

A context-dependent model of proximity in physically situated environments

anonymous
someplace

Abstract

The paper presents a computational model for a context-dependent analysis of a physical environment in terms of spatial proximity. The model provides a basis for grounding linguistic analyses of spatial expressions in visual perception. The model uses potential fields to model spatial proximity. It has been implemented, and when combined with a handcrafted grammar, is used to enable a conversational robot to carry out a situated dialogue with a human. The key concept in our approach is defining the region that is proximal to a landmark based on the spatial configuration of other objects in the scene. The model extends existing approaches to proximity by including object salience (visual, discourse) and interference effects between multiple objects that could act as landmarks. Theoretically, the model can help motivate the choice between topological and projective prepositions, and provides a basis for defining regions with vague spatial extent.

1 Introduction

Our long-term goal is to develop embodied conversational robots that are capable of natural, fluent situated dialog with one or more interlocutors. An inherent aspect of situated dialog is reference to aspects of the physical environment that the interlocutors are situated in. In this paper, we present a computational model which provides a context-dependent analysis of the environment in terms of *spatial proximity*. We show how we can use this model to ground utterances that use topological prepositions (“the ball near the box”) or nouns expressing spatial proximity (“the corner”).

Proximity is ubiquitous in situated dialog, but there are deeper “cognitive” reasons for why a context-dependent model of proximity is needed to facilitate fluent dialog with a conversational robot. This has to do with the cognitive load that processing proximity expressions involve. Pragmatically, the Principle of Minimal Cooperative Effort [Clark and Wilkes-Gibbs, 1986] states that both the speaker’s effort in producing an utterance, and the hearer’s effort in interpreting

it, should be minimal. In particular, the Principle of Sensitivity [Dale and Reiter, 1995] states that when producing a (spatial) referring expression, the speaker should prefer features which the hearer is known to be able to interpret and perceive. Psycholinguistic data indicates that a spatial proximity expression (1b) presents a heavier cognitive load than a referring expression which distinguishes an object purely on physical features (1a), yet is easier to process than a spatial projection expression (1c) [van der Sluis and Krahmer, 2004].

- (1) a. the blue ball
- b. the ball near the box
- c. the ball to the right of the box

One explanation for this preference is that feature-based descriptions are easier to resolve perceptually, with a further distinction among features as given in Figure 1: object type is the easiest to process, before absolute gradable predicates (e.g. color), which is still easier than relative gradable predicates (e.g. size) [Dale and Reiter, 1995]. On the other hand, the interpretation and realization of spatial expressions requires effort and attention [Logan, 1994; 1995].

Similarly we can distinguish between the cognitive loads of processing different forms of spatial relations. Focusing on static prepositions, topological prepositions have a lower cognitive load than projective prepositions. Topological prepositions (e.g. “at”, “near”) describe proximity to an object. Projective prepositions (e.g. “above”) describe a region in a particular direction from the object. The extra cognitive load of projective prepositions arises from the different frames of reference that are used in language and the consequent three-dimensional rotations and translations of coordinates systems that may be required to define the direction described by the preposition; cf. [Krahmer and Theune, 1999].

Unfortunately, most work on computationally modelling spatial relations [Andre *et al.*, 1989; Olivier and Tsujii, 1994; Fuhr *et al.*, 1998; Mukerjee *et al.*, 2000; Regier and Carlson,

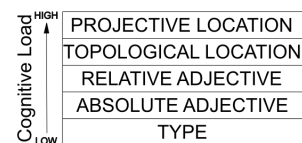


Figure 1: Cognitive load of different forms of reference

2001; Gorniak and Roy, 2004; Kelleher and van Genabith, 2005] focuses solely on projective prepositions. We lack a comprehensive model for topological prepositions. Without such a model, a conversational robot is not able to interpret spatial proximity expressions nor to motivate their contextually and pragmatically appropriate use – despite their ubiquity, and their positioning on the cognitive load hierarchy which makes them preferable over projective expressions.

Contributions In this paper, we address this problem. We present a computational approach to modelling proximity in physically situated contexts. The model includes the visual and discourse salience of objects as parameters, and uses energy functions to model how spatial templates associated with other landmarks may interfere to establish what are contextually appropriate ways to locate a trajectory relative to these landmarks. The resulting model enables a conversational robot to interpret and produce spatial proximity expressions that refer to objects in the environment. We focus on topological prepositions such as “near” or “at”. Furthermore, we show how this model of proximity enables us to model vague spatial regions, such a “the corner” or “the center”, which inherently require an understanding of proximity and our essential to orientation in many situated contexts.

Overview §2 presents effects we can observe in grounding spatial expressions, and discusses how they prove problematic for existing models. In §3 we discuss our model, and how we have integrated it with a handcrafted categorical grammar. §4 shows how the model captures examples involving the effects observed in §2. We end with conclusions.

2 Data

We already pointed out in §1 that people use spatial expressions to denote objects if they cannot distinguish them just by reference to easily perceivable physical properties [Dale and Reiter, 1995]. Furthermore, experimental data reveals that topological prepositions are easier to process cognitively than projective prepositions [van der Sluis and Krahmer, 2004].

In this section we discuss psycholinguistic experiments which argue that what is considered proximal is sensitive to the current visually situated context. Particularly, the experiments give rise to two hypotheses about context dependence: (1) interference of proximities of surrounding objects shrink the area considered to be close to an object, and (2) an increase (decrease) in salience of an object enlarges (shrinks) the area considered to be close to it. We argue below that existing models do not capture these hypotheses.

The psycholinguistic experiments reported in [Logan and Sadler, 1996] examined various topological prepositions. In these experiments, a human subject was shown sentences, each with a picture of a spatial configuration. Every sentence was of the form “The X is [relation] the O”. The accompanying picture contained an **O** in the center of an invisible 7-by-7 cell grid, and an **X** in one of the 48 surrounding positions. The subject then had to rate how well the sentence described the picture, on a scale from 1 (bad) to 9 (good).

Table 2 below gives the mean goodness rating for the relation “near to” as a function of the position occupied by the X, as reported in [Logan and Sadler, 1996]. If we plot the mean

goodness rating for “near to” against the distance between the trajectory X and the landmark O, we get the graph in Figure 3.

1.74	1.90	2.84	3.16	2.34	1.81	2.13
2.61	3.84	4.66	4.97	4.90	3.56	3.26
4.06	5.56	7.55	7.97	7.29	4.80	3.91
4.47	5.91	8.52	O	7.90	6.13	4.463
3.47	4.81	6.94	7.56	7.31	5.59	3.63
3.25	4.03	4.50	4.78	4.41	3.47	3.10
1.84	2.23	2.03	3.06	2.53	2.13	2.00

Figure 2: 7-by-7 cell grid with mean goodness ratings for the relation “near to” as a function of the position occupied by X

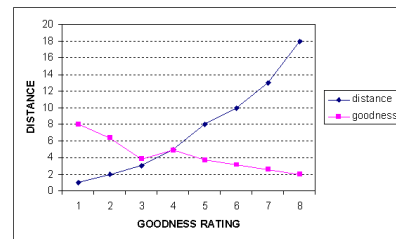


Figure 3: Mean goodness rating vs. distance between X, O

Both the table and the graph make it clear that the ratings diminish as we increase the distance between X and O. At the same time, we can observe that even at the extremes of the grid the ratings were still above 1 (the minimum rating). Indeed, in the four corners of the grid, the points most distant from the landmark, the mean ratings nearly average twice the minimum rating.

Hence, we have to further qualify the observed inverse relation between acceptability rating and distance. First, the observed drop in ratings does not evince that there is a maximum distance for proximity. This contradicts previous computational models of topological prepositions like [Gapp, 1994], which define a maximum distance as a parameter of the extension of the landmark. Second, although there is no maximum distance, we do observe a slope. The model of [Gapp, 1994] does accurately capture that contextual factors determine the steepness of this slope, in Gapp’s account the size of an object: E.g., given prototypical size, the region denoted by “near the building” is larger than that of “near the apple”. Author have also observed that, besides visual salience, discourse salience also influences spatial interpretation [Regier and Carlson, 2001; Roy, 2002] – but so far, only gaze-based attention has been included.

The final phenomenon we consider is the effect that other objects in the scene have. The location of other objects in the scene can inhibit locations being considered part of the focus space of the landmark. For example, consider the two scenes (side-view) given in Figure 2. In the scene on the left-hand side, we can use the description “the blue box near the black box” to refer to object (c), for the following reasons. First, we need to distinguish the boxes (a) and (c) from box (b). The cognitive load hierarchy in Figure 1 predicts that, given we

cannot use type to distinguish the objects, the use of an absolute gradable adjective presents the least load increase – hence “blue” to set (a) and (c) apart from (b). Next, to distinguish between (a) and (c), we have to use the proximity of (c) to (b) to set it apart from (a), which is (sufficiently) further from (b). However, consider now the scene on the right-hand side. In this context, the description “the blue box near the black box” seems inappropriate as an expression denoting (c). The placing of object (d) between (b) and (c) prevents us from using a proximal relation to locate (c) relative to (b). Although the absolute distance between (b) and (c) remains the same, the context no longer enables us to classify that distance as near due to the interference of (d).

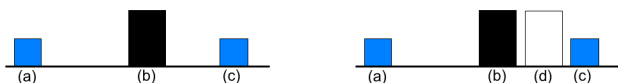


Figure 4: Proximity and distance

To recapitulate, we see that what counts as proximal to a landmark in a given situated context is based on several factors. Distance inversely affects proximity (i.e. downward slope), whereas salience positively affects the area surround the landmark that counts as proximal (i.e. the degree of the slope). The actual situated context may, however, inhibit the extension of that area in a particular direction, if there are interfering objects. Authors like [Gapp, 1994] and [Regier and Carlson, 2001] have proposed to model proximity as a function of distance and gaze-based attention, but this does not model possible inhibition effects nor the full range of salience. In the next section, we propose how we can capture these factors in a single, unified approach.

3 Approach

Below we discuss how we create a model of the situated environment in terms of proximity. We represent the environment at the level of objects and scene features, leaving the inclusion of events to future work. We take objects and scene features to function as possible landmarks within a scene. Then, for each possible landmark, the model establishes the region around the landmark that counts as proximal to it.

We use *potential fields* to model the gradation of applicability with distance as shown in [Logan and Sadler, 1996]. The fundamental component of the potential field model is a *potential function* that computes the cost of describing a spatial configuration using a proximal description. This cost can range from 0 to 1. The lower the cost at a point in the region around the landmark, the more appropriate it is to say that an object located at that point is near the landmark.

We create the model in two stages. First, for each landmark we create a potential field across the region of the scene that models the applicability of a point in the scene being described as proximal to that landmark. Second, we look where the potential fields of different landmarks overlap. For each point in the overlap between two or more fields, we then compute the difference between the potential of the landmark with the highest applicability (i.e., the lowest potential) at that

point, and the other potentials. If this difference is less than by a predefined *ambiguity factor* we mark the point as being ambiguous with respect to what landmark it is considered to be close to. Otherwise, we mark the point as being proximate to the landmark with the highest applicability.

In §2 we noted three factors effecting the appropriateness of describing the spatial relationship between two objects in a scene using a proximal spatial relation. These were: (1) the distance between the object, (2) the size and salience of the object functioning as the landmark, and (3) the location of other objects in the scene. We capture these factors as follows: (1) and (2) are modelled by the potential field model, which we discuss in §3.1, whereas (3) is captured by the overlaying of the potential fields, described in §3.2. Finally, in §3.3 we discuss how we can ground linguistic analyses of referring expressions involving spatial proximity in the model.

3.1 Computing the potential field of a landmark

At the first stage, we need to compute for each landmark the potential field that models proximity to that landmark. We compute these fields on the projection of the scene onto the two-dimensional plane, which we model as a two-dimensional array *ARRAY* of points.

Next, we consider a landmark *LM* at some arbitrary point *PL* in the scene. For each point *PT* in the point array *ARRAY*, we compute a potential value for the *LM* at that point using the following equation:

$$P_{prox} = dist_{normalised}(PL, PT, ARRAY) * salience(LM) \quad (1)$$

Equation 1 makes clear how we compute the potential value from the distance between the point *PT* and the location of the landmark *PL*, and the salience of the landmark.

As distance we use a normalised distance function $dist_{normalised}(PL, PT, ARRAY)$. This function returns a value between 0 and 1 to represent the normalised distance between points *PL* and *PT* within the scene.¹ The smaller the distance between *PT* and *PL*, the lower the value returned, i.e. the lower the cost the more acceptable it is to say that *PT* is close to *PL*. In this way, this component of the potential field captures the gradual gradation in applicability evident in [Logan and Sadler, 1996]. Table 5 illustrates this in detail, giving the normalised distances between each cell in a 7-by-7 cell grid and the location of the landmark *PL*.

We model the influence of visual and discourse salience on the area that is considered to be proximal to the landmark as a function $salience(LM)$ – the other component of Equation 1. The function returns a value between 0 and 1 that represents the relative salience of the landmark *LM* in the scene. The relative salience ascribed to an object by this function is dependent on its visual and discourse salience.

The visual salience component is computed using the visual saliency algorithm described in [Kelleher and van Genabith, 2004]. This algorithm computes a relative salience

¹We normalise by computing the distance between the two points, and then dividing this distance by the maximum distance between point *PL* and any point in the scene.

1.00	0.72	0.56	0.17	0.56	0.72	1.00
0.72	0.44	0.28	0.11	0.28	0.44	0.72
0.56	0.28	0.11	0.06	0.11	0.28	0.56
0.17	0.11	0.06	PL	0.06	0.11	0.17
0.56	0.28	0.11	0.06	0.11	0.28	0.56
0.72	0.44	0.28	0.11	0.28	0.44	0.72
1.00	0.72	0.56	0.17	0.56	0.72	1.00

Figure 5: 7-by-7 cell grid with normalised distances between the point PL and the coordinates of the other cells in the grid

for each object in a scene. The factors contributing to the salience of an object are its perceivable size and its centrality relative to the viewer focus of attention. The algorithm returns salience scores in the range of 0 to 1. The fact that the visual salience algorithm captures object size permits our framework to model the effect of landmark size on proximity through the salience component of the potential field. The discourse salience of an object is computed based on recency of mention as defined in [Hajičová, 1993] except we represent the maximum overall salience in the scene as 1, and use 0 to indicate that the landmark is not salient in the current context.

The $salience()$ function integrates these two components by summing them and dividing the result by 2. This again results in a range of salience values between 0 and 1. Summing and then dividing is preferred over multiplication as it avoids the problem of multiplying by 0. This problem would arise in a situation where we would have an object with a very high visual salience, but with a discourse salience 0 because it has not been mentioned yet. In this case the landmark would be ascribed an overall salience of 0, contrary to expectations.

We already remarked before (cf. §2) that when the salience of a landmark increases, the area which could be described as proximal to the landmark becomes larger. Consequently we want the potential function to decrease the potential computed at a point as the salience of the landmark increases. In order to achieve this we need to invert the salience ratings. We achieve this by taking the each objects salience value from 1 plus the minimum salience ascribed to an object in the scene. The motivation for adding the minimum salience value to 1 prior to the subtraction is that it prevents an object with a salience of 1 being ascribed an overall salience of 0. The following equation defines how the salience is computed:

$$salience(LM) = (1 + minimumSalienceInScene) - ((VisualSalience(LM) + DiscourseSalience(LM))/2) \quad (2)$$

The data in Tables 6 and 7 illustrates how the applicability ratings change under influence of an increase in salience. The lower values show a lower cost, i.e. it is more appropriate to call a point in that region (still) proximal to the landmark. Table 6 gives the potentials computed for each cell in a 7-by-7 cell grid using Equation 1 on the normalised distances in Table 5 for a landmark with an inverted salience of 0.5 as computed using Equation 2. Table 7 illustrates the potential field for a landmark with a higher salience (which results in a lower inverted salience). When we compare the grid cells

along the peripheries of the tables, we see that the cells in Table 7 have lower values than those in Table 6. This shows the salience effect: the higher salience enables us to take points further from the landmark and still call them “close”. The graph in Figure 8 makes the effect of salience on the computed potentials more visual: As the salience increases, the potentials decrease.

Algorithm 1 formally describes how the landmark potential fields are computed.

0.50	0.36	0.28	0.08	0.28	0.36	0.50
0.36	0.22	0.14	0.06	0.14	0.22	0.36
0.28	0.14	0.06	0.03	0.06	0.14	0.28
0.08	0.06	0.03	PL	0.03	0.06	0.08
0.28	0.14	0.06	0.03	0.06	0.14	0.28
0.36	0.22	0.14	0.06	0.14	0.22	0.36
0.50	0.36	0.28	0.08	0.28	0.36	0.50

Figure 6: 7-by-7 cell grid with potentials based on Equation 1 using the normalised distances in Table 5, and an inverted salience of 0.5 for the landmark

0.20	0.14	0.11	0.03	0.11	0.14	0.20
0.14	0.09	0.06	0.02	0.06	0.09	0.14
0.11	0.06	0.02	0.01	0.02	0.06	0.11
0.03	0.02	0.01	PL	0.01	0.02	0.03
0.11	0.06	0.02	0.01	0.02	0.06	0.11
0.14	0.09	0.06	0.02	0.06	0.09	0.14
0.20	0.14	0.11	0.03	0.11	0.14	0.20

Figure 7: 7-by-7 cell grid with potentials based on Equation 1 using the normalised distances in Table 5, and an inverted salience of 0.2 for the landmark

3.2 Overlaying the Landmark Potential Fields

Once we have constructed the potential fields for the landmarks the next stage in creating the proximity regions is to check for overlap between the potential fields. We do this by iterating over each point in the scene, and comparing the potentials of the different landmarks at each point. If the primary landmark’s (i.e., the landmark with the lowest potential the point) potential is less than the potentials of each of the other landmarks when they are divided by a predefined factor the point is deemed to be in the proximal area of the landmark. If not we take the point to be ambiguous, and not in

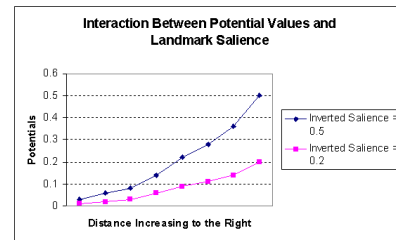


Figure 8: Interaction between potential field and salience

Algorithm 1 Computing the landmark potential fields.

Require: A set of candidate landmarks $CL = cl_1, cl_2, \dots, cl_n$ each with a point in space defining its center of mass $ct_i.center_of_mass$ and an associated salience $ct_i.salience$, ($salience \in 0 \dots 1$); and the set of points defining the region the energy landscape will be computed for $P = p_1, p_2, \dots, p_3$.

Ensure: A set of potential fields, one for each landmark in CL .

```
Let  $MAX\_DIST = 0$ 
Let  $distances[][] = array[|CL|][|P|]$ 
Let  $potentials[][] = array[|CL|][|P|]$ 
for  $i = 0$  to  $|P|$  do
  for  $j = 0$  to  $|CL|$  do
     $distance[i][j] = euclidean.distance(pt_i, ct_j.center\_of\_mass)$ 
    if  $MAX\_DIST < distances[i][j]$  then
       $MAX\_DIST = distances[i][j]$ 
    end if
  end for
end for
{normalise  $distances[][]$  by dividing by  $MAX\_DIST$ }
for  $i = 0$  to  $|P|$  do
  for  $j = 0$  to  $|CL|$  do
     $distances[i][j] = distances[i][j]/MAX\_DIST$ 
  end for
end for
{compute potential for each CL at each point in region}
for  $i = 0$  to  $|P|$  do
  for  $j = 0$  to  $|CL|$  do
     $potentials[i][j] = distances[i][j] * ct_j.salience$ 
  end for
end for
```

the proximal region of any of the landmarks. The motivation for this factoring of the potentials of the other landmarks is to capture situations where the difference in potential between the primary landmark and one or more of the other landmarks at a given point is relatively small. Algorithm 2 defines the procedure for overlaying the potential fields of the landmarks in a scene.

Figure 9 illustrates the overlaying of the potential fields of two landmarks, PL1 and PL2. The difference in the slopes of each of the landmarks' potential field is due to the different salient scores these landmarks were attributed. The potential field for PL1 was computed with an inverted salience of 0.2. The potential field for PL2 was computed with an inverted salience of 0.5. The ambiguous regions were computed using a factor of $\rho = 2.0$. The points that were deemed to be ambiguous are located at the positions where the ambiguous regions series is plotted at 1.00 on the Y-axis. The regions that are defined as proximate to a landmark are denoted by the horizontal extent of the box surrounding the label of the landmark. The proximate region for PL1 covers the area on the X-axis where the ambiguous region's plot is at 0 on the y-axis and PL1's potential field model is lower than PL2's. Similarly, the region of proximity defined for PL2 covers the area of the X-axis where the ambiguous region's plot is at 0 on the Y-axis and PL2's potential field is lower than PL1's.

There are two points evident in Figure 9 that are worth noting. First, the greater the salience ascribed to a landmark the larger the region of proximity associated with it. In Figure 9 the region of proximity associated with landmark PL1, which

Algorithm 2 Overlaying the landmarks' potential fields.

Require: A set of landmarks each with a potential field that defines $P_{potential}$, the potential of the landmark at the point P, and ρ a scaling factor.

Ensure: A energy landscape that defines for each point in the region the landmark it is considered proximate to or that it is ambiguous.

```
for each point P in the region do
  Primary $landmark$  = the landmark with the lowest potential at point P
  for each landmark $_i < Primarylandmark$  do
    if Primary $landmark.P_{potential} < (landmark_i.P_{potential}/\rho)$  then
      P  $\in$  proximal region of Primary $landmark$ 
    else
      P  $\in$  ambiguous region
    end if
  end for
end for
```

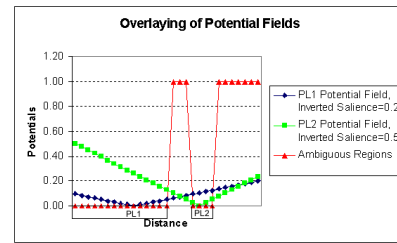


Figure 9: Graph showing overlaying of the potential fields for two landmarks PL1 and PL2. The landmark potential fields were computed with different inverted salience scores, 0.2 and 0.5 respectively. The locations of the landmarks are marked on the X-axis. The ambiguous regions were computed using a factor of 2.0 and are located where the ambiguous regions series is plotted at 1.00 on the Y-axis.

had an inverted salience of 0.2, is much larger than the proximate region associated with landmark PL2, which had an inverted salience of 0.5. Second, the overlaying of potential fields naturally defines the extent of the proximate regions. In Figure 9 this is particularly evident when we focus on the region to the right of PL2. This region is deemed as ambiguous because of the effect of the potential field of the more salient PL1. Following the cognitive load model described in §1, objects located in this region should be described with a projective relation, such as “to the right of PL2” rather than a proximal relation. We will return to this point in the discussion section of the paper, §4.

3.3 Linguistic analysis of spatial expressions

Below we briefly describe what requirements for linguistic analysis of spatial expressions we can derive from the literature, and how we can address these requirements in our approach.

One, a linguistic analysis should be able to make explicit the inherent *vagueness* of spatial gradable predicates like “close”, and spatial nouns like “corner”: They each specify proximity to a landmark, but leave the exact extent of that proximity *vague*. Vagueness is a pervasive feature

of natural language, and the truth-conditions for vagueness are inherently context-dependent, as various authors have argued [Graff, 2000; Kyburn and Morreau, 2000; Barker, 2002; Kennedy and McNally, 2004].

Two, we should be able to obtain an analysis *incrementally*. Visually situated contexts present a huge amount of information. Incremental processing of spatial expressions is an important means to focus attention already *while* processing, by drawing in environmental information to disambiguate [Schuler, 2001], and dynamically establishing contextual standards against which vague references are interpreted [Barker, 2002; DeVault and Stone, 2004].

Three, a linguistic analysis should provide enough information to establish the *visual grounding* of spatial expressions: How can the robot relate a logical form, obtained as a grammatical analysis of a spatial expression, and a scene it visually perceives, so that it can locate the objects or features in the scene which the expression applies to? Approaches presented in the literature agree on the need for ontologically rich representations, but differ in how these representations are subsequently grounded in vision. The literature presents a spectrum of approaches, ranging from “scruffy” ones which rely on machine learning methods to establish a statistical mapping between visual and linguistic features [Oates *et al.*, 2000; Roy, 2002]; to approaches that use manually constructed mappings between linguistic constructions and probabilistic functions which evaluate whether an object can act as referent on the basis of visual features [Gorniak and Roy, 2004]; to “neat” ones that employ constraint resolution over symbolic representations [DeVault and Stone, 2004].

We address the above requirements as follows. To enable the robot to communicate in natural language, we have developed a grammar in Combinatory Categorical Grammar (CCG) [Steedman, 2000; Baldrige and Kruijff, 2003], which we can parse incrementally [Steedman, 2000]. The grammar describes the compositional relation between the syntactic structure of an utterance and its semantics. We model semantics as an ontologically richly sorted, relational structure, formalized in a description logic-like framework called Hybrid Logic Dependency Semantics (HLDS) [Kruijff, 2001; Baldrige and Kruijff, 2002; White, 2004]. Parsing an utterance yields a representation of its semantics.² For space reasons, we focus below on how we represent and interpret the semantics of spatial expressions, and omit the syntactic analyses.

- (2) the box near to you
 $\textcircled{b:\text{phys-obj}}$ (**box**)
 & $\langle \text{Delimitation} \rangle$ **unique**
 & $\langle \text{Number} \rangle$ **singular**
 & $\langle \text{Quantification} \rangle$ **specific.singular**
 & $\textcircled{b:\text{phys-obj}}$ $\langle \text{Location} \rangle$ ($r : \text{region} \ \& \ \text{near}$)
 & $\langle \text{Proximity} \rangle$ **proximal**
 & $\langle \text{Positioning} \rangle$ **static**
 & $\textcircled{r:\text{region}}$ $\langle \text{FromWhere} \rangle$ ($y1 : \text{hearer} \ \& \ \text{you}$)
 & $\langle \text{Number} \rangle$ **singular**

²We use OpenCCG [White, 2004]: <http://www.sf.net/openccg/>

Example (2) illustrates the semantic analysis for “the box near you”. The representation consists of several, related *elementary predicates*. One type of elementary predicate represents a discourse referent as a proposition with a handle: $\textcircled{b:\text{phys-obj}}$ (**box**) means that the referent b is a physical object, namely a **box**. Another type of elementary predicate represents dependencies between referents as modal relations, e.g. $\textcircled{b:\text{phys-obj}}$ $\langle \text{Location} \rangle$ ($r : \text{region} \ \& \ \text{near}$) means that discourse referent b (the box) is located in a region r that is near to a landmark. We represent regions explicitly to enable later reference to the region using deictic reference (e.g. “there”). Within each elementary predicate we can additionally have semantic features. For example, the region r characterizes a static location of b , and –most importantly– it expresses *proximity* to a landmark. The example explicitly represents the hearer as being that landmark.

Elementary predicates provide a natural granularity for incremental semantic processing. We use the sorting information (e.g. *phys-obj*, *region*) to interpret the *linguistic* meaning of an utterance further using ontology-based spatial reasoning. This yields several inferences that need to hold for the scene, comparable to [DeVault and Stone, 2004] where reasoning can expand constraints that need to be satisfied. Where we differ from DeVault & Stone, however, is in how we check whether these inferences hold: like [Gorniak and Roy, 2004], we map these conditions onto the energy landscape computed by the potential field functions. This enables us to take into account inhibition effects arising in the actual situated context, which neither Gorniak & Roy nor DeVault & Stone do.

Finally, we can deal with vagueness at two levels. The literature suggests that vague expressions such as gradable predicates like “close” are interpreted on contextual standards, cf. [Kennedy, 2004]. Essentially, this means we have a measurement function on a scale, and it is this scale that is context-dependent: we need to (1) measure degrees of “closeness” against (2) what counts as close in the current context. We establish the latter aspect in our scene interpretation model using potential field functions. These functions are modulated by contextual factors (notably, salience), and –together with possible interference effects– give rise to an energy landscape that models proximity, i.e. what counts as close in the current situated context. This establishes the basis for a contextual standard. Next we can create contextual standards, following [DeVault and Stone, 2004]; this is, however, beyond the scope of the current paper.

4 Discussion

The background for this paper is the general problem of *symbol grounding* [Harnad, 1990]: how can we link symbols to perceptual input? In our current setting this means, how can the robot relate a logical form, obtained as a grammatical analysis of a spatial expression, and a scene it visually perceives, so that it can locate the objects or features in the scene which the expression applies to? What makes the problem difficult (beyond visual recognition and classification of objects), is that establishing the region that is proximal to a landmark depends on the situational, visual and dialogue context. As we explained in the preceding sections, we cannot

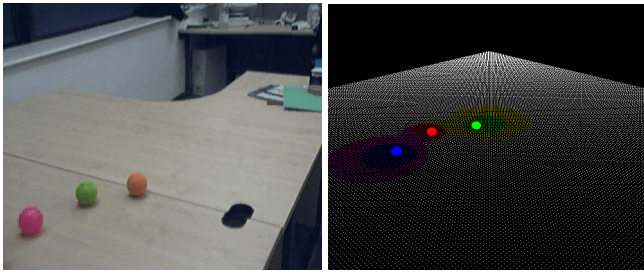


Figure 10: Scene analysis

define this region on a purely geometrical basis. The salience of LM within a particular context and the salience and location of other objects in the scene relative to LM determine to what extent it is acceptable to use a proximal spatial relation to locate a trajector relative to it.

Figure 10 shows a real scene on the left-hand side, and a scene analysis on the right-hand side. For the shown scene analysis we have assumed all objects to have an equal salience: on the left, the blue ball; in the middle, the red ball; and on the right, the green ball. As the analysis correctly shows, each object has a proximity potential field (shown in its own color) but, due to interference between potential fields, we see that proximity is usually ambiguous between at least two landmarks.

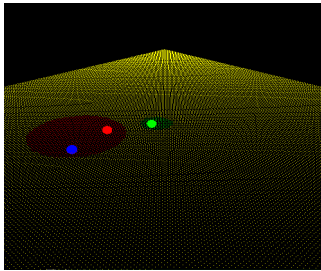


Figure 11: Locating the object in the blue position

The figure in Figure 11 changes this situation. For the shown scene analysis, we have assumed that only the green ball and the red ball are salient, to degrees of 0.6 and 1.0 respectively; leaving momentarily the blue ball out of the picture. We can observe an interference effect between the red ball and the green ball: the potential field representing proximity to the red ball forms an ellipsoid, being inhibited to the right through interference with the potential field of the green ball.

- (3) the [object] near the red ball

Now assume that we would place a ball in the position of the blue ball. As it does not have a potential field yet, we can refer to it as being near the red ball, using the utterance in (3). For (3) we obtain a semantic analysis similar to the logical form shown in (2). We then resolve the landmark description “red ball” to the red ball in the scene, and determine whether an object of description “[object]” is located at a po-

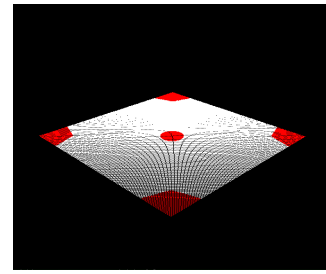


Figure 12: Corners and center of a scene

sition within the region of the potential field that represents unambiguous proximity of the red ball.

Finally, consider Figure 12. It illustrates how we can model nouns that express a vague spatial extent, like “corner” or “center”, using the potential functions we use to model proximity. We originate the potential functions in the geometrical absolutes, i.e. the dead center of the scene and its absolute corners. These potential fields again may interfere with potential fields spawned up by objects in the scene.

5 Conclusions

In this paper, we presented an approach to modelling proximity in situated environments, to be able to ground spatial expressions involving e.g. topological prepositions, or nouns expressing a spatial extent. We discussed available psycholinguistic data to substantiate the usefulness of having such a model for interpreting and generating natural, fluent situated dialogue between a human and a conversational robot; and that we need a context-dependent representation of what is (situationally) appropriate to consider proximal to a landmark. Context-dependence thereby involves salience of landmarks as well as inhibition effects between landmarks.

We argued that none of the main models model for interpreting spatial prepositions capture the effects we have observed in §2 e.g. the AVS model of [Regier and Carlson, 2001] (and used in [Roy, 2002; Gorniak and Roy, 2004]), or the distance-based model for topological prepositions proposed in [Gapp, 1994], the main reasons being that they (a) only include restricted forms of “attention”, and (b) do not account for inhibition effects. We presented a model in which we can address these issues, and we exemplified how logical forms representing semantic analyses of spatial proximity expressions can be grounded in this model.

One line of future work we want to consider how we can expand the available psycholinguistic evidence for the processing of spatial prepositions, to be able to explore further the interactions between salience and interference. Another line follows up on the observations we made regarding the relation between the way we model proximity, and vagueness. For one, we would like to investigate how the dynamics of vagueness [Barker, 2002] are affected when the robot moves through an environment: as the scene and the robot’s field of vision change, established contextual standards are likely to change as they are dependent on the spatial configuration of the observed scene. This raises the question whether contextual standards should also be accorded a salience measure.

References

- [Andre *et al.*, 1989] E. Andre, Herzog. G., and T. Rist. Natural language access to visual data: Dealing with space and movement. In *Report 63, Universitat des Saarlandes, SFB 314 (VITRA), Saarbrücken, 1989. Presented at the 1st Workshop on Logical Semantics of Time, Space and Movement in Natural Language, Toulouse, France.*, 1989.
- [Baldrige and Kruijff, 2002] Jason Baldrige and Geert-Jan M. Kruijff. Coupling CCG and hybrid logic dependency semantics. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, Pennsylvania, 2002.
- [Baldrige and Kruijff, 2003] Jason Baldrige and Geert-Jan M. Kruijff. Multi-modal combinatory categorial grammar. In *Proceedings of the Annual Meeting of the European Association for Computational Linguistics (EACL 2003)*, Budapest, Hungary, 2003.
- [Barker, 2002] Chris Barker. The dynamics of vagueness. *Linguistics and Philosophy*, 25(1):1–36, 2002.
- [Clark and Wilkes-Gibbs, 1986] H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22:1–39, 1986.
- [Dale and Reiter, 1995] R. Dale and E. Reiter. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263, 1995.
- [DeVault and Stone, 2004] David DeVault and Matthew Stone. Interpreting vague utterances in context. In *Proceedings of COLING 2004*, volume 2, pages 1247–1253, Geneva, Switzerland, 2004.
- [Fuhr *et al.*, 1998] T. Fuhr, G. Socher, C. Scheering, and G. Sagerer. A three-dimensional spatial model for the interpretation of image data. In P. Olivier and K.P. Gapp, editors, *Representation and Processing of Spatial Expressions*, pages 103–118. Lawrence Erlbaum Associates, 1998.
- [Gapp, 1994] K.P. Gapp. Basic meanings of spatial relations: Computation and evaluation in 3d space. In *National Conference on Artificial Intelligence (AAAI-94)*, pages 1393–1398, 1994.
- [Gorniak and Roy, 2004] Peter Gorniak and Deb Roy. Grounding semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470, 2004.
- [Graff, 2000] Delia Graff. Shifting sands: An interest-relative theory of vagueness. *Philosophical Topics*, 20:45–81, 2000.
- [Hajičová, 1993] Eva Hajičová. *Issues of sentence structure and discourse patterns*, volume 2 of *Theoretical and Computational Linguistics*. Charles University Press, Prague, Czech Republic, 1993.
- [Harnad, 1990] Stevan Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.
- [Kelleher and van Genabith, 2004] J. Kelleher and J. van Genabith. Visual salience and reference resolution in simulated 3d environments. *Artificial Intelligence Review*, 21(3):253–267, 2004.
- [Kelleher and van Genabith, 2005] J. Kelleher and J. van Genabith. In press: A computational model of the referential semantics of projective prepositions. In P. Saint-Dizier, editor, *Syntax and Semantics of Prepositions.*, Speech and Language Processing. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2005.
- [Kennedy and McNally, 2004] Christopher Kennedy and Louise McNally. Scale structure, degree modifications, and the semantics of gradable predicates. Unpublished manuscript, January 15 2004.
- [Kennedy, 2004] Christopher Kennedy. Towards a grammar of vagueness. Unpublished manuscript, April 5 2004.
- [Krahmer and Theune, 1999] E. Krahmer and M. Theune. Efficient generation of descriptions in context. In R. Kibble and K. van Deemter, editors, *Workshop on the Generation of Nominals, ESS-LLI'99*, Utrecht, The Netherlands, 1999.
- [Kruijff, 2001] Geert-Jan M. Kruijff. *A Categorial-Modal Logical Architecture of Informativity: Dependency Grammar Logic & Information Structure*. PhD thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic, April 2001.
- [Kyburn and Morreau, 2000] Alice Kyburn and Michael Morreau. Fitting words: Vague language in context. *Linguistics and Philosophy*, 23:577–597, 2000.
- [Logan and Sadler, 1996] Gordon D. Logan and Daniel D. Sadler. A computational analysis of the apprehension of spatial relations. In P. Bloom, M.A. Peterson, L. Nadel, and M.F. Garrett, editors, *Language and Space*, pages 493–529. The MIT Press, Cambridge MA, 1996.
- [Logan, 1994] Gordon D. Logan. Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, 20:1015–1036, 1994.
- [Logan, 1995] Gordon D. Logan. Linguistic and conceptual control of visual spatial attention. *Cognitive Psychology*, 28:103–174, 1995.
- [Mukerjee *et al.*, 2000] A. Mukerjee, K. Gupta, S. Nautiyal, P. Mukesh, M. Singh, and N. Mishra. Conceptual description of visual scenes from linguistic models. *Journal of Image and Vision Computing*, 18, 2000.
- [Oates *et al.*, 2000] Tim Oates, Zachary Eyer-Walker, and Paul R. Cohen. Toward natural language interfaces for robotic agents: Grounding linguistic meaning in sensors. In *Proceedings of the Fourth International Conference on Autonomous Agents*, pages 227–228, 2000.
- [Olivier and Tsujii, 1994] P. Olivier and J. Tsujii. Quantitative perceptual representation of prepositional semantics. *Artificial Intelligence Review*, 8(147-158), 1994.
- [Regier and Carlson, 2001] Terry Regier and Laura A. Carlson. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology*, 130(2):273–298, 2001.
- [Roy, 2002] Deb K. Roy. Learning words and syntax for a scene description task. *Computer Speech and Language*, 16(3), 2002.
- [Schuler, 2001] William Schuler. Computational properties of environment-based disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*, Toulouse, France, 2001.
- [Steedman, 2000] Mark Steedman. *The Syntactic Process*. The MIT Press, Cambridge MA, 2000.
- [van der Sluis and Krahmer, 2004] I.F. van der Sluis and E.J. Krahmer. The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. In R. Kibble and K. van Deemter, editors, *ICSLP04*, 2004.
- [White, 2004] Michael White. Efficient realizations of coordinate structures in combinatory categorial grammar. *Research on Language and Computation*, 2004.