer London. https://doi.org/https://doi.org/10.1007/1-84628-168-7

Dimitrova, D., Chu, M., Wang, L., Özyürek, A., & Hagoort, P. (2016). Beat that word: How listeners integrate beat gesture and focus in multimodal speech discourse. *Journal of Cognitive Neuroscience*, *28*(9), 1255–1269. https://doi.org/10.1162/jocn_a_00963

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., Kendrick, K. H., Levinson, S. C., Manrique, E., Rossi, G., & Enfield, N. J. (2015). Universal principles in the repair of communication problems. *PLOS ONE*, *10*, 1–15. https://doi.org/https://doi.org/10.1371/journal.pone.0136100

Eshghi, A., Howes, C., Gregoromichelaki, E., Hough, J., & Purver, M. (2015). Feedback in conversation as incremental semantic update. *Proceedings of the 11th International Conference on Computational Semantics*, 261–271.

Fisher, J. B., Lohmer, V., Kern, F., Barthlen, W., Gaus, S., & Rohlfing, K. (2022). Exploring monological and dialogical phases in naturally occurring explanations. *Künstliche Intelligenz*, *36*, 317–326. https://doi.org/doi.org/10.1007/s13218-022-00787-1

Fisher, J. B., Robrecht, A., Kopp, S., & Rohlfing, K. J. (2023). Exploring the semantic dialogue patterns of explanations – a case study of game explanations. *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue*.

Gander, A. J., & Gander, P. (2020). Micro-feedback as cues to understanding in communication. In C. Howes, S. Dobnik, & E. Breitholtz (Eds.), *Dialogue and Perception. Extended papers from DaP2018* (pp. 1–11). Gothenburg University.

Gardner, R. (2001). *When listeners talk: Response tokens and listener stance* (Vol. 92). John Benjamins. https://doi.org/10.1075/pbns.92

Glenberg, A. M., Schroeder, J. L., & Robertson, D. A. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition*, *26*, 651–658. https://doi.org/10.3758/bf03211385

Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearers*. Academic Press.

Goodwin, C. (1985). Notes on story structure and the organization of participation. In J. M. Atkinson (Ed.), *Structures of social action* (pp. 225–246). Cambridge University Press.

Goodwin, M. H., & Goodwin, C. (1986). Gesture and coparticipation in the activity of searching for a word. *Semiotica*, *62*(1–2), 51–75. https://doi.org/10.1515/semi.1986.62.1-2.51

Gravano, A., Hirschberg, J., & Beňuš, Š. (2012). Affirmative cue words in task-oriented dialogue. *Computational Linguistics*, *38*, 1–39. https://doi.org/10.1162/COLI_a_00083

Grimminger, A., Rohlfing, K. J., & Stenneken, P. (2010). Children's lexical skills and task demands affect gestural behavior in mothers of late-talking children and children with typical language development. *Gesture*, *10*(2–3), 251–278. https://doi.org/doi.org/10.1075/gest.10.2-3.07gri

Haas, J., Warnke, V., Niemann, H., Cettolo, M., Corazza, A., Falavigna, D., & Lazzari, G. (1999). Semantic boundaries in multiple languages. *6th European Conference on Speech Communication and Technology*, 535–538. https://doi.org/10.21437/Eurospeech.1999-138

Habets, B., Kita, S., Shao, Z., Özyürek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of Cognitive Neuroscience*, *23*(8), 1845–1854.

Hadar, U. and Steiner, T.J. and Grant, E.C. and Clifford Rose, F. (1983). Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, *2*(1), 35–46. https://doi.org/https://doi.org/10.1016/0167-9457(83)90004-0

Heller, V. (2021). Embodied displays of "Doing Thinking." epistemic and interactive functions of thinking displays in children's argumentative activities. *Frontiers in Psychology*, *12*, 636671. https://doi.org/10.3389/fpsyg.2021.636671

Hessels, R. S. (2020). How does gaze to faces support face-to-face interaction? a review and perspective. *Psychonomic Bulletin & Review*, *27*(5), 856–881. https://doi.org/10.3758/s13423-020-01715-w

Holler, J., & Wilkin, K. (2011). An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses. *Journal of Pragmatics*, *43*(14), 3522–3536. https://doi.org/110.1016/j.pragma.2011.08.002

Holler, J., & Stevens, R. (2007). The effect of common ground on how speakers use gesture and speech to represent size information. *Journal of Language and Social Psychology*, *26*(1), 4–27. https://doi.org/10.1177/0261927X06296428

Hömke, P., Holler, J., & Levinson, S. C. (2017). Eye blinking as addressee feedback in face-to-face conversation. *Research on Language and Social Interaction*, *50*(1), 54–70. https://doi.org/10.1080/08351813.2017.1262143

Ismail, N. M., & Syahputri, V. N. (2022). "i mean you can stop. i already understand you": Head tilts during conversations. *Lingua Didaktika: Jurnal Bahasa dan Pembelajaran Bahasa*, *16*(1), 1–11. https://doi.org/10.24036/ld.v16i1.116673

Jacobs, N., & Garnham, A. (2007). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language*, *56*(2), 291–303. https://doi.org/https://doi.org/10.1016/j.jml.2006.07.011

Jokinen, K., Harada, K., Nishida, M., & Yamamoto, S. (2010). Turn-alignment using eye gaze and speech in conversational interaction. *Proceedings of the European Conference on Speech Communication and Technology (INTERSPEECH'10)*, 2018–2021. https://doi.org/10.21437/Interspeech.2010-571

Jokinen, K., Nishida, M., & Yamamoto, S. (2010). On eye gaze and turn taking. *Proceedings of the International Conference on Intelligent User Interfaces Workshop on Eye Gaze in Intelligent Human-Machine Interaction*, 118–123. https://doi.org/10.1145/2002333.2002352

Kandana-Arachchige, K. G., Blekic, W., Loureiro, I. S., & Lefebvre, L. (2021). Covert attention to gestures is sufficient for information uptake. *Frontiers in Psychology*, *12*, 776867. https://doi.org/10.3389/fpsyg.2021.776867

Kang, S., Tversky, B., & and, J. B. B. (2015). Coordinating gesture, word, and diagram: Explanations for experts and novices. *Spatial Cognition & Computation*, *15*(1), 1–26. https://doi.org/10.1080/13875868.2014.958837

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260–267.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*(1), 22–63. https://doi.org/10.1016/0001-6918(67)90005-4

Kendon, A. (1970). Movement coordination in social interaction: Some examples described. *Acta Psychologica*, *32*(2), 100–125. https://doi.org/10.1016/0001-6918(70)90094-6

Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in southern italian conversation. *Journal of Pragmatics*, *23*(3), 247–279. https://doi.org/10.1016/0378-2166(94)00037-F

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge University Press. https://doi.org/10.1017/cbo9780511807572

Kinalzik, N., & Heller, V. (2020). Establishing joint imagined spaces in game explanations. *Research on Children and Social Interaction*, *4*(1), 28–50. https://doi.org/10.1558/rcsi.12417

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, *24*(2), 145–167. https://doi.org/10.1080/01690960802586188

Klein, J. (2009). Erklären-was, erklären-wie, erklären-warum. typologie und komplexität zentraler akte der welterschließung. In R. Vogt (Ed.), *Erklären. gesprächsanalytische und fachdidaktische perspektiven* (pp. 25–36). Stauffenburg-Verlag.

Kotthoff, H. (2009). Erklärende aktivitätstypen in alltags- und unterrichtskontexten. In J. Spreckels (Ed.), *Erklären im kontext. neue perspektiven aus der gesprächs- und unterrichtsforschung* (pp. 120–146). Schneider.

Kousidis, S., Malisz, S., Wagner, P., & Schlangen, D. (2013). Exploring annotation of head gesture forms in spontaneous human interaction. *Proceedings of the Tilburg Gesture Meeting (TiGeR2013)*, 1–4. https://pub.uni-bielefeld.de/record/2567303

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Departmental Papers (ASC)*, *30*(3), 1–16. https://doi.org/https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology (3rd ed)*. Sage Publications.

Krych, M., & Clark, H. H. (1997). Coordinating hands, eyes, and voice. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *19*. https://escholarship.org/uc/item/7qb8792n

Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *The American Journal of Psychology*, *113*, 387–404. https://doi.org/10.2307/1423365

Lai, C. (2009). Perceiving surprise on cue words: Prosody and semantics interact on right and really. *Proceeings of Interspeech 2009*, 1963–1966. https://doi.org/10.21437/Interspeech.2009-475

Lai, C. (2010). What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue. *Proceedings of Interspeech 2010*, 1413–1416. https://doi.org/10.21437/Interspeech.2010-429

Lazarov, S., Biermeier, K., & Grimminger, A. (2024). Changes in the topical structure of explanations are related to explainees' multimodal behaviour. *Interaction Studies*, *25*(3), 257–280. https://doi.org/10.1075/is.23033.laz

Lazarov, S., & Grimminger, A. (under review). How are mutual gaze and gaze withdrawals related to the changing topical structure of dyadic explanatory interactions? *Journal of Nonverbal Behavior*, *tba*, tba. https://doi.org/tba

Lazarov, S., Schaffer, M. E., GLadow, V., Buschmeier, H., Grimminger, A., & Buhl, H. (2025). Applications of video-recall for the assessment of understanding and knowledge in explanatory contexts [OSF Preprints]. https://doi.org/10.31219/osf.io/u24kz_v1

Lazarov, S., & Grimminger, A. (2024a). Variations in explainers' gesture deixis in explanations related to the monitoring of explainees' understanding. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*, 4805–4812. https://escholarship.org/uc/item/7dz8n8tf

Lazarov, S., & Grimminger, A. (2024b). Verbal signals of understanding do not predict a decrease of gesture deixis. *Proceedings of the 3rd International Multimodal Communication Symposium*, 54–55. http://mmsym.org/?page%20id=379.

Lazarov, S., & Grimminger, A. (2025). Different explanation topics, different gestural dimensions? [poster presentation] [July 9–11, 2025, Radboud University, Nijmegen, NL].

Lazarov, S., Türk, O., Grimminger, A., Wagner, P., & Buschmeier, H. (2025). Annotation schemes project A02 "Monitoring the understanding of explanations". osf.io/j2dha

Li, W., Wang, F., Mayer, R. E., & Liu, T. (2022). Animated pedagogical agents enhance learning outcomes and brain activity during learning. *Journal of Computer Assisted Learning*, *38*(3), 621–637. https://doi.org/https://doi.org/10.1111/jcal.12634

Malisz, Z., Włodarczak, M., Buschmeier, H., Skubisz, J., Kopp, S., & Wagner, P. (2016). The ALICO corpus: Analysing the active listener. *Language Resources and Evaluation*, *50*, 411–442. https://doi.org/10.1007/s10579-016-9355-6

Markson, L., & Paterson, K. B. (2009). Effects of gaze-aversion on visual-spatial imagination. *British Journal of Psychology*, *100*(3), 553–563. https://doi.org/https://doi.org/10.1348/000712608X371762

McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, *32*(7), 855–878. https://doi.org/10.1016/S0378-2166(99)00079-X

McKern, N., Dargue, N., Sweller, N., Sekine, K., & Austin, E. (2021). Lending a hand to storytelling: Gesture's effects on narrative comprehension moderated by task difficulty and cognitive ability. *Quarterly Journal of Experimental Psychology*, *74*(10), 1781–1895. https://doi.org/10.1177/17470218211024913

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

McNeill, D. (1998). Speech and gesture integration. In J. M. Iverson & S. Goldin-Meadow (Eds.), *The nature and functions of gesture in children's communication* (pp. 11–27). Jossey-Bass. https://doi.org/10.1002/cd.23219987902

McNeill, D. (2005). *Gesture and thought*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226514642.001.0001

McNeill, D. (2006). Gesture and communication. In K. Brown (Ed.), *Encyclopedia of Language & Linguistics* (2nd ed., pp. 60–66). Georgetown University Press. https://doi.org/10.1016/B0-08-044854-2/00798-7

Müller, C. (2013). 12. gestures as a medium of expression: The linguistic potential of gestures. In C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill, & S. Tessendorf (Eds.), *Volume 1* (pp. 202–217). De Gruyter Mouton. https://doi.org/doi:10.1515/9783110261318.202

Neiberg, D., Salvi, G., & Gustafson, J. (2013). Semi-supervised methods for exploring the acoustics of simple productive feedback. *Speech Communication*, *55*, 451–469. https://doi.org/10.1016/j.specom.2012.12.007

Nota, N., Trujillo, J. P., & Holler, J. (2021). Facial signals and social actions in multimodal face-to-face interaction. *Brain Sciences*, *11*(8), 1017. https://doi.org/10.3390/brainsci11081017

Nota, N., Trujillo, J. P., & Holler, J. (2023). Specific facial signals associate with categories of social actions conveyed through questions. *PLOS ONE*, *18*(7), 1–26. https://doi.org/10.1371/journal.pone.0288104

Nota, N., Trujillo, J. P., Jacobs, V., & Holler, J. (2023). Facilitating question identification through natural intensity eyebrow movements in virtual avatars. *Scientific Reports*, *13*. https://doi.org/https://doi.org/10.1038/s41598-023-48586-4

Park, H. W., Gelsomini, M., Lee, J. J., & Breazeal, C. (2017). Telling stories to robots: The effect of backchannelling on a child's storytelling. *Proceedings of the 2017 ACM/IEEE International Conference in Human-Robot Interaction*, 100–109. https://doi.org/10.1145/2909824.3020245

Pekarek Doehler, S. (2022). Multimodal action formats for managing preference: Chais pas 'dunno' plus gaze conduct in dispreferred responses to questions. *Journal of Pragmatics*, *197*, 81–99. https://doi.org/https://doi.org/10.1016/j.pragma.2022.05.010

Phelps, F. G., Doherty-Sneddon, G., & Warnock, H. (2006). Helping children think: Gaze aversion and teaching. *British Journal of Developmental Psychology*, *24*(3), 577–588. https://doi.org/10.1348/026151005X49872

Ping, R. M., Goldin-Meadow, S., & Beilock, S. L. (2013). Understanding gesture: Is the listener's motor system involved? *Journal of Experimental Psychology: General*, *143*(1), 195–204. https://doi.org/https://psycnet.apa.org/doi/10.1037/a0032246

Priesters, M. A., & Mittelberg, I. (2013). Individual differences in speakers' gesture spaces: Multi angle views from a motion capture study. *TiGeR Workshop*, 1–4.

Rohlfing, K., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H., Buschmeier, H., Grimminger, A., Hammer, B., Häb-Umbach, R., Horwath, I., Hüllermeier, E., Kern, F., Kopp, S., Thommes, K., Ngonga Ngomo, A.-C., Schulte, C., Wachsmuth, H., Wagner, P., & Wrede, B. (2021). Explanation as a social practice: Toward a conceptual framework for the social design of ai systems. *IEEE Transactions on Cognitive and Developmental Systems*, *13*, 717–728. https://doi.org/10.1109/TCDS.2020.3044366

Rohrer, P. L., Tütüncübasi, U., I, V.-G., Florit-Pons, J., Gibert, N. E., P, R., Shattuck-Hufnagel, S., & P., P. (2020). The multimodal multidimensional (m3d) labeling system. https://osf.io/ankdx/

Rohrer, P. L., Delais-Roussarie, E., & Prieto, P. (2020). Beat gestures for comprehension and recall: Differential effects of language learners and native listeners. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.575929

Roscoe, R., & Chi, M. T. H. (2008). Tutor learning: The role of explaining and responding to questions. *Instructional Science*, *36*(4), 321–350. https://doi.org/http://dx.doi.org/10.1007/s11251-007-9034-5

Rossano, F. (2013). Gaze in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 308–329). Wiley-Blackwell.

Rossano, F. (2012). *Gaze Behavior in Face-to-face Interaction* [Doctoral dissertation, Radboud University Nijmegen, Nijmegen].

RStudio Team. (2020). *Rstudio: Integrated development environment for r*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language*, *50*(4), 696–735. https://doi.org/10.1016/B978-0-12-623550-0.50008-2

Sasaki, J., & Sasaki, G. (2014). *Deep sea adventure (tabletop game)*. Oink Games.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh-huh' and other things that come between sentences. In D. Tannen (Ed.), *Analyzing discourse: Text and talk. 32nd georgetown round table on languages and linguistics 1981* (pp. 71–93). Georgetown University Press.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211–232. https://doi.org/10.1016/0010-0285(89)90008-X

Stojnic, U., Stone, M., & Lepore, E. (2013). Deixis (even without pointing). *Philosophical Perspectives*, *27*(1), 502–525. https://doi.org/https://doi.org/10.1111/phpe.12033

Streeck, J. (2008). Depicting by gesture. *Gesture*, *8*(3), 285–301. https://doi.org/https://doi.org/10.1075/gest.8.3.02str

Tang, B., Frye, H. A., Gelfand, A. E., & Silander, J. A. (2023). Zero-inflated beta distribution regression modeling. *Journal of Agricultural, Biological and Environmental Statistics*, *28*, 117–137. https://doi.org/https://doi.org/10.1007/s13253-022-00516-z

Teßendorf, S. (2013). Emblems, quotable gestures, or conventionalized body movements. In C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. David McNeill, & S. Teßendorf (Eds.), *Body – Language – Communication: An International Handbook on Multimodality in Human Interaction* (pp. 82–100). De Gruyter Mouton. https://doi.org/10.1515/9783110261318.82

Türk, O., Lazarov, S., Wang, Y., Buschmeier, H., Grimminger, A., & Wagner, P. (2024). Predictability of understanding in explanatory interactions based on multi-

modal cues. *Proceedings of the 26th International Conference on Multimodal Interaction*, 449–458. https://doi.org/10.1145/3678957.3685741

Türk, O., Wagner, P., Buschmeier, H., Grimminger, A., Wang, Y., & Lazarov, S. (2023). MUNDEX: A multimodal corpus for the study of the understanding of explanations. *Proceedings of the 1st International Multimodal Communication Symposium*, 63–64. http://nbn-resolving.de/urn:nbn:de:0070-pub-29805458

Vendler, Z. (1994). Understanding misunderstanding. In D. Jamieson (Ed.), *Language, mind, and art: Essays in appreciation and analysis in honor of Paul Ziff* (pp. 9–22). Springer. https://doi.org/10.1007/978-94-015-8313-8_2

Ward, N. (2006). Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, *14*, 129–182. https://doi.org/10.1075/pc.14.1.08war

Welsh, D. P., & Dickson, J. W. (2005). Video-recall procedures for examining subjective understanding in observational data. *Journal of Family Psychology*, *19*(1), 62–71. https://doi.org/10.1037/0893-3200.19.1.62

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. *Proceedings the 5th International Conference on Language Resources and Evaluation*, 1556–1559. https://aclanthology.org/L06-1082/

Włodarczak, M., Buschmeier, H., Malisz, S., Kopp, S., & Wagner, P. (2012). Listener head gestures and verbal feedback expressions in a distraction task. *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog, INTERSPEECH2012 Satellite Workshop*, 93–96. https://www.isca-archive.org/fbid_2012/wodarczak12_fbid.html

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100. https://doi.org/https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Wu, R.-J. R., & Heritage, J. (2025). The assertoric nod: Non-concordant uses in responses to polar questions in english conversation. *Discourse Studies*. https://doi.org/10.1177/14614456241310591

Yngve, V. H. (1970). On getting a word in edgewise. In M. A. Campbell et al. (Eds.), *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society* (pp. 567–577). Chicago Linguistic Society.

Appendix: Related Publications

This appendix contains the original publications cited in the empirical part of the dissertation: three articles (Lazarov & Grimminger, under review; Lazarov et al., 2024; Lazarov & Grimminger, 2024a), one conference poster presentation (Lazarov & Grimminger, 2025), and one conference abstract (Lazarov & Grimminger, 2024b). The publications are attached in the order of appearance in Chapters 6 and 7.

Statement about authors' contribution

The following publications attached to this cumulative thesis are the product of collaborations between the first author and other co-authors. In all attached works the first author is the same as the author of the cumulative thesis entitled "The reflection of interactional monitoring in the dynamics of verbal and nonverbal explaining behavior". Each co-author's contribution to the attached works is described below.

1. Lazarov, S., Biermeier, K., & Grimminger, A. (2024). Changes in the topical structure of explanations are related to explainees' multimodal behaviour. *Interaction Studies, 25*(3), 257–280. https://doi.org/10.1075/is.23033.laz.

Stefan Lazarov – Writing: original draft of the manuscript, Writing: Review and Editing

Kai Biermeier – Conceptualization of the analysis presented in section *3.4.1. Conditional probabilities*

Angela Grimminger – Writing: Review and Editing

2. Lazarov, S., & Grimminger, A. (under review). How are mutual gaze and gaze withdrawals related to the changing topical structure of dyadic explanatory interactions? *Journal of Nonverbal Behavior*, tba, tba. https://doi.org/tba

Stefan Lazarov – Writing: original draft of the manuscript, Writing: Review and Editing

Angela Grimminger – Writing: Review and Editing

3. Lazarov, S., & Grimminger, A. (2025). Different explanation topics, different gestural dimensions? [poster presentation] [July 9–11, 2025, Radboud University, Nijmegen, NL].

Stefan Lazarov – Writing: original draft of the manuscript, Writing: Review and Editing

Angela Grimminger – Writing: Review and Editing

4. Lazarov, S., & Grimminger, A. (2024a). Variations in explainers' gesture deixis in explanations related to the monitoring of explainees' understanding. Proceedings of the Annual Meeting of the Cognitive Science Society, 46, 4805–4812. https://escholarship.org/uc/item/7dz8n8tf.

Stefan Lazarov – Writing: original draft of the manuscript, Writing: Review and Editing

Angela Grimminger – Writing: Review and Editing

5. Lazarov, S., & Grimminger, A. (2024b). Verbal signals of understanding do not predict

a decrease of gesture deixis. Proceedings of the 3rd International Multimodal

Communication Symposium, 54–55. https://mmsym.org/?page%20id=379.

Stefan Lazarov – Writing: original draft of the manuscript, Writing: Review and Editing

Angela Grimminger – Writing: Review and Editing

I hereby declare all abovementioned authors' contributions as correct.

Paderborn, 24.06.2025

_____ _____

Changes in the topical structure of explanations are related to explainees' multimodal behaviour

Stefan Lazarov, Kai Biermeier, and Angela Grimminger
Paderborn University

Everyday explanations are interactive processes with the aim to provide a less knowledgeable person with reasonable information about other people, objects, or events. Because explanations are interactive communicative processes, the topical structure of an explanation may vary dynamically depending on the immediate feedback of the explainee. In this paper, we analyse topical transitions in medical explanations organised by different physicians (explainers) related to different forms of multimodal behaviour of caregivers (explainees) attending an explanation about the procedures of an upcoming surgery of a child. The analyses reveal that explainees' multimodal behaviour with gaze shifts (and particularly gaze aversion) can predict a transition from an elaborated topic to a new one, whereas explainees' forms of multimodal behaviour with static gaze cannot be related to changes of the topical structure.

Keywords: explanations, multimodal behaviour, elaborations, conditional probabilities

1. Introduction

Everyday explanations between humans are interactive processes pursuing the aim to provide a less knowledgeable person (henceforth explainee) with reasonable information about other people, objects, or events (e.g., Rohlfing et al., 2021). One example is the explanatory domain of health care in conversations between medical doctors (henceforth physicians) and patients. In such contexts, a physician takes the role of an explaining expert who has more knowledge than the patient (i.e., the explainee). The relation between an expert and a non-expert presupposes that the expert is the conversation partner who organises the explanation in terms of setting an explanandum or several explananda (i.e., the object(s)

https://doi.org/10.1075/is.23033.laz
Interaction Studies 25:3 (2024), pp. 257–280. ISSN 1572-0373 | E-ISSN 1572-0381

of the explanation), such as the diagnosis, reasons, medical procedures, and treatments, and also structuring the explanans (i.e., the way that an explanation is expressed). Because we view explanations as being co-constructed and therefore unfolding within the interaction (Rohlfing et al., 2021), the number of different topics and elaborations of them, most likely, changes and varies dynamically depending on the immediate feedback of the explainee (Clark & Krych, 2004).

In this study, we investigate ten naturalistic, videotaped dialogues between caregivers who are required to give their agreement to an upcoming surgery of their children, and physicians. In particular, we analyse the temporal relation between different forms of caregivers' multimodal behaviour (including gaze, head nodding and backchannelling) and different forms of topical transitions within explanations that were initiated by different physicians. The current research addresses the process of interactional monitoring, in which interlocutors track the process of understanding an explanandum (Clark & Krych, 2004). Explainers may feel required to make elaborations via additions, completions, or paraphrases, so that explainees are able to solve understanding-related issues (Dingemanse et al., 2015), but there is yet no account about what (non-)verbal signals of an explainee are interpreted as being either related to a request for an elaboration or as a sign of sufficient understanding.

2. Theoretical background

2.1 Monitoring listeners' behaviour

Explanatory dialogues such as those between physicians and patients (in this study, caregivers) support the so-called "bilateral accounts" of interactions, meaning that interaction partners inform each other about the current state of the explanation or understanding, and they *monitor* one another (Clark & Krych, 2004). More precisely, listeners' (non-)verbal behaviour is found to influence the length and form of the turns made by interlocutors (Sacks et al., 1974). For example, speakers tend to adapt the content of their utterances, e.g., make repetitions, in relation to listeners' disattending gaze behaviour, which may be interpreted as a request for further information (Goodwin, 1981). By monitoring listeners' (non-)verbal behaviour, explainers could adapt the structure of their explanations to (their interpretation of) the immediate behaviour of the listening and perceiving explainee. Thus, explainers have the organisational control over explanations, but the successful completion of the explanation also depends on the explainees.

According to Clark & Krych (2004), speakers pay attention at and respond to several levels during interactions with listeners, namely at the verbal level

(speech), the nonverbal level (e.g., faces, bodies and gestures), or also at shared scenes. For example, verbal contributions represent full content utterances or short backchannels as immediate feedback. The face area encompasses facial mimics and eye gaze behaviour by which interlocutors indicate attention to each other's verbal (Clark & Krych, 2004: 63; Goodwin, 1981; Kendon, 1967) and nonverbal behaviour (e.g., Bavelas et al., 1986). Further, facial signals such as blinking (Hömke et al., 2017) or eyebrow movements (Hömke et al., 2022) have been shown to be related to (mis)understanding. Body movements, such as head nods or tilts, also play an important role in signalling immediate feedback that is monitored by the speakers (Clark & Krych, 2004: 64), such as understanding (Ismail & Syahputri, 2022).

2.2 (Non-)verbal forms of behaviour

For our analysis, we did not consider the overall (non-)verbal behaviour of the caregivers during the entire interactions with the physicians, but only forms of such behaviour around specific temporal areas of interest, namely the transitions to elaborations and those from elaborations to new topics initiated by the physicians (see 3.3. for further details). An initial video inspection of those temporal areas indicated that caregiver's forms of multimodal feedback behaviour varied in different combinations of gaze behaviour, head nodding and backchannelling. Each of these modalities could be related to different interactional and cognitive processes, which could be interpreted by the interlocutors in different and ambiguous ways while attempting to complete the goal of an interaction (e.g., understanding). Other forms of behaviour, such as eyebrow movements or blinking are also relevant for analysing interactional processes, but they were excluded from this study because of the angle and quality of the video recordings.

2.2.1 *Gaze behaviour*

Explainees' gaze behaviour may serve as a relevant signal related to different interactional processes monitored by explainers (Clark & Krych, 2004), such as visual attention (Argyle & Cook, 1976, Goodwin, 1981; Kendon, 1967), cognitive processing (Glenberg et al., 1998; Goodwin, 1981; Phelps et al., 2006) or disengagement from a task (Doherty-Sneddon & Phelps, 2007). The present study focuses on two forms of explainees' gaze behaviour: changing gaze behaviour (gaze shifts) and non-changing (static) gaze behaviour. For the gaze shifts, we are additionally interested in one particular form, namely gaze aversions away from the explainers.

For interlocutors' gaze behaviour in dyadic human-human interactions, it has been reported that listeners gaze more often at speakers than vice-versa, and both participants gaze at each other briefly during moments of feedback elicitation by

the speakers and feedback giving by the listeners (Argyle & Cook, 1976; Bavelas et al., 2002; Kendon, 1967). Further and with respect to conversational management, interlocutors' gaze behaviour may influence the closure and the expansion of topics of an interaction (Rossano, 2005, 2013). However, the results regarding the function of gaze behaviour at different phases within turn-taking are not conclusive (Degutyte & Astell, 2021). In our analysis of explainees' gaze behaviour, their non-changing gaze at the explainers while listening to the explanation and their gaze at explanation-relevant objects are categorized as "static gaze".

Explainees do not constantly gaze at explainers, and the time in which explainees gaze at explainers varies individually (Argyle & Cook, 1976, Goodwin, 1981; Kendon, 1967). A shift in explainees' gaze direction may occur for different reasons: for example, in relation to attention drawing using pointing gestures by the explainers (Clark, 2003), the need to process and assimilate an explanandum and the explanans, or as an act of visual reference or attention to a third entity (Morency et al., 2006). Experimental studies on both adults' and children's gaze behaviour during solving cognitive tasks have demonstrated that averting the gaze from an interlocutor while thinking about an answer occurs more frequently for more challenging tasks, and that this behaviour boosts successful task performance with regard to knowledge retrieval, or analytical thinking such as arithmetic tasks and memory (Glenberg et al., 1998; Phelps et al., 2006). However, a study on teacher — student interactions has shown that speakers seem to misinterpret listeners' gaze aversions when asked to solve a certain task rather as a disengagement from the task than as cognitive processing (Doherty-Sneddon & Phelps, 2007).

Gaze shifts together with other nonverbal signals such as facial expressions might be less ambiguous. Nota et al. (2021) discovered that gaze shifts used with an eyebrow frown indicate questioning, whereas gaze shifts with an eyebrow raise signal thinking, e.g., about a correct response, or a lack of knowledge about the requested answer.

Less is known about the interpretation of non-changing gaze directions. Research shows that any type of non-changing gaze direction for a specific time period during an interaction indicates explainees' ongoing attention (Argyle & Cook, 1976; Bavelas et al., 2002; Clark, 2003). It is possible that explainees' gaze that is statically directed towards the explainer, the explanandum or away is interpreted as ongoing attention and engagement with the explainer and the current explanandum and explanans. However, it is still not clear whether explainees' engagement could be related to explainers' initiation of elaborations and topic changes. Therefore, we decided to investigate explainees' static gaze, gaze shifts and gaze aversions exploratively in relation to the elaborations and topic changes initiated by different explainers.

2.2.2 *Head nodding and backchannelling*

Both head nodding and linguistic backchannels are ambiguous signals, also when co-occurring. Head nodding is one of the most frequent gestural types of non-verbal feedback, even when the interaction partners do not look at each other (Allwood & Cerrato, 2003), and it can indicate either understanding or engagement of the addressee (Gander & Gander, 2020). When accompanied by linguistic backchannels (such as *okay, yes, aha, right*, etc.), head nods express either agreement or approval between interlocutors, depending on the type of linguistic backchannel (Allwood & Cerrato, 2003).

Linguistic backchannels are brief feedback responses which may be represented by lexical (such as *right, okay*) or non-lexical forms (such as *mhm, yeah*) (Allwood et al., 1992; Arnold, 2012). Linguistic backchannels are considered to serve various functions: they can signal a state of engagement or understanding without being a turn-taking signal (Arnold, 2012; Park et al., 2017; Yngve, 1970). They are used as acknowledgements (as continuer phrases such as *uh-huh, yeah*), as signals for continuous attention, perception or for unconditional understanding (Allwood et al., 1992; Clark & Brennan, 1991; Eshghi et al., 2015; Schegloff, 1982). Allwood et al. (1992) analysed the ambiguous meaning of the backchannels *yes, mhm* and *ok*, which also have been observed in the data of the present study. In contrast to the linguistic backchannel *no*, which straightforwardly expresses denial, refusal and rejection, other backchannel forms, such as *yes, mhm* and *ok* are related to multiple context dependent interpretations, such as acceptance, approval, understanding and attention (Allwood et al., 1992: 9; Arnold, 2012).

Because of the ambiguity of both signals, we cannot make any predictions regarding head nods and linguistic backchannels related to different forms of topical transitions in our study on physician-caregiver interactions. As a first step, we investigate co-occurrences of multiple signals and provide a sequential relation between these types of (non-)verbal feedback and changes of the explanation course to elaborations or new topics.

2.3 Explanation structure

A way to analyse the explanation structure, and thus, different forms of transitions, is to segment the explanation into explanation episodes (Roscoe & Chi, 2008). An explanation episode is defined as "a brief segment of the overall explanation [...] devoted to one particular topic" (ibid., p. 333). The particular topics within an explanation are different sub-explananda. Thus, an explanation of an upcoming surgery (which is the explanandum) contains different topics or sub-explananda such as diagnosis, reasons, medical procedures, or treatments. Each

sub-explanandum consists of newly introduced information pieces (new topics) which were most often elaborated by the physicians, e.g., via paraphrases or clarifications (elaborations). Two excerpts from different physician-caregiver interactions below provide examples for the transitions from newly introduced sub-topics to elaborations and vice-versa.

Example 1. Explanation about the medical procedure
[1] "*We are going here along. We do a small incision over the old scar, approximately a few centimeters here and here.*" — an introduction of a new sub-topic about the procedure
[1.1] "*With this small cut you can pull out the metal rail by one centimeter*" — an elaboration related to the sub-topic that was introduced in [1]
[2] *"So, the operation lasts approximately fifteen minutes, it's a small ambulant operation.*" — a change to a new sub-topic
[2.1] "*This means going home on the same day.*" — an elaboration related to the changed sub-topic [2]

Example 2. Explanation about the importance of the surgery
[1] *"At last, it's a blocking issue. We have to create a drain.*" — an introduction of a new sub-topic
[1.1] "*The gallbladder must be able to flow past the pancreas, so that the whole thing is not a problem.* — an elaboration of the sub-topic [1]
[2] *"And as I said, for the later quality of life, she can eat and drink normally.*" — a change to a new sub-topic

3. The current study

In the current study, naturalistic physician — caregiver explanations were analysed. The reason for these explanations was an upcoming surgery of individual children that required the caregivers' agreement. The goal of the current study is to investigate caregivers' forms of multimodal feedback and their relations to the two types of topical transitions of the explanation of the upcoming surgery, more specifically, transitions to elaborations (T-EL) and those from elaborations to new topics (T-NT).

3.1 Participants

Seven physicians and thirteen caregivers participated in the study consisting of eleven interactions. All participants signed a consent form, and the study was approved by the Ethics Board of the university. Four of the seven physicians participated each in two interactions. Also, two of the explanations were attended by

two caregivers (both caregivers' behaviour was considered in the analysis), and ten of the explanations were attended by the child which was supposed to undergo a medical intervention. However, children's behaviour was not analysed because most of them were at an age, at which they were not able to perceive or interpret the medical explanation. The language in ten of the eleven explanations was German, for one it was English. No socio-demographic data about the physicians and the caregivers was collected because the medical consultations were planned in short term. Due to technical reasons regarding the camera angles during data collection, causing inability to observe and analyse multimodal behaviour, one of the eleven interactions was excluded from the analysis because the multimodal behaviour could not be observed. Thus, ten interactions, seven physicians and twelve caregivers were analysed.

3.2 Procedure

All physician — caregiver interactions were recorded at a paediatric department of a hospital in Germany. In these explanatory dialogues, the physicians explained different relevant aspects regarding the surgery, e.g., the diagnosis, necessity, medical procedures and treatment.

Two of the eleven interactions took place in the chief physician's office, and they were each attended by two caregivers. As the environment in that room was quiet, the first two interactions lasted longer (approx. 25 minutes) compared to the rest. For the other nine interactions, different physicians each explained the surgery to one caregiver, and they were recorded in the outpatient department of the same hospital, ranging between 04:54 — 14:23 minutes.

The interactions were recorded with two cameras, one directed towards the physicians and another one directed towards the caregivers, capturing the face and torso area. No additional cameras or other techniques for separate voice recording or eye tracking were used in data collection because they would have disrupted the naturalness of the interactions and the attention of the physicians and the caregivers, who had a conversation on an important topic. The mean duration of the analysed ten interactions was 11:06 min. (*SD* = 7:10 min.).

3.3 Data coding

3.3.1 *Explainees' (non-)verbal behaviour*

The video recordings were transcribed and annotated in ELAN (Wittenburg et al., 2006). The spoken utterances were segmented into explanation episodes, and the explainees' gaze behaviour, head gestures and backchannels were coded.

An illustration of the coding procedure is shown in Figure 1, and a referential situation is provided in Example 3.

Gaze behaviour. The two camera perspectives allowed manual annotations of gaze directions, i.e., towards the interlocutor, the materials, and away. Although gaze was initially coded according to the gaze direction, the current study focuses on the type of gaze behaviour, not on the viewing target. For this purpose, two categories of gaze behaviour were distinguished: static gaze and gaze shift. These two categories were mutually exclusive.

- A gaze was coded as *static* if there was no change of the viewing direction of the caregivers before a transition.
- A change in caregivers' gaze direction before a transition was coded as a *shift*. Based on research on gaze aversion, this category was subdivided into *gaze shifts without aversion* and *gaze shift with aversion* (henceforth gaze aversion).
  - "gaze aversion": cases when the participant changed their viewing direction away from the other interaction partner, i.e., directed towards explanation-related materials or away from the shared referential space between them and the physicians.
  - gaze shifts without aversion: Any other change of the gaze direction, for example, from the explanation-related material away or vice versa.

Head gestures. The explainees used predominantly head nods during the segments of interest. Head shakes were observed only fourteen times in relation to the analysed episodic transitions across the analysed ten interactions. Therefore, only occurrences of head nods were included in the statistical analysis. For head nods, we considered single or multiple up and down head movements (Allwood & Cerrato, 2003).

Linguistic backchannels were identified as such based on the definition given in Yngve (1970), Allwood et al., (1992) and Arnold (2012) (e.g., *okay, yes, mhm, alright*), but were not coded for specific types as the most frequent types of backchannelling in the areas of interest, i.e., the analysed transitions, were continuers or short affirmations.

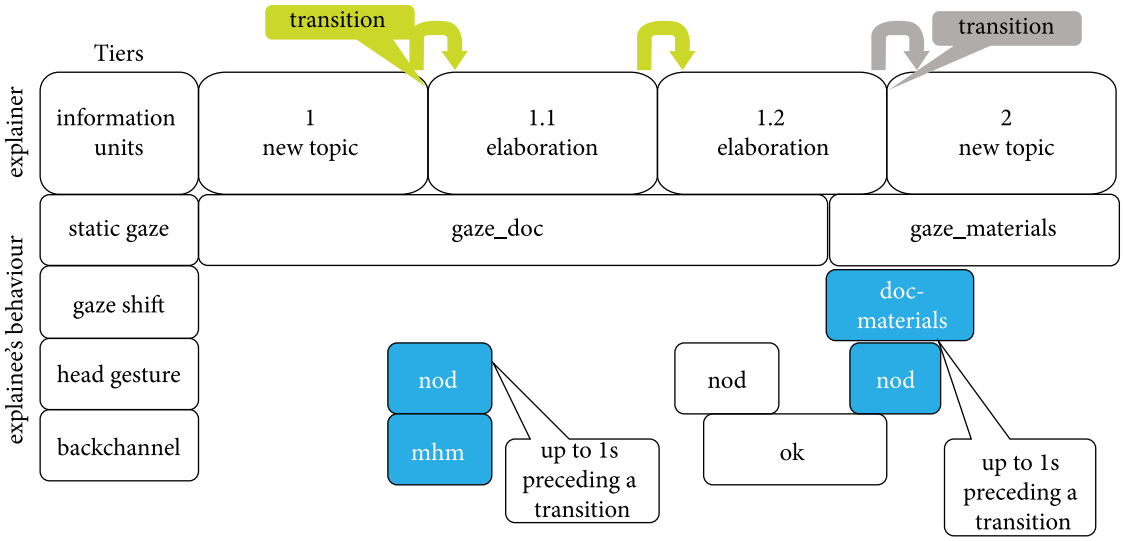

Figure 1. Coding model. The figure illustrates the annotation procedure in ELAN. The numbering of topics refers to the order of appearance. Full numbers represent a new topic, and decimal numbers represent the category "elaborations". The boundaries between the topical segmentations are the actual transition points. The green coloured transitions are those to elaborations, and the grey coloured transitions are those from elaborations to new topics. The light-blue coloured gaze shifts, head nods and backchannels are those co-occurring in the areas of interest, i.e., up to one second before transitions into elaborations or from elaborations to new topics. Modalities in other temporal relations were not analysed.

3.3.2 *Forms of multimodal behaviour*

Because either form of caregivers' gaze behaviour (static, shifted with aversion, and shifted without aversion) was always present and our research was focused on multimodality, we labelled the categories of caregivers' behaviour based on the occurring type of their gaze behaviour and co-occurring modalities (head nodding or backchannelling). For example, gaze shifts with no co-occurring modalities were labelled "unimodal gaze shift". Accordingly, gaze shifts either with co-occurring head nodding or with backchannelling were labelled "bimodal gaze shifts with head nodding or backchannelling", and gaze shifts with co-occurring head nodding and backchannelling were labelled "multimodal gaze shifts". Because there were three forms of gaze behaviour (see above), we observed in total twelve forms of multimodal behaviour of the attending caregivers: unimodal (one of the three forms of gaze behaviour only), bimodal (different forms of gaze behaviour together with head nodding or backchannelling, resulting in six different bimodal forms), multimodal (different forms of gaze, head nodding and backchannelling; resulting in three forms).

3.3.3 *Explainers' explanation structure*

Roscoe & Chi's (2008) definition of explanation episodes was applied for this analysis (Figure 1). We differentiated between *new topic* or *elaboration* of a previous topic. The category "elaborations" was defined as either further reasoning, additions, paraphrases or repetitions. A segment was labelled as n*ew topic* if the segment contained information not previously given. In the analyses, two different types of transitions between the segments were related to the multimodal behaviour of the explainees: transitions to elaborations (T-EL), which contained both transitions from a new topic to elaborations or transitions between elaborations, and transitions from elaborations to new topics (T-NT). These transitions were used for tracking the structural change of the explanations and for setting the temporal frame of collecting explainees' multimodal behaviour signals, which we analysed in relation to the types of explanation episodes. Note that this segmentation did not regard the semantic content of the episodes, i.e., the topics were not specified, because this level of detail was not necessary for our research question. At the first step, coders identified the new topics, i.e., different sub-explananda contained in the explanations. Example 3 provided below is related to the illustration in Figure 1, and it shows that a physician introduces the topic about the nutrition of a child after recovery, which is marked here as topic [1]. Then the physician makes two elaborations concerning the nutrition of the child, marked as [1.1] and [1.2]. After the elaborations about the child's nutrition, the physician changes the topic about regular check-ups, marked as [2]. The illustrated coding model in Figure 1 shows the numerical coding of explanation episodes, i.e., whether a segmentation was either a new topic or an elaboration, without providing labels to the semantic content of the explanation episodes. Full numbers (e.g., 1, 2, 3, etc.) corresponded to a new topic, and decimal numbers (e.g., 1.1., 1.2., 1.3., etc.) corresponded to elaborations of the respective topic. The caregivers' multimodal behaviour (gaze, head gestures and backchannels) was related to the different forms of transitions.

Example 3. Explanation about post-treatment care
[1] "*Later, as said, there will be no restriction regarding eating.*" — a new sub-topic
[1.1] *"She can eat completely normally, if you also want fast food, gummy bears or similar, right?"* — an elaboration of the sub-topic [1]
[1.2] *"You are not doing this. She can eat healthily. She has a normal gastrointestinal tract.*" — another elaboration of the sub-topic [1]
[2] *"And what we should do, of course, are check-ups with ultrasound.*" — a new sub-topic

3.3.4 *Temporal threshold*

For the analysis of caregivers' gaze behaviour, head nods and backchannels in the current study, we regarded those instances of (non-)verbal behaviour which were initiated by the caregivers within the timeframe of 1 second preceding a T-EL or a T-NT (see Figure 1). Because there is no conventionalised time frame for analysing reactions to multimodal feedback in the form of making an elaboration or changing the topic, we limited the time frame to 1 second as an exploratory approach. For our approach, the onset of each behaviour signal was considered. For example, the transition from topic 1 to elaboration 1.1 (Figure 1) was coded as a transition from a new topic to an elaboration with preceding caregiver's multimodal behaviour of *static gaze* + *nodding* + *backchanneling*; the transition from elaboration 1.1. to elaboration 1.2. was coded with only the caregiver's *static gaze* because the onset of the head nod and the backchannel occurred after the transition; the transition from elaboration 1.2 to the new topic 2 was annotated with a preceding caregiver's bimodal behaviour of *averted gaze* + *nodding*.

3.3.5 *Reliability*

The coding of caregivers' gaze behaviour, head gestures and backchannels was conducted in two steps: (1) Independent, trained pre-annotators did the initial transcription and coding, which was (2) checked and corrected by two other coders who were specifically and intensively trained for the coding in the current study. That means, all of the caregivers' behaviours were coded twice by two independent coders. The topical segments of the physicians' explanations were coded by two trained coders. After the training, we calculated Krippendorff's α as an inter-rater score between the pre-annotators and the trained coders. For measuring inter-rater reliability, the first three minutes from three of the ten physician-caregiver interactions were considered. Krippendorff's alpha (α) is suitable for small sample sizes and for different scale levels (Krippendorff, 2013). In addition, possible random agreement that may overestimate the actual agreement is factored out. Krippendorff (2004) provides the following guidelines for interpretation of the α value: a satisfactory intercoder reliability is given at a value above $\alpha = .800$, while values between $\alpha = .667$ and $\alpha = .800$ should be evaluated with caution. Reliability analyses revealed high agreement for gaze behaviour ($\alpha = 0.94$), head gestures ($\alpha = 0.81$), and segments of new topics and elaborations ($\alpha = 0.85$).

3.4 Data analysis

The goals of our study were (1) to reveal absolute frequencies and conditional probabilities of caregivers' forms of multimodal feedback occurring prior to the

transitions into elaborations and those from elaborations into new topics initiated by the physicians, and (2) to measure the effect of caregivers' forms of multimodal behaviour on the transitions initiated by the physicians.

All instances of transitions to elaborations (T-EL) and transitions from elaborations to new topics (T-NT) initiated by the physicians and caregivers' forms of multimodal feedback behaviour occurring within the time frame of 1 second preceding those transitions were extracted. The occurrences of T-EL and T-NT were treated as the response variable, and caregivers' forms of multimodal behaviour were treated as a fixed effect.

3.4.1 *Conditional probabilities*

For the analysis of the transitions' dependence on multimodal behaviour, we used the definition of conditional probability (Dekking et al., 2005). It formalises the probability of an event A, given that event B occurs:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{\#(A \cap B)}{\#\Omega}}{\frac{\#B}{\#\Omega}} = \frac{\#(A \cap B)}{\#(B)}$$

For example, to calculate conditional probabilities for T-EL following bimodal behaviour, we replaced the numerator and the denominator of the fraction with absolute frequencies:

- $\#(A \cap B)$ — the number of T-EL initiated after one form of multimodal behaviour, e.g., bimodal behaviour with gaze aversions and head nodding
- $\#B$ — the number all transitions initiated after bimodal behaviour with gaze aversions and head nodding

Here is an example related to the calculation of conditional probabilities for bimodal behaviour with gaze aversions and head nodding of the caregivers before transitions into elaborations initiated by the physicians in our study:

$$P(\text{transitions to elaborations} \mid \text{gaze aversions with head nodding}) = \frac{\text{number of transitions to elaborations after gaze aversions with head nodding}}{\text{number of all transitions after gaze aversions with head nodding}}$$

The example is interpreted as following: The probability that physicians initiate T-EL after caregivers' bimodal behaviour with gaze aversions and head nodding results from dividing the instances of T-EL after bimodal behaviour with gaze aversions and head nodding by all transitions after bimodal behaviour with gaze aversions and head nodding across the ten interactions. Conditional proba-

bilities are calculated separately for all twelve forms of multimodal behaviour and provide an insight into the relation between the transition types initiated by the physicians for each form of multimodal behaviour.

3.4.2 *Generalised linear mixed model (GLMM)*

The second research goal was to investigate the effect of caregivers' forms of multimodal behaviour on the transitions initiated by the physicians. Because a Shapiro-Wilk test indicated a non-normal distribution ($W = 0.82$, $p < 0.01$), we used a *Generalised linear mixed model* (GLMM) in our analysis. The GLMM was built with the forms of caregivers' multimodal behaviour and the two types of transitions initiated by the physicians as interacting fixed effects, as well as including the different physicians and caregivers as random effects. A GLMM allows a crossing modelling of the factorial levels, in which we specified physicians' transition type interacting with the different forms of caregivers' multimodal behaviour as the fixed factor. This contrastive interference was followed by pairwise comparisons between the two transition types with respect to the twelve forms of multimodal behaviour. Before the statistical analysis, the frequencies of transitions for each form of caregivers' (non-)verbal behaviour were converted into proportions, for each interaction separately. By that, the differences between the behaviour of the different physicians interacting with different caregivers were normalised.

In some of the analysed interactions, there were 0 instances of T-EL or T-NT preceded by some of the forms of caregivers' multimodal behaviour, turning some proportions into 0% (0.00) or 100% (1.00). To avoid errors in our statistical model related to confusion with binominal regression, we applied a zero-inflation (Tang et al., 2023) to the data of the response variable, turning proportions of 0.00 into 0.000001 and proportions of 1.00 into 0.999999 (using *glmmTMB* package, Brooks et al., 2017). Thus, we built our statistical model as follows:

$$glmmTMB(PROP\_adjusted \sim BEHAVIOUR * TRANSITION + (1 \mid EX/EE), \\ data = dataframe, family = beta\_family())$$

In the model, *PROP_adjusted* are the adjusted proportions of the response variable, caregivers' forms of multimodal behaviour and the two types of transitions initiated by the physicians were defined as two interacting fixed effects, and the random effect represented the different caregivers nested within the different physicians.

4. Results

4.1 Conditional probabilities across participants

We observed 502 cases of transitions between explanation episodes. As the relevant topical transitions initiated by the physicians included those from a new topic to an elaboration, and one elaboration into another, the total number of T-EL ($n = 289$) was higher than the number of T-NT ($n = 213$). Results for conditional probabilities are illustrated in Figure 2, and absolute frequencies are given in brackets.

First insights from the frequency analysis suggest that the most frequently initiated transitions by the physicians occur after caregivers' static gaze and after combinations of static gaze with head nodding and backchannelling. The lowest numbers of transitions are observed regarding caregivers' multimodal gaze shifts (without aversions).

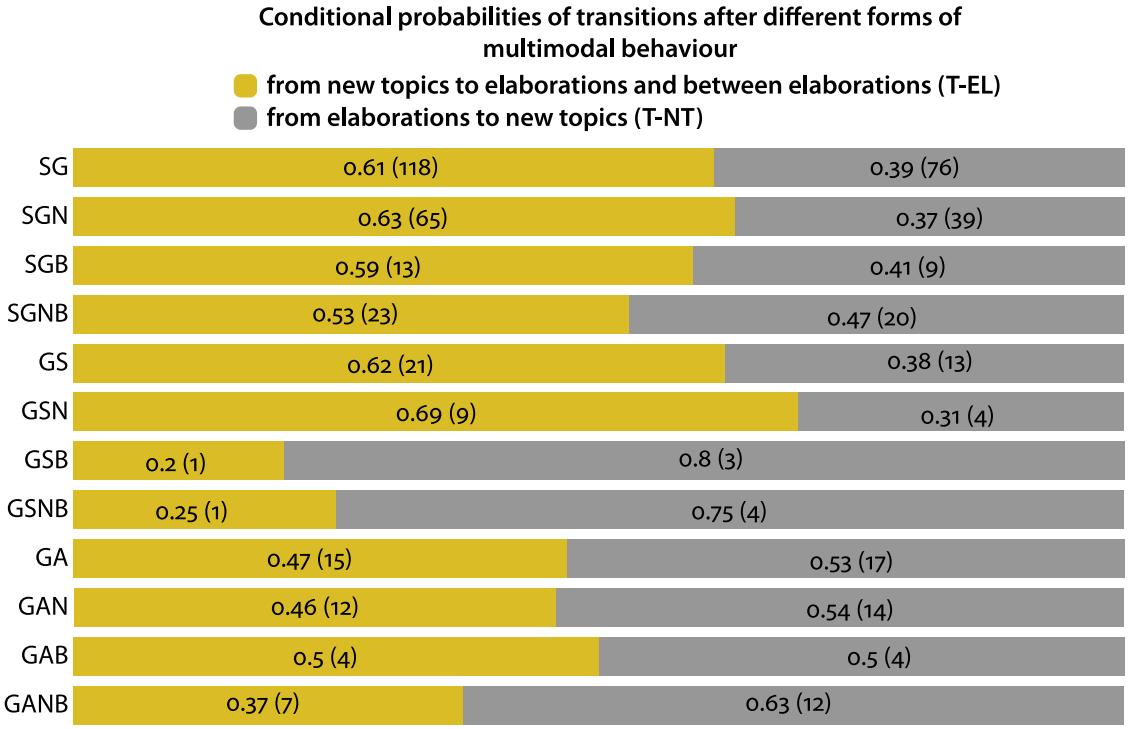

Figure 2. Conditional probabilities and absolute frequencies. SG — static gaze; GS — gaze shift (without aversion); GA — gaze aversion; N — head nodding; B — backchannelling

With respect to caregivers' static gaze behaviour, it was more likely that physicians initiated a T-EL than T-NT afterwards (see Figure 2). The same tendency was observed for caregivers' unimodal gaze shifts (without aversions) and

bimodal gaze shifts (without aversions) and head nodding. Caregivers' gaze shifts (without aversions) bimodally with backchannelling or multimodally with both backchannelling and head nodding was more likely to be followed by T-NT than by T-EL. For caregivers' gaze aversions, either unimodally or bimodally with either backchannelling or head nodding, the probabilities for both types of transitions were nearly the same. Only for multimodal gaze aversions, a higher probability for T-NT compared to T-EL was observed.

4.2 Analysis of fixed and random effects

In addition to our exploration of frequencies and conditional probabilities, we were also interested in the effect of caregivers' forms of multimodal behaviour on the two types of transitions initiated by the physicians. To address the second research goal, we conducted generalised linear mixed effects models using *glmmTMB* (see 3.4.2.).

Overall, the glmmTMB model indicated a slightly better fit of our data ($AIC = -2292$, $BIC = -2198$) compared to a null model (without the fixed effect), and also a high proportion of variance including the fixed and the random effects across all participants (Conditional $R^2 = 0.95$). However, at the level of the caregivers nested within the different physicians we found little variability in the response variable within each nested combination ($\sigma^2 = <0.001$, $SD < 0.001$). A higher variability in the response variable was found at the level of the different physicians ($\sigma^2 = 0.03$, $SD = 0.17$). The random effects summary suggests that the proportional variations of transitions across the different physician-caregiver interactions were higher than proportional variations within each of the interactions.

The fixed effects summary of the glmmTMB in Table 1 contains only results for those forms of caregivers' multimodal behaviour that indicated a significant effect on the transitions initiated by the caregivers. A significant effect on the T-EL was indicated by caregivers' bimodal behaviour with gaze shifts (without aversions) co-occurring either with head nodding or with backchannelling, multimodal behaviour with gaze shifts (without aversions), bimodal behaviour with gaze aversions co-occurring with backchannels and multimodal behaviour with gaze aversions. Only caregivers' multimodal behaviour with gaze aversions indicated a significant effect also on the T-NT. The parameter estimates (β) of the forms of multimodal behaviour related to the parameter estimate of the intercept (unimodal static gaze) in Table 1 indicate that physicians initiated T-EL more likely after caregivers' multimodal behaviour with gaze shifts (without aversions), and also that physicians initiated T-NT after caregivers' multimodal behaviour with gaze aversions. The likelihood that T-EL are initiated after bimodal behav-

iour with gaze shifts (without aversions) co-occurring with head nodding or backchannelling, and that T-EL are initiated after bimodal behaviour with gaze aversions and backchannelling decreases from the intercept in a negative direction. Thus, only caregivers' multimodal behaviour with gaze shifts (without aversions) predicted an increase of T-EL whereas caregivers' multimodal behaviour with gaze aversions predicted an increase of T-NT.

Table 1. A fixed effect summary for caregivers' multimodal behaviour with significant effect on the transitions initiated by the physicians

Behaviour	Transition	*M*	*SD*	*β*	*S.E.*	*z*	*p*
SG (Int.)	T-EL	0.63	0.09	0.11	0.44	0.25	> 0.05
GSN	T-EL	0.22	0.35	−1.41	0.59	−2.40	< 0.05
GSB	T-EL	0.03	0.10	−1.69	0.58	−2.92	< 0.01
GSNB	T-EL	0.05	0.16	1.71	0.58	2.94	< 0.01
GAB	T-EL	0.08	0.25	−1.68	0.58	−2.89	< 0.01
GANB	T-EL	0.15	0.27	−1.45	0.59	−2.48	< 0.05
GANB	T-NT	0.63	0.43	2.24	0.85	2.64	< 0.01

Note: SG — static gaze; GS — gaze shift (without aversion); GA — gaze aversion; N — head nodding; B — backchannelling; (Combinations of the abbreviations refer to bimodal or multimodal behaviour.); T-EL — transitions to elaborations; T-NT — transitions from elaborations to new topics

For comparing the proportions of T-EL with the proportions T-NT for each form of multimodal behaviour, we ran a post-hoc Tukey test. Table 2 summarises the estimated means (β) for the pairs of proportions that significantly differed. Significant differences between the proportions of T-EL and the proportions of T-NT were found for caregivers' unimodal gaze shifts (without aversions) ($\beta = 1.42$, $S.E. = 0.60$, $z = 2.37$, $p < 0.05$) and for caregivers' multimodal behaviour with gaze aversion, head nodding and backchannelling ($\beta = -2.04$, $S.E. = 0.57$, $z = -3.55$, $p < 0.001$). These findings suggest that T-EL demonstrated a stronger relation to unimodal behaviour with gaze shifts (without aversions) than T-NT. On the other hand, T-NT demonstrated a stronger relation to multimodal behaviour with gaze aversions, compared to T-EL.

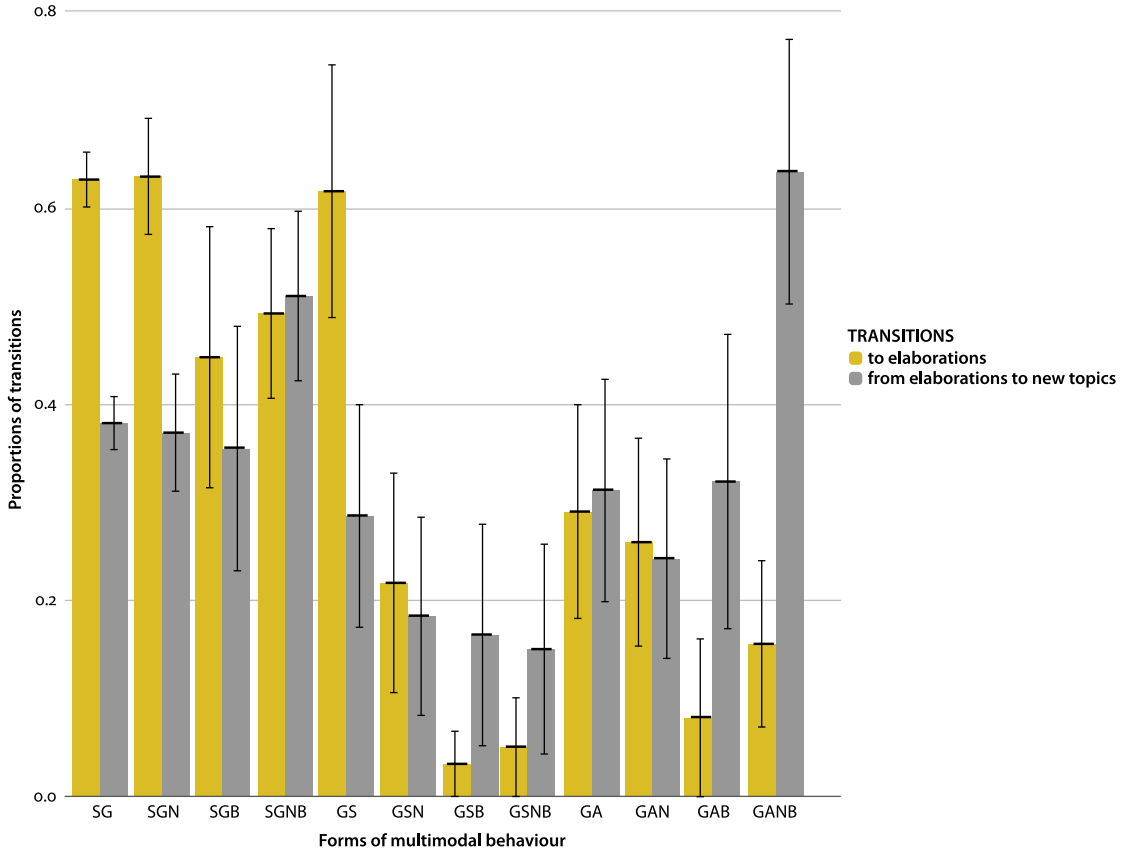

Figure 3. Proportions of transitions (MEAN & S.E.) initiated by the physicians for each form of caregivers' multimodal behaviour. SG — static gaze; GS — gaze shift (without aversion); GA — gaze aversion; N — head nodding; B — backchannelling; Combinations of the abbreviations refer to bimodal or multimodal behaviour

Table 2. A summary of estimated means for proportions of transitions which indicated significant differences per form of multimodal behaviour

Behaviour	Transition	*M*	*SD*	*β*	*S.E.*	LCL	UCL
GS	T-EL	0.61	0.41	0.53	0.44	−0.33	1.39
GS	T-NT	0.28	0.36	−0.89	0.41	−1.70	−0.08
GANB	T-EL	0.15	0.27	−1.34	−0.39	−2.12	−0.57
GANB	T-NT	0.63	0.43	0.70	0.42	−0.13	1.53

Note: GS — gaze shift (without aversion); GA — gaze aversion; N — head nodding; B — backchannelling; (Combinations of the abbreviations refer to bimodal or multimodal behaviour.); T-EL — transitions to elaborations; T-NT — transitions from elaborations to new topics

4.3 Summary of results

In the first part of the analysis, we compared the conditional probabilities between T-EL and T-NT initiated by the different physicians following twelve forms of caregivers' multimodal behaviour. This analysis revealed that all forms of caregivers' behaviour with static gaze, as well as unimodal gaze shifts and bimodal gaze shifts (without aversions) with head nodding were more likely to be followed by T-EL than T-NT. For the other six forms of multimodal behaviour, the analysis revealed higher probabilities for T-NT than for T-EL. The application of a GLMM on proportional data indicated positive estimates with significant effect on the transitions only for two forms of multimodal behaviour: (1) multimodal behaviour with gaze shifts (without aversions) on T-EL, and (2) multimodal behaviour with gaze aversions on T-NT. The analysis of pairwise comparisons also supported the relation between multimodal behaviour with gaze aversions and T-NT.

5. Discussion

In this paper, we analysed the relations between transitions to elaborations (T-EL) and transitions from elaborations to new topics (T-NT), which were initiated by explainers, and different forms of explainees' multimodal feedback behaviour. The analysis was based on a sample of ten videotaped interactions between explaining physicians and attending caregivers, who were asked to agree to an upcoming surgery of their child. For our analysis, we coded caregivers' gaze behaviour (static, shifting, and averting from the physicians), head nodding and backchannelling, which resulted in twelve forms of multimodal feedback. To relate the topical structure of the explanations to multimodal forms of behaviour, we segmented the physicians' explanations into episodes (Roscoe & Chi, 2008) indicating the different topics and their elaborations. The data analysis pursued two research goals.

First, we analysed frequencies and conditional probabilities of T-EL and T-NT initiated after each form of caregivers' multimodal behaviour. Second, we analysed the effect of caregivers' forms of multimodal behaviour on the frequencies of T-EL and T-NT. Because of non-normal data distribution, we normalised the frequencies of transitions into proportions for each form of caregivers' behaviour. The proportional types of transitions were analysed in a zero-inflated glmmTMB model integrating the interaction of two fixed effects, (1) the transitions initiated by the physicians and caregivers' forms of multimodal behaviour and (2) a random effect with a nested condition — each caregiver interacting with one of the seven physicians.

5.1 Caregivers' static gaze behaviour with or without other modalities

Although the analysis of conditional probabilities suggested that there was a higher probability of initiating T-EL than T-NT following all forms of multimodal behaviour with static gaze, the statistical model did not indicate a significant effect of any form of behaviour with static gaze on the transitions. Thus, based on our dataset, we cannot conclude that non-changing gaze behaviour with or without other modalities, such as head nodding and backchannelling, could be certainly related to T-EL or T-NT.

According to Kendon (1967) and Bavelas et al. (2002), listeners look at speakers for longer periods with short intervals of aversions from speakers. Furthermore, head nodding and linguistic backchannel responses, such as *mhm* or *yeah* are evoked when speakers establish mutual gaze with listeners (Bavelas et al., 2002). In the analysed areas of interest in our data, the caregivers were gazing at the explaining physicians who produced long utterances. In many cases, the caregivers' static gaze directed at the physicians or at the presented documentation materials co-occurred with head nodding and backchannels such as *mhm, ok*, and *yes*. Both modalities are also reported to have multiple and ambiguous interpretations. Head nodding is one of the most frequent forms of gestural feedback, even when listeners' gaze is not directed toward the speaker (Allwood & Cerrato, 2003), and their interpretation varies between understanding, attention, or agreement between interlocutors (Allwood & Cerrato, 2003; Gander & Gander, 2020). Similarly, different forms of backchannels indicate continuous attention, acceptance or agreement (Allwood et al., 1992; Arnold, 2012). From our analysis, we can only conclude that static gaze behaviour, with or without head nodding and backchannelling, can only be related to explainees' visual attention at the explainers (Argyle & Cook, 1976; Goodwin, 1981; Kendon, 1967). Future analyses may include the kinematics for the head gesture, or prosodic features for linguistic backchannels (Gravano et al., 2007; Lai, 2010; Ward & Tsukahara, 2000), to resolve the ambiguity of head nodding and backchannelling.

5.2 Caregivers' gaze shifts and gaze aversions with or without other modalities

In contrast to our findings regarding caregivers' forms of static gaze behaviour, we found that unimodal gaze shifts and bimodal gaze shifts (without aversions) co-occurring with head nodding precede T-EL with higher probability than T-NT. The other forms of gaze shifts (without aversions) and gaze aversions revealed a higher tendency to precede T-NL than T-EL. Gaze aversions are a special form of gaze shifts, i.e., a change of caregivers' viewing direction away from the physi-

cian to the materials or aside. The statistical model indicated that T-EL can be predicted by caregivers' multimodal behaviour with gaze shifts (without aversions). Only caregivers' multimodal behaviour with gaze aversions demonstrated a stronger relation to T-NT, based on the analysis of fixed effects and pairwise comparisons. Why are only aversions from the explaining physicians to another entity or away that co-occur with head nodding and backchannelling associated with a T-NT and all other forms of gaze shifts and gaze aversions not?

Gaze shifts are related to either visual attention drawn by pointing gestures (Clark, 2003), an act of visual attention to a third entity, or cognitive processes, such as thinking about the explanandum (Morency et al., 2006). Also, if gaze shifts co-occur with head nodding or backchanneling, gaze shifts may also be related only to visual attention because head nodding and backchannelling evoke ambiguous interpretations (see 5.1).

In contrast to the relation between caregivers' gaze shifts and T-EL, our statistical analysis suggested that only caregivers' multimodal behaviour with gaze aversions, head nodding and backchannelling has a stronger relation to T-NT, compared to T-EL. Thus, multimodal feedback behaviour may resolve ambiguous interpretations of (non-)verbal signals, such as head nodding and backchannelling (Allwood et al., 1992; Allwood & Cerrato, 2003; Arnold, 2012; Gander & Gander, 2020). Also, gaze aversions from the explainer were previously related to the reduction of cognitive load during thinking about the answer for demanding cognitive tasks (Doherty-Sneddon & Phelps, 2007; Glenberg et al., 1998; Phelps et al., 2006). Rossano (2005; 2013) reports that interlocutors' gaze withdrawals from each other are related to a completion of a topical sequence, in comparison to interlocutors' mutual gaze which in 95% of the cases leads to an expansion of a topical sequence. Our results are in line to what Rossano (2005; 2013) observed. Also, our finding that caregivers' multimodal behaviour with gaze aversions is related to T-NT could be addressed in further research on multimodal signals of cognitive processing to discover whether averted gaze from explainers co-occurring with head nodding and backchannelling signals a closure of a topical sequence.

5.3 Concluding remarks

The presented research provides new insights into the relation between explainers' T-EL and T-NT and explainees' multimodal behaviour. Despite the many open questions regarding some forms of multimodal behaviour, our findings contribute to research on multimodal behaviour and its relation to the dynamic structures of explanations. Our analysis in this study did not provide insights about the frequencies of mutual gaze between the physicians and the caregivers

before the transitions. Based on the theoretical background on continuous interactional monitoring of each other's multimodal behaviour (Clark & Krych, 2004), analysing the physicians' gaze directions in the future could provide more concrete insights about the focus of monitoring the caregivers' behaviour. Therefore, we suggest that future studies relate interlocutors' mutual gaze to the same types of transitions. Future works could also include other forms of behaviour, such as eyebrow movements or prosodic features of backchannels and conduct analyses on a larger data set.

Funding

This work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research foundation) TRR 318/1 2021 — 438445824.

Acknowledgements

We thank all participants for supporting this research and our student assistants for their help in transcribing and annotating the video data.

References

Allwood, J., Nivre, J., & Ahlsén, E. (1992). On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics*, *9*(1), 1–26.

Allwood, J. & Cerrato, L. (2003). A study of gestural feedback expressions. *Proceedings of the 1st Nordic Symposium on Multimodal Communication*. Copenhagen, 7–22.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Arnold, K. (2012). Humming along. *Contemporary Psychoanalysis*, *48*(1), 100–117.

Bavelas, J. B., Black, A., Lemery, C. R., & Mullett, J. (1986). "I show you how you feel": Motor mimicry as a communicative act. *Journal of Personality and Social Psychology*, *50*, 322–329.

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: the role of gaze. *Journal of Communication*, *52*(3), 566–580.

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnússon, Á., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). GLMMTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R Journal*, *9*(2), 378–400.

Clark, H. & Brennan, S.A. (1991). Grounding in Communication. In Resnick, L.B., Levine, J.M. & Teasley, S.D. (Eds.) *Perspectives on socially shared cognition* (pp. 127–142). APA Books.

Clark, H.H. (2003). Pointing and placing. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet* (pp. 243–268). Lawrence Erlbaum.

Clark, H.H., & Krych, M.A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*, 62–81.

Degutyte, Z., & Astell, A. (2021). The role of eye gaze in regulating turn taking in conversations: A systematized review of methods and findings. *Frontiers in Psychology*, *12*, 616471.

Dekking, F.M., Kraaikamp, C., Lopuhaä, H.P., & Meester, L.E. (2005). *A modern introduction to probability and statistics: Understanding why and how*. Springer London.

Dingemanse, M., Roberts, S.G., Baranova, J., Blythe, J., Drew, P., Floyd, S., et al. (2015). Universal Principles in the Repair of Communication Problems. *PLoS ONE 10*(9): e0136100.

Doherty-Sneddon, G., & Phelps, F.G. (2007). Teacher's responses to children's eye gaze. *Educational Psychology*, *27*(1), 93–109.

Eshghi, A., Howes, C., Gregoromichelaki, E., Hough, J., & Purver, M. (2015). Feedback in conversation as incremental semantic update. *Proceedings of the 11th International Conference on Computational Semantics*. London, 261–271.

Gander, A.G., & Gander, P. (2020). Micro-feedback as cues to understanding in communication. *Dialogue and Perception — Extended Papers from DaP2018*. In C. Howes, S. Dobnik, & E. Breitholtz (Eds.) *CLASP Papers in Computational Linguistics* (pp. 1–11). Gothenburg University.

Glenberg, A.M., Schroeder, J.L., & Robertson, D.A. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory & cognition*, *26*(4), 651–658.

Goodwin, Ch. (1981). *Conversational organisation between speakers and hearers*. Academic Press.

Gravano, A., Benus, S., Chávez, H., Hirschberg, J., & Wilcox, L. (2007). On the role of context and prosody in the interpretation of 'okay'. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, 800–807.

Hömke, P., Holler, J., & Levinson, S.C. (2017). Eye blinking as addressee feedback in face-to-face conversation. *Research on Language and Social Interaction*, *50*(1), 54–70.

Hömke, P., Levinson, S.C., & Holler, J. (2022). Eyebrow movements as signals of communicative problems in human face-to-face interaction. *PsyArXiv*, https://osf.io/preprints/psyarxiv/3jnmt.

Ismail, N.M., & Syahputri, V.N. (2022). "I Mean You Can Stop. I Already Understand You": Head Tilts during Conversations. *Lingua Didaktika: Jurnal Bahasa dan Pembelajaran Bahasa*, *16*(1), 1–11.

Kendon, A. (1967). Some functions of gaze-direction in social attention. *Acta Psychologica*, *26*, 22–63.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Departmental Papers (ASC)*, *30*(3), 1–16.

Krippendorff, K. (2013). *Content Analysis: An Introduction to Its Methodology* (3rd ed). Sage Publications.

Lai, C. (2010). What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue. *Proceedings of Interspeech 2010*, Makuhari, 1413–1416.

Morency, L. P., Christoudias, C. M., & Darell, T. (2006). Recognizing gaze aversion gestures in embodied conversational discourse. *Proceedings of the 8th International Conference on Multimodal Interfaces*, 298–294.

Nota, N., Trujillo, J. P., & Holler, J. (2021). Facial signals and social actions in multimodal face-to-face interaction. *Brain Sciences*, *11*(8), 1017.

Park, H. W., Gelsomini, M., Lee, J. J., & Breazeal, C. (2017). Telling Stories to Robots: The Effect of Backchannelling on a Child's Storytelling. *Proceedings of the 2017 ACM/IEEE International Conference in Human-Robot Interaction*. Vienna, Austria, 100–108.

Phelps, F. G., Doherty-Sneddon, G., & Warnock, H. (2006). Helping children think: Gaze aversion and teaching. *British Journal of Developmental Psychology*, *24*(3), 577–588.

Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H. M., Buschmeier, H., Esposito, E., Grimminger, A., Hammer, B., Häb-Umbach, R., Horwath, I., Hüllermeier, E., Kern, F., Kopp, S., Thommes, K., Ngonga Ngomo, A.-C., Schulte, C., Wachsmuth, H., Wagner, P., & Wrede, B. (2021). Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(3), 717–728.

Roscoe, R. & Chi, M. T. H. (2008). Tutor learning: the role of explaining and responding to questions. *Instructional Science*, *36*(4), 321–350.

Rossano, F. (2005). When it's over is it really over? On the effects of sustained gaze vs. gaze withdrawal at sequence possible completion. *International Pragmatic Association, Riva del Garda, July*.

Rossano, F. (2013). Gaze in conversation. In J. Sidnell, & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 308–329). Malden, MA: Wiley-Blackwell.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn taking in conversation. *Language*, *50*, 696–735.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In Tannen, D. (Ed.) *Analysing discourse: Text and talk. Georgetown University Roundtable on Languages and Linguistics 1981* (pp. 71–93). Georgetown University Press.

Tang, B., Frye, H. A., Gelfand, A. E. et al. (2023). Zero-Inflated Beta Distribution Regression Modeling. *Journal of Agricultural, Biological and Environmental Statistics*, *28*, 117–137.

Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, *32*(8), 1177–1207.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. *Proceedings of the 5th international conference on language resources and evaluation (LREC)*, 1556–1559.

Yngve, V. H. (1970). On getting a word in edgewise. *Papers from the sixth regional meeting Chicago Linguistic Society*, April 16–18, 1970, Chicago Linguistic Society, Chicago (pp. 567–578).

Address for correspondence

Stefan Lazarov
Transregional Collaborative Research Centre 318 "Constructing Explainability"
Paderborn University
33098 Paderborn
Germany

stefan.lazarov@uni-paderborn.de

Biographical notes

Stefan Lazarov is a PhD student at Paderborn University and a researcher at the Transregional Collaborative Research Centre 318 (TRR 318) "Constructing Explainability" in project A02 "Monitoring the Understanding of Explanations". His research is focused on the relation between different explanatory phenomena and human multimodal behaviour.

https://orcid.org/0009-0009-0892-9483

Kai Biermeier completed his master's degree in Computer Science and works as a lab technician within the TRR 318.

https://orcid.org/0000-0002-2879-2359

kai.biermeier@uni-paderborn.de

Angela Grimminger is a senior researcher at the Psycholinguistics research group at Paderborn University, Head of the SprachSpielLabor, and a principal investigator in Project A02 at the TRR 318.

https://orcid.org/0000-0002-5749-9362

angela.grimminger@upb.de

Publication history

Date received: 13 October 2023
Date accepted: 16 November 2024

How are mutual gaze and gaze withdrawals related to the changing topical structure of dyadic explanatory interactions?

Stefan Lazarov [ORCID: 0009-0009-0892-9483][*1,2] and Angela Grimminger [ORCID: 0000-0002-5749-9362][1,2]

[1]TRR 318 'Constructing Explainability', Paderborn University, Warburger Str. 100, 33098 Paderborn, Germany
[2]Faculty of Arts and Humanities, Psycholinguistics Research Group, Paderborn University, Warburger Str. 100, 33098 Paderborn, Germany
[*]Corresponding author: stefan.lazarov@uni-paderborn.de

Abstract

Gaze behavior, being continuously accessible to interlocutors in face-to-face interactions, serves as a cue managing turn-taking, regulating the duration of topical sequences, and supporting cognitive processing in various everyday conversational contexts. The present study seeks to enhance the understanding of the relation between two forms of interactive gaze behavior – mutual gaze and gaze withdrawals – and the topical structure in the explanatory discourse. To do so, we analyzed 24 dyadic board game explanations in which one explainer subsequently explained a board game to three different explainees, and the board game was physically absent from the shared space. The present study pursues two objectives: (1) to compare the proportional distribution of mutual gaze and gaze withdrawals immediately preceding topical shifts across the 24 explanations, and (2) to determine whether gaze withdrawals are initiated more frequently by the interlocutor who introduces the topic change, by the listening interlocutor, or by both interlocutors simultaneously. Based on previous research (Lazarov et al., 2025; Rossano, 2012), we hypothesized that (1) topic changes in explanations are more frequently preceded by gaze withdrawals than by mutual gaze, and (2) gaze withdrawals are initiated more frequently by the interlocutor who initiates a topic change than by the other interlocutor. Both hypotheses were verified by our analysis. Furthermore, our findings indicated that the relation between gaze withdrawals and topic changes is particularly salient at the level of different explainers, as compared to the intra-individual differences observed for each of the explainers interacting with different explainees.

Keywords: gaze behavior, mutual gaze, gaze withdrawals, explanations, topical structure

Declarations

Authors' contributions **S.L.**: Conceptualization, Writing – Original Draft, Writing – Review & Editing. **A.G.**: Writing, Review & Editing

Competing interests The authors report no conflicts of interest.

Non-financial interests The authors declare that they have no financial or personal relationships that could inappropriately influence or bias the content of this paper.

Funding This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824.

Ethics approval All studies have been approved by the Ethics Committee of the University of Paderborn on January 11, 2021.

Consent for publication All participants signed a consent form before participating in the studies, which allows researchers in the projects to collect and analyze their sociodemographic, audio, and video data, as well as to publish the results of the analyzes in scientific publications.

Availability of data and materials The annotation manual, materials, and statistical codes used for conducting the analysis are available at https://osf.io/duynt/?view_only=fa46b8df67fb4d21a5fc5c4dbd26e839. Following the data protection agreement between the researchers and the participants, the audio and video data of the presented study are not publicly available. For special requests, please contact the corresponding author.

1 Introduction

In human–human interaction, the interlocutors' eye gaze behavior serves multiple functions, such as managing turn-taking (Argyle & Cook, 1976; C. Goodwin, 1981; Jokinen, Harada, et al., 2010; Jokinen, Nishida, & Yamamoto, 2010; Kendon, 1967), signaling cognitive (Glenberg et al., 1998; Phelps et al., 2006) or language processing (Beattie, 1981), as well as a cue used for feedback elicitation from the addressees (Argyle & Cook, 1976; Bavelas, Black, et al., 2002; Kendon, 1967). Furthermore, maintaining eye contact with each other, referred to as mutual gaze (Argyle & Cook, 1976; Cook, 1977), and breaking the eye contact, referred to as gaze withdrawal (C. Goodwin, 1981, 1985; Rossano, 2012, 2013), have both been shown to indicate the interlocutors' engagement within an interaction. Specifically, mutual gaze has been associated with the expansion of topical sequences, whereas gaze withdrawals have been linked to the closure of topical sequences across various contexts of everyday interactions (Rossano, 2012, 2013).

The relation between the interlocutors' interactive gaze behavior and the topical structure in conversational contexts is also relevant in the context of explanatory interactions. Explanatory interactions about an entity or a process (explanandum) are maintained by an explainer (the more knowledgeable person) and an explainee (the less knowledgeable person) (Rohlfing et al., 2021). The explainer's task is to increase the explainee's knowledge and understanding about the explanandum, which is associated with further processes, such as interactional monitoring, scaffolding and co-constructions (Buschmeier et al., 2023; Rohlfing et al., 2021). While co-constructions emerge from the bidirectional (non-)verbal interaction between interlocutors (Rohlfing et al., 2021), scaffolding refers to a process in which a more knowledgeable partner adjusts the explanation to the explainee's abilities and cognitive processing (Wood et al., 1976). By continuously monitoring each other's behavior over the course of an explanation, the explainers and the explainees inform each other about the development of an explanation, for example by their verbal and nonverbal forms of behavior, such as their speech, faces, bodies, and gestures (Clark & Krych, 2004).

Explainers employ different explanatory structures, such as elaborations of a previous topic or introductions of new topics. The topical structure may be related to different forms of the interlocutors' multimodal behavior, such as their eye gaze behavior, head gestures and vocal backchannels (Lazarov et al., 2025; Rossano, 2012, 2013). However, this previous research on interactive gaze behavior in relation to the topical structure of the discourse accounts for some limitations: For example, Lazarov et al. (2025) focused solely on the explainees' eye gaze behavior, head gestures, and backchanneling without addressing the explainers' eye gaze behavior (and thus whether the explainers monitored the explainees' gaze behavior) and its relation to the topical structure of medical explanations. Further, Rossano (2012, 2013) investigated the interlocutors' interactive gaze behavior in the context of spontaneous everyday conversational contexts other than interactions that specifically target a certain goal, which for explanations is the increase in an explainee's knowledge and understanding (Buschmeier et al., 2023).

In this article, we address these limitations by examining two forms of interactive gaze behavior – mutual gaze and gaze withdrawals – in dyadic human–human explanatory interactions about a board game in relation

to changes of explanation topics (i.e., different sub-explananda). In addition, the present study incorporates a nested study design in which one explainer explained a board game to three different explainees subsequently. By addressing this aspect, the present study seeks to investigate how the gaze withdrawals of each interlocutor, the explainer, the explainee or both, are related to topic changes initiated by either party, and whether this relation becomes more pronounced only at the level of different explainers, or depending on interacting with different explainees.

2 Theoretical background

2.1 Interactive eye gaze behavior

Eye gaze behavior is one of the most continuously pronounced forms of nonverbal communication, as it serves two primary functions: (a) facilitating information uptake, for example, on interlocutors' visual attention and emotional expressions, and (b) remaining continuously accessible to others in face-to-face interactions (for a review, see Hessels, 2020). However, even in face-to-face interactions, interlocutors do not maintain constant eye contact throughout the conversation. The duration of the intervals during which both interlocutors gaze at each other (i.e., maintain mutual gaze) over the course of an interaction varies individually and depends on whether the interlocutor is speaking or listening (Argyle & Cook, 1976; C. Goodwin, 1981; Kendon, 1967). For instance, listeners tend to maintain prolonged gaze at speakers, interrupted with brief gaze aversions, whereas speakers tend to look at listeners less frequently, presumably in order to elicit feedback from the listeners (Bavelas, Coates, & Johnson, 2002; Brône et al., 2017; Kendon, 1967). In addition, speakers tend to briefly withdraw their gaze from listeners when they formulate utterances, which is assumed to be related to language planning (Allen & Guy, 1977; Beattie, 1981; Kendon, 1967). When speakers gaze at listeners, they often do so to elicit verbal or nonverbal feedback, e.g., in the form of head gestures and backchannel responses (e.g., *mhm, uh-huh, okay*) (Yngve, 1970), providing insight into the listeners' attention and cognitive processing (Argyle & Cook, 1976; Bavelas, Coates, & Johnson, 2002; Kendon, 1967). Moreover, speakers tend to seek feedback by briefly gazing at the listeners toward the end of their utterances, and after that they withdraw their gaze from the listeners, which is often related to the conclusion of their turns (Brône et al., 2017; Degutyte & Astell, 2021; Kendon, 1967). In sum, mutual gaze between speakers and listeners creates a temporal window that enables two key interactional possibilities: (1) a shift in interaction roles (speaker–listener) between interlocutors (Argyle & Cook, 1976; Kendon, 1967); and (2) the elicitation of feedback responses from the listeners by the speakers (Bavelas, Coates, & Johnson, 2002).

With respect to the topic management in conversations, gaze behavior influences the duration and continuity of conversational topics (Rossano, 2012, 2013). According to Rossano (2012), topics are extended to 95% of the cases when mutual gaze is established between speakers and listeners. Conversely, topics tend to get discontinued in 84% of the cases when the interlocutors withdraw their gaze from each other. Rossano (2012) analyzed gaze behavior in everyday conversations in various dyadic (and occasionally triadic) human–human interactions

using a qualitative, conversation analysis-based approach to observe the interlocutors' gaze behavior moment by moment. In the present study, we investigate the occurrence of mutual gaze and gaze withdrawals prior to topic changes in dyadic explanatory interactions, by applying a quantitative statistical approach. Additionally, the conversations analyzed in Rossano's (2012) work include competing multiple activities during the conversations, such as dining together, playing cards, or traveling in a car, which require some physical and mental resources to be distributed between the speaker and the competing task (C. Goodwin, 1985). In contrast, the present study on explainers' and explainees' mutual gaze and gaze withdrawals in explanations does not include unrelated competing activities apart from the explainees' participation in the explanations organized by the explainers. In a more recent study analyzing ten physician–caregiver explanations about a child's upcoming surgery, Lazarov et al. (2025) found that the explaining physicians more likely initiated topic changes in the explanations after the caregivers (i.e., the explainees) averted their gaze while signaling attention through co-occurring head nodding and vocal backchanneling. By comparison, when the explainees maintain a consistent gaze direction (e.g., toward the explainers) with or without other multimodal behaviors, the explainer was more likely to initiate an elaboration of a previous topic (Lazarov et al., 2025). In their research, Lazarov et al. (2025) did not consider the explainers' eye gaze behavior, leaving the presence of mutual gaze between the explainers and the explainees unaddressed.

Beyond the conversational management, gaze withdrawals (also called gaze aversions) are further associated with cognitive processing. For example, gaze aversion during interaction is linked to cognitive processing in adults (Abeles & Yuval-Greenberg, 2017; Allen & Guy, 1977; Glenberg et al., 1998) and children (Phelps et al., 2006). Experimental studies have shown that a person's gaze aversions while thinking, e.g., about a solution of challenging arithmetical or verbal tasks, enhance a person's overall task performance (Glenberg et al., 1998; Phelps et al., 2006). Further, gaze aversions have been observed to co-form the so-called "thinking face" in language processing together with other modalities, such as facial mimics and body posture (Bavelas & Chovil, 2018; M. H. Goodwin & Goodwin, 1986; Heller, 2021).

Further, gaze aversions aid the mental imagination of invisible objects, for example during matrix imagination tasks (Markson & Paterson, 2009). This suggests that gaze aversion may facilitate cognitive processing even when external visual referents are unavailable. This is relevant for the present study because the explanandum is absent from the shared space between explainers and explainees.

2.2 The topical structure of explanations

In explanations, the explainers and the explainees engage in resolving questions such as *what*, *how*, and *why* (Klein, 2009). An explanatory interaction about an explanandum can be structured through multiple sub-explananda, each corresponding to a specific explanation topic (or sub-topic), such as physical objects, abstract concepts, and processes. According to Roscoe and Chi (2008), each topic represents a brief episodic segment within the overall explanation.

The episodic segmentation of explanation topics was applied by Lazarov et al. (2025), who examined the relation between the explainees' multimodal behavior and transitions between different explanation topics in physician-initiated explanations concerning an upcoming pediatric surgery of a child. In their study, explanation topics corresponded to sub-explananda, including diagnosis, reasons for surgery, medical procedures, and treatments. Regarding the interactional responsibility for the topical organization of explanations, Fisher, Lohmer, et al. (2023) found that the topics of medical explanations are predominantly and naturally introduced by the explainers (the physicians) and to a minimum by the explainees, for example, by asking questions.

In the context of board game explanations, the board game itself serves as the overarching explanandum, which the explainers convey to the explainees. However, board games are organized by logically prescribed manuals consisting of pre-defined rules. Consequently, a board game explanation typically follows the instructions outlined in the manual. Thus, in the present study, different rules represent distinct sub-explananda or sub-topics, constituting the topical structure of board game explanations.

2.3 Hypotheses

In the present study, we investigated the relation between the topical structure of board game explanations and the explainers' and explainees' mutual gaze, or gaze withdrawals, respectively. To address our research question more specifically, we hypothesized that **the proportion of gaze withdrawals occurring prior to topical changes is higher than the proportion of mutual gaze**. Our first hypothesis is motivated by previous research by (Rossano, 2012, 2013), which suggests that the interlocutors' gaze withdrawals are associated with the closure of topical sequences in everyday conversational contexts. According to Rossano (2012), only 5% of the analyzed cases of topical closures were associated with mutual gaze. However, we also predicted that mutual gaze may occur prior to topic changes because (1) speakers often gaze at the end of their utterances toward their interlocutor in order to elicit short feedback responses from them (Argyle & Cook, 1976; Bavelas, Coates, & Johnson, 2002; Brône et al., 2017; Kendon, 1967); and (2) the explainees, who are the more listening part during the explanation (Fisher, Lohmer, et al., 2023), would maintain prolonged gaze toward the explainers (Argyle & Cook, 1976; Bavelas, Coates, & Johnson, 2002; C. Goodwin, 1981; Kendon, 1967), which presumably also includes the end of utterance.

In addition to our first hypothesis, we sought to deepen our analysis of gaze withdrawals that occur prior to topic changes. Based on previous research on the relation between interactive gaze behavior and management of turn-taking in human–human interaction (Allen & Guy, 1977; Argyle & Cook, 1976; Bavelas, Coates, & Johnson, 2002; C. Goodwin, 1981; Kendon, 1967), and specifically the results showing that interlocutors' briefly avert their gaze prior to formulating an utterance (Allen & Guy, 1977; Kendon, 1967), we generated our second hypothesis: **Gaze withdrawals are initiated more frequently by the interlocutor who initiates a topic change than by the other interlocutor who does not initiate a topic change. For example, we assume that before the explainers introduce a new topic, they briefly withdraw their gaze from the explainees**. We assume

the same for the explainees. Because the introduction of topics is related to the formulation of utterances, we assumed that either the explainers or the explainees will withdraw their gaze before introducing a new topic. However, gaze withdrawals prior topic changes may be also observed in the other interlocutor, for example the explainee, who does not initiate a topic change. In a recent study, Lazarov et al. (2025) demonstrated that the explainees' withdraw their gaze from the explainers before the explainer initiated topic changes, which could have occurred as a signal of cognitive processing (Abeles & Yuval-Greenberg, 2017; Glenberg et al., 1998).

3 Methods

3.1 Participants

In the present study, the data from 32 participants was analyzed. Among them, eight were explainers, and the other 24 were explainees. All explainers were German native-speaking adults (age: $M = 23.6$, $SD = 3.38$), two of which were male, and six were female. Only 18 of the 24 explainees provided socio-demographic data about age ($M = 26.0$, $SD = 9.75$), gender (7 male and 11 female) and native language (all German). However, this was not problematic for the present study because the research question does not address participants' gender or age. All participants signed a consent form prior to the study, which had been approved by the Ethics Board of the university.

3.2 Materials

For the data analysis in the present study, we randomly selected a subsample of 24 explanations from the video-corpus [BLINDED]. The corpus was compiled to investigate the relation between the explainees' multimodal behavior and their levels of understanding (ranging between understanding and non-understanding) moment by moment. The corpus contains 87 dyadic, explanatory interactions between explainers and explainees about a collaborative and competitive board game named 'Deep Sea Adventure' (Sasaki & Sasaki, 2014) in German language. According to the corpus design, one explainer explains the board game to three (or two) different explainees subsequently. All explanations consist of three phases: 1) the board game was physically absent, 2) the board game was physically present, and 3) an interactive game play. The game phases vary in length. For the present study, we analyzed only the phase in which the board game was absent from the shared space. The mean duration of all 24 explanations overall (incl. all three phases) was 26:49 min ($SD = 05:30$ min). The mean duration of the analyzed phase with the board game absent was 07:04 min ($SD = 03:44$ min).

The explanatory interactions were recorded from six camera perspectives: two cameras directed at participants' face area, two cameras directed at participants' torso area, one side-positioned camera, and one camera positioned at the top. Additionally, participants' speech was audio-recorded for speech analysis.

3.3 Procedure

All participants were assigned either the role of explainer, who was responsible for conveying the rules of a board game, or explainee, who received an explanation. The explainers were given the opportunity to learn and practice the board game on their own a few days prior to the study. In contrast, the explainees were not informed in advance about which board game would be explained to them. Both, the explainers and the explainees, should not have been familiar with each other prior to the study. Further, they were not informed about the main objective of the study prior to it to ensure that their interactions would remain as natural as possible, mirroring everyday communicative exchanges. The only instructions provided to the explainers were (1) the sequence in which the board game should be introduced to the explainees (i.e., the game phases), and (2) and that their explanations are sufficiently comprehensive to enable the explainees to play the game independently afterward.

3.4 Data coding

For the present study, annotators segmented the explanations into explanation episodes, and annotated explainers' (EX) and explainees' (EE) verbal and nonverbal behavior by following an annotation manual. All video data was annotated using the software ELAN (Wittenburg et al., 2006). The application of the annotation procedures described below is illustrated in Figure 2.

3.4.1 Explanation topics and topic changes

Explanation topics The segmentation of the board game explanations into distinct explanation topics was conducted based on previous methodological approaches outlined by Klein (2009) and Roscoe and Chi (2008) (see 2.2). The list of various sub-explananda anticipated to emerge within the explanations was derived from the official game instructions, which served as the primary preparatory material for the explainers prior to the study. Thus, the list of explanation topics was arranged according to the following structure:

- *Introduction* - Introducing the main explanandum, i.e., the name of the board game and the type of the game, e.g., a collaborative and competitive.
- *Preparation* - Explaining the overall structure and the components constituting the entire game, e.g., a submarine, oxygen bottle, treasure chips, dices, etc., and their physical and functional features.
- *Goal* - Explaining the main goal of the game, i.e., to collect as many treasures as possible and return successfully to the submarine.
- *Turn progressions* - Explaining the players' turn in the game, including rolling the dice, collecting treasures, and the related following consequences for the players, such as the reduction of oxygen and steps.
- *End of a round* - Explaining the conditions according to which each comes to an end.

- *End of the game* - Explaining the conditions according to which the entire game is announced over and which player wins or loses the game.

To represent this hierarchy more clearly, the different explanation topics (and sub-topics) were numbered similarly to the approach by Fisher, Robrecht, et al. (2023) (see Figure 1). As the schematic illustration of the annotation scheme shows, topic groups *2*, *4*, and *5* contain sub-topics. For these groups, annotators were asked to code the related sub-topics, e.g., *2.1*, *4.1*, *5.1*, etc. The segments of explanation topics span either full utterances or parts of utterances.

Examples of coded explanation topics introduced by the explainers:

(1) Introduction: "We are going to play a little game, and I hope you enjoy diving (pause) because we are going to dive now. (pause) And, I am going to explain you how the game works (pause) without you seeing how the game looks. (pause)"

(2.1) Game framing structure: "And, (pause) the game consists of (pause) three rounds which are overall one."

(2.2) Submarine: "And as mentioned, we are diving. It means, we will have a (pause) little (pause) submarine (pause)"

(2.3) Oxygen supply: "in which there is a certain amount of oxygen available. (pause) And this oxygen constrains our game turns."

Prior to the main annotation process, two annotators coded 10% of the data to ensure reliability (Fleiss' $\kappa = 0.79$).

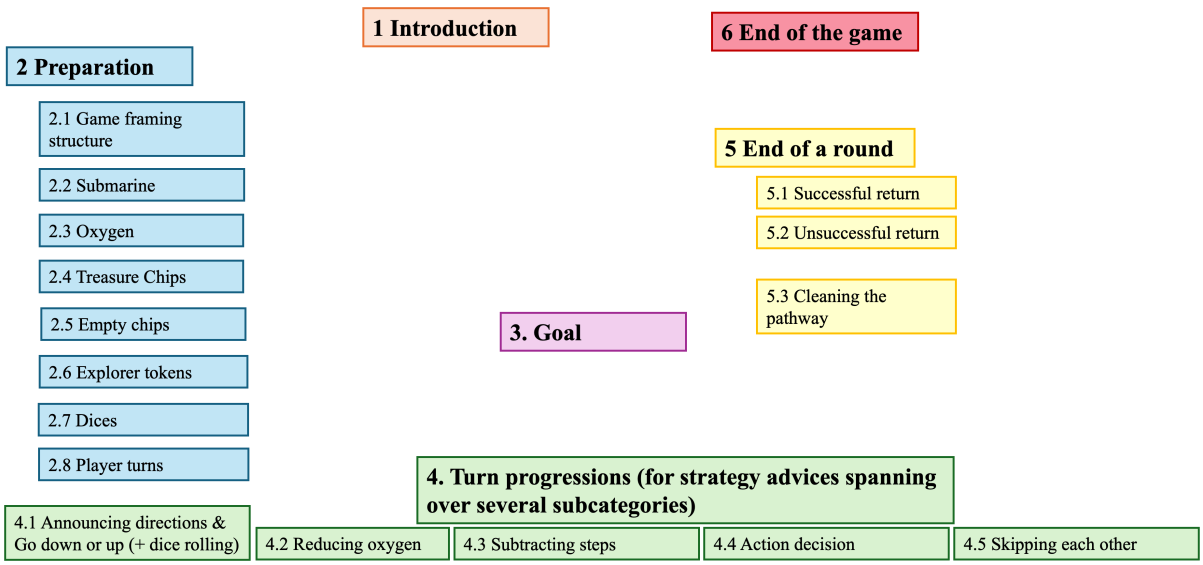

Figure 1: Annotation of explanation topics

Topic changes Following the annotation of the topics, topic changes (also between sub-topics) were annotated on a separate tier by creating very brief segments that contained annotations identifying the interlocutor who initiated the topic change – either the EX or the EE.

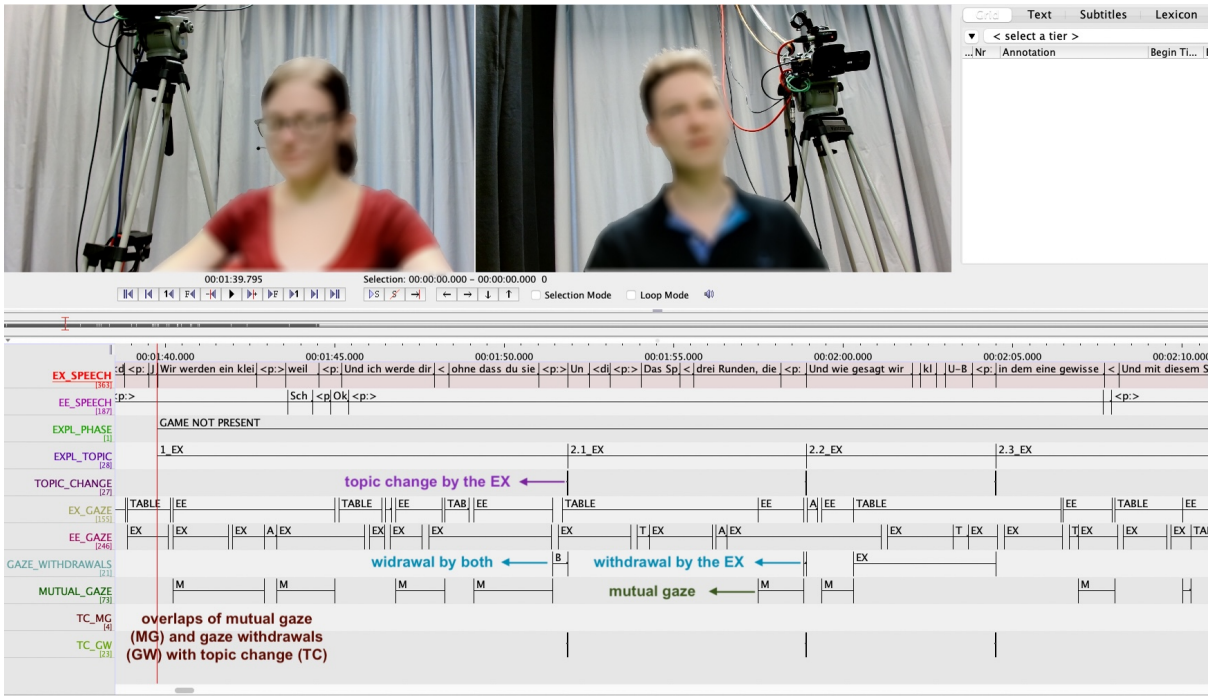

Figure 2: Annotation of explanation topics, topic changes and interactive gaze behavior in ELAN. Abbreviations: EX = explainer (left camera), EE = explainee (right camera), MG = mutual gaze, GW = gaze withdrawal, TC = topic change.

3.4.2 Eye gaze behavior

Eye gaze behavior was coded for both, the explainers and the explainees. First, participants' eye gaze was annotated according to three possible directions: (1) toward the interaction partner, (2) toward the table (the shared referential space between both participants), and (3) away (see Figure 3). One annotator coded the gaze behavior of both the explainers and the explainees using the video recordings captured by the cameras directed at the participants' facial areas. To ensure reliability, a second annotator coded 10% of the dataset during the training phase, yielding a Fleiss' κ of 0.82.

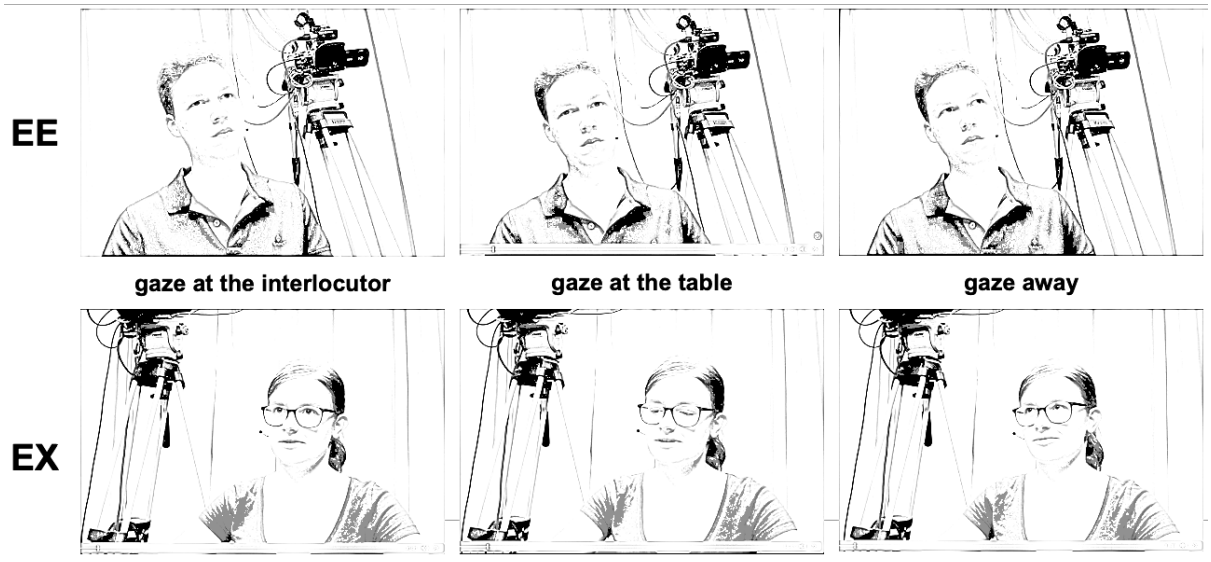

Figure 3: Annotation of gaze directions.

Second, mutual gaze and gaze withdrawals were coded on separate tiers based on the annotations of the gaze direction. *Mutual gaze* was coded when the gaze directions of both participants overlapped. *Gaze withdrawals* between explainers and explainees were determined by the instances in which the gaze direction shifted either toward the table or away. For the purpose of the present study, only those instances of the gaze withdrawals preceding topic changes were annotated. It was further specified within the segments of gaze withdrawals whether the explainer, the explainee, or both had initiated the withdrawal before a topic change. A gaze withdrawal by either the explainer or the explainee was segmented if one participant averted their gaze without subsequently redirecting it toward the other person before a topic change occurred. A gaze withdrawal by both participants was annotated if both, the explainer and the explainee, averted their gaze from one another simultaneously prior to the topic change.

Subsequently, the frequencies of the interlocutors' mutual gaze and gaze withdrawals preceding the topic changes were determined by generating overlapping annotations of topic changes with the annotations of mutual gaze and gaze withdrawals.

3.5 Data analysis

For the analysis of the frequencies of the participants' mutual gaze and gaze withdrawals that preceded the topic changes, random and fixed effects were considered. The random effect represented the nested design of data collection, according to which each of the eight explainers interacted with three different explainees subsequently. Additionally, two fixed effects were included: (1) the initiator of the topic change – either the explainer or the explainee, and (2) the interactive gaze behavior – mutual gaze or gaze withdrawal, the latter initiated by either only the explainer, or the explainee, or by both simultaneously. A Shaphiro-Wilk test of normality indicated that the frequencies of the participants' mutual gaze and gaze withdrawals preceding the topic changes were not normally distributed ($W = 0.68$, $p < 0.001$). In accordance with the structure of the dataset, the observed non-normal distribution, and the addressed research objectives in each hypothesis, we conducted two separate Generalized Linear Models (GLM) using *R Studio* (RStudio Team, 2020).

Analysis of Hypothesis 1 (1) To test our first hypothesis investigating the overall proportional occurrences of mutual gaze and gaze withdrawals prior to topic changes (regardless which interlocutor initiated the topic change), we compared the proportions of mutual gaze with the proportions of gaze withdrawals occurring prior to topic changes applying a simple Generalized Linear Model (GLM) without the random effect.

Model:

```
glm(cbind(successes, failures) ~ GAZE_GROUPED, data = binomial_data, family =
binomial())
```

The response variable in the GLM is represented by the proportional occurrences of each form of interactive gaze behavior preceding topic changes initiated by either the explainers or the explainees. The fixed effect

consisting of two levels (mutual gaze and gaze withdrawals) was represented by the variable "GAZE GROUPED". The model excludes the random effect as the first hypothesis addressed the overall occurrences of mutual gaze and gaze withdrawals across the dataset.

Analysis of Hypothesis 2 To test our second hypothesis, which focused on a more detailed analysis of the relation between gaze withdrawals and the initiation of topic changes in relation to the participants' role in the explanatory discourse – explainer or explainee – we conducted a Generalized Linear Mixed Effects Model (GLMM). By adding the random effect to the statistical model, we provided further insights into the individual differences between the explainers, as well as the intra-individual variations for each explainer (in relation to the nested design of our dataset—one explainer interacting with three explainees).

```
glmer(FREQUENCY ~ INTERACTIVE_GAZE * TOPIC_INITIATED_BY + (1|EX/EE), data = dataset,
family = poisson())
```

In the second model, FREQUENCY of the interlocutors' gaze behavior occurring prior to topic changes was the response variable. The fixed effect was represented by the interaction between INTERACTIVE_GAZE (filtered for gaze withdrawals only: by the explainers, the explainees, or both) and TOPIC_INITIATED_BY (the initiator of topic change: the explainers or the explainees). In this model, the random effect (each explainer interacting with three different explainees) was included.

4 Results

The analysis of the first hypothesis which focused on comparing the overall proportional occurrences of mutual gaze and gaze withdrawals prior to topic changes indicated that the proportion of gaze withdrawals was higher ($M = 0.80$, $SD = 0.18$) than the proportion of mutual gaze ($M = 0.20$, $SD = 0.18$). Further, the results revealed a considerable variation across the analyzed subsample (see Figure 4). The GLM demonstrated a better model fit ($AIC = 341.69$) compared to a null-model that includes only the random effect ($AIC = 923.39$), as well as a significant effect of both forms of interactive gaze behavior: mutual gaze ($\beta = -1.55$, $S.E. = 0.10$, $z = -14.84$, $p =< 0.001$) and gaze withdrawals ($\beta = 3.11$, $S.E. = 0.15$, $z = 20.99$, $p =< 0.001$).

In addition to the statistical model, we calculated the predicted probabilities for each form of interactive gaze behavior based on the parameter estimates of each predictor:

$$p_{\text{mutual}} = \text{plogis}(-1.5572) = \frac{1}{1 + e^{1.5572}} \approx 0.174$$

$$p_{\text{withdrawal}} = \text{plogis}(1.5572) = \frac{1}{1 + e^{-1.5572}} \approx 0.826$$

According to the results, the probability of mutual gaze preceding topic changes was 0.174, whereas the

probability of gaze withdrawals preceding topic changes was 0.827. With respect to the results from the statistical analysis and the post hoc probabilistic analysis, our first hypothesis could be verified.

Figure 4: Proportional distribution of mutual gaze and gaze withdrawals prior to topic changes.

For our second hypothesis, we investigated the relation between the initiator of gaze withdrawals and the initiator of topic changes. We found that the interaction partner who initiated a topic change was the one who more often withdrew their gaze prior to it compared to the frequency of gaze withdrawal of their interlocutor. More specifically, when the explainers initiated a topic change, they also withdrew their gaze more often ($M = 9.75$, $SD = 7.55$) compared to the explainee's gaze withdrawals ($M = 2.58$, $SD = 3.56$). When the explainees initiated a topic change, they withdrew their gaze more often ($M = 2.29$, $SD = 2.77$) compared to the explainer's gaze withdrawals ($M = 0.54$, $SD = 1.18$) (see Figure 5). The high standard deviations of all related mean values suggest a considerable intra-individual variability across the observed dataset.

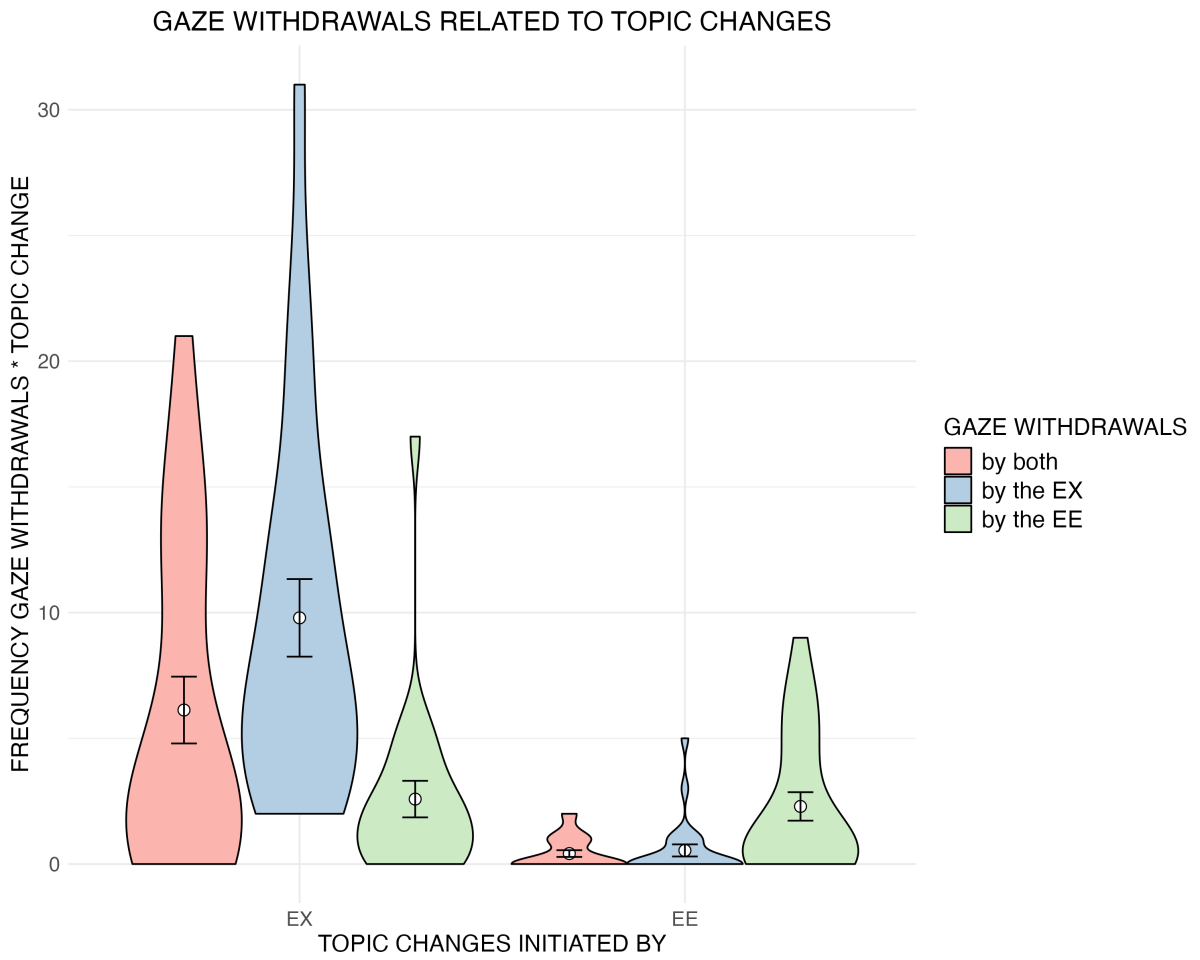

Figure 5: Gaze withdrawals solely by the explainers, solely by the explainees, or by both before topic changes initiated by either party.

In some explainer–explainee interactions, there were zero occurrences of gaze withdrawals either by the explainers, the explainees, or both observed before the topic changes. In order to avoid overdispersion in the data that may comprise the reliability of the results, we adjusted the statistical model to a negative binomial GLMM using the glmmTMB package (Brooks et al., 2017). Thus, the statistical model was modified as follows:

```
glmmTMB(FREQUENCY ~ INTERACTIVE_GAZE * TOPIC_INITIATED_BY + (1|EX/EE), data =
dataset, family = nbinom2())
```

The glmmTMB demonstrated a better model fit ($AIC = 592.13$) than the initial Poisson-GLMM ($AIC = 714.56$). Regarding the fixed effects, all factorial levels, except for the explainers' gaze withdrawals preceding the topic changes initiated by the explainees, showed a significant effect on the outcome variable (see Table 1). The results outlined that all three forms of gaze withdrawals, i.e., solely by the explainers, by the explainees, or both interlocutors simultaneously, predict topic changes initiated by the explainers.

To further test our second hypothesis, we conducted pairwise comparisons using simple contrasts. The results indicated significant differences between the explainers' and the explainees' gaze withdrawals prior to the topic changes initiated by the explainers ($\beta = 3.85$, $S.E. = 1.15$, $z = 4.49$, $p < 0.001$). Similarly, for the topic

changes initiated by the explainees, the pairwise comparisons revealed significant differences between the gaze withdrawals by the explainers and the explainees ($\beta = 0.23$, $S.E. = 0.09$, $z = -3.64$, $p < 0.001$), as well as between the withdrawals by the explainees and both interlocutors together ($\beta = 0.18$, $S.E. = 0.08$, $z = -4.01$, $p < 0.001$). In sum, the results from the statistical model indicate that our second hypothesis could also be verified; that is, gaze withdrawals are initiated more frequently by the interlocutor who initiated a topic change than by the other interlocutor.

Table 1: Model Summary

Gaze withdrawal by	Topic initiation by	*Estimate*	*SE*	*z*	*p*
EX & EE	EX	1.67	0.26	6.34	< .001
EX	EX	0.55	0.27	2.03	.043
EE	EX	-0.80	0.30	-2.69	.007
EX & EE	EE	-2.73	0.42	-6.57	< .001
EX	EE	-0.30	0.56	-0.54	.591
EE	EE	2.51	0.52	4.85	< .001

Note. EX = Explainer; EE = Explainee.

Regarding the random effects, we observed a lower variation at the level of individual explainers interacting with three different explainees ($\sigma^2 = 0.06$, $SD = 0.24$), and higher variation at the level of the individual explainers, regardless of the presence of different explainees ($\sigma^2 = 0.23$, $SD = 0.48$). However, the standard deviation for the random effect that included the interaction with different explainees suggested that, for some of the dyads, the presence of different explainees is also related to distribution of the interlocutors' gaze withdrawals preceding the topic changes initiated by either interlocutor. This observation was also supported by the explained proportion of variance when both the fixed and the random effects were included in the model (Conditional $R^2 = 0.72$), compared to the proportion of variance explained alone by the fixed effects (Marginal $R^2 = 0.60$).

5 Discussion

The current study examined the relation between mutual gaze and gaze withdrawals and changes of the topical structure across 24 dyadic board game explanations, in which the explanandum (i.e., the board game) was absent from the shared space between the explainers and the explainees. The analysis of these two forms of interactive gaze behavior followed a top-down approach, by which (1) the overall proportional distribution of mutual gaze and gaze withdrawals preceding topic changes was analyzed using a Generalized Linear Model (GLM); and based on the results of analysis 1, (2) the relation between the initiator of gaze withdrawals (the explainer, the explainee, or both) and the initiator of topic changes (the explainer or the explainee) was analyzed applying a negative binominal Generalized Linear Mixed Effects Model (glmmTMB).

Our first hypothesis that states that the proportion of gaze withdrawals occurring prior to topic change is higher than the proportion of mutual gaze could be verified by the results of the statistical analysis. Specifically, the proportion of gaze withdrawals was significantly higher than the proportion of mutual gaze across the 24

board game explanations. These findings were consistent with the previous research by Rossano (2012, 2013), that demonstrated that the interlocutors' gaze withdrawals are related to the closure of topical sequences. However, in contrast to Rossano's (2012, 2013) investigation on naturalistic, everyday dyadic (and triadic) conversations, the present study focused specifically on the context of dyadic explanations, and particularly on the explanation phase in which the explanandum was absent from the shared space. In doing so, the current study contributes to a deeper understanding of the relation between gaze withdrawals and the topical sequencing in human–human interactions. In addition, we also observed instances of mutual gaze preceding topic changes, and the statistical analysis indicated a probability of 0.17 for mutual gaze to precede topic changes. Our finding that a minimal amount of the mutual gaze occurrences is associated with the closure of topical sequences is also in line with the results from Rossano's (2012) study on spontaneous conversations. This is not surprising given the different functions that gaze behavior in human interactions serves: Interlocutors who are the less verbally active part at certain moments during the interaction (in our case the explainees) are more likely to maintain prolonged gaze toward a speaker (in our case the explainers) (Degutyte & Astell, 2021). In addition, the explainers may have also directed their gaze at the explainees for brief moments to elicit feedback from them (Argyle & Cook, 1976; Bavelas, Coates, & Johnson, 2002; Brône et al., 2017; Kendon, 1967), which led to the establishment of mutual gaze prior to the topic changes. In this regard, the relation between mutual gaze and the changes of the explanation topics accounts for continuous interactional monitoring (Clark & Krych, 2004) in explanatory interactions.

Our second hypothesis that states that gaze withdrawals are initiated more frequently by the interlocutor who initiates a topic change than by the other interlocutor could be also verified by the statistical model and the pairwise comparisons. For the topic changes initiated by the explainers, the pairwise comparisons indicated a significantly higher frequency of the gaze withdrawals by the explainers compared to gaze withdrawals by the explainees. The same significant difference was found for the topic changes initiated by the explainees who withdrew their gaze from the explainers more frequently than the explainers. These results were in line with the assumption that the gaze withdrawals are related to the topic initiation via formulating an utterance (Allen & Guy, 1977; Beattie, 1981; Kendon, 1967).

One of the major findings in our analysis is that the explainers were the interlocutors who initiated topic changes more frequently than the explainees. Therefore, we will discuss the initiation of gaze withdrawals prior to the topic changes initiated by the explainers more deeply. In this relation, the model indicated that the initiation of topic changes by the explainers can be predicted by the gaze withdrawals solely by the explainers, by the explainees, and by both interlocutors simultaneously. For each predictor there may be different reasons related to the dynamic of face-to-face interactions.

First, the explainers' gaze withdrawals preceding the topic changes initiated by themselves could be related to initiation of a topic change as an act represented by the introduction of a new idea through the formulation of an utterance. According to (Allen & Guy, 1977; Beattie, 1981; Kendon, 1967) the speakers tend to withdraw their gaze before formulating an utterance in order to free cognitive resources for language planning.

Second, the explainees' gaze withdrawals occurring before topic changes initiated by the explainers could be related to previous research on the cognitive function of gaze aversions. Gaze aversions have been previously discussed as being a signal to cognitive processing for children and adults, especially regarding challenging tasks involving thinking (Abeles & Yuval-Greenberg, 2017; Bavelas & Chovil, 2018; Glenberg et al., 1998; M. H. Goodwin & Goodwin, 1986; Heller, 2021). In the context of research on explanatory interactions, the finding from the current study is in line with the recent research on the relation between the explainees' multimodal behavior and the topical changes in medical explanations by Lazarov et al. (2025). In their research, Lazarov et al. (2025) found that the explainees' gaze aversions from the explainers, head gestures and backchannels were significantly related to transitions from elaborations to new topics. In the current study, the explainees' gaze withdrawals were analyzed as gaze aversions, and thus, the explainees' gaze withdrawals from the explainers while the explainers were gazing at the explainees and initiating a topic change could be related to the cognitive processing of the explanation. Supporting the assumption about the explainees' cognitive processing, their gaze withdrawals from the explainers could be related to the notion that withdrawals from the interlocutor aid the mental imagination of invisible objects (Markson & Paterson, 2009). In our study, the explanandum, i.e., the board game, was physically absent during the analyzed explanation phase.

Third, the simultaneous gaze withdrawals by both the explainers and the explainees also appeared to be a significant predictor for the topic changes initiated by the explainers. This finding was also in line with the previous research by Rossano (2012, 2013). This type of joint gaze withdrawals was analyzed in an exploratory manner. However, this suggests that both the explainers and the explainees reach a point of topical closure, but withdraw their gaze for different reasons: Given what we know from previous research, the explainees may have averted their gaze in order to cognitively process the explanation, and meanwhile the explainers may have completed the current topic and prepared the utterance formulation marking the beginning of the next topic.

Lastly, the second statistical model explored the variance of the explainers' and the explainees' interactive gaze behavior related to the initiation of topic changes with respect to the design of the dataset, in which each explainer sequentially interacted with three different explainees. The random effect indicated that the variance of explaining behavior in the form of shaping the topical structure of explanations was more noticeable between the individual explainers than within each dyad (that is one explainer interacting with three different explainees). This finding suggested that the topical structure of explanations is determined to a higher extent by the explainers' individual explaining behavior and to a lower extent by the presence of different explainees and their feedback behavior. One possible reason for this could be that the explainees as the less-knowledgeable participants than the explainers are also verbally less active in the explanations (Fisher, Lohmer, et al., 2023), and therefore giving the explainers (the more knowledgeable participants) more space for organizing the topical structure of the explanations.

6 Conclusion

The present study related two forms of interactive gaze behavior – mutual gaze and gaze withdrawals – to changes in the topical structure of 24 board game explanations, in which eight explainers each subsequently explained a board game to three different explainees while the board game was absent from the shared space. The revealed results support the assumption that gaze withdrawals are associated with changes of the explanation topics. In-depth analysis showed that the explainers shape the topical structure of explanations to a greater extent, and thus their gaze withdrawals from the explainees are an inherent and predictable part of their explaining behavior. However, the explainees' gaze withdrawals from the explainers as a form of nonverbal feedback behavior also predict topic changes initiated by the explainers. Thus, the interactive gaze behavior of the explainers and the explainees related to the dynamically changing topical structure of explanations contributes to the general understanding of interactional monitoring.

7 Limitations

While the present study focused solely on the relation between mutual gaze, gaze withdrawals, and topical changes in explanations, it did not investigate the relation between the interlocutors' mutual gaze and the expansion of topics as explored in the research by Rossano (2012, 2013). Additionally, regarding the relation between the interlocutors' gaze behavior and the topic changes in explanations, future research could address other forms of the explainees' feedback behavior, such as head gestures and backchannels, as well as references to explainees' understanding or interpreted understanding by the explainers. Including such factors would contribute to a more comprehensive understanding of the changing topical structure of explanations as a result of the feedback elicitation function of mutual gaze (Argyle & Cook, 1976; Bavelas, Coates, & Johnson, 2002; Kendon, 1967), which itself contributes to the entire picture of interactional monitoring (Clark & Krych, 2004). Lastly, a future comparative study including explanation phases in which the board game is present on the shared space could investigate whether the presence of the explanandum is a contextual factor which influences the relation between interactive gaze behavior and the initiation of topic changes.

References

Abeles, D., & Yuval-Greenberg, S. (2017). Just look away: Gaze aversions as an overt attentional disengagement mechanism. *Cognition*, *168*, 99–109. https://doi.org/10.1016/j.cognition.2017.06.021

Allen, D. E., & Guy, R. F. (1977). Ocular breaks and verbal output. *Sociometry*, *40*(1), 90–96. https://doi.org/10.2307/3033550

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge University Press.

Bavelas, J. B., Black, A., Lemery, C. R., & Mullet, J. (2002). “i show you how you feel”: Motor mimicry as a communicative act. *Journal of Personality and Social Psychology*, *50*, 322–329. https://doi.org/10.1037/0022-3514.50.2.322

Bavelas, J. B., & Chovil, N. (2018). Some pragmatic functions of conversational facial gestures. *Gesture*, *17*(1), 98–127. https://doi.org/10.1075/gest.00012.bav

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, *52*(3), 566–580. https://doi.org/10.1111/j.1460-2466.2002.tb02562.x

Beattie, G. W. (1981). A further investigation of the cognitive interference hypothesis of gaze patterns during conversation. *British Journal of Social Psychology*, *20*(4), 243–248. https://doi.org/10.1111/j.2044-8309.1981.tb00493.x

Brône, G., Oben, B., Jehoul, A., Vranjes, J., & Feyaerts, K. (2017). Eye gaze and viewpoint in multimodal interaction management. *Cognitive Linguistics*, *28*(3), 449–483. https://doi.org/10.1515/cog-2016-0119

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Buschmeier, H., Buhl, H. M., Kern, F., Grimminger, A., Beierling, H., Fischer, J. B., Groß, A., Horwath, I., Klowait, N., Lazarov, S., Lenke, M., Lohmer, V., Rohlfing, K. J., Scharlau, I., Singh, A., Terfloth, L., Vollmer, A.-L., Wang, Y., Wilmes, A., & Wrede, B. (2023). Forms of understanding of xai-explanations. https://arxiv.org/abs/2311.08760

Clark, H. H., & Krych, M. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*, 62–81. https://doi.org/10.1016/j.jml.2003.08.004

Cook, M. (1977). Gaze and mutual gaze in social encounters: How long—and when—we look others ”in the eye” is one of the main signals in nonverbal communication. *American Scientist*, *65*(3), 328–333. https://www.jstor.org/stable/27847843

Degutyte, Z., & Astell, A. (2021). The role of eye gaze in regulating turn taking in conversations: A systematized review of methods and findings. *Frontiers in Psychology*, *12*, 616471. https://doi.org/10.3389/fpsyg.2021.616471

Fisher, J. B., Lohmer, V., Kern, F., Barthlen, W., Gaus, S., & Rohlfing, K. (2023). Exploring monological and dialogical phases in naturally occurring explanations. *Künstliche Intelligenz*, *36*, 317–326. https://doi.org/10.1007/s13218-022-00787-1

Fisher, J. B., Robrecht, A., Kopp, S., & Rohlfing, K. J. (2023). Exploring the semantic dialogue patterns of explanations – a case study of game explanations. *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue*. https://www.semdial.org/anthology/Z23-Fisher_semdial_0007.pdf

Glenberg, A. M., Schroeder, J. L., & Robertson, D. A. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition*, *26*, 651–658. https://doi.org/10.3758/bf03211385

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. Academic Press.

Goodwin, C. (1985). Notes on story structure and the organization of participation. In J. M. Atkinson (Ed.), *Structures of social action* (pp. 225–246). Cambridge University Press.

Goodwin, M. H., & Goodwin, C. (1986). Gesture and coparticipation in the activity of searching for a word. *Semiotica*, *62*(1–2), 51–75. https://doi.org/10.1515/semi.1986.62.1-2.51

Heller, V. (2021). Embodied displays of "Doing Thinking." epistemic and interactive functions of thinking displays in children's argumentative activities. *Frontiers in Psychology*, *12*, 636671. https://doi.org/10.3389/fpsyg.2021.636671

Hessels, R. S. (2020). How does gaze to faces support face-to-face interaction? a review and perspective. *Psychonomic Bulletin & Review*, *27*(5), 856–881. https://doi.org/10.3758/s13423-020-01715-w

Jokinen, K., Harada, K., Nishida, M., & Yamamoto, S. (2010). Turn-alignment using eye gaze and speech in conversational interaction. *Proceedings of Interspeech 2010*, 2018–2021. https://doi.org/10.21437/Interspeech.2010-571

Jokinen, K., Nishida, M., & Yamamoto, S. (2010). On eye gaze and turn taking. *Proceedings of the International Conference on Intelligent User Interfaces Workshop on Eye Gaze in Intelligent Human–Machine Interaction*, 118–123. https://doi.org/10.1145/2002333.2002352

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*(1), 22–63. https://doi.org/10.1016/0001-6918(67)90005-4

Klein, J. (2009). Erklären-was, erklären-wie, erklären-warum. typologie und komplexität zentraler akte der welterschließung. In R. Vogt (Ed.), *Erklären. gesprächsanalytische und fachdidaktische perspektiven* (pp. 25–36). Stauffenburg-Verlag.

Lazarov, S., Biermeier, K., & Grimminger, A. (2025). Changes in the topical structure of explanations are related to explainees' multimodal behaviour. *Interaction Studies*, *25*(3), 257–280. https://doi.org/10.1075/is.23033.laz

Markson, L., & Paterson, K. B. (2009). Effects of gaze-aversion on visual-spatial imagination. *British Journal of Psychology*, *100*(3), 553–563. https://doi.org/10.1348/000712608X371762

Phelps, F. G., Doherty-Sneddon, G., & Warnock, H. (2006). Helping children think: Gaze aversion and teaching. *British Journal of Developmental Psychology*, *24*(3), 577–588. https://doi.org/10.1348/026151005X49872

Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H., Buschmeier, H., Grimminger, A., Hammer, B., Häb-Umbach, R., Horwath, I., Hüllermeier, E., Kern, F., Kopp, S., Thommes, K., Ngonga Ngomo, A.-C., Schulte, C., Wachsmuth, H., Wagner, P., & Wrede, B. (2021). Explanation as a social practice: Toward a conceptual framework for the social design of AI systems. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(3), 717–728. https://doi.org/10.1109/TCDS.2020.3044366

Roscoe, R., & Chi, M. T. H. (2008). Tutor learning: The role of explaining and responding to questions. *Instructional Science*, *36*(4), 321–350. https://doi.org/0.1007/s11251-007-9034-5

Rossano, F. (2012). *Gaze behavior in face-to-face interaction* [Doctoral dissertation, Radboud University Nijmegen, Nijmegen].

Rossano, F. (2013). Gaze in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 308–329). Wiley-Blackwell.

RStudio Team. (2020). *Rstudio: Integrated development environment for R*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

Sasaki, J., & Sasaki, G. (2014). *Deep sea adventure (tabletop game)*. Oink Games.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. *Proceedings the 5th International Conference on Language Resources and Evaluation*, 1556–1559. https://aclanthology.org/L06-1082/

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Yngve, V. H. (1970). On getting a word in edgewise. In M. A. Campbell et al. (Eds.), *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society* (pp. 567–577). Chicago Linguistic Society.

CONSTRUCTING EXPLAINABILITY
TRR 318
Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation
UNIVERSITÄT PADERBORN

Different explanation topics, different gestural dimensions?

Stefan Lazarov & Angela Grimminger, Paderborn University
e-mail: stefan.lazarov@uni-paderborn.de

10th ISGS 2025, July 9-11, Nijmegen

Background

- When the explanandum is absent from the shared space,
  - explainers rely on co-speech gestures to construct imagined spaces and provide the explainees with spatial orientation [1,2]
  - by employing gesture dimensions, such as deixis, iconicity, and temporal highlighting [3].
- Gesture deixis does not decrease even when explainers monitor explainees' understanding [4].
- **How are gesture iconicity and temporal highlighting along with gesture deixis distributed within different categories of explanation topics?**

Hypothesis

Along the dimension of gesture deixis, gesture iconicity is expected to dominate in topics concerning object features, whereas temporal highlighting is expected to dominate in topics concerning action processes and conditional rules.

Motivation

- Iconicity is used to depict object features and actions [3,7,8,9].
- Temporal highlighting is used to put emphasis on important syntactic / semantic content [3,10].

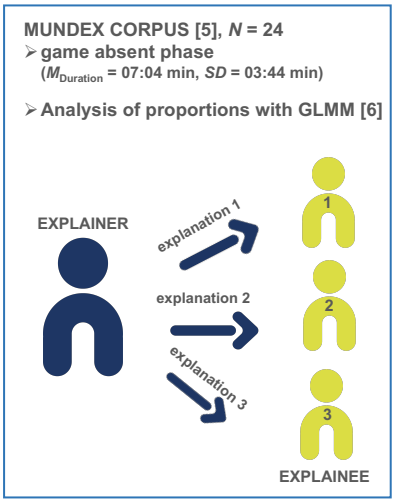

Fig 1. Data collection design

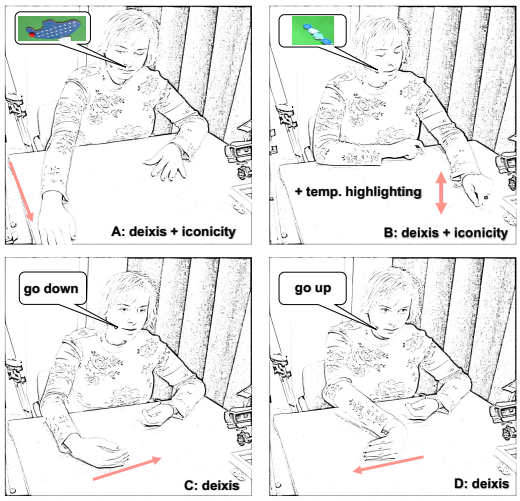

Fig 2. Annotation of co-speech gestures (κ = 0.94).

Object features	Action processes	Conditional rules
Game preparation - Submarine - Oxygen - Treasure chips - Empty chips - Explorer tokens - Dices	**Turn Progressions** - Announcing directions - Reducing the oxygen - Subtracting steps - Action decision - Skip each other	**End of a round** - Successful return - Unsuccessful return - Cleaning the pathway **End of the game**

Tab 1. Annotation of game-specific explanation topics (κ = 0.79).

Results

- No significant difference between iconicity and temporal highlighting in topics about object features.
- Temporal highlighting occurred significantly more often than iconicity in topics about action processes and conditional rules.
- **The hypothesis could be partly verified.**
- Higher variation within each explainer interacting with different explainees than across the 8 explainers.

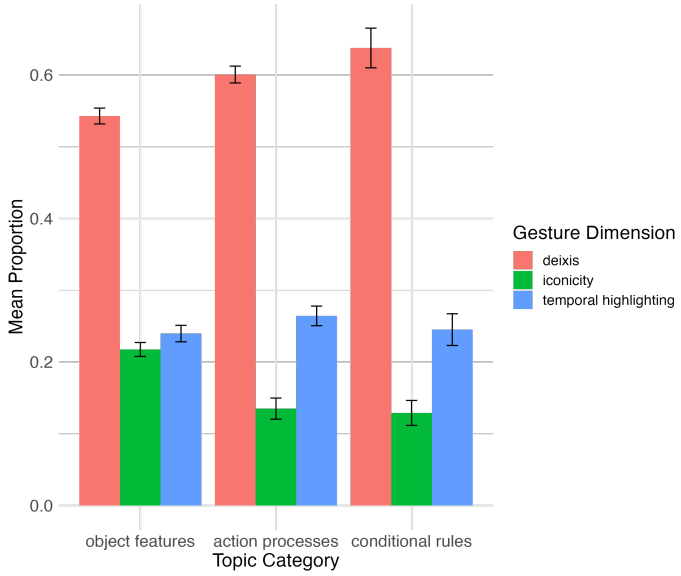

Fig 3. Proportional frequencies of gesture dimensions within categories of explanation topics.

contrast	Est	SE	z	p
Deixis - Iconicity	1.28	.2	6.51	<.0001
Deixis – Temp. highlighting	1.18	.19	6.06	<.0001
Iconicity – Temp. highlighting	-.1	.2	-.49	.877

Tab 2. Pairwise comparisons (Tukey) within **object features**.

contrast	Est	SE	z	p
Deixis - Iconicity	2.36	.22	10.72	<.0001
Deixis – Temp. highlighting	1.28	.19	6.61	<.0001
Iconicity – Temp. highlighting	-1.07	.22	-4.82	<.0001

Tab 3. Pairwise comparisons (Tukey) within **action processes**.

contrast	Est	SE	z	p
Deixis - Iconicity	3.34	.24	14.08	<.0001
Deixis – Temp. highlighting	2.49	.21	11.63	<.0001
Iconicity – Temp. highlighting	-.86	.24	-3.58	.001

Tab 4. Pairwise comparisons (Tukey) within **conditional rules**.

Discussion

- The the continuous use of deixis is related to the absence of an explanandum [2,4,11].
- Spatial references and temporal highlighting are performed more frequently than object depictions.

SCAN FOR SUPPLEMENTARY DATA AND REFERENCES

This work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) TRR 318/1 2021 - 438445824.

Variations in explainers' gesture deixis in explanations related to the monitoring of explainees' understanding

Stefan Lazarov (stefan.lazarov@uni-paderborn.de)
Faculty of Arts and Humanities, Psycholinguistics Group
& TRR 318 "Constructing Explainability", Paderborn University
33098 Paderborn, Germany

Angela Grimminger (angela.grimminger@uni-paderborn.de)
Faculty of Arts and Humanities, Psycholinguistics Group
& TRR 318 "Constructing Explainability", Paderborn University
33098 Paderborn, Germany

Abstract

In this study on the use of gesture deixis during explanations, a sample of 24 videorecorded dyadic interactions of a board game explanation was analyzed. The relation between the use of gesture deixis by different explainers and their interpretation of explainees' understanding was investigated. In addition, we describe explainers' intra-individual variations related to their interactions with three different explainees consecutively. While we did not find a relation between interpretations of explainees' complete understanding and a decrease in explainers' use of gesture deixis, we demonstrated that the overall use of gesture deixis is related to the process of interactional monitoring and the attendance of a different explainee.

Keywords: explanation; gesture deixis; monitoring; understanding

Introduction

Explanations are co-constructive interactions in which an explainer provides a less-knowledgeable person (*explainee*) with information about an entity or a process (*explanandum*) to increase their knowledge and understanding (Rohlfing et al., 2021). To increase explainees' knowledge and understanding, and to resolve understanding-related problems, explainers use verbal and non-verbal modes of communication, such as speech and gestures, simultaneously. Both modalities form an integrated system, which becomes apparent in the tight temporal and semantic coupling (Kendon, 2004; Kita, 2009; McNeill, 2005). Co-speech gestures, which express semantically related content to the spoken parts of utterances, can provide interactional guidance and support understanding via pointing, representing and highlighting certain aspects (de Ruiter, 2000). Although previous empirical research has provided findings about gestures' role in contributing to addressees' understanding (Congdon et al., 2017; Habets et al., 2011; Kelly et al., 2004; Kelly et al., 2010), it is not yet known how explainers use particular gestural functions in relation to their interpretations of explainees' different levels of understanding.

In this paper, we address this open question and focus particularly on explainers' use of gesture deixis in board game explanations in the physical absence of an explanandum, i.e., the board game. Because the game was not present first, the explainers organized the interaction by applying skills of memories and constructive imagination (Bühler, 1982; West, 2014). The goals of our study are to discover 1) how gesture deixis is used by different explainers in the temporal relation to their interpretations of explainees' understanding (assessed retrospectively), and 2) how this relation could be explained by explorations of explainers' intra-individual gestural behavior during interactions with different explainees. For this purpose, we analyzed the behavior of eight explainers, each of them explaining a board game to three different explainees consecutively (in total, 24 explanatory dialogues). We want to clarify that even though the present study focusses on gestures, the analyzed gestural forms co-occurred with speech in a natural dialogue situation.

The different dimensions of gestures

The absence of a physical explanandum may hamper addressees' comprehension of the spatial organization of unknown objects. Speech and gesture deixis play an essential role in solving the problem of spatial orientation (Bühler, 1965). Co-speech gestures may serve different functions such as highlighting or drawing on a surface shared by the interlocutors. For example, drawing invisible objects by performing gestures is essential in the successful establishment of joint imagined spaces (Kinalzik & Heller, 2020). In cases when the explanandum is absent from the shared referential space, an imaginary presentation of the explanandum by explainers' pointing and drawing behavior may be required. Therefore, applying McNeill's (2006) assumption that gestures represent multidimensional functions ("iconicity", "metaphoricity", deixis, "temporal highlighting" (for beats), "social interactivity", p. 301) to our coding seems to be more appropriate than an application of McNeill's (1992) classical formal and functional categorization of the different gesture types. The current study focuses concretely on the dimension of deixis together with other dimensions (e.g., deixis and iconicity, or deixis and highlighting) and includes hybrid gestural forms.

In general, deictic gestures represent behavior (such as pointing using extensible body parts) which establishes an

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

indexical link between a reference and a referent (McNeill, 1992; de Ruiter, 2000). They aim at attracting interlocutors' attention and at contributing to the understanding of spoken references (Clark, 2003; Stojnic et al., 2013). Regarding McNeill's dimensions, gesture iconicity represents certain features of a referent and is semantically related to the co-occurring speech (de Ruiter, 2000; McNeill, 1992; Poggi, 2008). Like gesture deixis, gesture iconicity contributes to the attraction of addressee's attention, and to their memory recall and comprehension (Dargue et al., 2021; Kandana-Arachchige et al., 2021; McKern et al., 2021). In contrast to the dimensions of deixis and iconicity, temporal highlighting, realized by beat gestures, does not convey semantic information, but it emphasizes information by being temporally aligned with a related part of a spoken utterance and with prosodic marking (Beege et al., 2020; Dimitrova et al., 2016). Some research has reported that beat gestures may contribute to understanding, however at a much lower degree than deixis or iconicity do in native speaking contexts (Austin & Sweller, 2014; Dimitrova et al., 2016; Rohrer et al., 2020). Thus, we investigate the dimension of gesture deixis in observable hybrid forms of co-speech gesture categories to account for the multidimensionality of gestures. Together with other dimensions, such as iconicity and highlighting, we relate explainers' use of gesture deixis to their interpretations of explainees' understanding.

Gestures and comprehension

Previous research relating speakers' co-speech gestures to addressees' understanding has shown that gestures have a general positive effect on understanding (Congdon et al., 2017; Grimminger et al., 2010). As mentioned in the previous section, gesture deixis and iconicity bear semantic information, i.e., they convey meaning (McNeill, 1992; 2006). The semantic congruency between gestures and speech has been also related to a faster reaction time and gesture interpretation, during addressees' observation of speakers' gestures (Habets et al., 2011; Kelly et al., 2010; Ping et al., 2013). Furthermore, observing gestures has been shown to reduce learners' cognitive load and foster social engagement (Li et al., 2021). Although studies have found that co-speech gestures can increase understanding, it is not yet clear how the use of deictic gestures (and hybrid forms) by explainers is related to the dynamics of explainers' interpretations of explainees' understanding, also with respect to interacting with different explainees consecutively.

Monitoring explainees' understanding

Understanding is defined as a cognitive process with gradual qualities (levels) ranging between non-understanding, partial understanding and complete understanding (Bazzanella & Damiano, 1999; Vendler, 1994). In addition to the levels of non-understanding and partial understanding, there is another state, misunderstanding, which refers to an incorrect reception of information. Misunderstandings could be resolved after a detection of the problem and the initiation of a repair by the explainer (Vendler, 1994).

In interaction processes, interlocutors monitor each other's (non-)verbal behavior continuously and elicit information about the achieved level of understanding of an explanandum (Clark & Krych, 2004). Because levels of understanding are gradually changing, monitoring explainees' (non-)verbal signals by explainers could lead to a dynamic variation of explainers' strategies of explaining, including variations in gesturing. Following this assumption, a dynamic variation in gesturing may be observed in relation to explainers' interpretations of explainees' understanding. Monitoring explainees' understanding could be a challenging task for explainers due to the possibility of misinterpretations of explainees' (non-)verbal feedback. Previous research on the interpretations of (non-)verbal feedback signals has shown that (non-)lexical backchannels (Allwood et al., 1992; Arnold, 2012; Bavelas et al., 2000; Ward & Tsukahara, 2000; Yngve, 1970) and head nods (Allwood & Cerrato, 2003; Gander & Gander, 2020) evoke ambiguous interpretations towards either unconditional understanding or solely attention. Furthermore, gaze aversions from an explaining interlocutor can be misinterpreted by explainers as disengagement from a task (Doherty-Sneddon & Phelps, 2007; Jongerius et al., 2022) rather than as a signal of ongoing cognitive processing (Glenberg et al., 1998). Even though explainees' various multimodal signals may lead to misinterpretations because they have been reported to be ambiguous, it is yet interesting how explainers' interpretations of different levels of understanding may be related to characteristics of gesture use on the dimension of deixis.

One way of documenting explainers' interpretations of explainees' understanding in explanatory dialogues moment by moment is the collection of protocolled retrospective accounts from the explainers (Kuusela & Paul, 2000). An applicable related procedure is the conduction of video-recall. Video-recall is a post-test procedure after the main interaction study which aims at stimulating interaction partners' short-term memory of an interaction that has already taken place. Video-recalls can be conducted, for example, by presenting a videorecording of an interaction to the interaction partners and providing the participants with instructions about the demanded focus on specific aspects and events of an explanation (see Methods for a detailed description of the video-recall procedure in this study).

The individuality of gestural behavior

In addition to the relation between explainers' gesture deixis and their interpretations of explainees' understanding, we are also interested in the individual behavior of each explainer towards three different explainees. Previous research on formal gesture features, such as form and path, has shown that gesturing is idiosyncratic, i.e., speaker-individual (Bergmann & Kopp, 2009; Priesters & Mittelberg, 2013). However, Bergmann & Kopp (2009) suggest that the idiosyncratic gesture production by different speakers may also vary in relation to the dialogue situation and the presence of a different addressee. Further, individuals' higher gesture

rates have been reported when there is a greater the degree of expertise between interlocutors (Holler & Stevens, 2007; Jacobs & Garnham, 2007; Kang et al., 2015), or when the explanandum is not present during an explanation (Holler & Stevens, 2007). Based on the previous findings on individual gesturing behavior, we would like to extend the research on this topic by describing the intra-individual variations of different explainers' gesture deixis related to their interpretations of the levels of understanding of three different explainees.

Hypotheses

In the present study, we investigate the dynamics in explainers' gesture deixis in relation to the monitoring of explainees' levels of understanding. Because explainers were required to organize their explanations in the physical absence of an explanandum, also drawing on memories and imagination (Bühler, 1982; West, 2014) about the spatial organization of the board game, we expected the occurrence of hybrid gesture forms combining gesture deixis with iconicity (e.g., drawing) or highlighting.

Based on previous studies on the comprehension providing function of gestures (Congdon et al., 2017; Grimminger et al., 2010; Kang et al., 2015), and specifically on deictic (Clark, 2003; Stojnic et al., 2013) and iconic gestures (Dargue et al., 2021; Kandana-Arachchige et al., 2021; McKern et al., 2021), we assume that explainers change the frequency of their gestures based on their interpretations of explainees' understanding. Although previous studies have analyzed gesture use in experimental conditions in which speech is less accessible or less informative, we assume that this also accounts for naturalistic conversation settings, such as those in the present study. We hypothesized that:

(1) Following explainers' interpretation of explainees' complete understanding, explainers' gesture deixis decreases while following explainers' interpretations of explainees' non-, partial or misunderstanding, explainers' gesture deixis increases.

Second, we are interested in intra-individual differences in forms of gesture deixis. Because of the scarcity of empirical work on speakers' gesturing related to interpretations of addressees' (levels of) understanding, this is addressed in an exploratory manner. Based on the previous findings on the individual use of gestures by different speakers (Bergmann & Kopp, 2009; Priesters & Mittelberg, 2013) and variations depending on the addressee (Holler & Stevens, 2007; Jacobs & Garnham, 2007; Kang et al., 2015), we hypothesized that:

(2) The gesture deixis of individual explainers varies depending on the attendance of a different explainee.

We will explore the effect of three different explainees on the gesture deixis of one explainer.

Methods

Data Corpus and Procedure

The sample analyzed in the present study has been randomly selected from the MUNDEX corpus ("Multimodal understanding of explanations") (Türk et al., 2023). MUNDEX is a large video-corpus which contains 87 dyadic, explanatory interactions about the board game *Deep Sea Adventure* in German language. It has been collected to investigate the monitoring of multimodal signals of understanding of explanations.

Dyadic interactions The interactions were videorecorded from six different camera angles (two at each participant's face area, two directed towards each participant's torso, hands and head, one side angle, and one top angle over both interaction partners). The speech of both interlocutors was additionally audio-recorded with individual headsets. In the dyadic interactions, an explainer explained a board game either to three or two explainees consecutively. The interlocutors were unknown to one another. The game was given to the explainers one or two days prior to the study, so that they could learn it on their own. No guided instructions as how to learn the game or additional instructions of the game were provided to them by the experimenters in order to avoid modeling a way of explaining the game during the study. All explainers were thus free to organize the explanations by themselves without any guidance by the experimenters because the study focused on explanatory phenomena natural conversations. The only guidance that the explainers received was to begin the explanations without presenting the board game to the explainees, then to freely choose the moment at which they present the board game to the explainees, and finally to play the game interactively. Thus, each interaction consists of three timely varying phases: game absent, game present and a game play. For the present analysis, we randomly selected eight different explainers, resulting in 24 explanations in total. The mean duration of all 24 explanations overall (incl. all three phases) was 26:49 min (*SD* = 05:30 min). The mean duration of the analyzed phases with the board game absent was 07:04 min (*SD* = 03:44 min).

Video-recall task Following the dyadic interactions, each explainer and each explainee took part in a video-recall task, in which they individually watched the recorded dyadic interaction (side angle camera). Before this task, both interaction partners were instructed to comment on any moment from the interaction for which they recognize explainees' (for the explainers) or their own (for the explainees) different levels of understanding, and to use the key terms *understanding, partial understanding, non-understanding,* and *misunderstanding*. For this analysis, only explainers' comments were used. Each explainer participated three (or two) times in the video-recall task, depending on the number of explainees to whom they explained the game.

Participants

The subsample used for the current analysis consists of eight explainers, who were German native speaking adults (*M* = 23.6, *SD* = 3.38). Among them, two were males, and six were females. Only 18 of the 24 explainees provided socio-

demographic information about age (M = 26.0, SD = 9.75), gender (7 male and 11 female) and native language (also German). All participants signed a consent form. The study had been approved by the Ethics Board of the university.

Data coding

All data analyzed in the present study were annotated using *ELAN* software (Max Plank Institute for Psycholinguistics, The Language Archive). Three coders annotated the data. Coder A annotated explainers' hand gestures in the dyadic interactions and explainers' comments on explainees' understanding from the video-recall task. Coders B and C annotated 10% of the data, respectively, to assess reliability.

Hand gestures For annotating explainers' hand gestures, coder A segmented and annotated gesture phrases (McNeill, 1992), that is, gestural movements constituted of gesture strokes and optional preparation and retraction phases of the arm and hand. The recordings from the camera perspective directed towards the torso, hand and head of the explainers were used (with the audio turned on) because it allowed observing explainers' hand shapes and movements over the shared referential space. To ensure reliability, 10% of the data were annotated by coder B (κ = 0.94). Coders identified first explainers' pointing behavior based on explainers' hand / finger shape, and then they annotated the relevant gesture functions according to the feature definitions provided by McNeill (1992, 2006). We observed the dimension of gesture deixis not only in the one-dimensional form of deictic gestures, but also in hybrid forms including iconicity or beats, i.e., deictic-iconic or deictic-beat gestures. *Deictic gestures* were coded based on a single pointing towards a direction or a location where an invisible object would be placed, and co-occurring with the related spoken reference. *Deictic-iconic gestures* were coded based on the criteria for categorical deictic gestures complemented by hand or finger shapes or movements depicting an object, features of an object, or a path. The explainers from our study were observed to point at locations while depicting objects by either positioning the index finger and the thumb in an object related form or drawing objects on the shared referential space by the index finger (Streeck, 2008). *Deictic-beat gestures* were coded based on the criteria for categorical deictic gestures, complemented by (repetitive) biphasic rhythmic hand / finger movements in the presence of prosodic highlighting.

Levels of understanding Coder A annotated the explainers' comments during the video-recall task into the four levels of understanding (Vendler, 1994) that the participants were given as key terms: *understanding, partial understanding, non-understanding*, and *misunderstanding*. Many of the comments could be directly coded based on the presence of these key terms. However, there were other types of comments which did not contain the provided key terms for understanding from the instructions, but rather synonymous or colloquial expressions, for example "to make click" (coll. German for understanding) or "to be unable to visualize" (for non-understanding). Those expressions were coded as one of the levels of understanding. Also, there were comments which were not directly related to explainees' understanding, but rather to the quality of explanation, and such unrelated comments were not considered in the analysis. Coders were trained to sort and decode the relevant information related to explainees' level of understanding. Coder C annotated 10% of the data for a reliability check (κ = 0.85).

Data analysis

For the analysis, all forms of deictic gestures (deictic, deictic-iconic, and deictic-beat) were collapsed into a single variable (gesture deixis). The number of all forms of deictic gestures produced in the gaps between the annotated levels of understanding were counted. The gaps represented the time between two documented levels of explainees' understanding by the explainers. The number of explainers' reports on explainees' levels of understanding varied between the individual dyadic interactions (Table 1).

Table 1. Number of reported levels of understanding across the analyzed subsample of 24 dyadic interactions.

reported levels of:	sum	range	*M*	*SD*
understanding	89	1-22	4.94	5.30
partial understanding	58	1-9	3.41	2.53
non-understanding	61	1-10	2.54	2.10
misunderstanding	18	1-8	2.00	2.34

The data frame was structured according to the nested design of data collection, i.e., the random effect was structured hierarchically in two columns (explainer and explainee). Before choosing the appropriate statistical model, we ran Shapiro-Wilk normality test, which indicated a non-normal distribution of explainers' gestures across the 24 interactions (W = 0.90, p < 0.05). Because of non-normal distribution, we ran a Generalized Linear Mixed Effects Model (GLMM) in *Rstudio* (Rstudio Team, 2020), using the *lme4* package (Bates et al., 2015) with the function:

glmer <- GEST_FREQ ~ UNDERSTAND + (1 | EX/EE)

The frequencies of explainers' different forms of deictic gestures were used as the response variable. The monitored levels of understanding (four-level) were the fixed effect applying a simple contrast, comparing the levels of partial understanding, non-understanding and misunderstanding to the reference level of understanding. The random effect was defined by the nested study design representing each explainer interacting with a different explainee.

Results

Our statistical model indicated a balanced good fit (AIC = 1271.2; BIC = 1283.9) compared to a null model without the fixed effect (AIC = 1384.8; BIC = 1391.2), a low proportional variance based on the fixed effect (marginal R^2 = 0.165), but

a higher proportional variance in combination with the random effect (conditional R^2 = 0.943). The nested random effect indicated a greater variance of individual explainers' gesture deixis across interacting with different explainees (σ^2 = 0.21, *SD* = 0.45) compared to the variance across the eight different explainers regardless the attendance of three different explainees (σ^2 = 0.08, *SD* = 0.29). The fixed effects summary (Table 2 and Figure 1) suggests that the levels *understanding*, *partial understanding* and *misunderstanding* have a significant effect on the variations of the frequencies of gesture deixis across the explainers.

Table 2. Explainers' frequency of gesture deixis related to interpretations of explainees' understanding.

effect	*M*	*SD*	*β*	*SE*	*z*	*p*
U (int.)	46.19	32.59	3.95	0.14	27.59	***
PU	38.0	23.69	-0.33	0.06	-5.83	***
NU	61.36	44.94	0.05	0.05	0.97	ns
MU	27.28	26.48	-0.67	0.09	-7.63	***

*** ($p < 0.001$), ns ($p > 0.05$)
U = understanding (intercept), PU = partial understanding, NU = non-understanding, MU = misunderstanding

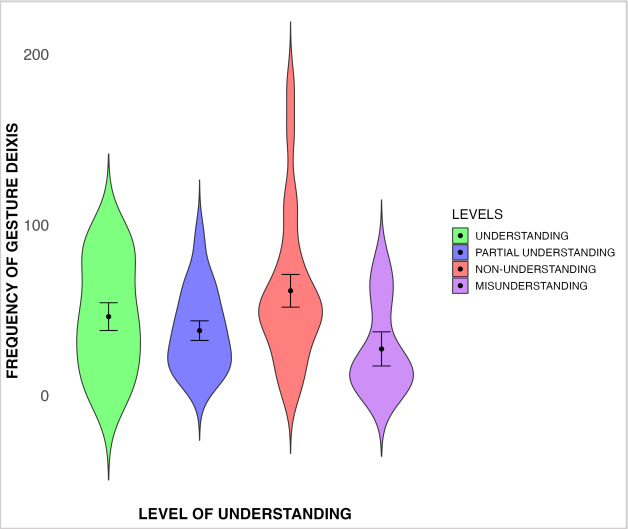

Figure 1: Explainers' gesture deixis related to interpretations of explainees' understanding.

For testing our first hypothesis whether explainers' gesture deixis decreases after monitoring complete understanding or increases after monitored partial, non- and misunderstanding, we looked at the estimated means and conducted post-hoc pairwise comparisons for significant differences. The results are summarized in Table 3. Overall, the results do not suggest that explainers' gesture deixis decreases following the interpretation of complete understanding in explainees' behavior. Gesture deixis after monitoring non-understanding increases slightly compared to gesture deixis after monitoring complete understanding. However, the difference between both extremes is not significant (β = -0.05, *SE* = 0.05, *z* = 0.97, $p > 0.05$). We observed that gesture deixis decreases in relation to explainers' reports of explainees' partial understanding and misunderstanding. The statistical model indicated significant differences for the comparison between understanding and partial understanding (β = 0.33, *SE* = 0.06, *z* = 5.83, $p < 0.001$), as well as for the comparison between understanding and misunderstanding (β = 0.67, *SE* = 0.09, *z* = 7.63, $p < 0.001$).

Table 3. Explainers' frequency of gesture deixis related to interpretations of explainees' understanding: Estimated means and *SE*.

Understanding	*EM*	*SE*	*LCL*	*UCL*
U	3.95	0.14	3.67	4.23
PU	3.62	0.14	3.33	3.90
NU	4.00	0.14	3.73	4.28
MU	3.28	0.16	2.97	3.59

Although the results indicated that monitoring explainees' understanding, partial understanding and misunderstanding is related to variations of explainers' gesture deixis, hypothesis 1 could not be verified. The frequency of explainers' gesture deixis following interpretations of explainees' complete understanding is not significantly different than the frequency of gesture deixis following interpretations of explainees' non-understanding, and it decreases significantly following interpretations of explainees' partial and misunderstanding.

For hypothesis 2, we explored intra-individual differences in explainers' gesture deixis to reveal the random effect variations from our statistical model in a descriptive manner. The first part of our analysis indicated higher intra-individual variations of explainers' gesture deixis regarding the three different explainees compared to inter-individual variations between the eight explainers. The individual charts in Figure 2 illustrate normalized proportions derived from the absolute frequencies of each explainer's gesture deixis related to the reported levels of understanding of each explainee. The variance of monitored levels of understanding for each of the interactions between an explainer (EX) and an explainee (EE) is immediately visible: Explainers have not reported on monitoring all four levels of understanding in each interaction with a different explainee. Thus, we can compare the use of gesture deixis only for non-understanding, partial understanding and understanding. Regarding the level of non-understanding, we observed intra-individual differences in the proportions of gesture deixis for EX12, EX13 and EX16. All explainers who monitored explainees' partial understanding used gesture deixis differently when interacting with a different explainee. Comparable differences between the proportions of explainers' gesture deixis related to monitoring explainees' understanding were observed in EX7, EX9, EX11, EX13 and EX19. Our results indicate that the use of gesture deixis is related not only to the

attendance of a different explainee but also to the monitored level of explainees' understanding by the explainers. The results on the variance at the level of different explainees and the influence of the monitored levels of explainees' understanding on the frequencies of gesture deixis, support hypothesis 2 that explainers exhibit intra-individual variations in gesture deixis regarding the monitored levels of understanding.

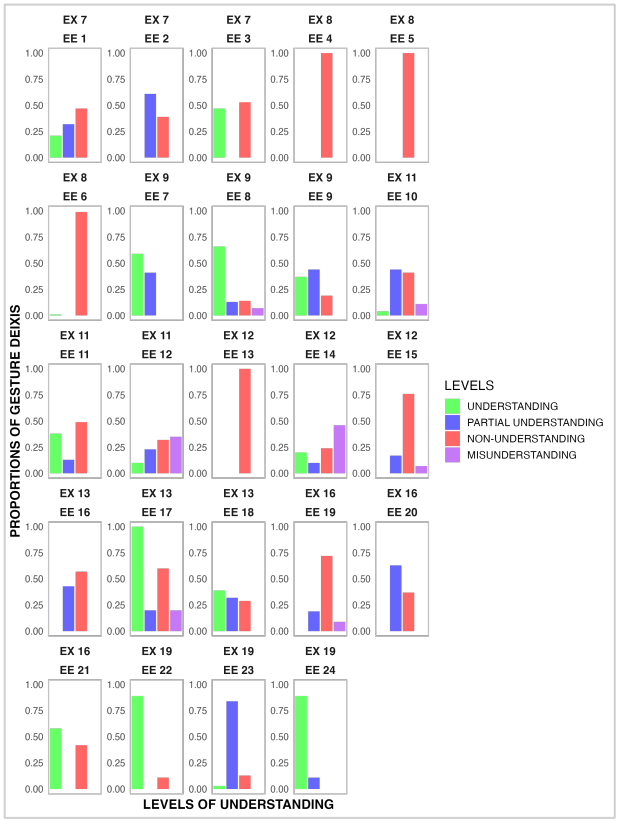

Figure 2: Individual proportional variations of explainers' gesture deixis related to interpretations of explainees' understanding.

Discussion

In this study, explainers' gesture deixis in relation to their interpretations of explainees' levels of understanding when explaining a board game was analyzed. Further, explainers' intra-individual variations of gesture deixis when interacting with different explainees were addressed. Other than hypothesized, the results indicated that monitoring explainees' complete understanding is not followed by a decrease in explainers' gesture deixis. Also, the exploration of individual explainer's gestures revealed that explainers adapted their deictic gestures within each interaction with a different explainee and their interpretation of the level of understanding.

Based on previous research on addressees' increasing comprehension when observing co-speech gestures (Clark, 2003; Congdon et al., 2017; Dargue et al., 2021; Kandana-Arachchige et al., 2021; McKern et al., 2021; Stojnic et al., 2013), we assumed that explainers' interpretations of explainees' complete understanding would be associated with a decrease in their pointing behavior in the interaction. In our analysis, we did not find support for this assumption. Explainers' use of gesture deixis during the explanations in the absence of the board game remained stable, even when interpreting explainees' complete understanding. One possible reason for explainers' continuous use of gesture deixis could be that the absence of the board game required the establishment of joint imagined spaces (Kinalzik & Heller, 2020), also by pointing to invisible locations and referents. This might have been especially pronounced because the explainers familiarized themselves with the game instructions before the study, and thus they had become experts of the board game, in comparison to the explainees who were novices. Because of this knowledge gap during the interaction and the physical explanandum being absent, explainers may have expected a continuous high demand for a visual presentation of the board game components and their spatial organization on the imagined space by the less knowledgeable explainees (Kang et al., 2015).

Our results on explainers' individual use of gesture deixis when interacting with different explainees could be related to previous findings on speakers' individual behavior (Priesters & Mittelberg, 2013) and possible variations depending on the attendance of different addressees (Bergmann & Kopp, 2009; Jacobs & Garnham, 2007). Regarding the findings from the current study, we conclude that gesture deixis is related not only to different explainees, but also to explainers' monitoring of explainees' understanding.

In this paper, we focused only on explainers' retrospective reports on explainees' understanding without considering other forms of dynamics, such as explainees' verbal and nonverbal behavior in the interactions or the topical organization of the explanations. This is a limitation of the study. Therefore, in future analyses we aim to expand our research to explore explainers' gesture deixis within certain topics from the explanations, such as specific game rules. Thus, the consideration of the topical organization (i.e., openings and closures of topics, as well as elaborations of topics) would also allow the analysis of explainers' gesture deixis and their relation to dynamics of explainees' understanding within specific explanation episodes.

Further research will consider more fine-grained statistical analyses including hybrid gestural forms (i.e., deictic-iconic, deictic-beat) and different forms of explainees verbal and nonverbal forms of feedback behavior (e.g., gaze behavior, head gestures, and linguistic backchannels).

Acknowledgments

This work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) TRR 318/1 2021 - 438445824. We thank all participants for supporting this research and our research assistants for their help in transcribing and annotating the video data.

References

Allwood, J., Nivre, J., & Ahlsén, E. (1992). On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics, 9*(1), 1–26. https://doi.org/10.1093/jos/9.1.1

Allwood, J. & Cerrato, L. (2003). A Study of Gestural Feedback Expressions. In P. Paggio, K. Jokinen & A. Jönsson. (Eds.), *First Nordic Symposium on Multimodal Communication* (pp. 7-22).

Arnold, K. (2012). Humming along. *Contemporary Psychoanalysis, 48*(1), 100–117. https://doi.org/10.1080/00107530.2012.10746491

Austin, E. E., & Sweller, N. (2014). Presentation and production: the role of gesture in spatial communication. *Journal of Experimental Child Psychology, 122*, 92-103. https://doi.org/10.1016/j.jecp.2013.12.008

Bates, D. M., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting Linear Mixed-Effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co narrators. *Journal of Personality and Social Psychology, 79*(6), 941–952. https://doi.org/10.1037/0022-3514.79.6.941

Bazzanella, C., & Damiano, R. (1999). The interactional handling of misunderstanding in everyday conversations. *Journal of Pragmatics, 31*(6), 817-836. https://doi.org/10.1016/S0378-2166(98)00058-7

Beege, M., Ninaus, M., Schneider, S., Nebel S., Schlemmel, J., Weidenmüller, J., Moeller, K. & Rey G. D. (2020). Investigating the effects of beat and deictic gestures of a lecturer in educational videos. *Computers & Education, 156*, Article 103955. https://doi.org/10.1016/j.compedu.2020.103955

Bergmann, K., & Kopp, S. (2009). Systematicity and idiosyncrasy in iconic gesture use: empirical analysis and computational modeling. In S. Kopp & I. Wachsmuth (Eds.), *Lecture Notes in Computer Science* (pp. 182-194). https://doi.org/10.1007/978-3-642-12553-9_16

Bühler, K. (1965). *Sprachtheorie: Die Darstellungsfunktion der Sprache*. Gustav Fischer Verlag.

Bühler, K. (1982). The deictic field of language and deictic words. In R. J. Jarvella & W. Klein (Eds.), *Speech, place and action: Studies in Deixis and related topic* (pp. 9-30). John Wiley & Sons, Ltd.

Clark, H. H. (2003). Pointing and placing. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet* (pp. 243–268). Lawrence Erlbaum. https://doi.org/10.4324/9781410607744

Clark, H. H. & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50*, 62-81. https://doi.org/10.1016/j.jml.2003.08.004

Congdon, E., Novack, M. A., Brooks, N., Hemani-Lopez, N., O'Keefe, L., & Goldin-Meadow, S. (2017). Better together: Simultaneous presentation of speech and gesture in math instruction supports generalization and retention. *Learning and Instruction, 50*, 65–74. https://doi.org/10.1016/j.learninstruc.2017.03.005

Dargue, N., Phillips, M. & Sweller, N. (2021). Filling the gaps: observing gestures conveying additional information can compensate for missing verbal content. *Instructional Science, 49*, 637-659. https://doi.org/10.1007/s11251-021-09549-2

de Ruiter, J. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and Gesture (Language Culture and Cognition)* (pp. 284-311). Cambridge University Press. https://doi.org/10.1017/CBO9780511620850.018

Dimitrova, D., Chu, M., Wang, L., Özyürek, A., Hagoort, P. (2016). Beat That Word: How Listeners Integrate Beat Gesture and Focus in Multimodal Speech Discourse. *Journal of Cognitive Neuroscience, 28*(9), 1255-1269. https://doi.org/10.1162/jocn_a_00963

Doherty-Sneddon, G., & Phelps, F. G. (2007). Teacher's responses to children's eye gaze. *Educational Psychology, 27*(1), 93-109. https://doi.org/10.1080/01443410601061488

ELAN (Version 6.2.) [Computer Software]. (2021). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. https://www.archive.mpi.nl/tla/elan

Gander, A. G., & Gander, P. (2020). Micro-feedback as cues to understanding in communication. Dialogue and Perception – Extended Papers from DaP2018. In C. Howes, S. Dobnik, & E. Breitholtz (Eds.), *CLASP Papers in Computational Linguistics* (pp. 1-11). Gothenburg University.

Glenberg, A. M., Schroeder, J. L., & Robertson, D. A. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory & cognition, 26*(4), 651–658. https://doi.org/10.3758/bf03211385

Grimminger, A., Rohlfing, K. J., & Stenneken, P. (2010). Children's lexical skills and task demands affect gestural behavior in mothers of late-talking children and children with typical language development. *Gesture*, *10*(2–3), 251–278. https://doi.org/10.1075/gest.10.2-3.07gri

Habets, B., Kita, S., Shao, Z., Özyürek, A., & Hagoort, P. (2011). The Role of Synchrony and Ambiguity in Speech–Gesture Integration during Comprehension. *Journal of Cognitive Neuroscience, 23*(8), 1845–1854. https://doi.org/10.1162/jocn.2010.21462

Holler, J. & Stevens, R. (2007). The effect of common ground on how speakers use gesture and speech to represent size information. *Journal of Language and Social Psychology, 26,* 4–27. https://doi.org/10.1177/0261927X06296428

Jacobs, N. & Garnham, A. (2007). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language, 56,* 291–303. https://doi.org/10.1016/j.jml.2006.07.011

Jongerius, C., Hillen, M. A., Romijn. J., A., Smets, E. M. A., & Koole, T. (2022). Physician gaze shifts in patient physician interactions: functions, accounts and responses. *Patient Education and Counceling, 105*(7), 1-14. https://doi.org/10.1016/j.pec.2022.02.018

Kandana-Arachchige, K. G., Blekic, W., Simoes Loureiro, I., & Lefebvre, L. (2021). Covert attention to gestures is sufficient for information uptake. *Frontiers in Psychology, 12*, 776867. https://doi.org/10.3389/fpsyg.2021.776867

Kang, S., Tversky, B., & Black, J. B. (2015). Coordinating gesture, word, and diagram: Explanations for experts and novices. *Spatial Cognition & Computation*, *15*(1), 1–26. https://doi.org/10.1080/13875868.2014.958837

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language, 89*(1), 253–260. https://doi.org/10.1016/S0093-934X(03)00335-3

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two Sides of the Same Coin: Speech and Gesture Mutually Interact to Enhance Comprehension. *Psychological Science, 21*(2), 260-267. https://doi.org/10.1177/0956797609357327

Kendon, A. (2004). *Gesture: Visible Action as Utterance.* Cambridge University Press. https://doi.org/10.1017/cbo9780511807572

Kinalzik, N., & Heller, V. (2020). Establishing joint imagined spaces in game explanations: Differences in the use of embodied resources among primary school children. *Research on Children and Social Interaction, 4*(1), 28-50. https://doi.org/10.1558/rcsi.12417

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes, 24*(2), 145–167. https://doi.org/10.1080/01690960802586188

Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology, 113*(3), 387-404. https://doi.org/10.2307/1423365

Li, W., Wang, F., Mayer, R. E., & Liu, T. (2021). Animated pedagogical agents enhance learning outcomes and brain activity during learning. *Journal of Computer Assisted Learning, 38*(3), 621–637. https://doi.org/10.1111/jcal.12634

McKern, N., Dargue, N., Sweller, N., Sekine, K. & Austin, E. (2021). Lending a hand to storytelling: Gesture's effects on narrative comprehension moderated by task difficulty and cognitive ability. *Quarterly Journal of Experimental Psychology, 74*(10), 1781-1895. https://doi.org/10.1177/17470218211024913

McNeill, D. (1992). *Hand in Mind: What Gestures Reveal about Thought.* The University of Chicago Press.

McNeill, D. (2005). *Gesture and Thought.* University of Chicago Press. https://10.7208/chicago/9780226514642.001.0001

McNeill, D. (2006). Gesture and Communication. In K. Brown (Ed.), *Encyclopedia of Language & Linguistics (Second Edition)* (pp. 60-66). Elsevier. https://doi.org/10.1016/B0-08-044854-2/00798-7

Ping, R. M., Goldin-Meadow, S., & Beilock, S. L. (2013). Understanding gesture: Is the listener's motor system involved? *Journal of Experimental Psychology: General, 143*(1), 195–204. https://doi.org/10.1037/a0032246

Poggi, I. (2008). Iconicity in different types of gestures. *Gesture, 8*(1), 45–61. https://doi.org/10.1075/gest.8.1.05pog

Priesters, M. A., & Mittelberg, I. (2013). Individual differences in speakers' gesture spaces: Multi angle views from a motion capture study. *TiGeR Workshop*, Tilburg, NL. https://tiger.uvt.nl/pdf/papers/priesters.pdf

Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H. M., Buschmeier, H., Esposito, E., Grimminger, A., Hammer, B., Häb-Umbach, R., Horwath, I., Hüllermeier, E., Kern, F., Kopp, S., Thommes, K., Ngonga Ngomo, A.-C., Schulte, C., Wachsmuth, H., Wagner, P., & Wrede, B. (2021). Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems. *IEEE Transactions on Cognitive and Developmental Systems, 13*(3), 717-728. https://doi.org/10.1109/TCDS.2020.3044366

Rohrer, P. L., Delais-Roussarie, E. & Prieto, P. (2020). Beat Gestures for Comprehension and Recall: Differential Effects of Language Learners and Native Listeners. *Frontiers in Psychology, 11*, Article: 575929. https://doi.org/10.3389/fpsyg.2020.575929

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. http://www.rstudio.com/.

Stojnic, U., Stone, M., & Lepore, E. (2013). Deixis (even without pointing). *Philosophical Perspectives, 27*(1), 502–525. https://doi.org/10.1111/phpe.12033

Streeck, J. (2008). Depicting by gesture. *Gesture, 8*(3), 285-301. https://doi.org/10.1075/gest.8.3.02str

Türk, O., Wagner, P., Buschmeier, H., Grimminger, A., Wang, Y., & Lazarov, S. (2023). MUNDEX: A multimodal corpus for the study of the understanding of explanations. In P. Paggio & P. Prieto (Eds.), *Book of Abstracts of the 1st International Multimodal Communication Symposium* (pp. 63–64).

Vendler, Z. (1994). Understanding Misunderstanding. In D. Jamieson (Ed.), *Language, Mind, and Art* (pp. 9-22). Springer. https://doi.org/10.1007/978-94-015-8313-8_2

Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics, 32*(8), 1177–1207. https://doi.org/10.1016/s0378-2166(99)00109-5

West, D. E. (2014). *Deictic Imaginings: Semiosis at Work and at Play*. Springer. https://doi.org/10.1007/978-3-642-39443-0

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting Chicago Linguistic Society*, April 16-18, 1970, Chicago Linguistic Society, Chicago (pp. 567-578).

Verbal signals of understanding do not predict a decrease of gesture deixis

Stefan Lazarov[1], Angela Grimminger[1]
Faculty of Arts and Humanities, TRR-318 "Constructing Explainability"
[1]*Paderborn University*
stefan.lazarov@uni-paderborn.de

In explanatory dialogues, a more experienced interlocutor (explainer, henceforth EX) aims at increasing the understanding of a less experienced interlocutor (explainee, henceforth EE) about an entity or a process (i.e., explanandum) via co-constructions and scaffolding [1]. There are situations in which the explanandum is absent from the shared referential space between the EX and the EE, and EXs need to provide EEs with additional spatial orientation by using co-speech gestures indicating certain locations or the shape of invisible objects [2, 3].

In the present study, we investigated the relation between the EXs' gesture deixis and EEs' verbal signals of understanding in dyadic explanations of a board game, in a sample of 5 German-speaking adult EXs, each explaining a board game to 3 different adult EEs individually, resulting in 15 explanatory dialogues. The analyzed explanations are constituted by three phases (game absent, game present and game play) [4]. Initial observations of the video data indicated an increased gestural behavior by the different EXs during the game absent phase (i.e., the explanandum is not visible) compared to the other phases; therefore, only this phase was analyzed here. Also based on initial observations of the video data, we followed McNeill's multidimensional view on gestures, including the dimension of deixis [5].

Our research question and hypothesis are motivated by previous research: In general, it was shown that co-speech gestures enhance addressees' understanding [6]. More specifically, speakers' deictic and iconic gesture rates were found to decrease significantly after addresses' feedback of understanding [7]. Further, it was reported that teachers' deictic and iconic gestures increase after detecting spots of students non-understanding [8]. Based on this, we hypothesized that EXs' gesture deixis would decrease after EEs' verbal signals of understanding and increase after EEs' verbal signals of partial and non-understanding.

EEs' verbal utterances were coded in relation to EEs' understanding (e.g., backchannels *ok, yes, alright,* and also repetitions of EXs' utterances), partial understanding (e.g., polar and tag questions), and non-understanding (e.g., open questions, corrections), based on a discourse annotation scheme (k = 0.89). EXs' gesture phrases were coded based on the occurrence of gesture strokes and with respect to the dimension of gesture deixis, being observed in deictic, deictic-iconic, or deictic-beat gestures (k = 0.94). To incorporate the study design of 1 EX interacting with 3 different EEs and considering a non-normal data distribution, we conducted a Generalized Linear Mixed Effects Regression analyzing EXs' raw frequencies of gesture deixis during the game absent phase.

The results (Tab.1) indicate a significant effect of the three levels of understanding signaled by EEs on the frequencies of EXs' gesture deixis following these signals. Post-hoc comparisons (Tab. 2) reveal that the frequency of EXs' gesture deixis after EEs' signals of understanding is significantly higher than after EEs' signals of partial and non-understanding (Fig. 1). Contrary to our hypothesis, our findings are not in line with previous research. One possible reason for the high frequencies of EXs' gesture deixis after EEs' signals of understanding could be related to the existing knowledge gap between the more experienced EXs and the novice EEs, who were not familiar with the physical appearance of the game components and their placement on the shared referential space. Another related reason could be that the EXs may have noticed a continuous high demand for spatial orientation on the invisible shared referential space in EEs' (non-)verbal behavior during the game absent phase.

EX	gesture deixis after:	*M*	*SD*	*β*	*SE*	*z*	*p*
EE	understanding (Int.)	87.47	41.44	4.37	0.15	28.36	< 0.001
	partial understanding	21.13	23.10	-1.42	0.16	-22.74	< 0.001
	non-understanding	5.00	13.37	-2.86	0.12	-24.15	< 0.001

Table 1: *A summary of descriptive statistics and fixed effects.*

pairwise comparison:	*β*	*SE*	*z*	*p*
understanding – partial understanding	1.42	0.06	22.74	< 0.001
understanding – non-understanding	2.86	0.12	24.15	< 0.001
partial-understanding – non-understanding	1.44	0.13	11.25	< 0.001

Table 2: *Post-hoc pairwise comparisons*

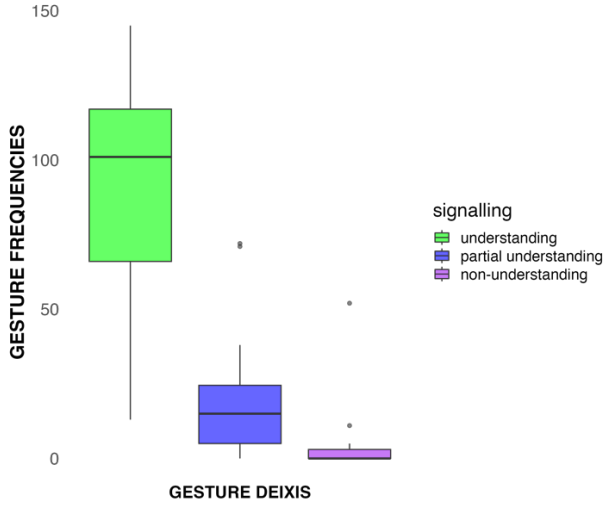

Figure 1: *EXs' gesture deixis related to levels of EEs' understanding.*

References

[1] K. J. Rohlfing et al., "Explanation as a social practice: toward a conceptual framework for the social design of AI systems," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 3, pp. 717–728, Sep. 2021, doi: 10.1109/tcds.2020.3044366.

[2] H. H. Clark, "Pointing and placing," in *Pointing: Where language, culture, and cognition meet*, S. Kita, Ed. Lawrence Erlbaum, 2003, pp. 251–276. doi: 10.4324/9781410607744-14.

[3] N. McKern, N. Dargue, N. Sweller, K. Sekine, and E. Austin, "Lending a hand to storytelling: Gesture's effects on narrative comprehension moderated by task difficulty and cognitive ability," Quarterly Journal of Experimental Psychology, vol. 74, no. 10, pp. 1791-1805, Jun. 2021, doi:10.1177/17470218211024913.

[4] O. Türk., P. Wagner, H. Buschmeier, A. Grimminger, Y. Wang, and S. Lazarov, "MUNDEX: A multimodal corpus for the study of the understanding of explanations" in *Book of Abstracts of the 1st International Multimodal Communication Symposium*, P. Paggio and P. Prieto, Eds., 2023, pp. 63–64.

[5] D. McNeill, "Gesture and communication," in *Encyclopedia of Language & Linguistics*, K. Brown, Ed. Elsevier, 2006, pp. 58–66. doi: 10.1016/b0-08-044854-2/00798-7.

[6] S. D. Kelly, A. Özyürek, and E. Maris, "Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension," *Psychological Science*, vol. 21, no. 2, pp. 260–267, Dec. 2010, doi: 10.1177/0956797609357327.

[7] J. Holler and K. Wilkin, "An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses," *Journal of Pragmatics*, vol. 43, no. 14, pp. 3522–3536, Nov. 2011, doi: 10.1016/j.pragma.2011.08.002.

[8] M. W. Alibali, M. J. Nathan, R. B. Church, M. S. Wolfgram, S. Kim and E. J. Knuth, "Teachers' gestures and speech in mathematics lessons: forging common ground by resolving trouble spots", *ZDM Mathematics Education*, vol. 45, pp. 425-440, Jan 2013, doi: 10.1007/s11858-012-0476-0.