Generating a Novel Dataset of Multimodal Referring Expressions

Nikhil Krishnaswamy
Brandeis University
nkrishna@brandeis.edu

James Pustejovsky
Brandeis University
jamesp@brandeis.edu

Abstract

Referring expressions and definite descriptions of objects in space exploit information about both object characteristics and locations. Linguistic referencing strategies can rely on increasingly high-level abstractions to distinguish an object in a given location from similar ones elsewhere, yet the description of the intended location may still be unnatural or difficult to interpret. Modalities like gesture may communicate spatial information like locations in a more concise manner. When communicating with each other, humans mix language and gesture to reference entities, changing modalities as needed. Recent progress in AI and human-computer interaction has created systems where a human can interact with a computer multimodally, but computers often lack the capacity to intelligently mix modalities when generating referring expressions. We present a novel dataset of referring expressions combining natural language and gesture, describe its creation and evaluation, and its uses to train models for generating and interpreting multimodal referring expressions.

1 Introduction

Psychological studies suggest that gesture serves as a bridge between understanding actions situated in the world and linguistic descriptions, such as symbolic references to entity classes and attributes (Butterworth, 2003; Capirci et al., 2005). Many researchers (e.g., Clark et al. (1983); Volterra et al. (2005)), view gesture as a common mode of reference vis-à-vis *common ground*. Gesture is well-suited to directly grounding spatial information; pointing can bind to a location or be coerced to object(s) in that location (Ballard et al., 1997). Demonstrative or attributive language (e.g., size, shape, qualitative relations), can specify entities by binding those characteristics to information received via gesture. Thus, language affords abstract strategies to distinguish an object in a given location from similar ones elsewhere (e.g., *the chair closest to that door*—with pointing, or *the green block at the right side of the table*).

As an environment becomes more complex, so does the language used to give directions or single out specific items in it (Skubic et al., 2004; Moratz and Tenbrink, 2006). An object indicated by deixis is usually also the topic of discussion (Brooks and Breazeal, 2006), but deixis may be ambiguous depending on distance from agent to target object, or other objects close to the target object (Latoschik and Wachsmuth, 1997), while language can supplement it for more useful definite descriptions (Bangerter, 2004). Co-temporal/overlapping speech and gesture (or an "ensemble" (Pustejovsky, 2018)) often involves deixis to ground the location, and language to specify further information (Sluis and Krahmer, 2004). As a task's natural language requirements grow more complex, subjects rely on other modalities to carry semantic load, particularly as the need for immediate interpretation grows (Whitney et al., 2016).

Studies in this area have a long history in computational linguistics/semantics (e.g., Claassen (1992); Krahmer and van der Sluis (2003)), human-robot interaction (e.g., Kelleher and Kruijff (2006); Foster et al. (2008)), and computational and human discourse studies (e.g., Bortfeld and Brennan (1997); Funakoshi et al. (2004); Viethen and Dale (2008)). Following these, we seek to build models for generating, recognizing, and classifying referring expressions that are both *natural* and *useful* to the human interlocutors of computational dialogue systems. Here, we present a novel dataset of Embodied Multimodal Referring Expressions (EMRE), blending gesture and natural language (English text-to-speech), used by an avatar in a human-computer interaction (HCI) scenario. We describe raw data generation, annotation and evaluation, preliminary analysis, and expected uses in training machine learning models for generating referring expressions in real-time that are appropriate, salient, and natural in context.

2 Data Gathering

As our goal is to train models which a system can use to generate and interpret naturalistic multimodal referring expressions during interaction with a human, we gathered data using such a system—specifically VoxSim, a semantically-driven visual event simulator based on the VoxML semantic modeling language (Pustejovsky and Krishnaswamy, 2016), that facilitates data gathering using Monte-Carlo parameter setting to simulate motion predicates in 3D space (Krishnaswamy and Pustejovsky, 2016). We created a variant on the *Human-Avatar-Blocks World* (HAB) system (Krishnaswamy et al., 2017; Narayana et al., 2018), in which VoxSim visualizes the actions taken by an avatar in the 3D world as she interprets gestural and spoken input from a human interlocutor.[1] A shortcoming of the HAB system is the asymmetry between the language that the system's avatar is capable of recognizing and interpreting, and the English utterances it can generate (Krishnaswamy and Pustejovsky, 2018). Specifically, the avatar can 1) produce complete sentences of structures that it cannot entirely parse and 2) properly interpret spatial terms and relations between objects, but cannot fluently use them to refer to objects or the relations between them. Improvements to the first asymmetry are under development separately, and here we present data for creating a robust model of referring techniques in all available modalities, to help rectify the second asymmetry, for more fluent interaction in this and other HCI systems.

The gesture semantics in VoxSim are largely based on the formalisms of Lascarides and Stone (2006; 2009a; 2009b). Multimodal information in a multimodal system cannot be assumed to follow the same format as unimodal information (Oviatt, 1999). Language in an ensemble cannot be assumed to be identical to language used alone. A reference to an object may be grounded in gesture, natural language, or both, subject to constraints that vary per modality. We therefore generated a dataset that can be examined for where these constraints occur, and under which circumstances human evaluators, as proxies for interlocutors with the avatar in a live interaction, prefer one referring modality to another, and with what descriptive detail.

$$
\left[\begin{array}{l}
\textbf{point} \\
\text{TYPE} = \left[\begin{array}{ll}
\text{HEAD} = & \textbf{assignment} \\
\text{ARGS} = & \left[\begin{array}{l}
\text{A}_1 = \textbf{x:agent} \\
\text{A}_2 = \textbf{y:finger} \\
\text{A}_3 = \textbf{z:location} \\
\text{A}_4 = \textbf{w:physobj location}
\end{array}\right] \\
\text{BODY} = & \left[\begin{array}{l}
\text{E}_1 = extend(x, y) \\
\text{E}_2 = def(vec(x \;\; ! \;\; y \quad z), as(w))
\end{array}\right]
\end{array}\right]
\end{array}\right]
$$

Figure 1: VoxML typing of [[POINT]] (Pustejovsky and Krishnaswamy, 2016). E_2 defines the target of deixis as the intersection of the vector extended in e_1 with a location, and reifies that point as a variable w. A_4, shows the compound binding of w to the indicated region and objects within that region.

2.1 Video and Quantitative Data

In our test scenario, there are six equally-sized target blocks on a table, for which the avatar generates referring expressions; two each are *red*, *green*, or *purple*. This gives each block an identifiable, non-unique characteristic that requires disambiguation. They may also be used in the definite descriptions of other blocks. There are three unique objects on the table: a *plate*, a *knife*, and a *cup*. These "landmark" objects will never be the object of a referring expression, but may be used in referring to the target block.

In all scenes, we store the spatial relations between all objects. We used qualitative relations as defined in a subset of the Region Connection Calculus (RCC8) (Randell et al., 1992) and Ternary Point Configuration Calculus (TPCC) (Moratz et al., 2002), and included in the library QSRLib (Gatsoulis et al., 2016). Where calculi in QSRLib only cover 2D spatial relations, VoxSim uses extensions such as RCC-3D (Albath et al., 2010) or computes axial overlap with the Separating Hyperplane Theorem (Schneider, 2014). All spatial relations used were mapped to a linguistic term, such that the RCC8 relation EC (Externally Connected) would be referred to as *touching* in the language generation phase.

For generating utterances, we explore 2 variables: number of relational adjuncts, and type of demonstratives used. Conditions on proximal vs. distal demonstratives have been explored in multiple studies (Botley and McEnery, 2001; Strauss, 2002) and the boundaries between proximal and distal egocentric

[1] https://github.com/VoxML/VoxSim

space are regularly shown to be flexible (Coventry et al., 2014). The distributions of demonstratives vary across languages (Proulx, 1988; Meira, 2003; Hayashi, 2004; Piwek et al., 2008) but seem to be consistently conditioned on distance (spatial, textual, or grammatical) between all indexes involved in an utterance and not just the object of focus. This data is only for English definite descriptions but VoxSim provides a platform to create multimodally grounded data for any language, in principle.

This data comprises a set of approximately 10-second videos, each showing the avatar referring to one object. Blocks and landmark objects were placed randomly in the scene, and each block was referred to in turn. All videos consist of two segments: **1**) The target object is encircled in pink to draw attention to it; **2**) the avatar indicates the target object with either an animated deictic gesture (pointing), with spoken English, or an ensemble containing both. The camera through which the 3D virtual world is rendered is placed at the coordinates of the avatar's head, so directions in her linguistic descriptions are consistent with the viewer's perspective. *In front of* x means closer to the agent than x and *behind* x is further away. The avatar referred to each object five times: once with gesture only, twice with spoken language only, and twice with the ensemble. Where gesture was involved, the avatar pointed to the object with the closer hand as measured by Euclidean distance, with an extended index finger (see Fig. 2). The extended finger (the *stroke phase* per Kendon (2004)) was held for 2 seconds.

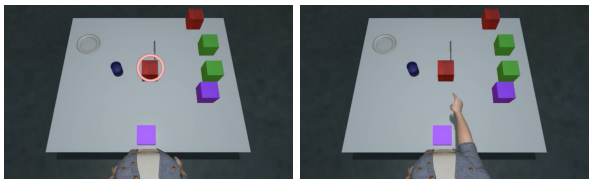

Figure 2: Sample frames. The pink circle (L) indicates the red block in the center as the target object. (R) shows the avatar pointing to it. The block might be described as "this red block" or "that red block in front of the knife", or, without deixis, "the red block right of the cup" or "the red block in front of the knife, right of the cup, and left of the green block."

We generated linguistic descriptions using the following process: **1)** If the referring modality was solely linguistic, the article "the" was used. For ensembles, a demonstrative "that" or "this" was used, and use of a distance distinction was randomly `false`—where the demonstrative was always "that"—or `true`. If `true`, the distinction type was randomly **relative** or **absolute**. For a *relative* distinction, "this" object was the closer of two identically-colored blocks to the agent, and "that" was the further of the two. For *absolute*, "this" was any object in the half of the table closer to the agent, and "that," any object in the other half. **2)** A set of n (randomly selected $2\ f0..3g$) relational descriptors were constructed describing the target object's current relations to other objects, ordered randomly. See Fig. 2 for multiple ways of describing a target object.

Each of the 6 blocks was referred to in turn, after which the objects on the table were moved to new random locations, and references started over. We captured 50 different object configurations, for a total of 1,500 video object references (5 references 6 blocks 50 configurations).[2] For each video, the referring modality, distance distinction and type, full descriptive phrase if any, and relational descriptors were stored in a database, with the full set of all object coordinates in the configuration depicted, the full relation set describing the configuration, and the Euclidean distance from the target object to the agent.

2.2 Annotation

Videos, grouped by configuration, were posted to Amazon Mechanical Turk (MTurk) as Human Intelligence Tasks (HITs). Each HIT contained the 5 videos showing references to one target in one configuration, and was completed by 8 Workers, for a total of 2,400 HITs. Workers were paid $0.10 per HIT and were given a maximum of 30 minutes to complete each. Each Worker viewed the 5 videos and ranked them on a Likert-type scale by how *natural* they considered the reference method in the video, 1 being least natural and 5 most. Workers could optionally add how they would have made definite reference to the target object. If Workers ranked the videos 1-2-3-4-5 or 5-4-3-2-1, we asked them to textually confirm this intent, to limit bots or workers not actually performing the task. We rejected answers that tied more than three videos, to limit bots or workers automatically ranking all the same. This process resulted in 1,500 videos depicting referring methods for objects in various configurations with quantitative values

[2] A sample video can be viewed at https://s3.amazonaws.com/emre-videos/emre_vid/EMRE-2019-01-07-095844.mp4

describing each, and 2,228 assessments of the naturalness of procedurally-generated references. Of the 2,400 HITs, 172 were rejected for not following instructions (providing rankings outside the 1-5 range or using non-numerical values), or for being judged as trying to game the system.[3] Workers on this task were limited to English speakers, and had an average lifetime approval rate of 93.25%. Over the entire batch, Workers took an average of 12 minutes, 11.5 seconds to complete each HIT.

3 Analysis and Discussion

We analyzed the probability distributions of typically high- and low-ranked referring expressions relative to various conditions in the video containing them. For instance, if "ensemble" referring expressions have a higher probability of a high rank than purely gestural references, this would demonstrate evaluators' preference for them. If, however, ensemble referring expressions are only more likely to receive a high ranking compared to gestural references when the target object is far from the agent, this would suggest that distance is a factor in using language to disambiguate. Since we used a Likert-type scale to rank the videos, leading to the possibility that evaluators would rank all videos as relatively good or bad but some better/worse than others, we not only assessed the probability of a video generated under a certain set of conditions receiving a particular score 1–5, but also the probability of a video receiving a score worse/better (2) than the median score of all that evaluator's rankings on that individual task. Below we present some of the strongest predictors and most interesting dependencies uncovered.

Fig. 3 shows the relative probability of score conditioned on *modality*. It is very clear that there is a strong preference for the *ensemble* (in yellow) compared to the others, and that the gesture only method (in blue) was routinely ranked worst while language only was more likely average in terms of naturalness. From the graph on the right, we can see that while the ensemble method was still most likely to achieve ratings above the median, this was not always *far* (i.e., +2) above the median, suggesting that either most referencing methods were considered adequate (the percentages at $X = 0$ also suggest this), or the ensemble method itself could be bettered in some way (likely by clearer or more detailed language—see Fig. 5). This indicates that while language alone suffices for definite reference but leaves room for improvement, and gesture alone is often insufficient, the combination is usually more natural, perhaps due to semantic content that is redundant in context and further reduces ambiguity (Gatt et al., 2011).

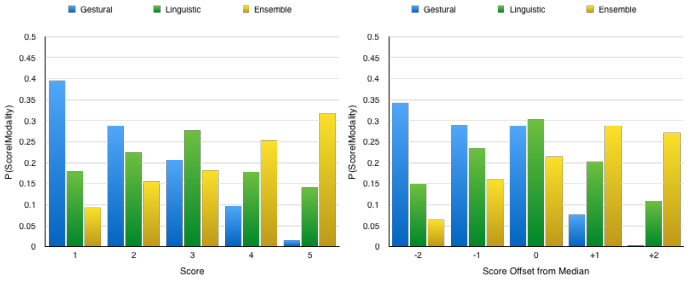

Figure 3: P(Score*j*Modality) [L]; P(Diff from median*j*Modality) [R]

Fig. 4 shows the probability of a referring method (in the *ensemble* modality only) receiving a score given the type of distance distinction used. The *absolute* distance distinction (shown in blue), is somewhat more likely than the *relative* distinction type to score highly suggesting either a relatively static demarcation between points considered "proximal" to the agent and "distal" points, or some role for the table, the surface relative to which the distance distinction was calculated. Conditioning on distance from object to agent showed no significant difference in probabilities.

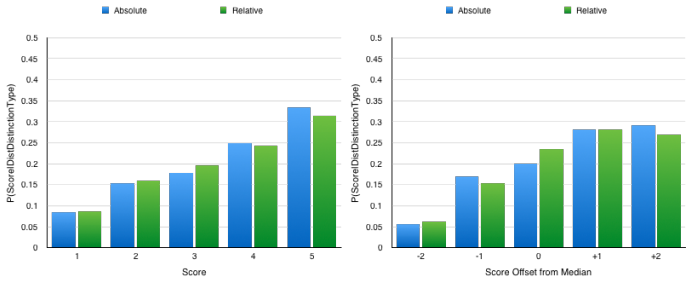

Figure 4: P(Score*j*Dist. distinction type) [L]; P(Diff from median*j*Dist. distinction type) [R]

Fig. 5 shows the probability of a particular rank given the number of relational descriptors used, for

[3] Some initial rejections on the basis of gaming the system were reversed upon subsequent communication with the worker, and these were included in the 2,228 figure.

the linguistic (L) and ensemble (R) modalities. In all cases evaluators slightly preferred 3 descriptors, and often 1 descriptor over 2 in the ensemble modality. This suggests something of a conflict between a clear if unwieldy use of 3 descriptors, and a concise single descriptor used with gesture.

Many more parameters can be analyzed for dependencies, and we have released evaluation scripts along with along with the fully-annotated dataset.[4] These initial results show clear preference for the ensemble referring method, a slight preference for absolute distance distinction as opposed to relative, and for more relational descriptors used in ensemble with gesture. We will also examine the data for dependencies between preference for number of descriptors, or distance from the agent to the target object, and the total set of relations in the scene. Modeling these will allow better ability to assess the entire scene context when generating natural referring expressions.

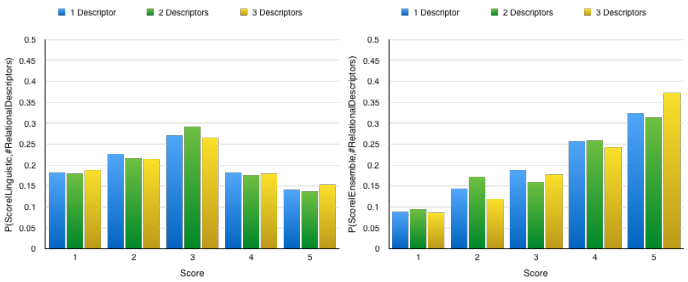

Figure 5: P(Score*j*Language only, # Descriptors) [L]; P(Score*j*Ensemble, # Descriptors) [R]

4 Conclusions and Future Work

Early analysis shows that evaluators, as a proxy for the human interlocutor, perceive references using the gesture-speech *ensemble* as more natural than unimodal referring methods. Not only does this make a convincing case for the computer to incorporate gestural output for fluent HCI, but we have also uncovered circumstances under which humans are likely to perceive the computer as referring fluently and naturally to objects in the interaction. The strongest predictor of high naturalness is the expressive modality, but there are many dependencies and we have provided techniques for uncovering those.

Going forward, we seek to use the evaluated data to train a model that can be deployed within this and other HCI systems to generate natural multimodal referring expressions in real-time, and to not only capture the strong predictors from Sec. 3, but also more subtle dependencies between contextual parameters. Some technical issues and solutions we anticipate are: **1)** Dependencies between multiple paramaters may arise from a particular configuration (e.g., two similar objects close to each other but too far from the agent to distinguish with deixis) that requires choosing a modality or level of specificity at runtime. We would suggest a convolutional neural net approach to assess relations in the scene, with gradients weighted by the information gained or lost by the addition of a particular relational descriptor for the target object; **2)** There may be cases when the avatar cannot use her hands for deixis (e.g., while holding other objects)—in this case she would need an intelligent model of linguistic-only reference to adequately single out an object in context; **3)** To capture the context of prior actions (e.g., *the green block next to the red block I just put down*), we would recommend a sequential model trained on the object configuration relation sets in the EMRE dataset, with an Approximate Nearest Neighbor (ANN) classifier between configurations in a live interaction and configurations in the data.

We have presented a novel dataset of referring techniques for definite objects in multiple configurations, with a goal of varying and combining multiple modalities available in a human-computer interaction system. As the dataset is relatively small, it should be expanded and linked to other multimodal corpora before training a publicly-deployable model, but previously we have shown that simulated data using qualitative relations is suitable for learning over smaller sample sizes (Krishnaswamy et al., 2019), and so we believe it is appropriate for training an initial model. Data like this should be of great use to researchers developing intelligent referring strategies in multimodal systems and to researchers studying multimodal semantics and referring expressions in general. After analysis, we have proposed some techniques for training models for its reuse and are currently developing experiments in which to deploy them.

[4] https://github.com/VoxML/public-data/tree/master/EMRE/HIT

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.

Ballard, D. H., M. M. Hayhoe, P. K. Pook, and R. P. Rao (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences 20*(4), 723–742.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science 15*(6), 415–419.

Bortfeld, H. and S. E. Brennan (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes 23*(2), 119–147.

Botley, S. and T. McEnery (2001). Proximal and distal demonstratives: A corpus-based study. *Journal of English Linguistics 29*(3), 214–233.

Brooks, A. G. and C. Breazeal (2006). Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pp. 297–304. ACM.

Butterworth, G. (2003). Pointing is the royal road to language for babies. In *Pointing*, pp. 17–42. Psychology Press.

Capirci, O., A. Contaldo, M. C. Caselli, and V. Volterra (2005). From action to language through gesture: A longitudinal perspective. *Gesture 5*(1), 155–177.

Claassen, W. (1992). Generating referring expressions in a multimodal environment. In *Aspects of automated natural language generation*, pp. 247–262. Springer.

Clark, H. H., R. Schreuder, and S. Buttrick (1983). Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior 22*(2), 245–258.

Coventry, K. R., D. Griffiths, and C. J. Hamilton (2014). Spatial demonstratives and perceptual space: Describing and remembering object location. *Cognitive Psychology 69*, 46–70.

Foster, M. E., E. G. Bard, M. Guhe, R. L. Hill, J. Oberlander, and A. Knoll (2008). The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pp. 295–302. ACM.

Funakoshi, K., S. Watanabe, N. Kuriyama, and T. Tokunaga (2004). Generating referring expressions using perceptual groups. In *International Conference on Natural Language Generation*, pp. 51–60. Springer.

Gatsoulis, Y., M. Alomari, C. Burbridge, C. Dondrup, P. Duckworth, P. Lightbody, M. Hanheide, N. Hawes, D. Hogg, A. Cohn, et al. (2016). Qsrlib: a software library for online acquisition of qualitative spatial relations from video.

Gatt, A., R. van Gompel, E. Krahmer, and K. van Deemter (2011). Non-deterministic attribute selection in reference production. In *Proceedings of the 2nd PRE-Cog Sci Workshop (Boston, MA)*.

Hayashi, M. (2004). Projection and grammar: notes on the action-projectinguse of the distal demonstrative are in japanese. *Journal of pragmatics 36*(8), 1337–1374.

Kelleher, J. D. and G.-J. M. Kruijff (2006). Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 1041–1048. Association for Computational Linguistics.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Krahmer, E. and I. van der Sluis (2003). A new model for generating multimodal referring expressions. In *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003) at EACL 2003*.

Krishnaswamy, N., S. Friedman, and J. Pustejovsky (2019). Combining Deep Learning and Qualitative Spatial Reasoning to Learn Complex Structures from Sparse Examples with Noise. In *AAAI Conference on Artificial Intelligence (AAAI)*. AAAI.

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 (2016). Multimodal semantic simulations of linguistically underspecified motion events. In *Spatial Cognition X: International Conference on Spatial Cognition*. Springer.

Krishnaswamy, N. and J. Pustejovsky (2018). An evaluation framework for multimodal interaction. *Proceedings of LREC*.

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 (2009a). Discourse coherence and gesture interpretation. *Gesture 9*(2), 147–180.

Lascarides, A. and M. Stone (2009b). A formal semantic analysis of gesture. *Journal of Semantics*, ffp004.

Latoschik, M. E. and I. Wachsmuth (1997). Exploiting distant pointing gestures for object selection in a virtual environment. In *International Gesture Workshop*, pp. 185–196. Springer.

Meira, S. (2003). addressee effects in demonstrative systems. *Deictic conceptualisation of space, time, and person 112*, 3.

Moratz, R., B. Nebel, and C. Freksa (2002). Qualitative spatial reasoning about relative position. In *International Conference on Spatial Cognition*, pp. 385–400. Springer.

Moratz, R. and T. Tenbrink (2006). Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial cognition and computation 6*(1), 63–107.

Narayana, P., N. Krishnaswamy, I. Wang, R. Bangar, D. Patil, G. Mulay, K. Rim, R. Beveridge, J. Ruiz, J. Pustejovsky, and B. Draper (2018). Cooperating with avatars through gesture, language and action. In *Intelligent Systems Conference (IntelliSys)*.

Oviatt, S. (1999). Ten myths of multimodal interaction. *Communications of the ACM 42*(11), 74–81.

Piwek, P., R.-J. Beun, and A. Cremers (2008). proximal and distal in language and cognition: Evidence from deictic demonstratives in dutch. *Journal of Pragmatics 40*(4), 694–718.

Proulx, P. (1988). The demonstrative pronouns of proto-algonquian. *International journal of American linguistics 54*(3), 309–330.

Pustejovsky, J. (2018). From actions to events. *Interaction Studies 19*(1-2), 289–317.

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, A. Cohn, B. Nebel, C. Rich, and W. Swartout (1992). A spatial logic based on regions and connection. In *KR'92. Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference*, San Mateo, pp. 165–176. Morgan Kaufmann.

Schneider, R. (2014). *Convex bodies: the Brunn–Minkowski theory*. Number 151. Cambridge university press.

Skubic, M., D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock (2004). Spatial language for human-robot dialogs. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 34*(2), 154–167.

Sluis, I. v. d. and E. Krahmer (2004). The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. In *Eighth International Conference on Spoken Language Processing*.

Strauss, S. (2002). This, that, and it in spoken american english: a demonstrative system of gradient focus. *Language Sciences 24*(2), 131–152.

Viethen, J. and R. Dale (2008). The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, pp. 59–67. Association for Computational Linguistics.

Volterra, V., M. C. Caselli, O. Capirci, and E. Pizzuto (2005). Gesture and the emergence and development of language. *Beyond nature-nurture: Essays in honor of Elizabeth Bates*, 3–40.

Whitney, D., M. Eldon, J. Oberlin, and S. Tellex (2016). Interpreting multimodal referring expressions in real time. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pp. 3331–3338. IEEE.