Using Relations To Describe Three-Dimensional Scenes: A Model of Spatial Relation Apprehension and Interference

Sebastien Ouellet (sebouel@gmail.com)

Jim Davies (jim@jimdavies.org)

Institute of Cognitive Science, Carleton University
1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6 Canada

Abstract

Understanding spatial information between the objects visible in a scene is crucial to tasks such as describing relevant features of the scene, navigating it based on a description of its features, and for creating novel imagined scenes based on linguistic input. We present a model of spatial relation apprehension able to map geometric information from 3D scenes containing multiple objects to English prepositions and verbs in terms of direction and distance, as well as topological relations. We created an implementation of the model to evaluate its effectiveness.

Keywords: Imagination; 3D scenes; English prepositions; Spatial cognition; Spatial representation; Spatial perception; Artificial intelligence

Introduction

When people perceive scenes, they understand spatial relations between objects. For example, when viewing a dining room, someone might perceive that the table is *in the center of* several chairs. Several competing models of spatial relation apprehension exist, and are supported in different ways and under different situations(Gorniak & Deb, 2004; Mukerjee, Gupta, Nautiyal, Singh, & Mighra, 2000; Lockwood, Forbus, & Usher, 2005; Regier & Carlson, 2001). A common situation that lacks an applicable model is a 3D scene composed of many objects (Kojima & Kusumi, 2007). The goal of this paper is to present a model able to describe a scene in a cognitively plausible way through the use of spatial terms related to distance, direction, and topological relations. The spatial relations supported by our model are deictic, where a target object is related to a reference object from the point of view of that reference object. An example of a possible scene description is the following: *The desk is in front of the couch. The desk is far from the couch. The pillow is to the right of the couch and close to it. The lamp is above the pillow.*

To represent the locations of relevant objects in a scene, the model needs to estimate how they are related to each other in a spatial manner. This step is inspired by the theory of spatial relation apprehension developed by Logan and Sadler (1996), who performed experiments to determine how people use English sentences to describe the relationship between two objects in a 2D scene. They suggested that people have a number of spatial templates corresponding to propositions about spatial properties, such as *in front of*. The current research was also inspired by an implemented model of 2D spatial relation detection by (Smith et al., 2010).

A spatial template is defined as a fuzzy membership function that returns a value indicating the applicability of the template for a pair of objects, behaving rather like a receptive field. For example, if a person is to determine whether a located object is in front of a reference object, the person would try to apply the *in front of* template in the scene being studied. Depending on how well the two objects fit the template, the person would then decide the degree to which the proposition indeed applies to the scene.

Our model advances from their work through the description of algorithms corresponding to the spatial templates discussed in Logan's study, and creating 3D versions of the implementations of Smith et al. Developing our model involved defining those spatial templates and the processes that allow the model to apply them in scenes. In our implementation, these algorithms are applied to objects in human-designed 3D environments created with computer-aided design (CAD) software, in the X3D format.

The next sections describe how the model is able to represent eleven spatial relations: *in front of, behind, right of, left of, above, below, contains, contained by, protrudes from, close to, far from*. Our model also accounts for the interference effect found in scenes that present a large number of objects (Kojima & Kusumi, 2006).

Spatial relation apprehension

Computing the spatial relations in a given scene is done according to the following steps: the assessment of the reference object's location in the scene, the selection and alignment of a reference frame based on the reference object, and the application of a spatial template on the located object (Carlson-Radvansky & Logan, 1997).

For the purpose of our model, we assume that the absolute locations, orientations, and sizes of objects in the scene are given as input. As for the reference frame manipulations, the model uses an environment-centered reference frame, where the coordinate axes are oriented according to longest dimension of the scene, i.e., the horizontal axis, or x axis, points toward the length of the scene and the vertical axis, or y axis, points against gravity. Since the model focuses on deictic relations, the reference frame's point of origin corresponds to the location of the reference object. The spatial templates are then applied to the pair of objects according to the reference frame.

To produce descriptions from a pair of objects, all known

Ouellet, S. & Davies, J. (2013). Using prepositions to describe three-dimensional scenes: A model of spatial relation apprehension and interference. In R. West & T. Stewart (Eds.), Proceedings of the 12th International Conference on Cognitive Modeling, Ottawa: Carleton University (unnumbered six page online proceedings)

spatial templates are applied iteratively to the pair and the ones with a high degree of membership are kept as good descriptions of the relations between those objects. This procedure is similar to the one described by Logan and Sadler (1996), in the case where viewers are asked to judge relations between objects and no specific relation is mentioned, e.g. *where is the ceiling with respect to the floor?*

The spatial relations are divided into three categories, defined by the geometrical parameters required for their apprehension. Directional relations depend on the orientation of the reference frame and are assessed for each axis, while distance relations are independent of the reference frame's orientation. As for the topological relations, which concerns containment, the result depends on the detail of the geometrical shape, which can be approximated through the use of bounding boxes.

Topological relations

Topological relations are considered binary (i.e., non-fuzzy) for the purpose of our model, such that a relation will be either be applicable or not to a pair of objects. For each pair of objects there is only one applicable relation among all topological relations. Three spatial templates are supported by our model: *contains, contained by, protrudes from*.

Two objects are said to intersect if, along any axis, the boundaries of the objects cross each other. The exact calculation depends on the representation of the geometrical shape, and will be discussed in the Method section. The acceptability rating of all topological relations becomes 0 if no intersection is found between the two objects.

The located object is said to *contain* the reference object when the boundaries of the located object extend farther than the reference object's boundaries in all dimensions. The *contained by* relation applies to the reverse case, where the reference object's boundaries have a larger volume. As soon as one part of the located object extends outside the reference object's boundaries, the located object is said to be *protruding from* the reference object. The relation is symmetrical, such that the reference objects is also considered to be *protruding from* the located object.

Distance relations

The two spatial templates *close to* and *far from* are dependent on the distance between the two objects and the size of the scene in which they are found. The following equations describe the computation involved:

$$rating = 1 - \frac{1}{1+30e^{-7\cdot distance}} \tag{1}$$

$$rating = \frac{1}{1+30e^{-7\cdot distance}} \tag{2}$$

where (1) is used for the *close to* template and (2) is used for the *far from* template. The *distance* value is calculated from the Euclidean distance between the two objects divided by the estimated maximum distance within the constraints of the scene, reflecting the tendency to perceive two things as being closer if they are in a smaller space. The equations were built to model the empirical findings described by Logan and Sadler (1996) on a relative domain, where a distance described as minimally close to is equal to 0 and the most close to is equal to 1.

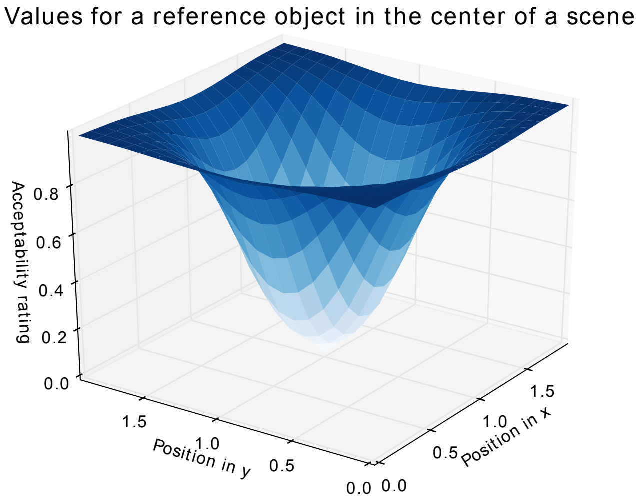

Figure 1: Surface plot of the acceptability ratings for all possible locations around a reference object for the spatial template *far from*, as defined in (2), restricted to a 2D plane

Directional relations

The model includes the following six spatial directional relations: *in front of, behind, right of, left of, above, below*. Each of these relations are associated with a acceptability rating computed by the following equation:

$$rating = \frac{located_j - reference_j}{\sqrt{\sum_{i=1}^{3}(located_i - reference_i)^2}} \tag{3}$$

where *located* and *reference* refers to coordinates of the center of the located object and the reference object on a single axis. The Euclidean distance between the two objects is calculated and divides the distance between the objects along the corresponding axis: x for *left of* and *right of*, y for *above* and *below*, and z for *in front of* and *behind*.

The function was designed to approximate empirical findings from Logan and Sadler (1996) and Hayward and Tarr (1995). However, these studies involved experiments where only the two objects of interest were present in the scenes shown to participants, and interference effects were reported by Kojima and Kusumi (2006) for scenes presenting multiple objects. These interference effects are accounted for by an operation that will modify the acceptability of relations during a later phase of the computation.

In addition, a boundary constraint regarding the size of the objects can negate a directional relation. If the boundaries of one of the two objects are large enough on a given axis, such

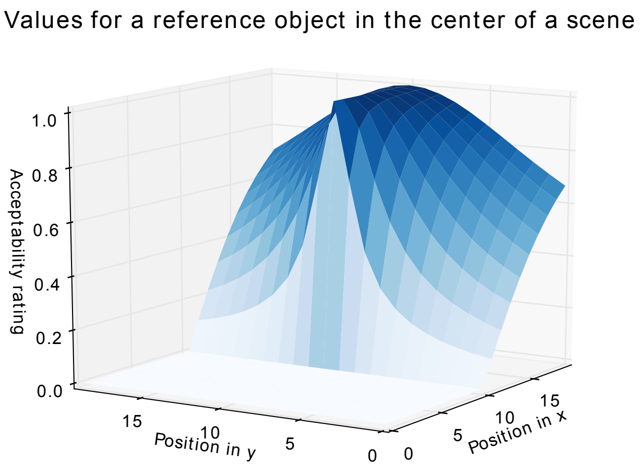

Figure 2: Surface plot of the acceptability ratings for all possible locations around a reference object for the spatial template *right of*, as defined in (3), restricted to a 2D plane

that the other object is within those boundaries, the relation corresponding to that axis is deemed irrelevant.

Interference effects

In a scene populated with many objects, the spatial templates previously described produce a high number of candidate objects for a given type of relation and reference object. According to Kojima and Kusumi (2006), there is an interference effect dependant on the proximity between those objects that change their acceptability ratings. This provides for the overall model an apprehension of spatial relations that is dependent on the context of the scene rather than depending only on the pairs of objects themselves.

For a given pair of objects, the presence of other objects will alter the calculation of spatial relations based on their proximity to the reference object. The presence of a closer object, might, in the case that it also satisfies the given spatial relation, decrease the acceptability of objects that are far from itself and increase slightly the acceptability of the local area. An example of this phenomenon can be seen in Figure 3 between the cube, the sphere, and the cone. The relation *the cone is to the left of the cube* has a lower acceptability rating than the relation *the sphere is the left of the cube* but it is deemed more relevant due to the interference of the proximity of the cone to the cube. A consequence of this effect is the reduction of the large set of candidate objects to a small set of more relevant ones, where the best matches for a given relation will supplant the others.

To model this effect for a given type of relation, such as *in front of*, the following steps are done once the acceptability ratings have been calculated for the relation between a given reference object and the other objects in the scene. For every object, the ratings given by the directional relation and distance relation are first combined into a product. The object with the strongest combined rating is considered a temporary reference object, and the spatial template for distance is then applied to the other objects, where the farthest object determines the range used for the relative domain calculations. This operation determines the proximity between the candidate objects from the strongest candidate, and all objects that get a rating above 0.90 are considered close enough such that their relation with the initial reference object is relevant. This last step models the effect that an object that previously was considered less acceptable for a given spatial relation might become more acceptable in a situation where it appears to be part of a cluster with a better candidate object for the spatial relation. An example of what might be considered a cluster is found in Figure 4, where the large cube and the cone are near each other. If we remove the cylinder and the small cube, both the large cube and the cone would be considered to the left of the sphere, which would not happen if the cone was only decreasing the acceptability of other objects, due to its greater proximity to the sphere.

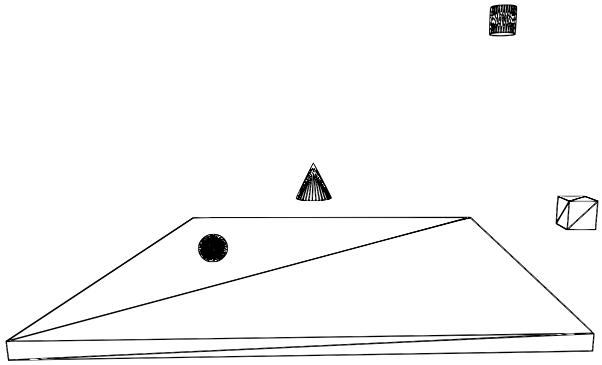

Figure 3: Front view of a sample scene containing five objects, including a large plane. In our model, the size of the plane inhibits the relevance of specific spatial relations. Interference comes into play when determining what is to the right of the cone, for example. The cylinder's location would be considered acceptable in an otherwise empty scene, but the presence of the cube, a much better candidate, makes the relation less relevant. A top view of the same scene is shown in Figure 5.

This interference operation is also applied to find the most relevant distance relations, and the procedure is identical with the exception of the use of a combined rating, as the acceptability ratings are directly used to find the closest or farthest object.

Method

We implemented our model as a Python 2.7 software package, available online[1], that parses XML files written according to the X3D standard. X3D is an ISO standard for representing 3D scenes, well known for its Web integration ca-

[1] http://github.com/science-of-imagination/mist

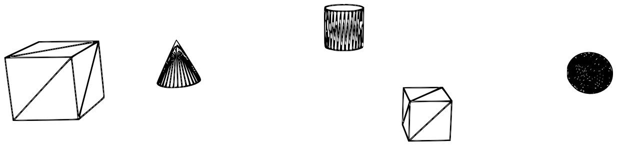

Figure 4: Front view of a sample scene presenting five objects in a row. The existence of objects between the reference object and the located object can make their relation irrelevant, which is the case for the sphere being to the right of the large cube. The row is oriented along the x axis to test the interference computation.

pabilities and for its compatibility with open source software (Hetherington, Farrimond, & Presland, 2006).

Our software takes an X3D file as input and identifies the objects described as either X3D shapes or sets of vertices. A shape, as defined in X3D, is a primitive object, such as a cone or a sphere, while sets of vertices are used to build complex meshes out of polygons described individually, a single set of vertices being associated with a polygon. The software then constructs an axis-aligned minimum bounding box for each object in the scene. A bounding box is also computed for the whole scene from the set of all objects. Bounding boxes are then used for all calculations.

While directional and distance templates can be directly implemented, the apprehension of intersections, necessary for topological relations, require computations specific to the geometrical representation of the objects, defined in the following equation:

$$rating = \begin{cases} 1, & \text{if } \dfrac{located_i + reference_i}{2} > distance \\ 0, & \text{if otherwise} \end{cases} \tag{4}$$

where *located* and *reference* are the dimensions of the corresponding bounding boxes along a specific axis and *distance* is the Euclidean distance between the center of the bounding boxes. The computation is done for each of the three axes, confirming an intersection as soon as a rating of 1 is returned.

For input we produced a set of hand-made files that represent scenes with geometrical shapes in the X3D format to qualitatively evaluate the effectiveness of the model. The files can be imported in Blender 2.5, an open source 3D content creation software, for the purpose of visualization. The scenes were created to expose the model to cases that would demonstrate its ability to perceive relevance among the set of all possible relations, which would test the effectiveness of both the interference computation and spatial templates.

Results and Discussion

A subset of the spatial relations found in the scene displayed in Figure 3 is shown in Table 1, as well as the acceptability ratings associated with them. The scale used for the acceptability ratings ranges from 0 to 1, where 1 represents a perfect degree of membership. The subset presented here is restricted to the relations found for the cube and the sphere. Table 2 presents a number of spatial relations found in the scene illustrated in Figure 4, the subset being restricted to the relations found for the cone and the cylinder.

Table 1: Spatial relations for the scene shown in Figure 3 and the associated acceptability ratings

Spatial relation	Rating
The cube is to the right of the cone	0.94
The cube is below the cylinder	0.87
The sphere is to the left of the cone	0.64
The sphere is above the plane	1.00
The sphere is below the cylinder	0.60
The sphere is behind the cone	0.60
The sphere is close to the plane	0.85
The sphere is far from the cylinder	0.87
The sphere is far from the cube	0.82

Figure 4 shows a situation where other models, the Attentional Vector Sum model and the Proximal and Center of Mass-Bounding Box model (Regier & Carlson, 2001), would, were the scene projected to a 2D plane along the horizontal axis, qualify the statement *the cone is to the left of the sphere* as a good spatial relation. The models ignore the presence of other objects to compute a spatial relation's acceptability rating, providing identical results for pairs of objects whether they occur in a crowded scene or alone. However, as described in (Kojima & Kusumi, 2006), the acceptability of the sphere as a located object that corresponds to *is left of* should be reduced by the presence of distractor objects, such as the cylinder and the small cube. Our model avoids such a pitfall thanks to the interference computation, and only returns that the cone is to the left of the cylinder. This shows that the model is sensible to all objects in the scene, producing results that are context-dependent.

Table 2: Spatial relations for the scene shown in Figure 4 and the associated acceptability ratings

Spatial relation	Rating
The cone is to the right of the large cube	0.98
The cone is to the left of the cylinder	0.94
The cone is far from the sphere	0.95
The cone is close to the large cube	0.81
The cylinder is to the left of the small cube	0.67
The cylinder is above the small cube	0.67
The cylinder is far from the large cube	0.93
The cylinder is far from the sphere	0.85

A situation in which the model developed by Smith et al. (2010) produces irrelevant or incorrect descriptions is the case of large objects appearing in a scene, such as the one present in Figure 3 and 5. The scene would lead this model to assert that the cone, the lowermost object in Figure 5, is in front of the large plane. This is the case because the z value of the centroid of the objects are compared in this model, without accounting the boundaries of the objects. Therefore, since the center of the cone is more forward than the center of the plane, the cone would be considered to be in front of the plane. To improve upon this model and in an effort to prune away the weakest spatial relations, our model disqualifies the sphere and the cone for some directional relations, based on the boundaries of the objects. In Figure 3, for example, the spatial relation *above* is the only one considered relevant by our model, which is analogous to the intuition that people rarely locate themselves in a room through the location of the floor's geometric center.

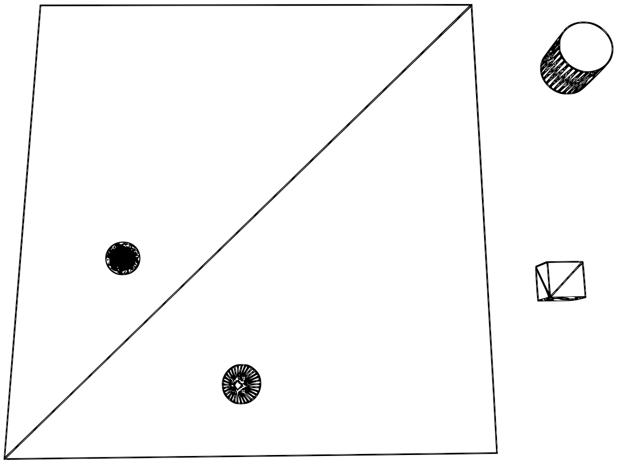

Figure 5: Top view of the sample scene shown in Figure 3, illustrating the position of the objects along the z axis. The cone is located at the bottom of the figure, near the edge of the large plane.

Applications

Our model describes scenes in a way that is similar and relevant to the human cognition of spatial relations. Representing space in that manner can be applied in a number of situations, as described below.

One of these applications can be found in imagination research, where a set of instructions are given to a drawing system and a visual scene is produced. Suppose you are reading a book and room is described with only the following sentence: *In the living room, a couch was close to the fireplace, with a long table to its left*. This will still give you, as a reader imagining the scene, a complete picture of a room, even though many other elements, including absolute distances, are not mentioned. A typical scene description obtained through the model would offer plausible spatial relations to accompany the elements in this sentence, as well as defining the mentioned spatial relations.

WordsEye is such a system, and is able to parse a paragraph to create a scene with its database of 3D models (Coyne & Sproat, 2001). However, it uses, for each object in its database, hand-coded ranges of spatial coordinates associated with all possible spatial terms parsed by WordsEye. The model could therefore expand WordsEye's capabilities, providing a way to automatically produce plausible ranges of spatial coordinates for any 3D object added to its database. The model will also dynamically change those ranges of spatial coordinates based on the scene in which the object is located, according to its context-sensitive capabilities offered by the interference computations and the frame-dependent spatial templates.

Designers of intelligent agents in tactical simulations, as those used in firefighting training systems, could benefit from realistic and automatic spatial understanding, our model allowing them, for example, to use instructions such as *walk in front of the table* directly without needing them first translated into absolute spatial coordinates. Without a model able to discern relevant relations, the designers would have to specify the location of the region that is considered in front of the table to the agent, forcing the designers to attribute a set of spatial coordinates for that location. However, protocols and human communications are encoded in qualitative terms, such as "in front of," and not in terms coordinates.

Extending the parsing abilities of our software implementation would also offers the possibility of taking in any number of virtual scenes and assessing their spatial properties, uncovering patterns useful to designers, especially in the case of environments created by a multitude of people such as ones found in public databases of assets. The production of scenes based on those patterns would also offer those designers a way to guarantee that their content is similar to what a human would expect it to be like. A benefit of this approach is that it would inform virtual and real world designers alike, in the case of reconstructed scenes, on features that make an environment desirable to users, such as the ease of navigation.

Conclusion

We presented a model of spatial relation apprehension and its software implementation that is able to parse scenes in three dimensions and return descriptions of the scene using spatial terms related to distance, direction, and containment. Our model was developed to reproduce how human viewers would perform when they are asked to describe a scene composed of abstract objects. To improve the relevance of the descriptions, as to reduce the number of statements returned by the system, cognitively plausible spatial templates and interference effects were modeled. The interference effects, in particular, produce results that are context-dependent, as the presence of other objects in the scene modify the acceptability ratings of the spatial relations. Future work includes the evaluation of the model with experiments involving participants, the integration of the model to simulations where agents perform tasks involving spatial cognition, and the development of design-oriented capabilities.

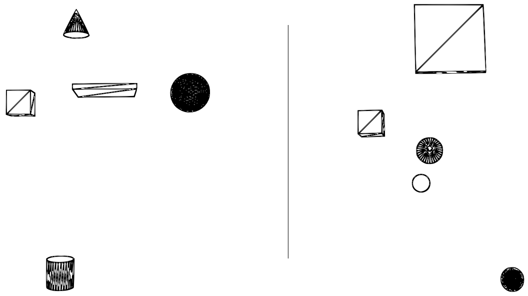

Figure 6: Another scene, containing five objects, tested with our model. A front view is shown to the left and a top view is shown to the left. The locations were randomly generated.

Table 3: Spatial relations output by our model for the scene pictured above and the associated acceptability ratings

Spatial relation	Rating
The plane is to the right of the cube	0.69
The plane is above the cylinder	0.77
The plane is behind the cone	0.91
The plane is far from the cylinder	0.80
The plane is far from the sphere	0.87
The cone is to the right of the cube	0.59
The cone is above the cube	0.71
The cone is behind the sphere	0.74
The cone is far from the cylinder	0.87
The cylinder is below the cube	0.89
The cylinder is in front of the plane	0.91
The cylinder is behind the sphere	0.52
The cube is in front of the plane	0.69
The cube is behind the sphere	0.76
The cube is far from the sphere	0.87
The sphere is to the right of the cube	0.65
The sphere is above the cylinder	0.70

References

Carlson-Radvansky, L. A., & Logan, G. D. (1997). The influence of reference frame selection on spatial template construction. *Journal of Memory and Language*, *37*(3), 411–437.

Coyne, B., & Sproat, R. (2001). Wordseye: an automatic text-to-scene conversion system. In L. Pocock (Ed.), *Proceedings of the 28th annual conference on computer graphics and interactive techniques* (pp. 487–496).

Gorniak, P., & Deb, R. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research (JAIR)*, *21*(2), 429–470.

Hayward, W. G., & Tarr, M. J. (1995). Spatial language and spatial representation. *Cognition*, *55*(1), 39–84.

Hetherington, R., Farrimond, B., & Presland, S. (2006). Information rich temporal virtual models using x3d. *Computers and Graphics*, *30*(2), 287–298.

Kojima, T., & Kusumi, T. (2006). The effect of an extra object on the linguistic apprehension of the spatial relationship between two objects. *Spatial Cognition and Computation*, *6*(2), 145–160.

Kojima, T., & Kusumi, T. (2007). Computing positions indicated by spatial terms in three-dimensional space. *Psychologia*, *50*(3), 203–223.

Lockwood, K., Forbus, K., & Usher, J. (2005). Spacecase: A model of spatial preposition use. In *Proceedings of the 27th annual conference of the cognitive science society* (pp. 1313–1318).

Logan, G., & Sadler, D. (1996). A computational analysis of the apprehension of spatial relations. In *Language and space language speech and communication.* The MIT Press.

Mukerjee, A., Gupta, K., Nautiyal, S., Singh, M. P., & Mighra, N. (2000). Conceptual description of visual scenes from linguistic models. *Image and Vision Computing*, *18*, 173–187.

Regier, T., & Carlson, L. A. (2001). Grounding spatial language in perception: am empirical and computational investigation. *Journal of Experimental Psychology: General*, *130*(2), 273.

Smith, C., Kathrine, V. B., Nuttall, J., Musca, J.-M., MacDougall, K., Miller, X., … Davies, J. (2010). Modeling english spatial preposition detectors. In A. K. Goel, J. Matega, & N. H. Narayanan (Eds.), *Proceedings of theory and application of diagrams (diagrams-2010)lecture notes on artificial intelligence 6170, springer.* (pp. 328–330). Berlin.