Creating Common Ground through Multimodal Simulations

James Pustejovsky[1], Nikhil Krishnaswamy[1], Bruce Draper[2], Pradyumna Narayana[2], and Rahul Bangar[2]

[1]Department of Computer Science, Brandeis University, Waltham, MA, USA
[2]Department of Computer Science, Colorado State University, Fort Collins, CO, USA
{jamesp,nkrishna}@brandeis.edu
{draper,prady,rahul.bangar}@colostate.edu

Abstract

The demand for more sophisticated human-computer interactions is rapidly increasing, as users become more accustomed to conversation-like interactions with their devices. In this paper, we examine this changing landscape in the context of human-machine interaction in a shared workspace to achieve a common goal. In our prototype system, people and avatars cooperate to build blocks world structures through the interaction of language, gesture, vision, and action. This provides a platform to study computational issues involved in multimodal communication. In order to establish elements of the common ground in discourse between speakers, we have created an embodied 3D simulation, enabling both the generation and interpretation of multiple modalities, including: language, gesture, and the visualization of objects moving and agents acting in their environment. The simulation is built on the modeling language VoxML, that encodes objects with rich semantic typing and action affordances, and actions themselves as multimodal programs, enabling contextually salient inferences and decisions in the environment. We illustrate this with a walk-through of multimodal communication in a shared task.

1 Introduction

In this paper, we discuss a developing approach towards modeling peer-to-peer communication using multiple modalities, e.g., language, gesture, vision, and action. This platform integrates a multimodal model of semantics (*Multimodal Semantic Simulations, MSS)* (Pustejovsky and Krishnaswamy, 2016; Krishnaswamy et al., 2017) with a realtime vision system for recognizing human gestures (Wang et al., 2017). This framework assumes both a richer formal model of events and their participants, as well as a modeling language for constructing 3D visualizations of objects and events denoted by linguistic expressions. We position this approach in the context of the questions posed by the workshop organizers, and provide more detail into the architecture that integrates the multiple sources of knowledge within the shared context of communication.

To begin with, let us distinguish between experience and action: two individuals who share an experience, such as witnessing a natural event, hearing a clap of thunder, or feeling the earth tremor, are jointly "co-perceiving an event". Hence, they are *co-situated* and *co-perceptive*. If these two beings are communicating in order to carry out a shared task, such as building a structure, moving objects, or clearing a space, then they can be considered "agents", who are not only co-perceiving the present situation and subsequent situations as they change, but are also acting, together or individually, as a result of communicative interactions. Hence, what is being shared in the latter is considerably richer and more complex in character. Namely, there is agreement, acceptance, or recognition of a common goal between the agents, what can be called "co-intent". These combined factors constitute the first aspects of is called "common ground": namely, co-situatedness, co-perception, and co-intent. The theory of common ground has a rich and diverse literature concerning what is shared or presupposed in human communication (Clark et al., 1991; Gilbert, 1992; Stalnaker, 2002; Asher, 1998; Tomasello and Carpenter, 2007).

When engaged in accomplishing a task jointly, agents share one additional anchoring strategy that greatly enhances the expressiveness of common ground: namely, the ability to "co-attend". Because of the inherently directed nature of attention and co-attention, we will, rather, speak of a "shared situated reference" in the discussion that follows. This ability will emerge as central to determining the denotations of participants in shared events. Events as we experience them are distinct from the way we refer to them with language. The mechanisms in language allow us to package, quantify, measure, and order our experiences, creating rich conceptual reifications and semantic differentiations. The surface realization of this ability is mostly manifest through our linguistic utterances, but is also witnessed through gestures. By examining the nature of the common ground assumed in communication, we can study the conceptual expressiveness of these systems (Pustejovsky, 2018).

We believe that simulation can play a crucial role in human-computer communication; it creates a shared epistemic model of the environment inhabited by a human and an artificial agent, and demonstrates the knowledge held by the agent publicly. Demonstrating knowledge is needed to ensure a shared understanding with the humans involved in the activity, but why create a simulation model, if the goal is to interact with an avatar or robot? If a robotic agent is able to receive linguistic information from a human commander or collaborator and interpret that relative to its current physical circumstances, it can create an epistemic representation of that same information. However, without a modality to express that representation independently, the human is unable to verify or query what the robotic agent is perceiving or how that perception is being interpreted. In a simulation environment the human and robot share an epistemic space, and any modality of communication that can be expressed within that space (e.g., linguistic, visual, gestural) enriches the number of ways that a human and a robotic agent can communicate within object and situation-based tasks, such as those investigated by Hsiao et al. (2008), Dzifcak et al. (2009), Cangelosi (2010).

The simulation environment provided by our model includes the perceptual domain of objects, properties, and events. In addition, propositional content in the model is accessible to the discourse, allowing access to beliefs, desires, and intentions (BDI), and for them to be distinguished by the agents to act and communicate appropriately. This provides the non-linguistic visual and action modalities, which are augmented by the inherently non-linguistic gestural modality enacted within the visual context.

As gesture is intended for visual interpretation, it is directly interpretable in the co-visual context if and only if the denotation of its interpretation function is directly available in the simulation through visual inspection (Lascarides and Stone (2006, 2009); Clair et al. (2010); Matuszek et al. (2014)). Direction or vector information encoded in a gesture hooks into direct deixis and satisfiable actions by projecting the gesture onto a distinct area of the simulation environment. For entity descriptions, speech allows for a randomly accessible and indexable descriptions, while gesture allows for physically and spatially grounded regions, objects, or configurations. Speech can express nominal, qualitative, tangible, or non-tangible attributes, while gesture is limited to orientation and direction, shape, size, relative distance, and manner of motion. For communicative events, where speech allows for any mood (declarative, interrogative, imperative), gesture is usually imperative, except for speech acts. Speech allows for an essentially unconstrained expression of propositional content, while gesture is limited to the propositional content of a speech act, and the content of a deixis or action description.

For explicit spatial grounding, gesture is computationally advantageous as spatially-encoded information in the gesture (e.g., direction, vector, etc.) can be directly grounded in the scene without having to process an additional linguistic layer of indirection. In some cases, spatially-denoting manner is easier to model with gesture than language. This is also the case with technology-driven multimodal display interfaces (Johnston, 2009), but here we are concerned with a co-situated situation where the interlocutors, human and virtual, share an epistemic space.

We assume interlocutors have explicit and distinct modes of interpretation as accessed through perception. Input from the environment (our embedding space) evidences various beliefs that we establish as knowledge. Our model of synthetic vision provides distinct but mutually coherent views on the shared environment. An example involving object occlusion is discussed in Section 3.

Any of the different modalities in our model can be used to disambiguate new information introduced

into the context by the other participant(s) by enumerating possible options, pruning question/answer steps based on the number of options, and interaction between interlocutors until an unambiguous interpretation is achieved. Examples of this are discussed in Section 4.

2 Multimodal Communication

2.1 Language

Our system uses VoxSim, an open-source language-driven event simulator (Krishnaswamy and Pustejovsky, 2016a,b) to operationalize events in real time, mapping from natural language input using a dynamic semantics and the modeling language VoxML (Pustejovsky and Krishnaswamy, 2016). VoxML semantic typing extends that provided within Generative Lexicon Theory (Pustejovsky, 1995), as well as the approach taken in Asher (2011), to enable a semantics of embodiment (Pustejovsky, 2013; Pustejovsky and Krishnaswamy, 2014). Embodiment, a requirement on and facilitated by the shared virtual environment, allows us to operationalize gestures as programs enacted within the 3D environment with a direct mapping to linguistic description. This enactment is performed by a virtual avatar, in the vein of prior work studying human-computer communication (Kopp et al., 2003; Sowa and Kopp, 2003; Kopp et al., 2005; Krämer et al., 2007; Weitnauer et al., 2008).

2.2 Gesture

Gestures and actions are the primary modes of communication between a person and an avatar in our system. The human user stands at one end of a table, facing a monitor at the other end showing a virtual continuation of the table, and the avatar. The user's motions are captured using a Microsoft Kinect v2 RGBD sensor. Gestures are detected in real time using in-house gesture recognition software (in press) and sent to the avatar via VoxSim. The avatar's gestures are displayed to the human on the monitor.

The real-time gesture detection system can recognize 34 hand poses trained from a publicly available dataset of naturally occurring gestures between human subjects working on blocks world tasks (Wang et al., 2017). It can also detect head nods, head shakes, stepping up to or back from the table, as well as the directions of arm motions. Other body poses not classified by the recognizer as one or the above are labeled "other," and not considered to be semantic gestures in this context.

Our prototype scenarios restrict the gestural lexicon to only eight gestural voxemes (visual lexemes), four of which pertain to actions and four to the state of discourse. The action-based gestural voxemes are: *point* (i.e., deixis), either at a block or open space; *grab*, the user mimicking grabbing a block; *carry*, mimicking moving a block they have grabbed; and *push*. *Point, carry* and *push* include directionality (e.g., "left" or "away"). The dialogue-based voxemes are: *engage* and *disengage*, to begin and end a dialogue; and *positive* and *negative acknowledge*, to signal agreement or disagreement.

Gestural voxemes, as we use the term here, are more general than physical gestures. For example, users can signal positive acknowledgement by nodding their heads or by giving a "thumbs up" sign with either or both hands. There are therefore seven different motions to signify positive acknowledgement (including combinations of nodding and thumbs up with either or both hands). Each voxeme therefore corresponds to a set of synonym gestures, often including but not limited to **enantiomorphs**. As the semantics of gesture are often culturally and situationally conditioned (Müller, 2004), we should note that the gesture set used here comes from a dataset elicited from United States university students (i.e., WEIRD (Henrich et al., 2010)), and used in a context designed to achieve a particular construction goal, and that should be taken into account and considering their descriptive properties, creation, and enactment (Streeck, 2008, 2009).

2.3 Synthetic Vision

In the context of peer-to-peer communication, we are concerned not only with co-belief associated with presupposed and shared communal knowledge (Grosz and Sidner, 1990; Cohen et al., 1990; Asher and Lascarides, 2008; Lascarides and Asher, 2009), but also that knowledge arising through the shared experiences of co-situated (but distinct) perceptions. For this (and other) reasons, perception reports have

often been modeled modally as a conventional epistemic operator (Shoham and Del Val, 1991), yet while this is a formal convenience it is not cognitively or computationally realistic (Musto and Konolige, 1993; Bell and Huang, 1998). This has been a gap in approaches to perception through modal logic, reducing the act of perception to a extramission theory of vision. Alternatively, while computational approaches to vision explicitly operationalize the recognition of features and objects, this is far from our goal of extracting epistemic relations from a scene. Some formal approaches have moved beyond these views, by acknowledging the role of sensory capture as distinct from the propositional content that results from this act (Wooldridge and Lomuscio, 1999), as well as (Schwarzentruber, 2010, 2011). We share this view, but aim to go further, by embedding the modal interpretation of the knowledge associated with perception as the result of a signal transmission from sensory data in the environment.

To this end, we adopt the strategies and modeling techniques of synthetic vision (Noser et al., 1995; Peters and O'Sullivan, 2002) combined with work in dynamic epistemic logic (Goranko and Otto, 2007; Plaza, 2007; Van Benthem, 2011). We distinguish between egocentric (viewer-based) and allocentric (externally-based) perspectives. Within a 3D environment, the problem of visibility is introduced because objects may be occluded from the frame of reference of the viewer. Further, because of distinct egocentric frames of reference (FoR) for each interlocutor in a co-situated, co-perceptual task, we need to track the interpretations (valuations) of qualitative spatial relations that rely on FoR, such as *in front of*, *to the left*, *behind*, etc. (Zimmermann and Freksa, 1996; Ligozat, 1993).

We implement an evidential-based model of perception, where knowledge is arrived at or "evidenced through" inferences over properties and relations from sensory input, with which we define $\mathcal{S}_a\phi$ (agent a sees that ϕ). Synthetic vision, $\mathcal{S}_a\phi$, for a computational agent within VoxSim is defined, in part, as a relation of *Evidences* ($\mathcal{E}$), given informally as follows: $\psi\mathcal{E}_a\phi$ states that a sensory input relation, ψ, provides enough evidence for agent a to know that ϕ. The set of evidencing propositions includes the following object attribute classes: *color*, *shape*, *size*, and *orientation*. For two or more objects, the qualitative spatial relations from RCC8 and RCC-3D (Randell et al., 1992; Albath et al., 2010) are used, along with orientational (directional) relations.

2.4 Action

At any point in the human-avatar interaction where the avatar's internal logic has extracted a meaning from the linguistic or gestural input, the avatar then communicates this understanding to the human by performing the associated action, such as pushing a block adjacent to another block, putting a block on top of another block, or moving a block into a specified region of the table.

Action can demonstrate clear, unambiguous understanding, such as moving a distinct block to a definite location, or can be a request for more meaning, as when the avatar can gesture toward a block it thinks *might* have been indicated by the human, in order to request confirmation.

Action therefore communicates the avatar's understanding of the current situational context, providing an epistemic basis that each party can perform inference over to bring their understanding of the other's vision of the external world into line.

3 Example Scenario and Walk-through

Figure 1 illustrates the same configuration of blocks as seen from the perspectives of the human and avatar. From the avatar's point of view (POV), the green block is occluded by the larger red block.

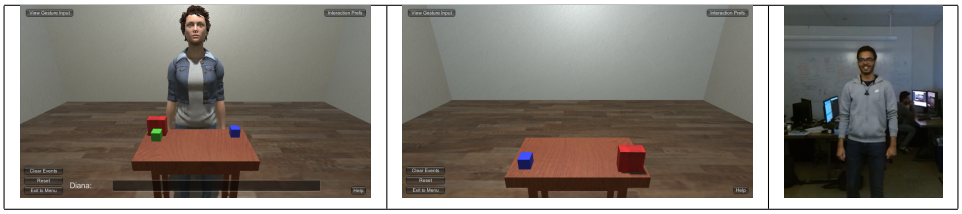

Figure 1: L: Human's view of scene; C: Avatar's view of blocks, table; R: Avatar's view of human

Using this configuration, an example scenario follows as depicted in Figure 2, with individual interactions indicated by the labeled frames which show the human's action in the left half of the frame and the avatar's action on the right.

When the human begins by pointing to his left (Frame A), this is interpreted unambiguously as indicating the red block, as the avatar cannot see the green one.

The human moves his left hand, palm open, from his left to his right, indicating "push to right" (Frame B). Here, the avatar enters a disambiguation loop, a state machine of which is shown in Figure 3. Using both text and audio output in natural language, the avatar requests a clarification: Does the human intend her to push the indicated (red) block adjacent to the (blue) block on the right side of the table? This is confirmed by a thumbs-up gesture from the human, and the avatar moves the red block to the left side of the blue block (Frame C).

This action has an effect unintended by the avatar: moving the red block has exposed the green one, the existence of which was previously unknown to the avatar. Before this point, the human and avatar are co-situated in and co-perceptive of the scene, but their respective visions of the external world do not cohere due to their differing beliefs regarding the existence of the green block. Now, when the green block is revealed to the avatar, they have the same beliefs about the three blocks in the scene, and the avatar must update her internal model of the scene to include the existence of the green block. The prior existence of the green block to this point is also included in the update axiomatically. Having global knowledge of the scene, we can see that the two agents' models of the world are now coherent, through an update facilitated by multimodal interaction.

Now that knowledge of the existence and location of the green block is common to both parties, when the human indicates (his) left side of the table, the avatar can unambiguously understand this to mean that the green block is indicated (Frame D). By then gesturing *grab* (hand in a clawlike position, as shown in Frame E), then moving his hand up, over to the red block, and down, the human can indicate to the avatar that he wants her to pick up the green block and move it either on top of the red block, or into the same region of the table, an ambiguity that the avatar can resolve though some questions, answered and acknowledged by the human. The result is a four-step staircase made of three blocks (Frame F).

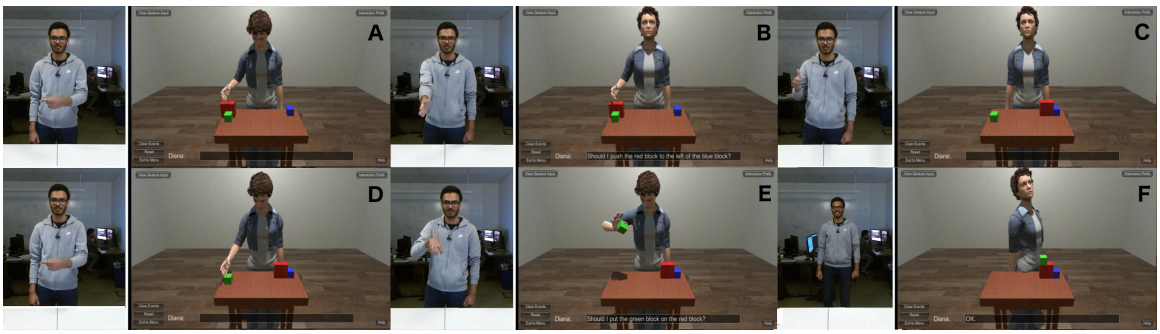

Figure 2: Example staircase construction scenario

Our model of communication allows for multiple parameters to be altered in order to make changes to the nature of the interaction. For example, the avatar can behave more proactively or less when disambiguating gestures, which can affect the action taken. Examples of this are discussed in Section 4.

4 Implications for Human-Computer Communication

Co-situatedness and co-perception of the participants give rise to some interesting considerations in their interaction. When, in the example from Section 3, the avatar begins moving the red block and reveals the green block, she might pause her move before completion and clarify that moving the red block was actually what the human intended; the introduction of the green block into her present model of the world causes her to update her past model, as discussed in Section 3, and the propagation of that dynamic epistemic model to the present creates a choice-point where none existed before.

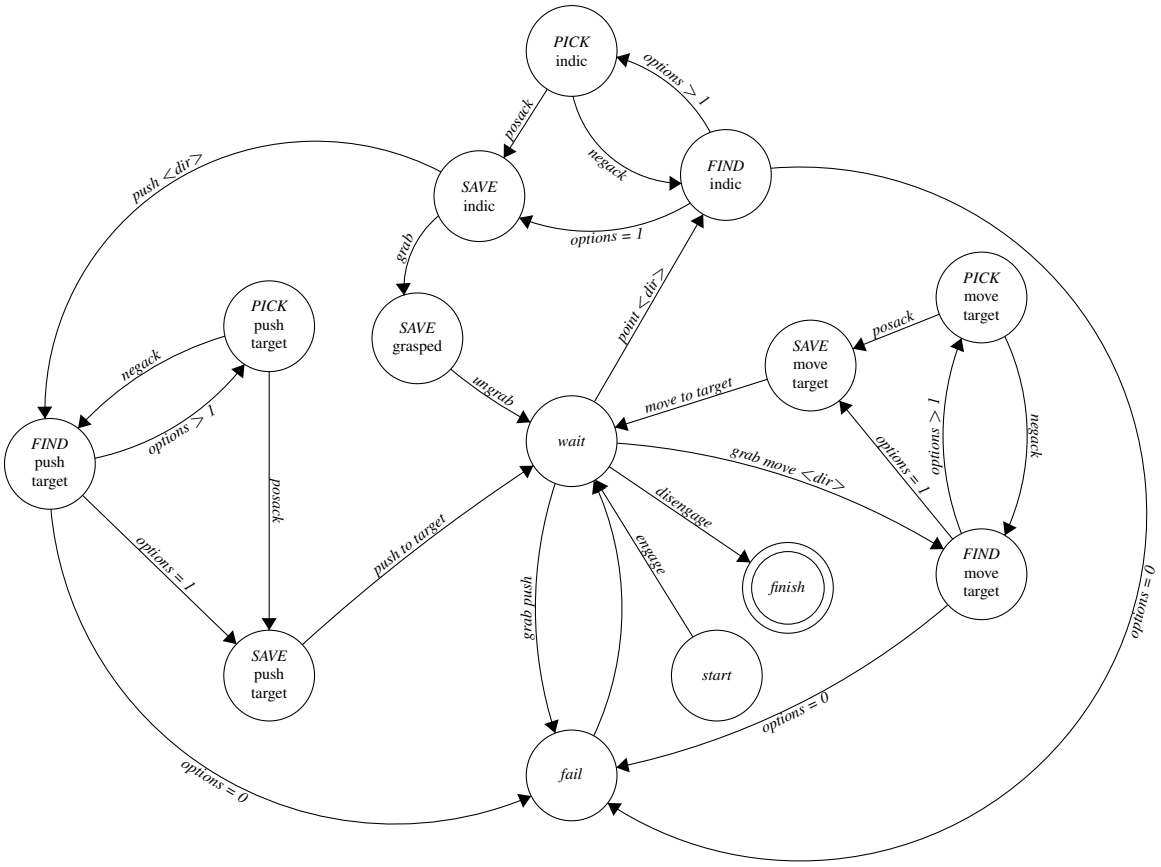

Figure 3: Finite state machine for dialogue management

Alternately, when the avatar reaches for or starts moving the red block before the green block is revealed, the human might interrupt with a "stop" gesture or utterance, indicating that the avatar is doing something divergent from the human's plans. Evidenced by the positions of the blocks and the actions being taken with them, the avatar can infer that the indicated block is not the one she is supposed to be affecting; since no other blocks are visible in the originally-indicated region of the table, this introduces the possibility that there exists another block to be manipulated. The avatar's model of the scenario changes from a two-block model $\vDash r, b$ at t_0 to a three-block model $\vDash r, g, b$ at t_n, where new information is introduced either by evidencing or inference in a logic of synthetic vision. This way we can represent psychological phenomena like object permanence in a simulation environment as changes in the entropy of the information available to the virtual agent through its (synthetic) vision of the scene.

5 Conclusion

Multimodal peer-to-peer interfaces require robust integration of conversational modalities in a naturalistic fashion. Here we outline the first steps toward such integration, based on the logic of a multimodal simulation semantics and 3D environment as the platform for shared common ground. We provide our computational agent with a framework for some of the faculties natively available to humans using sophisticated computer vision techniques to recognize gesture and by laying the groundwork for a modal logic of synthetic vision. The result is a framework and platform that interweaves linguistic and non-linguistic modalities to facilitate the completion of a shared task by exploiting the relative strengths of linguistic and non-linguistic context to exchange information in a situated communication.

Acknowledgements

We would like to thank the reviewers for their insightful comments. We have tried to incorporate their suggestions where possible. This work was supported by Contract W911NF-15-C-0238 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO). Approved for Public Release, Distribution Unlimited. The views expressed herein are ours and do not reflect the official policy or position of the Department of Defense or the U.S. Government. All errors and mistakes are, of course, the responsibilities of the authors.

References

Albath, J., J. L. Leopold, C. L. Sabharwal, and A. M. Maglia (2010). RCC-3D: Qualitative spatial reasoning in 3D. In *CAINE*, pp. 74–79.

Asher, N. (1998). Common ground, corrections and coordination. *Journal of Semantics*.

Asher, N. (2011). *Lexical meaning in context: A web of words*. Cambridge University Press.

Asher, N. and A. Lascarides (2008). Commitments, beliefs and intentions in dialogue. *Proceedings of Londial*, 35–42.

Bell, J. and Z. Huang (1998). Seeing is believing. In *4th Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense 98)*.

Cangelosi, A. (2010). Grounding language in action and perception: from cognitive agents to humanoid robots. *Physics of life reviews 7*(2), 139–151.

Clair, A. S., R. Mead, M. J. Matarić, et al. (2010). Monitoring and guiding user attention and intention in human-robot interaction. In *ICRA-ICAIR Workshop, Anchorage, AK, USA*, Volume 1025.

Clark, H. H., S. E. Brennan, et al. (1991). Grounding in communication. *Perspectives on socially shared cognition 13*(1991), 127–149.

Cohen, P. R., J. L. Morgan, and M. E. Pollack (1990). *Intentions in communication*. MIT press.

Dzifcak, J., M. Scheutz, C. Baral, and P. Schermerhorn (2009). What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pp. 4163–4168. IEEE.

Gilbert, M. (1992). *On social facts*. Princeton University Press.

Goranko, V. and M. Otto (2007). 5 model theory of modal logic. *Studies in Logic and Practical Reasoning 3*, 249–329.

Grosz, B. J. and C. L. Sidner (1990). Plans for discourse. In *Intentions in communication*. MIT Press.

Henrich, J., S. J. Heine, and A. Norenzayan (2010). Most people are not weird. *Nature 466*(7302), 29–29.

Hsiao, K.-Y., S. Tellex, S. Vosoughi, R. Kubat, and D. Roy (2008). Object schemas for grounding language in a responsive robot. *Connection Science 20*(4), 253–276.

Johnston, M. (2009). Building multimodal applications with emma. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pp. 47–54. ACM.

Kopp, S., L. Gesellensetter, N. C. Krämer, and I. Wachsmuth (2005). A conversational agent as museum guide–design and evaluation of a real-world application. In *International Workshop on Intelligent Virtual Agents*, pp. 329–343. Springer.

Kopp, S., T. Sowa, and I. Wachsmuth (2003). Imitation games with an artificial agent: From mimicking to understanding shape-related iconic gestures. In *International Gesture Workshop*, pp. 436–447. Springer.

Krämer, N. C., N. Simons, and S. Kopp (2007). The effects of an embodied conversational agents nonverbal behavior on users evaluation and behavioral mimicry. In *International Workshop on Intelligent Virtual Agents*, pp. 238–251. Springer.

Krishnaswamy, N., P. Narayana, I. Wang, K. Rim, R. Bangar, D. Patil, G. Mulay, J. Ruiz, R. Beveridge, B. Draper, and J. Pustejovsky (2017). Communicating and acting: Understanding gesture in simulation semantics. In *12th International Workshop on Computational Semantics*.

Krishnaswamy, N. and J. Pustejovsky (2016a). Multimodal semantic simulations of linguistically underspecified motion events. In *Spatial Cognition X: International Conference on Spatial Cognition*. Springer.

Krishnaswamy, N. and J. Pustejovsky (2016b). VoxSim: A visual platform for modeling motion language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. ACL.

Lascarides, A. and N. Asher (2009). Agreement, disputes and commitments in dialogue. *Journal of Semantics 26*(2), 109–158.

Lascarides, A. and M. Stone (2006). Formal semantics for iconic gesture. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (BRANDIAL)*, pp. 64–71.

Lascarides, A. and M. Stone (2009). A formal semantic analysis of gesture. *Journal of Semantics 26*(4), 393–449.

Ligozat, G. (1993). Qualitative triangulation for spatial reasoning. *Spatial Information Theory A Theoretical Basis for GIS*, 54–68.

Matuszek, C., L. Bo, L. Zettlemoyer, and D. Fox (2014). Learning from unscripted deictic gesture and language for human-robot interactions. In *AAAI*, pp. 2556–2563.

Müller, C. (2004). Forms and uses of the palm up open hand: A case of a gesture family. *The semantics and pragmatics of everyday gestures 9*, 233–256.

Musto, D. and K. Konolige (1993). Reasoning about perception. *AI Communications 6*(3-4), 207–212.

Noser, H., O. Renault, D. Thalmann, and N. M. Thalmann (1995). Navigation for digital actors based on synthetic vision, memory, and learning. *Computers & graphics 19*(1), 7–19.

Peters, C. and C. O'Sullivan (2002). Synthetic vision and memory for autonomous virtual humans. In *Computer Graphics Forum*, Volume 21, pp. 743–752. Wiley Online Library.

Plaza, J. (2007). Logics of public communications. *Synthese 158*(2), 165.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press, Cambridge, MA.

Pustejovsky, J. (2013). Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pp. 1–10. ACL.

Pustejovsky, J. (2018). From experiencing events in the action-perception cycle to representing events in language. *Interaction Studies 19*.

Pustejovsky, J. and N. Krishnaswamy (2014). Generating simulations of motion events from verbal descriptions. *Lexical and Computational Semantics (* SEM 2014)*, 99.

Pustejovsky, J. and N. Krishnaswamy (2016, May). VoxML: A visualization modeling language. In N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Randell, D., Z. Cui, and A. Cohn (1992). A spatial logic based on regions and connections. In M. Kaufmann (Ed.), *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning*, San Mateo, pp. 165–176.

Schwarzentruber, F. (2010). *Seeing, knowing, doing: case studies in modal logic*. Ph. D. thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier.

Schwarzentruber, F. (2011). Seeing, knowledge and common knowledge. In *LORI*, pp. 258–271.

Shoham, Y. and A. Del Val (1991). *A logic for perception and belief*. Department of Computer Science, Stanford University.

Sowa, T. and S. Kopp (2003). A cognitive model for the representation and processing of shape-related gestures. In *Proceedings of the European Cognitive Science Conference (EuroCogSci03)*, pp. 441.

Stalnaker, R. (2002). Common ground. *Linguistics and philosophy 25*(5), 701–721.

Streeck, J. (2008). Depicting by gesture. *Gesture 8*(3), 285–301.

Streeck, J. (2009). *Gesturecraft: The manu-facture of meaning*, Volume 2. John Benjamins Publishing.

Tomasello, M. and M. Carpenter (2007). Shared intentionality. *Developmental science 10*(1), 121–125.

Van Benthem, J. (2011). *Logical dynamics of information and interaction*. Cambridge University Press.

Wang, I., M. Ben Fraj, P. Narayana, D. Patil, G. Mulay, R. Bangar, R. Beveridge, B. Draper, and J. Ruiz (2017). Eggnog: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In *Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition*.

Wang, I., P. Narayana, D. Patil, G. Mulay, R. Bangar, B. Draper, R. Beveridge, and J. Ruiz (2017). Exploring the use of gesture in collaborative tasks. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, New York, NY, USA, pp. 2990–2997. ACM.

Weitnauer, E., N. M. Thomas, F. Rabe, and S. Kopp (2008). Intelligent agents living in social virtual environments–bringing max into second life. In *International Workshop on Intelligent Virtual Agents*, pp. 552–553. Springer.

Wooldridge, M. and A. Lomuscio (1999). Reasoning about visibility, perception, and knowledge. In *International Workshop on Agent Theories, Architectures, and Languages*, pp. 1–12. Springer.

Zimmermann, K. and C. Freksa (1996). Qualitative spatial reasoning using orientation, distance, and path knowledge. *Applied intelligence 6*(1), 49–58.