Do You See What I See? Effects of POV on Spatial Relation Specifications

Nikhil Krishnaswamy, James Pustejovsky
Department of Computer Science
Brandeis University
Waltham, MA 02453 USA
{nkrishna,jamesp}@brandeis.edu

Abstract

In this paper, we examine a set of object interactions generated with a 3D natural language simulation and visualization platform, VoxSim (Krishnaswamy and Pustejovsky 2016b). These simulations all realize the natural language relations "touching" and "near" over a test set of various objects within a 3-dimensional world that interprets descriptions of motion events and renders their visual instantiations from the perspective of an embodied virtual agent. These object interactions were evaluated by human judges using Amazon Mechanical Turk and we examine some of the qualitative interpretations provided by humans over these computer-generated interpretations of underspecified relations, conditioned on the frame of reference (agent's point of view) and object position relative to that point of view (POV). Through analysis of the human evaluations, we find that average evaluator satisfaction with many specifications for these relations appears to strongly depend on the relationship between the two objects and between the objects and the POV.

Introduction

While the idea of simulation in cognitive linguistics has become more popular over the past decade (Bergen 2012; Lakoff 2009), it has not been adopted into broad usage, due in part to arguments against the efficacy of simulation in explaining natural language understanding (Davis and Marcus 2016), particularly regarding linguistic phenomena involving continuous ranges or underspecified values. Here, we argue that simulation, when modeled within a dynamic qualitative spatial and temporal semantics, can provide a robust environment for examining the interpretation of linguistic behaviors, including those described qualitatively.

Qualitative representation is well-suited to handling questions of linguistic underspecification, and such questions occur frequently in the process of natural language understanding, as for any linguistic predicate, the amount and nature of information it provides varies. "Put the ball near the plate" is a perfectly valid sentence, one that can be "mentally simulated," and while values such as distance between the objects and relative orientation *can* be further specified, the minimal model of the event does not require it. However, for any such event enacted in the real world, these unstated or underspecified parameters have values, even if they are not measured. Thus there may be a set of values $[a]$ for a parameter underspecified in a sentence s for which the resulting event represents a proposition p such that $\mathcal{M} \vDash p_s[a]$ and another set of values $[b]$ which results in a proposition p such that $\mathcal{M} \nvDash p_s[b]$. Attempting to separate the two sets computationally entails two tasks:

- Building a computationally coherent model of a world that can be evaluated from an embodied human perspective;
- Determining which values result in an enactment of an event that satisfies a human judge's notion of that event.

Such an approach allows us to directly map a linguistic modality of expression through a dynamic semantics (Mani and Pustejovsky 2012) to a visual modality.

A verbal or relational predicate may impose constraints on the event or relation it describes. These constraints may be on the path or manner of motion (Pustejovsky and Moszkowicz 2011), or alternately on the resultant state of the event. For instance, the relations "touching" and "near," applied to objects, impose different constraints on the distance between them, while leaving relative orientation completely open, in principle.

A wealth of prior work exists on the role of orientation in Qualitative Spatial Reasoning (QSR) (Freksa 1992; Moratz, Renz, and Wolter 2000; Dylla and Moratz 2004; Renz and Nebel 2007), on QR as an information-bearer in situations with underspecified or incomplete knowledge (Kuipers 1994; Joskowicz and Sacks 1991), or on using cardinal directions or path knowledge in QSR (Frank 1996; Zimmermann and Freksa 1996). More recent work leverages QSR to improve machine learning (Falomir and Kluth 2017), object manipulation or environment navigation (Thrun et al. 2000; Rusu et al. 2008), all areas which this work can inform.

Kuipers (2000) contains formalisms for melding local geometric frames of reference into global frames, however most research into the intersection of frames of reference and QSR, e.g., Frank (1992), are appropriately proposed as algebraic formalisms, fitting with existing usage of qualitative representations to handle linguistic underspecification. In this application paper, we reverse the approaches used

in some of this work, inferring the effects of POV on human judgments from observed data, and believe this paper is well-situated to introduce a method of using multimodal simulation to address some of these evaluation issues. Our methodology implements and tests QR theories in real time and our results show that variations in relative point of view often introduce uncertainties in evaluators' satisfaction with the visualizations of these relational predicates, suggesting that the most satisfactory or prototypical instantiations of these QSR relations are ones that are minimally-dependent on point of view.

Perspective and Embodiment in a Multimodal Simulation

Central to understanding the role of perspective (point of view) in a multimodal simulation is the notion of *embodiment* (Gibbs Jr. 2005; Ziemke 2003; Kiela et al. 2016). Embodiment has many interpretations, but here we view it as a model generation constraint. Meaning centrally involves the activation of perceptual, motor, social, and affective knowledge that characterizes the content of utterances. Understanding a piece of language is hypothesized to entail performing mental perceptual and motor simulations of its content. Thus the requirements on a "multimodal simulation semantics" include, but are not limited to, the following components (Pustejovsky and Krishnaswamy 2016a):

- A minimal embedding space (MES) for the simulation must be determined. This is the 3D region within which the state is configured or the event unfolds (Pustejovsky and Krishnaswamy 2014);
- Object-based attributes for participants in a situation or event need to be specified; e.g., orientation, relative size, default position or pose, etc. (Pustejovsky 2013);
- An epistemic condition on the object and event rendering, imposing an implicit point of view (POV) (Levinson 2003);
- Agent-dependent embodiment; this determines the relative scaling of an agent and its event participants and their surroundings, as it engages in the environment (Krishnaswamy and Pustejovsky 2016a).

We expect these assumptions to apply in a broad sense across simulation-based approaches with the current technology.

In a visualization, the agent embodiment and associated point of view is represented at least minimally by the in-world camera, the point relative to which all graphical rendering is evaluated. When visualizations are presented to human judges, the established camera perspective then becomes their point of view, making the point of view a crucial variable in the effective representation of relative spatial relations.

The camera in a simulated visual world may be allowed to freely roam, be restricted to a fixed position, or be attached to a virtual agent. These variations in perspective allow a human interacting with or watching the simulation to share the same visual context as the computer, and to view it and the objects contained therein from a variable relative frame. If constrained to an agent, this provides a way of looking inside the agent's "brain" and perceiving what it perceives.

In order to control for the variable frame of reference while exercising a variety of qualitative spatial relations over the course of the event simulations, the camera was constrained to a fixed absolute perspective, while the objects being simulated were initialized in a grid pattern. This means that for any given object pair, the simulated motion event maintained the same relative relations between the respective objects and the camera, while transforming the relations between the objects relative to each other. Thus during evaluation, we could condition on object-to-object orientation and distance relative to the frame of reference.

VoxML and VoxSim Overview

Previously, we developed a concept modeling and visualization language, **VoxML**, to describe and encode qualitative and geometrical knowledge about objects and events that is presupposed in linguistic utterances but not made explicit in a visual modality (Pustejovsky and Krishnaswamy 2016b). This includes information about symmetry or concavity in an object's physical structure, the relations entailed by the occurrence of an event in a narrative, the qualitative relations described by a positional adjunct, or behaviors afforded by an object's *habitat* (Pustejovsky 2013; McDonald and Pustejovsky 2014).

Relational information in VoxML can be encoded using a variety of QSR calculi such as the situation calculus (Bhatt and Loke 2008), the Intersection Calculus (Kurata and Egenhofer 2007; Mark and Egenhofer 1995), or the Region Connection Calculus (RCC) (Randell et al. 1992) and 3D variants (Albath et al. 2010). Leveraging these spatial frameworks allow relations to be tested for at runtime using simple positional metrics, and then composed with object properties.

$$\left[\begin{array}{l}\textbf{on}\\ \text{LEX} = \left[\text{PRED} = \textbf{on}\right]\\ \text{TYPE} = \left[\begin{array}{l}\text{CLASS} = \textbf{config}\\ \text{VALUE} = \textbf{EC}\\ \text{ARGS} = \left[\begin{array}{l}\text{A}_1 = \textbf{x:3D}\\ \text{A}_2 = \textbf{y:3D}\end{array}\right]\\ \text{CONSTR} = \textbf{y}\rightarrow\text{HABITAT}\rightarrow\text{INTR}[align]\end{array}\right]\end{array}\right]$$

$$\left[\begin{array}{l}\textbf{in}\\ \text{LEX} = \left[\text{PRED} = \textbf{in}\right]\\ \text{TYPE} = \left[\begin{array}{l}\text{CLASS} = \textbf{config}\\ \text{VALUE} = \textbf{PO} \parallel \textbf{TPP} \parallel \textbf{NTPP}\\ \text{ARGS} = \left[\begin{array}{l}\text{A}_1 = \textbf{x:3D}\\ \text{A}_2 = \textbf{y:3D}\end{array}\right]\\ \text{CONSTR} = \textbf{y}\rightarrow\text{HABITAT}\rightarrow\text{INTR}[align]\textbf{?}\end{array}\right]\end{array}\right]$$

Figure 1: VoxML structures for "on" and "in" using qualitative RCC relations.

For example, from Fig. 1, knowing that "on" describes a relation of external connection, we can then semantically compose this with an object argument and test the conditions under which such a relation would be realized. For

example, if the object scoped by "on" is concave, another object placed in the concavity would be "on" that object, such as putting an apple on a plate, provided the plate is in an environment or *habitat* that *affords* that action.

Figure 2: Visualization of an apple being placed on a plate

Using a different object, such as a cup, would require the agent or simulation system to put the apple *in* the cup in order to achieve an analoguous configuration, due to differences in the object's structure, and hence its affordances. In this configuration, the object would at least partially interpenetrate the total bounds of the cup.

The examples of motion events referred to herein were generated using the software package **VoxSim**. VoxSim (Krishnaswamy and Pustejovsky 2016a; Krishnaswamy and Pustejovsky 2016b) is a semantically-informed 3D visual event simulator built on top of the VoxML framework using the Unity game engine (Goldstone 2009). Using Unity's capability in subsystems like graphics processing, UI, and physics, we built systems to handle language processing, theoretical reasoning, and AI in a real-time game or "game-like" environment, in the vein of work presented by Forbus et al. (2002) and Dill (2011).

VoxSim dynamically generates animated visualizations of motion events using real-world knowledge about said events and their object participants, marked up in VoxML. VoxSim exploits VoxML-encoded information about an object's *habitat*, or situational context that enables or disables certain *affordances* or actions that may be undertaken using the object. These affordances may be either *Gibsonian* or *telic* in nature (Gibson 1977; Gibson 1979; Pustejovsky 1995). Gibsonian affordances are those that emerge from an object's physical structure, such as "grasp" or "contain," while telic affordances are goal-directed affordances that emerge from the Gibsonian affordances (e.g., a cup can contain liquid, so it might have a telic affordance of "drink from"). Both types of affordances are important for determining what spatial relations a pair of objects must satisfy as the result of undergoing a transformation interpreted from a natural language command.

More information about VoxML, VoxSim, and implementational details of these considerations may be found at http://www.voxicon.net. The latest VoxSim source code is available at https://github.com/VoxML/VoxSim.

Data Acquisition

Objects		
block	book	banana
ball	blackboard	bowl
plate	bottle	knife
cup	grape	pencil
disc	apple	paper sheet

Table 1: Object test set

Using VoxSim, we generated 1210 individual videos of objects listed in Table 1 interacting in the qualitative, underspecified relations "touching" and "near." In each video, objects were moved through the scene without being affected by an agent, as shown in Fig. 4. Inputs to the simulator were all given in the form "put the x {touching, near} the y," such as "put the ball touching the block" or "put the apple near the bottle." As "touching" and "near" are both underspecified spatial relations, further value specification is required within the simulator to generate a distinct rendering of the event, such that the simulator must choose a specific relation for *touching(y)* before generating a "put x touching y" visualization, or must choose a location judged to be *near(y)* before generating a "put x near y" visualization.

Predicate	**Underspecified parameters**	**Possible values**
touching(x)	rel orientation	{*left(x)*, *right(x)*, *behind(x)*, *in_front(x)*, *on(x)*}
near(x)	transloc dir	$V \in \{\langle y\text{-x}(x), y\text{-y}(x), y\text{-z}(x)\rangle \mid \text{d}(x,y) < \text{d}(\text{edge}(\text{s}(y),y)), IN(\text{s}(y)), \neg IN(y)\}$

Table 2: Predicate value assignments

For "touching," we provided the system with a five-way choice between other relations that specified direction and orientation. The substitute predicates for "touching" were all assumed to be axis-aligned between the objects involved, and to include an EC connection per the RCC (that is, *left(x)* is operationalized as "left and touching"). The perspective-dependent relations *left*, *right*, *in_front* and *behind* were all computed relative to the camera's point of view, allowing us to condition evaluation of the results against this parameter.

For "near," d is a linear distance function and $s(x)$ is the object surface that supports the object x. Thus V represents a 3-vector denoting a point on the test surface that is closer to the target object (the argument of "near") than to the closest edge of the surface.

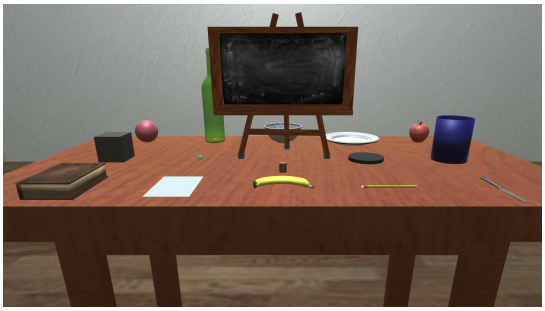

Figure 3: Test environment with all objects shown

Video was captured for each generated visualization, in which these parameters were randomly assigned using a Monte Carlo method, and logged to a database. The videos were submitted to Amazon Mechanical Turk for evaluation. In each human intelligence task (HIT), workers were asked to select which of three videos best depicts the input sentence that was used to generate all three. However, to allow for the possibility that the parameter being varied across

the three visualizations might actually be immaterial to the question of whether or not the visualization adequately depicts the utterance, we allowed evaluators to choose multiple answers if more than one video depicted the utterance equally well. We also allowed evaluators to choose "none" in situations where they thought none of the provided visualizations acceptably depicted the description. Evaluators often took advantage of these options. The raw results are therefore assumed to reflect the overall incidence of evaluators *accepting* a given visualization for the provided utterance. We therefore discuss the statistically evaluated results in terms of "acceptability judgments."

Each HIT was completed by 8 individual workers, totalling 3236 individual tasks evaluating visualizations of the predicates "touching" and "near."[1]

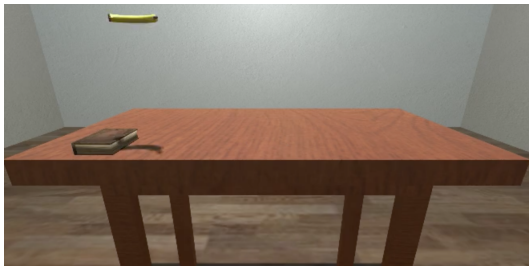

Figure 4: Sample object motion as might be seen during a visualization of "put the banana touching the book" or "put the banana near the book". During capture of an event, all objects not mentioned in the input sentence were removed.

Evaluation Metrics

Scene visualization work is not well-reflected in current evaluation, due to sparsity of datasets and lack of a general-domain gold standard (Johansson et al. 2005). Thus we have to determine *a priori* what metrics are most informative when assessing human judgments of events.

Human judgments of a visualization are given as "acceptable" or "unacceptable" relative to the event's linguistic description. While the precise values of the object coordinates and relative offsets are necessary for the computer to calculate and render the visuals, these are of less interest with regard to the viewer's acceptability judgment than the qualitative assessment of relations between the objects.

For "touching," we provided the system with five qualitative relations to choose from to specify the event, so we can assess the probability that for an arbitrary visualization, an arbitrary judge would judge it acceptable for its input sentence, conditioned on a) the relation between the objects at the end of the event, and b) the motion of the moving object through the event (e.g., the object moves from behind the stationary object to the right of it). These can be assessed relative to the established POV, as discussed above in the section on perspective and embodiment. When computing the relations that apply between the two objects at the start of the event, certain object pairs have two extant relations (e.g., "left" and "behind") since, as seen in Fig. 3, the objects are arranged in a grid pattern and not every object is necessarily axis-aligned with every other object at the event start. At the event's end, only one relation should apply.

For "near," we can condition on the same parameters but this misses the crucial question of distance, which is selected for in the linguistic predicate. As given in the value assignment table (Table 2), the resulting distance between the two objects in a "put near" event can fall in a continuous range subject to constraints (closer to the stationary object than to the edge of the table, not touching the stationary object). Unity generates the target placement of the moving object using a uniform distribution, in line with standard Monte Carlo methods (Sawilowsky 2003), although instances may be resampled if the location generated violates one of the aforementioned constraints. This allows us to plot all the distances that occur in the dataset for "put near" events as a probability density over the magnitude function of continuous random variable V (Table 2), partitioned into subsets. For this evaluation, we use quintiles (q = 5), although data using different quantiles can be easily generated by passing a different parameter to the evaluation script. We condition on ending orientation and ending distance between the moving object and stationary object, where the interval (0, $QU1$) represents the smallest 20% of distances in the dataset.

Results

"Touching"

QSR (start)	P(acc|QSR)	QSR (end)	P(acc|QSR)
behind(y)	0.5497	*behind(y)*	0.5474
in_front(y)	0.5692	*in_front(y)*	0.5816
left(y)	0.5753	*left(y)*	0.4995
right(y)	0.5725	*right(y)*	0.5560
on(y)	N/A	*on(y)*	0.6683
$\mu_{start} \approx 0.5667$		$\mu_{end} \approx 0.5725$	
$\sigma_{start} \approx 0.0116$		$\sigma_{end} \approx 0.0628$	

Table 3: Acceptability judgments and statistical metrics for "put *x* touching *y*" visualizations, conditioned on relations between *x* and *y* at event start and completion

Movement (M)	P(acc|M)
behind→behind(y)	0.5347
behind→in_front(y)	0.4758
behind→left(y)	0.5014
behind→right(y)	0.4888
behind→on(y)	0.7453
in_front→behind(y)	0.4523
in_front→in_front(y)	0.6447
in_front→left(y)	0.4601
in_front→right(y)	0.5756
in_front→on(y)	0.6234
left→behind(y)	0.5732
left→in_front(y)	0.5853
left→left(y)	0.5266
left→right(y)	0.5211
left→on(y)	0.6492
(cont'd)	(cont'd)

[1] A small number of responses were rejected due to evaluators failing to answer the required question.

(cont'd)	(cont'd)
Movement (M)	**P(acc|M)**
right→behind(y)	0.5406
right→in_front(y)	0.5786
right→left(y)	0.4777
right→right(y)	0.5847
right→on(y)	0.7081

$\mu_M \approx 0.5624$	$\sigma_M \approx 0.0811$
$\mu_{\rightarrow beh} \approx 0.5252$	$\sigma_{\rightarrow beh} \approx 0.0515$
$\mu_{\rightarrow fr} \approx 0.5711$	$\sigma_{\rightarrow fr} \approx 0.0701$
$\mu_{\rightarrow l} \approx 0.4911$	$\sigma_{\rightarrow l} \approx 0.0289$
$\mu_{\rightarrow r} \approx 0.5426$	$\sigma_{\rightarrow r} \approx 0.0455$
$\mu_{\rightarrow on} \approx 0.6815$	$\sigma_{\rightarrow on} \approx 0.0554$

Table 4: Acceptability judgments and statistical metrics for "put *x* touching *y*" visualizations, conditioned on *x* movement relative to *y*

"Near"

Dist (start)	**P(acc|QU)**	**Dist (end)**	**P(acc|QU)**
(0,$QU1$)	N/A	(0,$QU1$)	0.7523
($QU1$,$QU2$)	0.3542	($QU1$,$QU2$)	0.6207
($QU2$,$QU3$)	0.3829	($QU2$,$QU3$)	0.3890
($QU3$,$QU4$)	0.4444	($QU3$,$QU4$)	0.3655
($QU4$,∞)	0.4470	($QU4$,∞)	0.1295

$\mu_{start} \approx 0.4071$	$\mu_{end} \approx 0.4514$
$\sigma_{start} \approx 0.0461$	$\sigma_{end} \approx 0.2419$

Table 5: Acceptability judgments and statistical metrics for "put *x* near *y*" visualizations, conditioned on distance between *x* and *y* at event start and completion

Movement (M)	**P(acc|M)**
($QU1$,$QU2$)→(0,$QU1$)	0.7625
($QU1$,$QU2$)→($QU1$,$QU2$)	0.4044
($QU1$,$QU2$)→($QU2$,$QU3$)	0.2232
($QU1$,$QU2$)→($QU3$,$QU4$)	0.1667
($QU1$,$QU2$)→($QU4$,∞)	0.0682
($QU2$,$QU3$)→(0,$QU1$)	0.6848
($QU2$,$QU3$)→($QU1$,$QU2$)	0.5703
($QU2$,$QU3$)→($QU2$,$QU3$)	0.3750
($QU2$,$QU3$)→($QU3$,$QU4$)	0.2788
($QU2$,$QU3$)→($QU4$,∞)	0.1488
($QU3$,$QU4$)→(0,$QU1$)	1.000
($QU3$,$QU4$)→($QU1$,$QU2$)	0.3750
($QU3$,$QU4$)→($QU2$,$QU3$)	0.3750
($QU3$,$QU4$)→($QU3$,$QU4$)	0.5417
($QU3$,$QU4$)→($QU4$,∞)	0.2083
($QU4$,∞)→(0,$QU1$)	0.7698
($QU4$,∞)→($QU1$,$QU2$)	0.6863
($QU4$,∞)→($QU2$,$QU3$)	0.4217
($QU4$,∞)→($QU3$,$QU4$)	0.4162
($QU4$,∞)→($QU4$,∞)	0.1300

$\mu_M \approx 0.4303$	$\sigma_M \approx 0.2521$
$\mu_{\rightarrow (0,QU1)} \approx 0.8043$	$\sigma_{\rightarrow (0,QU1)} \approx 0.1360$
$\mu_{\rightarrow (QU1,QU2)} \approx 0.5090$	$\sigma_{\rightarrow (QU1,QU2)} \approx 0.1462$
$\mu_{\rightarrow (QU2,QU3)} \approx 0.3487$	$\sigma_{\rightarrow (QU2,QU3)} \approx 0.0865$
$\mu_{\rightarrow (QU3,QU4)} \approx 0.3509$	$\sigma_{\rightarrow (QU3,QU4)} \approx 0.1631$
$\mu_{\rightarrow (QU4,QU5)} \approx 0.1388$	$\sigma_{\rightarrow (QU4,QU5)} \approx 0.0577$

Table 6: Acceptability judgments and statistical metrics for "put *x* near *y*" visualizations, conditioned on start and end distance intervals between *x* and *y*

Dist (end)	**QSR**	**P(acc|QU,QSR)**
(0,$QU1$)	*behind(y)*	0.7730
(0,$QU1$)	*in_front(y)*	0.7349
(0,$QU1$)	*left(y)*	0.7338
(0,$QU1$)	*right(y)*	0.7712
($QU1$,$QU2$)	*behind(y)*	0.6701
($QU1$,$QU2$)	*in_front(y)*	0.5797
($QU1$,$QU2$)	*left(y)*	0.6675
($QU1$,$QU2$)	*right(y)*	0.5819
($QU2$,$QU3$)	*behind(y)*	0.4151
($QU2$,$QU3$)	*in_front(y)*	0.3644
($QU2$,$QU3$)	*left(y)*	0.3945
($QU2$,$QU3$)	*right(y)*	0.3825
($QU3$,$QU4$)	*behind(y)*	0.1713
($QU3$,$QU4$)	*in_front(y)*	0.4308
($QU3$,$QU4$)	*left(y)*	0.2093
($QU3$,$QU4$)	*right(y)*	0.4699
($QU4$,∞)	*behind(y)*	0.0972
($QU4$,∞)	*in_front(y)*	0.1401
($QU4$,∞)	*left(y)*	0.1250
($QU4$,∞)	*right(y)*	0.1348

$\mu_{end,qsr} \approx 0.4424$	$\sigma_{end,qsr} \approx 0.2380$
$\mu_{end=(0,QU1)} \approx 0.7532$	$\sigma_{end=(0,QU1)} \approx 0.0218$
$\mu_{end=(QU1,QU2)} \approx 0.6248$	$\sigma_{end=(QU1,QU2)} \approx 0.0508$
$\mu_{end=(QU2,QU3)} \approx 0.3891$	$\sigma_{end=(QU2,QU3)} \approx 0.0213$
$\mu_{end=(QU3,QU4)} \approx 0.3203$	$\sigma_{end=(QU3,QU4)} \approx 0.1518$
$\mu_{end=(QU4,QU5)} \approx 0.1243$	$\sigma_{end=(QU4,QU5)} \approx 0.0191$
$\mu_{qsr=beh)} \approx 0.4253$	$\sigma_{qsr=beh} \approx 0.2971$
$\mu_{qsr=fr} \approx 0.4500$	$\sigma_{qsr=fr} \approx 0.2246$
$\mu_{qsr=l} \approx 0.4260$	$\sigma_{qsr=l} \approx 0.2700$
$\mu_{qsr=r} \approx 0.4681$	$\sigma_{qsr=r} \approx 0.2362$

Table 7: Acceptability judgments and statistical metrics for "put *x* near *y*" visualizations, conditioned on distance between *x* and *y* and POV-relative orientation at event completion

Discussion

"Touching"

We observe a lower likelihood for visualizations to be judged acceptable when the moving object moves from behind the stationary object to in front of it, and vice versa. P(accept|*behind→in_front(y)*) is approximately 0.4758, which is approximately 1.07 standard deviations below the mean of the population for all starting/ending QSR relation pairs. This may be explained as an effect of the point of view imposed by the camera position, which may make it difficult to see if an object behind another object is actually making contact and satisfying the EC relation required by "touching," especially if a larger object is occluding a smaller object.

Visualizations where the moving object ends to the left of the stationary object were also less likely to be judged acceptable. P(accept|*left(y)*) is approximately 1.16 standard deviations below the mean likelihood of acceptance over the population for all event-end QSR relations. This is apparently independent of the moving object's starting location relative to the stationary object, but the dispreference is more significant for objects that start in front of, or to the right of, their destination.

- P(acc|*in_front→left(y)*) $\approx 0.4601 \approx \mu_M$ - $1.26\sigma_M$

- P(acc|*right→left(y)*) ≈ 0.4777 ≈ μ_M - 1.04σ_M

This could also be explained as an effect of the POV, in particular the distortion it causes in cases where larger objects closer to the camera (including laterally) may occlude objects further away, making it difficult to assess the satisfaction of the EC relation. Therefore, some objects that move from the right of another object to the left of it also move away from the camera, meaning that this effect is analogous to that seen in the *behind(y)* relations, and explains the similar result seen for *in_front→left(y)* motions. However, this hypothesis would not explain the absence of a symmetric inclination against *right(y)* relations so more experimentation or analysis is needed. Some of this may be related to features of the objects themselves, which are not strongly controlled for (discussed further in section on future directions).

There is a strong preference for the *on(y)* specification of *touching(y)* over all others, which matches linguistic intuition. "On" necessarily implies an EC relation, which is expressed in the VoxML (Fig. 1). P(accept|*on(y)*) falls approximately 1.52 standard deviations above the mean probability of acceptability of the population for all event-end relations. The strongest preference is for motion from *behind(y)* to *on(y)*, where P(accept|*behind→on(y)*) is approximately *2.25* standard deviations above the mean likelihood for acceptability over the whole population conditioned on start-to-end motion. In terms of point of view effects, this may be due to an occluded object being brought into view and very obviously made to touch its destination in a visualization with no obstructed view. Where "touching" is an underspecified predicate, the relations entailed by "on," while arguably somewhat overspecified as an interpretation of "touching" alone, seem to most clearly satisfy the qualitative specification of "touching" out of the options available. Notably, it is the only one *not* dependent on the relative point of view, suggesting that the relative point of view introduces some noise or confusion into the human judgeents, potentially for the reasons discussed above, among others.

"Near"

Unsurprisingly, evaluators preferred visualizations where the two objects ended up close to each other to those where the objects ended further apart.

- P(acc|$(0, QU1)$) ≈ μ_{end} + 1.24σ_{end}
- P(acc|$(QU1, QU2)$) ≈ μ_{end} + 0.70σ_{end}

In the first three distance intervals, we observe a slight preference for events where the moving object finishes the event behind the stationary object.

- P(acc|$(0, QU1)$,*behind(y)*) ≈ $\mu_{end=(0,QU1),qsr}$ + 0.90$\sigma_{end=(0,QU1),qsr}$
- P(acc|$(QU1, QU2)$,*behind(y)*) ≈ $\mu_{end=(QU1,QU2),qsr}$ + 0.89$\sigma_{end=(QU1,QU2),qsr}$
- P(acc|$(QU2, QU3)$,*behind(y)*) ≈ $\mu_{end=(QU2,QU3),qsr}$ + 1.22$\sigma_{end=(QU2,QU3),qsr}$

This may be an effect of foreshortening caused by the point of view, as with some of the "touching" specifications, which causes an object *x* which is *behind(y)* to appear closer to *y* than it actually is.

When conditioning on the joint distribution of the distance interval and the QSR relation, as shown in Table 7, there is some apparent confusion in judgments of events in the fourth distance interval, where σ for the population of P(accept|QSR) is greater than .15, where in all other intervals σ for P(accept|QSR) falls between .019 and .051. This is possibly a factor of workers being unable to judge purely from the visuals whether an object that began its movement from a position in the fourth distance interval relative to the stationary object actually ended the event nearer than it began, whereas in preceding intervals, the resulting location was more likely to be unambiguously "near" regardless of starting location.

Table 6 shows the judges' preferences for objects that moved between the different distance intervals, independent of direction or orientation. The quintiles were calculated based on the distributions of distances between objects at the *end* of the "put near" event, which is why Tables 5 and 6 show no objects beginning the event in the lowest distance interval. There is a clear preference for objects that move from a far interval to a near one, and the inverse is also true, with very low proportions of "acceptable" judgments for visualizations where the object moved from a near distance interval to a farther one. This reinforces the intuition that a qualitative term like "near" is understood to be inherently relative (Peters 2007).

Conclusions and Future Directions

As we opted for a focus on relational predicates, our current evaluation does not control for features of the individual objects, such as size. These features may be signaled in some of the feature vectors used to generate the simulations, but the signal is likely too weak to emerge without explicitly conditioning on them. Object features may explain some of the results above, particularly with regard to the occlusion of objects by other objects and the effects that has on the human judgments of the visualization's acceptability with respect to the input sentence. As object occlusion is a direct effect of both point of view and object size, such additional controlling may be able to further inform the relevance of POV to QSR judgments. Additionally, some the questions that are currently unanswered about this dataset, such as asymmetries in human judgments (such as the dispreference for "left" as a specification of "touching" without a similar dispreference for "right"), object size, and visual occlusion, can be subject to similar evaulation methods that specifically target those parameters as this paper targeted underspecified predicates using a simulation method, perhaps with a larger sample of evaluators.

Some of the POV-related effects on the acceptability probabilities shown may also be somewhat affected by the simulation environment. Viewing any 3D scene on a flat screen introduces small distortions in perception due the stereoscopic effects of rendering three dimensions in a 2D space (Wann, Rushton, and Mon-Williams 1995), but many of these, such as foreshortening, are also properties of nat-

ural binocular vision. These may simply be exaggerated by the virtual environment rather than introduced by it entirely. Using an immersive virtual reality display may or may not alleviate these issues, as VR technology is currently wrestling with its own stereoscopic artifacts, but evaluation using methodology like this, done with a VR headset rather than a flat screen, would provide an interesting point to test this hypothesis against.

Simulation and visualization of events provides a method of rapidly generating data for human evaluation, enabling broader investigation into human spatial cognition and reasoning. A natural language interface like that provided by VoxSim allows the generation of such experimental data without specialized skillsets. This has been a goal of many text-to-scene systems (Coyne and Sproat 2001; Seversky and Yin 2006; Chang et al. 2015), but VoxSim's implementation of motion semantics additionally allows it to be used as a platform to conduct experiments on the observables of motion events. The shared visual context between human and computer forces the handling of object embodiment (Pustejovsky, Krishnaswamy, and Do 2017), allowing researchers to examine the effects of nonlinguistic qualitative parameters such as point of view, as we have done.

This line of research has potential applications to QR theory in the realm of robotics, following in the vein of Moratz et al. (2001), by using a multimodal simulation environment to create a dynamic internal representation of the qualitative relations between objects that exist in a real-world scene. With an embodied virtual agent whose structure is isomorphic to the structure of a physical robotic agent, simulation provides a platform to conduct qualitative and probabilistic reasoning "live" as it moves through the world represented by its internal scene, able to condition on and control for effects of POV like those revealed in this paper.

In conclusion, we believe simulation-based approaches can serve the QR community as a means for computationally implementing QR theory in real-time applications. They also provide a method of testing QR frameworks and establishing boundaries on the types of parameter spaces, such as relative point of view and linguistic underspecification, where qualitative approaches might best their quantitative counterparts in informativity and robustness.

Acknowledgements

We would like to thank the reviewers for their insightful comments. 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

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

[2012] Bergen, B. K. 2012. *Louder than words: The new science of how the mind makes meaning*. Basic Books.

[2008] Bhatt, M., and Loke, S. 2008. Modelling dynamic spatial systems in the situation calculus. *Spatial Cognition and Computation* 86–130.

[2015] Chang, A.; Monroe, W.; Savva, M.; Potts, C.; and Manning, C. D. 2015. Text to 3D scene generation with rich lexical grounding. *arXiv preprint arXiv:1505.06289*.

[2001] Coyne, B., and Sproat, R. 2001. WordsEye: an automatic text-to-scene conversion system. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 487–496. ACM.

[2016] Davis, E., and Marcus, G. 2016. The scope and limits of simulation in automated reasoning. *Artificial Intelligence* 233:60–72.

[2011] Dill, K. 2011. A game AI approach to autonomous control of virtual characters. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*.

[2004] Dylla, F., and Moratz, R. 2004. Exploiting qualitative spatial neighborhoods in the situation calculus. In *International Conference on Spatial Cognition*, 304–322. Springer.

[2017] Falomir, Z., and Kluth, T. 2017. Qualitative spatial logic descriptors from 3d indoor scenes to generate explanations in natural language. *Cognitive Processing* 1–20.

[2002] Forbus, K. D.; Mahoney, J. V.; and Dill, K. 2002. How qualitative spatial reasoning can improve strategy game AIs. *IEEE Intelligent Systems* 17(4):25–30.

[1992] Frank, A. U. 1992. Qualitative spatial reasoning about distances and directions in geographic space. *Journal of Visual Languages & Computing* 3(4):343–371.

[1996] Frank, A. U. 1996. Qualitative spatial reasoning: Cardinal directions as an example. *International Journal of Geographical Information Science* 10(3):269–290.

[1992] Freksa, C. 1992. *Using orientation information for qualitative spatial reasoning*. Springer.

[2005] Gibbs Jr., R. W. 2005. *Embodiment and cognitive science*. Cambridge University Press.

[1977] Gibson, J. J. 1977. The theory of affordances. *Perceiving, Acting, and Knowing: Toward an ecological psychology* 67–82.

[1979] Gibson, J. J. 1979. *The Ecology Approach to Visual Perception: Classic Edition*. Psychology Press.

[2009] Goldstone, W. 2009. *Unity Game Development Essentials*. Packt Publishing Ltd.

[2005] Johansson, R.; Berglund, A.; Danielsson, M.; and Nugues, P. 2005. Automatic text-to-scene conversion in the traffic accident domain. In *IJCAI*, volume 5, 1073–1078.

[1991] Joskowicz, L., and Sacks, E. P. 1991. Computational kinematics. *Artificial Intelligence* 51(1-3):381–416.

[2016] Kiela, D.; Bulat, L.; Vero, A. L.; and Clark, S. 2016. Virtual embodiment: A scalable long-term strategy for artificial intelligence research. *arXiv preprint arXiv:1610.07432*.

[2016a] Krishnaswamy, N., and Pustejovsky, J. 2016a. Multimodal semantic simulations of linguistically underspeci-

fied motion events. In *Spatial Cognition X: International Conference on Spatial Cognition*, In press. Springer.

[2016b] Krishnaswamy, N., and Pustejovsky, J. 2016b. VoxSim: A visual platform for modeling motion language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 54–58. ACL.

[1994] Kuipers, B. 1994. *Qualitative reasoning: modeling and simulation with incomplete knowledge*. MIT press.

[2000] Kuipers, B. 2000. The spatial semantic hierarchy. *Artificial Intelligence* 119(1):191–233.

[2007] Kurata, Y., and Egenhofer, M. 2007. The 9+ intersection for topological relations between a directed line segment and a region. In Gottfried, B., ed., *Workshop on Behaviour and Monitoring Interpretation*, 62–76.

[2009] Lakoff, G. 2009. The neural theory of metaphor. *Available at SSRN 1437794*.

[2003] Levinson, S. 2003. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Language, culture, and cognition. Cambridge University Press.

[2012] Mani, I., and Pustejovsky, J. 2012. *Interpreting Motion: Grounded Representations for Spatial Language*. Oxford University Press.

[1995] Mark, D., and Egenhofer, M. 1995. Topology of prototypical spatial relations between lines and regions in English and Spanish. In *Proceedings of the Twelfth International Symposium on Computer- Assisted Cartography*, volume 4, 245–254.

[2014] McDonald, D., and Pustejovsky, J. 2014. On the representation of inferences and their lexicalization. In *Proceedings of the Second Annual Conference on Advances in Cognitive Systems*, volume 3, 135–152.

[2001] Moratz, R.; Fischer, K.; and Tenbrink, T. 2001. Cognitive modeling of spatial reference for human-robot interaction. *International Journal on Artificial Intelligence Tools* 10(04):589–611.

[2000] Moratz, R.; Renz, J.; and Wolter, D. 2000. Qualitative spatial reasoning about line segments. In *Proceedings of the 14th European Conference on Artificial Intelligence*, 234–238. IOS Press.

[2007] Peters, J. F. 2007. Near sets. Special theory about nearness of objects. *Fundamenta Informaticae* 75(1-4):407–433.

[2014] Pustejovsky, J., and Krishnaswamy, N. 2014. Generating simulations of motion events from verbal descriptions. *Lexical and Computational Semantics (* SEM 2014)* 99–109.

[2016a] Pustejovsky, J., and Krishnaswamy, N. 2016a. Visualizing events: Simulating meaning in language. In Papafragou, A.; Grodner, D.; Mirman, D.; and Trueswell, J., eds., *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 2841–2842. Cognitive Science Society.

[2016b] Pustejovsky, J., and Krishnaswamy, N. 2016b. VoxML: A visualization modeling language. In Calzolari, N.; Choukri, K.; Declerck, T.; Goggi, S.; Grobelnik, M.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 4606–4613. European Language Resources Association (ELRA).

[2011] Pustejovsky, J., and Moszkowicz, J. 2011. The qualitative spatial dynamics of motion. *The Journal of Spatial Cognition and Computation* 15–44.

[2017] Pustejovsky, J.; Krishnaswamy, N.; and Do, T. 2017. Object embodiment in a multimodal simulation. *AAAI Spring Symposium: Interactive Multisensory Object Perception for Embodied Agents*.

[1995] Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

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

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

[2007] Renz, J., and Nebel, B. 2007. Qualitative spatial reasoning using constraint calculi. *Handbook of spatial logics* 161–215.

[2008] Rusu, R. B.; Marton, Z. C.; Blodow, N.; Dolha, M.; and Beetz, M. 2008. Towards 3D point cloud based object maps for household environments. *Robotics and Autonomous Systems* 56(11):927–941.

[2003] Sawilowsky, S. S. 2003. You think youve got trivials? *Journal of Modern Applied Statistical Methods* 2(1):21.

[2006] Seversky, L. M., and Yin, L. 2006. Real-time automatic 3D scene generation from natural language voice and text descriptions. In *Proceedings of the 14th ACM international conference on Multimedia*, 61–64. ACM.

[2000] Thrun, S.; Beetz, M.; Bennewitz, M.; Burgard, W.; Cremers, A. B.; Dellaert, F.; Fox, D.; Haehnel, D.; Rosenberg, C.; Roy, N.; et al. 2000. Probabilistic algorithms and the interactive museum tour-guide robot Minerva. *The International Journal of Robotics Research* 19(11):972–999.

[1995] Wann, J. P.; Rushton, S.; and Mon-Williams, M. 1995. Natural problems for stereoscopic depth perception in virtual environments. *Vision research* 35(19):2731–2736.

[2003] Ziemke, T. 2003. What's that thing called embodiment? In *Proceedings of the 25th Annual meeting of the Cognitive Science Society*, 1305–1310. Citeseer.

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