Lexical Event Models for Multimodal Dialogues

James Pustejovsky(✉) and Yifan Zhu

Brandeis University, Waltham, MA 02453, USA
jamesp@brandeis.edu

Abstract. In order to understand multimodal interactions between humans or humans and machine, it is minimally necessary to identify the content of the agents' communicative acts in the dialogue. This can involve either overt linguistic expressions (speech or writing), content-bearing gesture, or the integration of both. But this content must be interpreted relative to a deeper understanding of an agent's Theory of Mind (one's mental state, desires, and intentions) in the context of the dialogue as it dynamically unfolds. This, in turn, can require identifying and tracking nonverbal behaviors, such as gaze, body posture, facial expressions, and actions, all of which contribute to understanding how expressions are contextualized in the dialogue, and interpreted relative to the epistemic attitudes of each agent. In this paper, we adopt Generative Lexicon's approach to event structure to provide a lexical semantics for ontic and epistemic actions as used in Bolander's interpretation of Dynamic Epistemic Logic, called *Lexical Event Modeling (LEM)*. This allows for the compositional construction of epistemic models of a dialogue state. We demonstrate how veridical and false belief scenarios are treated compositionally within this model.

Keywords: Theory of Mind · HCI · Epistemic Updating · Common ground tracking · multimodal dialogue · Generative Lexicon · Event Semantics

1 Introduction

With the introduction of large language models (LLMs) in the user experience for dialogue-based search and QA within HCI, much recent research has focused on aspects of Dialogue State Tracking (DST), the ability to identify and update the user's needs at each stage in the interaction, by taking into account the past dialogue moves and history. Such interactions are largely unimodal, characterized by linguistic queries and prompts from the human and linguistic responses by the system. Discourse policies for appropriateness or correctness of the system response can be arrived at by modeling such unimodal interactions, given the constrained nature of the context of the dialogue. Hence, most papers benchmarking the performance of dialogue models are often biased towards reflecting such interactions [9,19,26].

M. Kurosu et al. (Eds.): HCII 2024, LNCS 15374, pp. 174–192, 2025.
https://doi.org/10.1007/978-3-031-76803-3_10

When we move into the area of multimodal HCI or HRI dialogues, where information is conveyed through language, gesture, visual cues, and situated reference, interpretation becomes much more difficult [25]. Further, if we attempt to extend such interactions to model dialogues with multiple participants, we need to track not only the dialogue state, but also the *epistemic state* of each participant as well as the common ground of the entire group, as it develops during the dialogue [5]. This involves identifying the beliefs, desires, and intentions (Theory of Mind) for each actor in the interaction, as well as each actor's attitudes towards the other participants. These are constructed from not only the linguistic expressions uttered by each speaker, but from other communicative modalities, such as content-bearing gesture, as well as nonverbal behaviors, such as gaze, body posture, facial expressions, and actions, all of which contribute to understanding how expressions are contextualized in the dialogue, and interpreted relative to the epistemic attitudes of each agent.

In order to account for such representations, Dynamic Epistemic Logic has recently been implemented in the context of HCI and HRI to identify shared and divergent beliefs between participants [5,17]. For example, [5] demonstrates how epistemic updating and false beliefs can be modeled in an HRI task, illustrating the alignment of diverse modalities for determining belief states.

However, in multimodal dialogues, one of the major difficulties is determining how to compositionally construct epistemic models for the participants. There are three main dimensions of knowledge that need to be accounted for in such situations:

- **Language and gesture**: the different sources for the information that is announced or introduced into the context;
- **Gaze, posture, facial expressions**: nonverbal behaviors that indicate attention, co-attention, engagement, boredom, other emotional states;
- **Actions and objects in the world**: physical events that occur with the objects in the context.

The challenge for multimodal dialogue understanding is to determine how to compositionally construct epistemic models with these diverse sources of knowledge. More concretely, the question is how such representationally diverse sources are integrated, aligned, and harmonized into an operational form that fits within the mechanisms of Dynamic Epistemic Logic (DEL).

Given this challenge, in this paper, we study the creation of common ground in multimodal task-oriented interactions, in order to develop computational strategies and their models for representing and updating epistemic states. Our investigation involved studying the multimodal dialogue between a triad of co-situated students collaborating to solve a weights task for five blocks, using only a balance scale. The task is particularly suited for our purpose, because the participants naturally engage in the different modalities that are so crucial for understanding multimodal HCI: namely, speech, gesture, gaze, and of course joint actions. From the perspectives of both dialogue state modeling as well as common ground updating, there are several distinct action types and their effects that need to be identified and tracked:

(1) a. **Ontic actions**; interactions with and movements of the objects in the shared space; i.e., blocks and the balance scale;
b. **Epistemic actions**; changes to the epistemic state of one or more of the participants in the interaction.

We assume the architecture of the Common Ground Tracking model developed in [22,38,52], where an Evidence-based Dynamic Epistemic Logic is deployed to track common ground in a shared task. This involves two steps: applying recognition algorithms over each modality: speech [48], gesture [8], gaze detection [29], and action recognition [44]; aligning and interpreting the model results to determine common ground for the group [22].

In this paper, we extend this model by providing a compositional strategy for interpretation of each agent's epistemic state, given the current context. We present an extension of Bolander's model of DEL [5] adapted to multi-party dialogues, involving task-oriented interactions using multiple modalities. This approach, called *Lexical Event Modeling (LEM)*, enables the compositional construction of an epistemic model for a dialogue state as well as the updating to next state, given new information. The overall goal is to identify the epistemic content in situated dialogues, by interpreting the verbal and nonverbal behaviors of each agent, as well as referenced objects and situational relations in the context.

We proceed as follows. We examine the contribution of four distinct modalities (speech, gesture, action, and visual attention) to the information in a dialogue state and the associated update operations for determining epistemic content and common ground. In our model (as in the underlying corpus), each channel is encoded as an AMR-like representation, including: S-AMR for spoken language AMR [2]; GAMR for gesture [8]; Act-AMR for actions and events [44]; and the perception verb subset of PropBank from Act-AMR for attention. The predicative core for each AMR is based on the lexical resource, VerbNet-GL [47], which incorporates Generative Lexicon's dynamic event structure, distinguishing an event's *pre-state* and *post-state*, and the *program* mapping between them.

We distinguish verbal predicates as denoting either public or private events, and then provide an appropriate *epistemic framing* for the verb semantics, based on the role that belief, knowledge, doubt, or perception, plays relative to carrying out or performing this event. Two kinds of epistemic framing are identified: lexical and contextual. For example, an agent performing an action will believe (or know) that they are engaged in the act. Hence, for any verb containing an AGENT participant role, we presuppose an epistemic frame of belief towards that act. Similarly, in a dialogue, we contextualize a speech act as introducing an epistemic frame of belief (or the appropriate epistemic attitude) toward the proposition being uttered. A *lexical event model* will be identified as that component of the resulting Kripke structure, derived from an epistemic framing operation.

To illustrate how epistemic framing is interpreted from distinct modal descriptions, consider the dialogue state shown in Fig. 1, from one of the Weights Task Dataset videos [21]. In this scene, the participants in the image are denoted

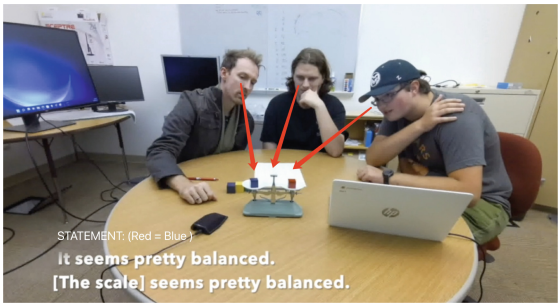


Fig. 1. Example of a multimodal interaction

as p_1, p_2, and p_3 from left to right, respectively. Participant p_1 says of the scale, "It seems pretty balanced." Let us refer to this as b. At the same time, both p_2 and p_3 are visually attending to the situation which b refers to.

This scene shows information conveyed through two modalities (speech and vision), each of which has epistemic consequences. For example, the utterance introduces a public announcement of proposition b, which presupposes p_1's belief in b, and the audience's belief that p_1 believes b. Similarly, the visual perception of b for the other two participants presupposes that they each either believe or at least have evidence for b. We will demonstrate how this information is contributed compositionally from the lexical event models for each modality.

2 Related Work

This work draws on research on common ground and Theory of Mind, multimodal HCI, Dialogue State Tracking, and the role of gesture in multimodal interactions. Multimodal dialogue involves having a shared understanding of both utterance meaning (content) and the speaker's meaning in a specific context (intent), involves the ability to link these two in the act of situationally grounding meaning to the local context, what is typically referred to as "establishing the common ground" between speakers [1,10,15,42,46]. The concept of common ground refers to the set of shared beliefs among participants in a Human-Human interaction (HHI) [16,28,45], as well as HCI [23,31] and HRI interactions [13,24,41].

The role of nonverbal behavior in multimodal communication has recently taken on new interest within CL and the broader AI community. Gesture AMR (GAMR) [8] considers gestures that convey the same propositional content and intentionality as speech acts. Gesture may have meaning on its own, or it may enhance the meaning provided by the verbal modality [14]. Also critical to multimodal dialogue is human action, which in addition to communicating deictic and bridging information can also make lasting changes to the world, affecting the common ground [44]. Additionaly, gaze as a non-verbal behavior, also serve a important role in communicating intent [20,29].

Dynamic Epistemic Logic (DEL) has been used extensively to model the manner in which epistemic state among agents is updated in dialogue, with the introduction of dynamic operators to represent changes in knowledge and beliefs resulting from informational events [33,49]. Two notable variants of DEL, formulated by Pacuit [3,4,32] and Bolander [5,6], differ in their treatment of information updates and underlying semantics. While both methodologies aim to capture how agents modify their epistemic states upon acquiring new information, they employ distinct frameworks and principles to achieve this objective.

Theory of Mind has also been encoded within the DEL framework as developed by Bolander [5] to tackle the problem of false belief. This framework tackles the epistemic perspectives held by multiple participants concerning the ongoing actions within the interaction. This model formalizes an agent's erroneous belief concerning a dynamic environment, as well as the capacity of other agents to identify and deliberate upon this agent's inaccurate epistemic condition. However, this necessitates the incorporation of linguistic resources to account for the lexical semantics of events within a dynamic epistemic model, elucidating how agents perceive and assimilate information as events transpire throughout a discourse.

3 Experimental Domains

3.1 The Sally-Anne Narrative

The Sally-Anne narrative is a classic psychological tool used to investigate the understanding of false beliefs in children, particularly in the context of Theory of Mind development [51]. The story involves two characters, Sally and Anne. Sally has a basket, while Anne has a box. Sally first places a marble into her basket, then leaves the scene. While Sally is away, Anne moves the marble from the basket to her own box. Sally finally returns to the scene. From the Theory of Mind perspective, the question asked to the child is: "Where will Sally look for her marble?"

More formally, the Sally and Anne narrative encompasses a sequential series of five steps, as in 2, representing five distinct situations. The initial scenario involves Sally and Anne, where Sally possesses a basket and a marble, and both participants are aware of these objects. Additionally, Anne possesses a box, which is also observed by both individuals. Subsequently, in the second scenario, Sally proceeds to place the marble inside her basket, with both participants witnessing this action. The third situation involves Sally's departure, which is observed by Anne. Moving on to the fourth scenario, Anne proceeds to remove the marble from the basket and transfers it into the box. However, it is important to note that Sally does not perceive this action. Finally, in the fifth and final scenario, Sally returns, which is acknowledged by Anne's observation (Fig. 2).

3.2 The Weights Task Dataset

The Weights Task, explored in [21], entails collaborative problem-solving task among groups of three participants. Participants are provided with the weight

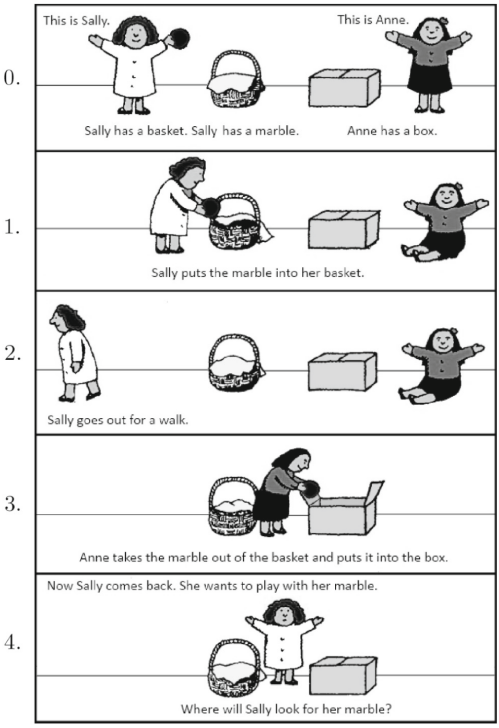


Fig. 2. Sally-Anne Experiment

of one block and are tasked with determining the weights of the remaining blocks and identifying the algebraic relationship between them (the Fibonacci Sequence). Given the task's context-dependent nature involving physical objects and reasoning about their properties, the communication can be annotated in several ways: speech with dense paraphrasing [48], gesture [8], action [44], non-verbal behaviors such as gaze [29] and body postures [40] as well as collaborative problem-solving (CPS) indicators following the framework of [43].

4 Dynamic Epistemic Logic

Dynamic Epistemic Logic (DEL) is an extension of classical epistemic logic that integrates dynamic operators to represent knowledge and belief changes resulting from information events. Two prominent variants of DEL, proposed by Pacuit and Bolander, diverge in their treatment of information updates and the underlying semantics. While both approaches strive to capture how agents modify their epistemic states upon receiving new information, they employ distinct structures and principles to achieve this objective.

Pacuit's approach to DEL introduces neighborhood models which employs a set of possible worlds (neighborhoods). In Pacuit's DEL framework, updates to

an agent's epistemic state are captured through evidence models, which represent the information contained in an update and is used to modify the neighborhood models, resulting in an updated representation of the agent's epistemic state. The dynamic semantics of Pacuit's DEL are established by applying evidence models to neighborhood models, thereby transforming the neighborhoods to reflect the new information. This process involves mechanism to filter or adjust the neighborhoods based on the compatibility of the evidence with the agent's prior epistemic state. This model emphasizes the compatibility of new evidence and existing beliefs [3,4,32].

Bolander's approach to DEL differs from Pacuit's by predominantly employing traditional Kripke models that utilize possible worlds and accessibility relations to represent an agent's knowledge or beliefs. The update process in Bolander's DEL is performed by taking the product of the current epistemic model (Kripke model) with the action or event model, which results in a new Kripke model reflecting the updated epistemic state. The semantics of updates in Bolander's DEL is grounded in the transformation of Kripke models through the application of action or event models. This involves modifying the accessibility relations between possible worlds based on the information conveyed by the action or event. This model focuses on how events alter accessible worlds and relations [5,6].

Bolander's Evidence-based Dynamic Epistemic Logic (EB-DEL) in [5] provides a systematic framework for formalizing the understanding of Theory of Mind, elucidating how individuals interpret and attribute mental states to other agents. Within the DEL framework, states denote the epistemic updates occurring within agents, both within the actual world and within potential alternative realities accessible from the actual world. Additionally, an event model delineates the actions that trigger such epistemic change. The event model is defined as $\mathcal{E}$=$(E, Q, pre, post)$, which includes preconditions and postconditions to illustrate the event updates, where

- The domain, denoted as $\mathcal{E}$, is a finite non-empty set comprising events.
- The accessibility relation $\mathcal{Q}$ is a mapping from agents in $\mathcal{A}$ to subsets of event pairs in $E \times E$.
- Each event in E is associated with a precondition, denoted as *pre*, which can be any formula in the language $\mathcal{L}(P, \mathcal{A})$.
- Each event in E is assigned a postcondition, referred to as *post*. Postconditions are expressed as conjunctions of propositional literals, representing atomic propositions and their negations, including the constants $\top$ and $\bot$.

The above graph depicts an event in the context of the Sally-Anne experiment, which describes Anne's action of moving the marble from the basket to the box, shown in Fig. 3. The event is labeled as $\langle \top, \neg t \wedge x \rangle$, indicating that the precondition of the event is trivial, and the postcondition specifies that the marble is in the basket while not being in the box. Furthermore, the event is only accessible to Anne in the actual world (marked with $\odot$), as indicated by an edge labeled with the name of the relevant accessibility agent, A.

5 Dynamic Event Structure

In this section, we explore how richer models of event semantics as developed in lexical semantics can be adapted and integrated into DEL and the notion of event model as discussed above. For this purpose, we adopt the view of event structure as first developed within Generative Lexicon Theory [35,36] and in terms of a dynamic event semantics, Dynamic Interval Temporal Logic (DITL) [27,39].

Anne takes the marble out of the basket and puts it into the box.

$\hat{e}_i : \langle t \wedge \neg x, \neg t \wedge x \rangle$

Fig. 3. Top: Anne's action of moving the marble from the basket to the box. Bottom: Event model for this action.

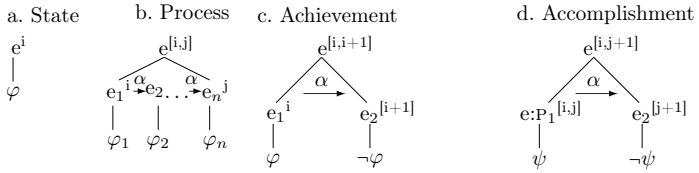
On this view, activities and events are interpreted as programs, π, transitioning between states or situations. A formula is interpreted as a propositional expression, with assignment of a truth value in a specific state in the model. For our purposes, a state is a set of propositions with assignments to variables at a specific time index. Atomic programs are relations from states to states, and hence interpreted over an input/output state-state pairing (cf. also [12,30]).

Following [39], we adopt the language of PDL to express an event structure enriched to dynamically track object attributes modified in the course of the event. All the events are represented as a sequence of states related by functions (programs) which go from state to state. The definitions of conventional Aktionsarten can be given as follows [11,34,50]:

(2) a. **State**: φ – *happy*, *tall*, *closed*, *in a box*.
b. **Process**: $\alpha; \alpha^*$ –*run*, *push*, *move*.
c. **Achievement**: $?\neg\varphi; \alpha; ?\varphi$ – *die*, *open*, *close*
d.**Accomplishment**: $(?\neg\varphi; \alpha)^+; ?\varphi$ – *build*, *put*, *write*.

Specifically, consider the dynamic event structures below. The structure in (a) below represents a **state**, e^i, at time i, with the propositional content, φ. The event structure in (c) illustrates how program α takes the world from e^i with content φ, to the adjacent state, e_2^{i+1}, where the propositional content has been negated, $\neg\varphi$. This corresponds directly to **achievements**. From these two types, the other two Vendlerian classes can be generated. **Processes** can be modeled

as an iteration of simple transitions, where two conditions hold: the transition is a change in the value of an identifiable attribute of the object; every iterated transition shares the same attribute being changed. This is illustrated in (b) below. Finally, **accomplishments** are built up by taking an underlying process event, e:P, denoting some change in an object's attribute, and synchronizing it with an achievement (simple transition): that is, e:P is unfolding while ψ is true, until one last step of the program α makes it the case that $\neg\psi$ is now true.



Of particular relevance to our present discussion, is GL's notion of opposition structure [36], and subsequent representations as pre-state and post-state conditions on event structures [18]. Simplified below, the transition from an initiating propositional content φ to its opposition $\neg\varphi$ is brought about by a program denoting the action inherent in the event.

Following the GL lexical representation for verbs presented in [7,37], we illustrate this with a simple transition predicate, such as *open*, (in "The door opened."), as shown below, where the CONST qualia role consists of *pre-state* and *post-state* components, encoding the opposition structure inherent in the change.

$$(3)\quad \lambda x \left[\begin{array}{l} \textbf{open} \\ \text{ARGSTR} = \left[\text{A1} = x : \textbf{phys}\right] \\ \text{EVENTSTR} = \left[\begin{array}{l} \text{E1} = \textbf{e}_1\textbf{:state} \\ \text{E2} = \textbf{e}_2\textbf{state} \\ \text{P1} = \textbf{p}_1\textbf{:program} \end{array}\right] \\ \text{QUALIA} = \left[\begin{array}{l} \text{CONST} = \left[\begin{array}{l} \text{PRE} = \neg open(e_1, x)) \\ \text{POST} = open(e_2, x) \end{array}\right] \\ \text{FORMAL} = \textbf{simple\_transition} \\ \text{AGENTIVE} = nil \end{array}\right] \end{array}\right]$$

To illustrate the dynamic encoding of state and action information in a DES representation, consider the lexical semantics for the accomplishment verb *put*, shown below.

$$(4)\quad \lambda z \lambda y \lambda x \left[\begin{array}{l} \textbf{put} \\ \text{ARGSTR} = \left[\begin{array}{l} \text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:physobj} \\ \text{A}_3 = \textbf{z:location} \end{array}\right] \\ \text{EVENTSTR} = \left[\begin{array}{l} \text{E1} = \textbf{e}_1\textbf{:state} \\ \text{E2} = \textbf{e}_2\textbf{state} \\ \text{P1} = \textbf{p}_1\textbf{:program} \end{array}\right] \\ \text{QUALIA} = \left[\begin{array}{l} \text{CONST} = \left[\begin{array}{l} \text{PRE} = \neg at(e_1, y, z)) \\ \text{POST} = at(e_2, y, z) \end{array}\right] \\ \text{FORMAL} = \textbf{accomp\_transition} \\ \text{AGENTIVE} = move(p_1, x, y) \end{array}\right] \end{array}\right]$$

In the next section, we show how the decompositional structure inherent in GL's event structure can be adapted to the event models as deployed in DEL's Kripke structures for agent epistemic modeling.

6 Lexical Event Models

In this section, we extend DEL's definition of event model to accommodate the event semantic information associated with specific predicates in the language. This is necessary if we are to compositionally create epistemic models for dialogue states, using data generated through automatic NLP and vision processing algorithms. This will involve two enhancements, described below:

- We retrieve the specific pre-state and post-state information for the verbal predicate associated with any action that has been recognized (annotated) within a dialogue state.
- We create an *epistemic framing* of an event or action, that can be lexically associated with a verbal predicate, encoded as part of a lexical resource.

A *lexical event model (LEM)* will be identified as that component of the resulting Kripke structure, derived from an epistemic framing operation.

Let's unpack each of these steps. The first step entails merely accessing the specific propositional content inherent in the oppposition structure for an action verb, as interpreted through the composition with its arguments. For example, consider the first agentive event in the Sally-Anne narrative, annotated as *Sally puts a marble in the basket.* Given the lexical semantics for the transition verb *put* shown in (4), the propositional content of *pre-state* and *post-state* after argument binding will be as follows (where m is marble and t is basket).

(5) a. *pre-state*: $\neg in(m, t)$
b. *post-state*: $in(m, t)$

Now consider the introduction of the epistemic framing for an event. We wish to position an ontic action from the perspective of the participants who are present during the event. We first distinguish verbal predicates as denoting either *public* or *private* events. While attitudes and beliefs are private to an agent, most actions performed by an agent are potentially public or witnessed by others. In this sense, they are *self-announcing*, known at least to the agent performing them.

Given this distinction, we define the *epistemic framing* for an event as a modal subordination of the event from the perspective (accessibility relations) of any cognitive participant in the event, in particular the agent. That is, we can recover the epistemic attitude based on the role that belief, knowledge, doubt, or perception, plays relative to carrying out or performing a particular event.

For any public agentive event, $\hat{e}_i$, where $V(e_i, \text{AG}, \ldots)$, we introduce a default "audience" role, AU. This will be the "other agent(s)" in an epistemic embedding: $V(e_i, \text{AG}, \ldots, \text{AU})$. Hence, we now have the following enriched lexical semantics for a public agentive event, such as *put*.

(6) a. $\lambda z \lambda y \lambda x \lambda e[put(e, x{:}\text{AG}, y{:}\text{TH}, z{:}\text{AU})]$

We can now introduce the epistemic framing for a public agentive (PA) event as follows (where $\hat{e}$ denotes the propositional content that event e occurs):

(7) a. Agentive Awareness: $\forall e[PA(e) \rightarrow K_{ag}\hat{e}$
b. Audience Witness: $\forall e[PA(e) \rightarrow B_{au}\hat{e}$

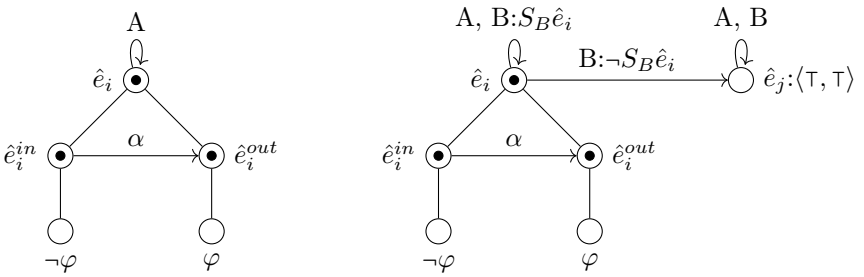
Similarly, in the context of a dialogue, we interpret a speech act as a public agentive event: hence it will fall under the application of epistemic framing, both for agent and audience. In particular, an agent performing a speech act, $V(e_{sa}, a, \varphi)$, will introduce an epistemic frame of belief (or the appropriate epistemic attitude) toward the proposition φ being uttered. For example, an agent a stating (a) will generate the epistemic frame shown in (b).

(8) a. The red block is the same weight as the blue block.
b. B_a *r=b.*

Given this discussion, we now define the concept of *Lexical Event Modeling.* Adapting Bolander [5] we assume a lexical event model of $\mathcal{L}(P, \mathcal{A})$ is $\mathcal{E}$ = *(E,Q,pre,post,in,out,ag,au)*, where

- The domain, denoted as $\mathcal{E}$, represents a finite non-empty set of events;
- The function $\mathcal{Q}$: $\mathcal{A} \rightarrow 2^{E \times E}$ assigns an "accessibility relation" $\mathcal{Q}(i)$ to each agent i in the set $\mathcal{A}$.
- The mapping *pre*: $E \rightarrow \mathcal{L}(P, \mathcal{A})$ associates a "precondition" with each event in E. The precondition can be formulated as any logical formula belonging to the language $\mathcal{L}(P, \mathcal{A})$.
- The mapping *post*: $E \rightarrow \mathcal{L}(P, \mathcal{A})$ assigns a "postcondition" to each event. Postconditions are expressed as conjunctions of propositional literals, specifically combinations of atomic propositions and their negations, which may include the logical constants $\top$ and $\bot$.
- *in*: $E \rightarrow \mathcal{L}(P, \mathcal{A})$ designates an "event input" for each event, indicating the input associated with the lexical event
- *out*: $E \rightarrow \mathcal{L}(P, \mathcal{A})$ assigns an "event output" to each event, representing the output produced by the lexical event.
- *ag*: for a public agentive event, denoted as e, for each $a \in \mathcal{A}$, a will possess the knowledge $K_a\hat{e}$.
- *au*: for a public agentive event e, for each $u \in \mathcal{A}$, u will hold the belief $B_u\hat{e}$.

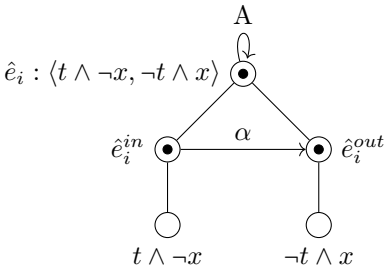
As in [5], $\mathcal{Q}_i$ will be used to denote the relation $\mathcal{Q}(i)$. Within the context of the framework, when considering an element e belonging to the set E, the pair $(\mathcal{E}, e)$ is denoted as an "action" (or "pointed event model") of $\mathcal{L}(P, \mathcal{A})$. In this context, e is specifically referred to as the "actual lexical event". The two diagrams below illustrate the basic structures of a Lexical Event Model.



The graph on the left represents the lexical event model associated with an achievement verbal predicate, such as *close* or *die*, with opposition structure and transition program α. In this world, the audience sees the event e_i and the event is achieved.

The graph on the right is the same lexicon event mode with the same agent and audience. However, it introduces a conditional element: that is, if the audience does not see the event, then the agent and the audience will be in a world which is non-veridical, and in that world, event e_i has never been achieved.

Now let's see how lexical event models provide richer propositional content to a standard event model. Consider again the event depicted in Fig. 3. The action, *transfer* is implied by the disparity between the precondition and the postcondition, but no explicit content is provided. The propositional content of the pre- and post-states for the *transfer* act are derived from the lexical event semantics associated with the verb. To illustrate this, consider the diagram below, where the same event of Anne's action of transferring the marble from the basket to the box is viewed as a Lexical Event Model.

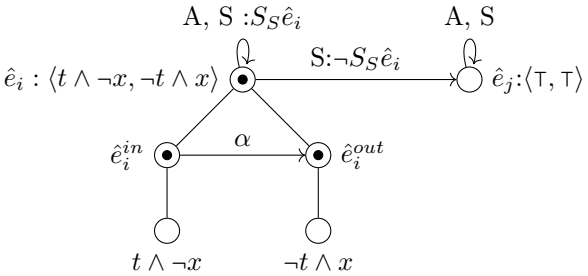


In this diagram, a lexical event has a pre-state and post-state designated, and a program α, carried out by an agent, A. The graph employs event input (designated as $\hat{e}_i^{in}$) and event output (referred to as $\hat{e}_i^{out}$) to denote the precondition and postcondition of an ontic alteration, where $t \wedge \neg x$ signifies "marble is in the basket and not in the box" and $\neg t \wedge x$ indicates "marble is not in the basket and in the box". Moreover, the manifestation of a transition is depicted by an edge labeled with the event α, which is *transfer* in this example. The relevant agent is identified by an edge labeled with their name, designated as A within this event.

Now consider the situated epistemic model, where we capture the visual attention or absence of attention by another agent, in this case Sally. We gloss this modal act below in (a) along with its modal form in (b).[1]

(9) a. Sally did not see the act of transfer.
b. $\neg S_S\varphi$

We compose the epistemic framing associated with this and get the following compositional event model:



Hence, the epistemic consequences of the Lexical Event Modeling strategy introduced here can be summarized as follows (where the modal B_a represents belief of an agent a, DO is an action, and SA is a speech act):

(10) a. **Acting is Believing:** $DO_a\varphi \rightarrow B_a\varphi$ (you believe your own actions) As an agent participant in an event, you believe it has happened.
b. **Saying is Believing:** $SA_a\varphi \rightarrow B_a\varphi$ (you believe what you say) As actor of a declarative speech act, you believe the proposition you express.
c. **Seeing is Believing:** $S_a\varphi \rightarrow B_a\varphi$ (you believe what you see) As witness to a situation or event, you believe it to have occurred.

7 Constructing Epistemic Models from Annotations

In order to demonstrate the use and composition of lexical event models, we consider a case of the Weight Task with false belief as described by the following procedural steps, shown in Figs. 4, 5 and 6:

1. Both participants, p_1 and p_2, as well as p_3, possess knowledge regarding the weight of the red block, which is determined to be 10 g. p_1 places the green block on the left scale, while positioning the red and blue blocks on the opposite side of the scale. Throughout this process, p_1 and p_2 focus their attention on the scale and the blocks, whereas p_3 directs his attention towards the laptop.
2. Subsequently, p_2 asserts that the scale is balanced. Both p_1 and p_2 continue to observe the scale and the blocks, while p_3 maintains his attention on the laptop.

[1] We normalize the distinction between knowledge and belief so that we maintain a KD45 logic.

Fig. 4. Participants get information about the green block. (Color figure online)

Fig. 5. Participants get information about the blue block. (Color figure online)

Fig. 6. Participants get information about the purple block.

3. At this stage, p_1 points at the blocks and the scales, drawing $p_2's$ attention towards him. In response, p_2 places the blue block on the left scale, and positions the red block on the opposing side. Throughout this process, all participants (p_1, p_2, and p_3) direct their attention towards the scale and the blocks.
4. Following the arrangement, p_3 declares that the scale is balanced. Consequently, all participants (p_1, p_2, and p_3) continue to focus their attention on the scale and the blocks.
5. In the subsequent step, p_1 places the purple block on the left scale, while positioning the red and green blocks on the opposite side. Throughout this process, p_1, p_2, and p_3 concentrate their attention on the scale.
6. Finally, p_1 claims that the weight of the purple block is 30 g, while p_3 disputes this claim, stating that the purple block does not weigh 30 g.

In this example, p_3 holds a false belief regarding the purple block due to his lack of awareness of the actions performed by participant p_1 in the first step and the balanced state of the scale in the second step. In the first step, only participants p_1 and p_2 observe the actions performed by p_1. Consequently, the visual annotations for this scenario indicate that p_1 and p_2 see the action ($S_{p_1,p_2}\hat{e}_i$), while p_3 does not ($S_{p_3}\neg\hat{e}_i$), where $\hat{e}_i$ represents the action performed by p_1. The corresponding action annotations for this scenario can be represented as follows:

```
(p / put-01 :ARG0 (p / p1) :ARG1 (g / green block) :ARG2 (l / left scale))
(p / put-01 :ARG0 (p / p1) :ARG1 (r / red block) :ARG2 (r1 / right scale))
(p / put-01 :ARG0 (p / p1) :ARG1 (b / blue block) :ARG2 (r / right scale))
```

In the second scenario, participant p_2 performs a speech act, indicating that the scale is balanced. Both p_1 and p_2 observe this balance, while p_3 does not. Therefore, the gaze annotations for this scenario indicate that p_1 and p_2 see the balanced scale ($S_{p_1,p_2}\hat{e}_j$), whereas p_3 does not ($\neg S_{p_3}\hat{e}_j$), where $\hat{e}_j$ represents the balanced state of the scale.

During the first step, all three participants are aware that the weight of the red block is 10 g. This knowledge is shared among them, leading to the belief:

$B_{p_1,p_2,p_3} red = 10g$. However, only participants p_1 and p_2 witness the actions performed by p_1. Based on the axiom "seeing is believing," we can infer that p_1 and p_2 believe in the occurrence of p_1's action ($B_{p_1,p_2}\hat{e}_i$), while p_3 holds the belief that it did not happen ($B_{p_3} \neg \hat{e}_i$). In the second step, drawing upon the axiom "you believe what you say" and the absence of objections from p_1, we can deduce that both p_1 and p_2 believe the scales to be balanced. Based on this perceived evidence, the participants conclude that the weight of the green block equals the combined weight of the red and blue blocks, expressed as $B_{p_1,p_2} green{=}red{+}blue$. However, p_3 lacks these beliefs and consequently does not hold the belief that $green = red + blue$. These two missing beliefs contribute to p_3's false belief in the sixth step.

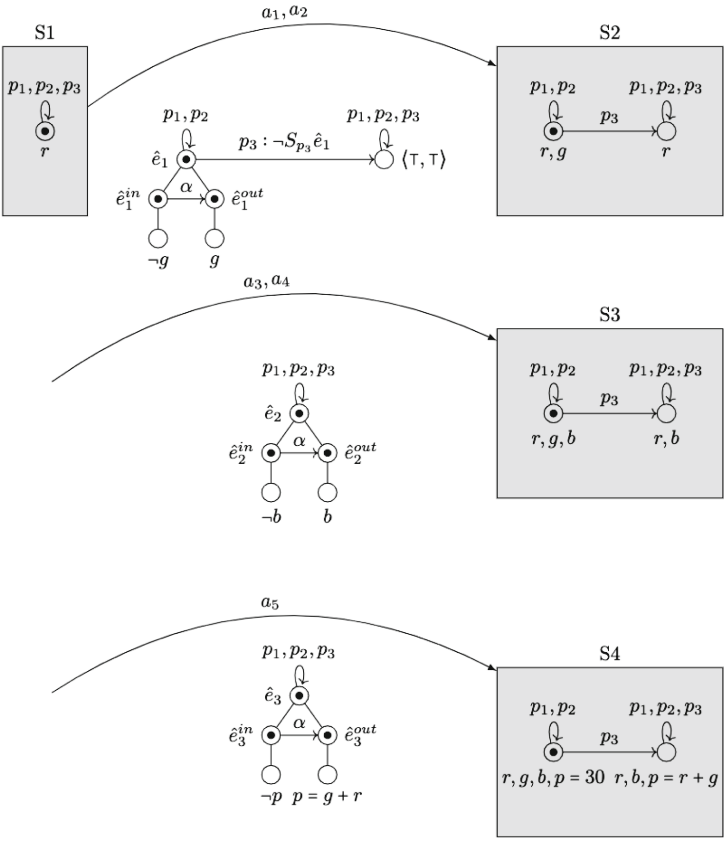


Fig. 7. The Weight Task Model

We employ Fig. 7 to illustrate the evidence-based epistemic updates of the exmaple. S_1 represents the participants' initial epistemic states wherein they all hold the belief that "red = 10 g". The actions performed by participant p_1

are represented by a_1, signifying the three put actions. The speech act from participant p_2 in the second step is represented by a_2. These actions are depicted as $\hat{e}_1$, with the output being that both p_1 and p_2 now believe that "green = red + blue" in S_2. As p_3 does not witness the put action and the balance of the scale, no changes occur for him, and consequently, he does not undergo any epistemic updates within the actual world S_2.

Moving on, a_3 and a_4 represent the third and fourth step, respectively. All three participants witness p_2 placing the blue and red blocks on opposite sides of the scale, observing that the scale is balanced. These actions are depicted as an event model $\hat{e}_2$. Based on this evidence, they all come to believe that the red block and the blue block possess equal weight. This update is reflected in S_3. It is worth noting that despite sharing the same knowledge regarding the blue block, S_3 still maintains two distinct epistemic worlds: only p_1 and p_2 possess comprehensive knowledge about the blocks in the actual world, while p_3 remains unaware of the green block due to their lack of previous epistemic updates.

The fifth step is represented by a_5. The action model illustrates that following the put action, observed by all participants, each individual revises their belief concerning the purple block. Specifically, they now believe that the weight of the purple block equals the combined weight of the red block and the green block. Drawing from this knowledge and the preceding epistemic updates regarding the green block, p_1 and p_2 further revise their beliefs concerning the weight of the purple block. However, since p_3 missed the belief updates regarding the green block, he is unable to perform the inference. Consequently, in S_4, p_3 lacks any knowledge updates regarding the actual weight of the purple block.

8 Conclusion and Future Work

In this paper, we examine the interpretation of multimodal dialogue and the contributions of ontic and epistemic events to the changes introduced in discourse. We introduce a technique for how lexical semantic information associated with verbal predicates can be integrated into epistemic event models as adopted in Dynamic Epistemic Logic, in order to faciliate a more compositional interpretation of epistemic state in dialogue modeling. To this end, we introduce the notion of Lexical Event Modeling (LEM), which encoded both the subeventual properties of events in a language, as well as the epistemic framing of agentive participants in these events. This is intended as the first step in the construction of a compositional procedure for computing an epistemic event model of a dialogue, by reading off the representations from multiple modalities in a discourse.

Acknowledgements. This work was supported in part by NSF grants DRL 2019805 and CNS 2033932 to Dr. Pustejovsky at Brandeis University. We would like to thank Nikhil Krishnaswamy, Ken Lai, Ricky Brutti, Chris Tam, and the reviewers for their comments and suggestions. The views expressed herein are ours alone.

References

1. Asher, N.: Common ground, corrections and coordination. J. Semantics **17**(4), 481–512 (1998)
2. Banarescu, L., et al.: Abstract meaning representation for sembanking. In: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pp. 178–186 (2013)
3. van Benthem, J., Fernández-Duque, D., Pacuit, E.: Evidence and plausibility in neighborhood structures. Ann. Pure Appl. Logic **165**(1), 106–133 (2014)
4. van Benthem, J., Pacuit, E.: Dynamic logics of evidence-based beliefs. Stud. Logica. **99**, 61–92 (2011)
5. Bolander, T.: Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. Jaakko Hintikka on knowledge and game-theoretical semantics, pp. 207–236 (2018)
6. Bolander, T., Andersen, M.B.: Epistemic planning for single-and multi-agent systems. J. Appl. Non-Classical Logics **21**(1), 9–34 (2011)
7. Brown, S.W., Bonn, J., Kazeminejad, G., Zaenen, A., Pustejovsky, J., Palmer, M.: Semantic representations for NLP using VerbNet and the generative lexicon. Front. Artif. Intell. **5**, 821697 (2022)
8. Brutti, R., Donatelli, L., Lai, K., Pustejovsky, J.: Abstract Meaning Representation for gesture. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 1576–1583. European Language Resources Association, Marseille, France (2022)
9. Budzianowski, P., et al.: MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 5016–5026. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/D18-1547, https://aclanthology.org/D18-1547
10. Clark, H.H., Brennan, S.E.: Grounding in communication. In: Resnick, L., B., L., John, M., Teasley, S., D (eds.) Perspectives on Socially Shared Cognition, pp. 13–1991. American Psychological Association (1991)
11. Dowty, D.R.: Word meaning and Montague grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ, vol. 7. Springer (1979)
12. Fernando, T.: Situations in LTL as strings. Inf. Comput. **207**(10), 980–999 (2009)
13. Fischer, K.: How people talk with robots: designing dialog to reduce user uncertainty. AI Mag. **32**(4), 31–38 (2011)
14. Goldin-Meadow, S.: Hearing Gesture: How Our Hands Help Us Think, vol. 14 (2003). https://doi.org/10.2307/j.ctv1w9m9ds
15. Grice, H.P.: Logic and conversation. In: Speech acts, pp. 41–58. Brill (1975)
16. Hadley, L.V., Naylor, G., Hamilton, A.F.d.C.: A review of theories and methods in the science of face-to-face social interaction. Nat. Rev. Psychol. **1**(1), 42–54 (2022). https://doi.org/10.1038/s44159-021-00008-w, https://www.nature.com/articles/s44159-021-00008-w, number: 1 Publisher: Nature Publishing Group
17. Hansen, L.D., Bolander, T.: Implementing theory of mind on a robot using dynamic epistemic logic. In: Twenty-Ninth International Joint Conference on Artificial Intelligence, pp. 1615–1621. International Joint Conference on Artificial Intelligence Organization (2020)
18. Im, S., Pustejovsky, J.: Annotating lexically entailed subevents for textual inference tasks. In: Twenty-Third International Flairs Conference (2010)

19. Jacqmin, L., Barahona, L.M.R., Favre, B.: Äúdo you follow me?Äú: A survey of recent approaches in dialogue state tracking. In: Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 336–350 (2022)
20. Kendrick, K.H., Holler, J., Levinson, S.C.: Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. Philos. Trans. R. Soc. B **378**(1875), 20210473 (2023)
21. Khebour, I., et al.: The weights task dataset: a multimodal dataset of collaboration in a situated task. J. Open Humanit. Data **10**(7), 1–7 (2024)
22. Khebour, I.K., et al.: Common ground tracking in multimodal dialogue. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 3587–3602 (2024)
23. Krishnaswamy, N., Pustejovsky, J.: Generating a novel dataset of multimodal referring expressions. In: Proceedings of the 13th International Conference on Computational Semantics-Short Papers, pp. 44–51 (2019)
24. Kruijff, G.J.M., et al.: Situated dialogue processing for human-robot interaction. In: Cognitive systems, pp. 311–364. Springer (2010)
25. Li, K., Li, J., Guo, D., Yang, X., Wang, M.: Transformer-based visual grounding with cross-modality interaction. ACM Trans. Multimed. Comput. Commun. Appl. **19**(6), 1–19 (2023)
26. Liao, L., Long, L.H., Ma, Y., Lei, W., Chua, T.S.: Dialogue state tracking with incremental reasoning. Trans. Assoc. Comput. Linguist. **9**, 557–569 (2021)
27. Mani, I., Pustejovsky, J.: Interpreting Motion: Grounded Representations for Spatial Language. Oxford University Press (2012)
28. Markowska, M., Soubki, A., Mar, G., Mirroshandel, S.A., Rambow, O., Wasilewska, A.: Formal representation of common ground in dialogue
29. Miller, P.W.: Body language in the classroom. Tech. Connecting Educ. Careers **80**(8), 28–30 (2005)
30. Naumann, R.: Aspects of changes: a dynamic event semantics. J. Semant. **18**, 27–81 (2001)
31. Ohmer, X., Duda, M., Bruni, E.: Emergence of hierarchical reference systems in multi-agent communication. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 5689–5706 (2022)
32. Pacuit, E.: Neighborhood semantics for modal logic. Springer (2017)
33. Plaza, J.: Logics of public communications. Synthese **158**(2), 165–179 (2007)
34. Pustejovsky, J.: The syntax of event structure. Cognition **1**(41), 47–81 (1991)
35. Pustejovsky, J.: The Generative Lexicon. MIT Press, Cambridge, MA (1995)
36. Pustejovsky, J.: Events and the semantics of opposition. Events as grammatical objects, pp. 445–482 (2000)
37. Pustejovsky, J.: Dynamic event structure and habitat theory. In: Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013), pp. 1–10. ACL (2013)
38. Pustejovsky, J., Krishnaswamy, N.: Embodied human computer interaction. KI-Künstliche Intelligenz **35**(3–4), 307–327 (2021)
39. Pustejovsky, J., Moszkowicz, J.: The qualitative spatial dynamics of motion. The Journal of Spatial Cognition and Computation (2011)
40. Radu, I., Tu, E., Schneider, B.: Relationships between body postures and collaborative learning states in an augmented reality study. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12164, pp. 257–262. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52240-7_47

41. Scheutz, M., Cantrell, R., Schermerhorn, P.: Toward humanlike task-based dialogue processing for human robot interaction. AI Mag. **32**(4), 77–84 (2011)
42. Stalnaker, R.: Common ground. Linguist. Philos. **25**(5-6), 701–721
43. Sun, C., Shute, V.J., Stewart, A., Yonehiro, J., Duran, N., D'Mello, S.: Towards a generalized competency model of collaborative problem solving. Comput. Educ. **143**, 103672 (2020)
44. Tam, C., Brutti, R., Lai, K., Pustejovsky, J.: Annotating situated actions in dialogue. In: Proceedings of the 4th International Workshop on Designing Meaning Representation (2023)
45. Traum, D.: A computational theory of grounding in natural language conversation. PhD thesis, University of Rochester (1994)
46. Traum, D.R., Larsson, S.: The information state approach to dialogue management. Current and new directions in discourse and dialogue, pp. 325–353 (2003)
47. Tu, J., et al.: GLAMR: augmenting AMR with GL-VerbNet event structure. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 7746–7759 (2024)
48. Tu, J., Rim, K., Pustejovsky, J.: Competence-based question generation. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 1521–1533 (2022)
49. Van Ditmarsch, H., van Der Hoek, W., Kooi, B.: Dynamic epistemic logic, vol. 337. Springer Science (2007)
50. Vendler, Z.: Verbs and times. The philosophical review pp. 143–160 (1957)
51. Wimmer, H., Perner, J.: Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition **13**(1), 103–128 (1983)
52. Zhu, Y., et al.: Modeling theory of mind in multimodal HCI. In: International Conference on Human-Computer Interaction