Size matters: Grounding quantifiers in spatial perception

Pauw, S.

Publication date
2013

Chapter 4

Grounded Categorization and Perceptual Deviation

Grounding language in sensorimotor spaces is an important and difficult task. In order for robots to be able to interpret and produce utterances about the real world, they have to link symbolic information to continuous perceptual spaces. This requires dealing with inherent vagueness, noise and differences in perspective in the perception of the real world. This chapter presents two case studies for spatial language and quantification that show how cognitive operations – the building blocks of grounded procedural semantics – can be efficiently grounded in sensorimotor spaces.

This chapter is a shortened version of a paper by Michael Spranger and myself on this subject: Spranger, M. and Pauw, S. (2012). Dealing with perceptual deviation: Vague semantics for spatial language and quantification. In Steels, L. and Hild, M., editors, *Language Grounding in Robots*, pages 173–192. Springer, New York

4.1 Introduction

Noisy sensor readings and algorithmic estimation errors make it difficult for autonomous systems to acquire stable, precise, and correct estimates of the environment. Moreover, language always happens between different individuals. When two interlocutors interact in a spatial scene, they will each see the world from their viewpoint and, consequently, estimate properties of objects in the world differently. We subsume such problems under the term *perceptual deviation* which denotes that two artificial agents in the same physical space estimate the properties of objects in their environment differently.

The problem of perceptual deviation is one that humans navigating the physical world face as well. For instance, people systematically estimate distance wrongly (Foley, 1980). Humans also have vastly varying sensor precision which has been observed, for instance, in color vision. Even people with average color vision, i.e. non color-blind subjects, have different retinal distribution of mid and long wave-length cones (Roorda and Williams, 1999). Lastly, humans interacting in spatial environments also perceive the world from their respective viewpoints.

Nevertheless, humans link symbolic information to noisy sensory information effortlessly. How this can be achieved for artificial systems has long been ignored. Traditional logic-based approaches to semantics focus almost entirely on the symbolic level and leave details of how to link semantics to sensorimotor spaces open. At the heart of such approaches is the notion of *strict* membership. A phrase such as "left blocks" is true for all objects which are blocks and to the left, in other words, all objects which are a member of the set of blocks and a member of the set of left objects. Consequently, each object in the world is either part of these sets or not. This idea can cause problems when being exposed to real-world problems such as perceptual deviation. An object might be to the left for one interlocutor but not to the left for another.

The classical approach has been criticized by psychologist and linguists alike. Rosch and Lloyd (1978), Lakoff (1987) and Langacker (1987) are examples of researchers who argue that human categorization is graded rather than strict. In their view, objects are more or less prototypical for a concept. Some objects are more `block` than others. They conclude that concepts are represented by prototypes, i.e., prototypical objects which allows other objects to be compared to them. Such a *lenient* view on the meaning of concepts has been used successfully to ground lexical language in sensorimotor streams (see Steels and Spranger, 2008; Bleys et al., 2009, for examples from action language and color). However, compositional semantics, the problem of how lexical items are combined into larger compositional semantic structures, has been mostly absent from these discussions. Furthermore, many of these proposals do not go far enough and fall back onto some version of strict membership. This chapter introduces a particularly strong version of lenient categorization that is exceptionally successful in dealing with problems of perceptual deviation and that is implemented in a larger framework for handling compositional semantics

To compare our proposal to traditional approaches, we operationalize the different ideas for a concrete piece of natural language: spatial language. We implemented spatial semantic primitives such as spatial categorization, perspective reversal and landmark processing, as well as quantifiers separately for the strict and the lenient approach in a formalism called Incremental Recruitment Language (IRL) (see Chapter 3 or Spranger et al., 2012b). We test each implementation in robot-robot interactions, called *spatial language games* (Spranger, 2011; Steels, 2012b) in which one robot is trying to draw attention to an object in the environment using spatial language. Subsequently, we can measure and quantify the success of these interactions and show why the lenient approach outperforms the classical approach.

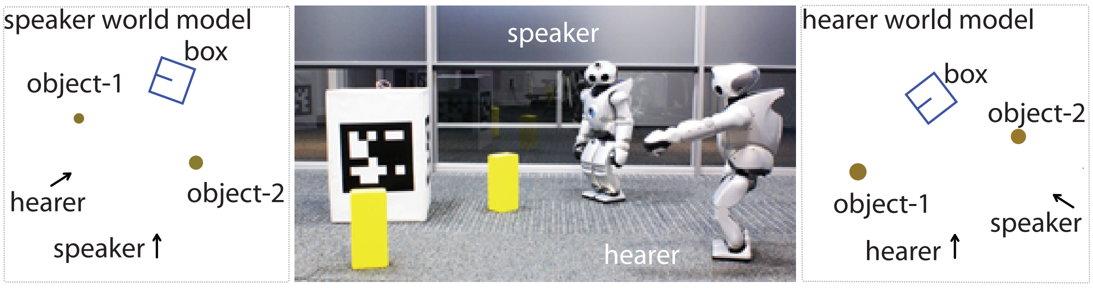

Figure 4.1: Experimental setup involving robots, blocks and a box.

4.1.1 Spatial Language Games

In order to study the effect of perceptual deviation we use an experimental setup in which two humanoid robots interact in a shared environment (spatial scene). One robot, the speaker, is trying to draw attention to an object in the environment using spatial language (see Figure 4.1). Here is the language game the robots play.

1. Both agents perceive the environment using their own camera. The vision system (Spranger et al., 2012a) computes a situation model (see Figure 4.1, left and right) which is comprised of blocks (circles), boxes (rectangle) and other robots (arrows). The perceiving robot is always the center of the coordinate system which is used to estimate distance and orientation of objects. The boxes have an inherent front which is visually marked and know by both robots.

2. The speaker picks an object from the context and conceptualizes a meaning for discriminating it. If he succeeds in finding an appropriate meaning, the structure is encoded in an utterance and passed to the hearer.

3. The hearer interprets the utterance by recovering the semantic structure and trying to find the object that the utterance refers to.

4. The hearer points to the object and the speaker confirms whether he pointed to the correct object.

Figure 4.1 shows real-world perceptual deviation problems. For the speaker, `object-1` is more to the left of the box (from the perspective of the box), whereas for the hearer the same object is more in front of the box (the front of the box is denoted by the small line in the rectangle). Figure 4.1 shows one scene[1] from close to 900 spatial scenes which we have recorded. Scenes differs in the number of objects and whether boxes are present or not. Some scenes have one box, some

[1] Each scene consists of two situation models, one for each robot.

feature/measure	average	stddev	min	max
distance deviation	7.2cm	6	0.002cm	59.4cm
angle deviation	8°	0.13	0.04°	51°

Table 4.1: Average, standard deviation (stddev), min and max values of angle and distance differences (angles in degrees) over 800 real-world spatial scenes.

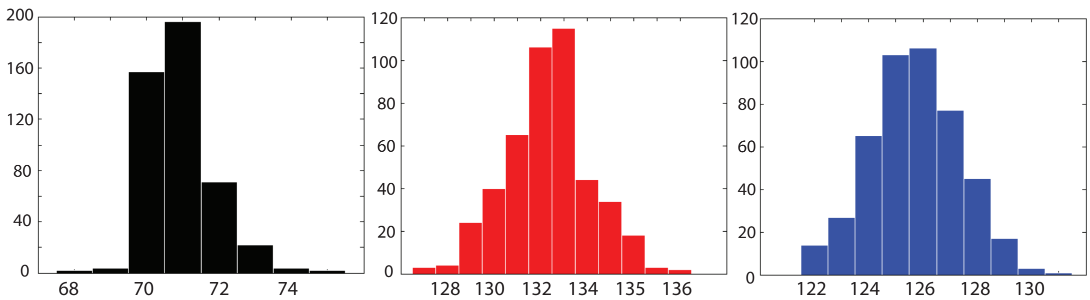

Figure 4.2: Camera noise histogram. The YCrCb values of a single pixel over time are recorded and analyzed using histograms (left - y-channel, middle - Cr-channel, right - Cb-channel

do not have any boxes. Some scenes have two objects, others up to 10 objects. For such scenes, we can precisely quantify the degree of perceptual deviation by measuring the differences in distances and angle for each perceiving robot. For example, `object-1` in the speaker world model (left image) has a distance of 81cm to the speaker. The hearer estimates the distance of the object to the speaker as being approx. 75cm. The estimation of the hearer is based on the distance he thinks the speaker has to the object. The following table shows the average differences in distance and angle measured for each robot and each object over 897 spatial scenes.

The table shows that the average perceptual deviation for the distance channel is 7cm with outliers that diverge up to 60cm. These are high numbers, but it is the sort of distance estimation problem one gets even with sophisticated computer vision systems. On average, angles diverge by around 8° with some going up to 51°. Based on these values we can conclude that perceptual deviation is always present and in some cases a quite severe problem.

4.1.2 Sources of Perceptual Deviation

There are four main sources for perceptual deviation: 1) sensor deviation 2) noisy or faulty sensors, 3) errors arising from algorithms used in estimating object properties, 4) differences in viewpoint on the scene.

Sensor deviation Sensors vary across individuals. We have already given an example for human color vision earlier. The same also holds for robots. For instance, CCD cameras from the same manufacturer have differences in light collection stats due to manufacturing margins.

Sensor noise Every sensor is noisy. Based on the type of sensor different sources of noise can be identified. For instance, CCD devices suffer from transfer inefficiency and shot noise (Healey and Kondepudy, 1994). Figure 4.2 shows color sensor readings taken by a digital camera in a static spatial scene. The graph shows the histograms of sensor readings from a single pixel for three different color sensors (brightness, red and blue channel). The histograms show that color readings vary over time.

Estimation errors Another source of errors and noise is related to algorithms used in object recognition and object feature extraction. For instance, the algorithm for the distance estimation of objects (see Spranger et al., 2012a) has distance estimation error properties shown in Figure 4.3. To estimate the position of objects the algorithm combines noisy sensor readings and integrates them over time and across different sources of information. In the process, noise and uncertainty from different sensor sources accumulate and potentially amplify.

Differences in perspectives Another source for perceptual deviation comes from the fact that agents perceiving the world from different bodies necessarily have different viewpoints on the scene. On the on hand, objects can look different from different angles and light conditions might vary across the environment. On the other hand, spatial properties are inherently egocentric. I can estimate the position of an object from my viewpoint, but my distance to the object is most likely different than from another person's point of view.

4.2 Strict Semantics

Consider an example of spatial language that highlights the problems that perceptual deviation causes for the strict approach. Suppose two robots interact in a spatial scene such as the one in Figure 4.4. The speaker says, "the block to the left of the box", to draw attention to `object-1`. For him this is an acceptable phrase for discriminating the object[2]. After all the object is the only block in the region to the left of the box. When the hearer interprets the phrase using the same mechanism, he fails. For him the object is to the right of the box and the set of blocks to the left of the box is actually empty. Obviously, the problem stems

[2] We assume an *intrinsic* interpretation of the phrase (Tenbrink, 2007).

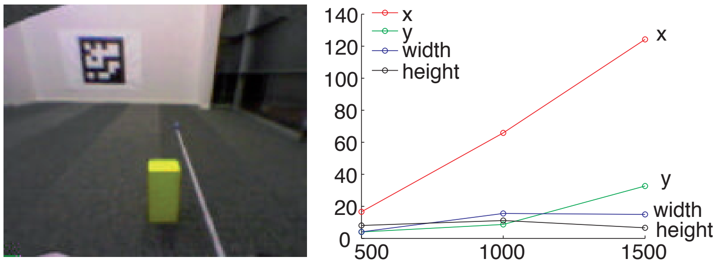

Figure 4.3: Measuring estimation errors. The block was put at 500, 1000 and 1500mm distance from the robot. Each time the vision system estimates the features (width, height, x and y) of the block. The graph to the right shows the root-mean-square-error (RMS) for each measurement. The x feature (x-axes runs towards the front) is most heavily affected by increasing distance.

from the fact that the hearer is applying a strict interpretation of the phrase. For him the region left has a fixed border and everything within the region is considered left.

Strict approaches can be implemented in different ways. For instance, a spatial relation can be characterized by the set of locations in space to which it applies (highly intractable in real-world scenarios), using regions (Kelleher and Costello, 2005), adaptive networks (Belpaeme, 2002), axioms (Eschenbach and Kulik, 1997), exemplars (Steels and Kaplan, 2002) and centroids (Bleys et al., 2009). Common to all attempts is that there are strict boundaries for category membership. An object either belongs to a certain category or not. For the sake of the argument, we only focus on centroids hereafter.

Centroids are the geometric center of convex regions in a particular sensorimotor space. For spatial relations such as `left` and `right`, for instance, centroids are points in the radial space around the robots. An object is considered to be `left` (or member of the category), when its angle is closest to the spatial point `left`, otherwise it is categorized as `right`. Consequently, every point in the sensorimotor space belongs to precisely one category from a particular set of categories and the complete sensorimotor space is decomposed into different sets of objects based on their category membership, a process known as Voronoi tesselation.

However, categorization is not enough. In order to refer to an object and try to draw attention to it, one has to *discriminate* the object from others in the set of objects. Therefore, a second condition is introduced. A category, say `left` is discriminating an object from the context, if the object is the only member of the category.

Here are the two conditions:

Strict category membership An object o is said to be a *strict member* of the category c, iff o is closer to c than to any other category from the repertoire

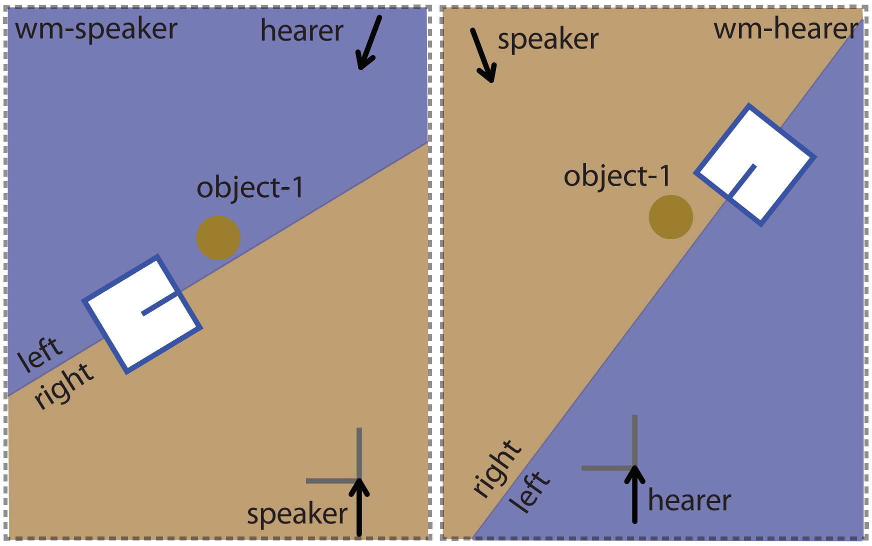

Figure 4.4: Impact of perceptual deviation. While for the robot to the left the object is left of the box, the same object is not left of the box for the robot on the right.

of categories C. This is known as categorization in machine learning.

Strict discriminating category A category c is said to be *strictly discriminating* when o, iff o is a strict member and the *only* member of the category.

Let us apply this to an example of compositional semantics say the meaning of the utterance "the left block". Figure 4.5 (left) shows the semantic representation of the utterance in IRL. The lexical items "the", "left" and "block" appear as so-called `bind`-statements which are pointers to the concept. All other nodes in the network are *cognitive operations* which denote how these concepts are processed. The referent of the phrase is computed (Figure 4.5, right) by going through every operation and executing it, a process known as *evaluation*.

`get-context` Introduces the situation model via the variable `?ctx`.

`filter-by-class` Applies the object class `block` by filtering objects in the context for those of type block, i.e. those who are strict members of the category. The result is available via the variable `?blocks`.

`filter-by-spatial-category-group-based` Applies the spatial relation `left` to further constrain the set of objects in the context for all objects that are to the left. The result is published in the variable `?left-blocks`[3].

[3]The precise implementation of this filter operation is based on the meaning of projective adverbs in English (Tenbrink and Moratz, 2003)

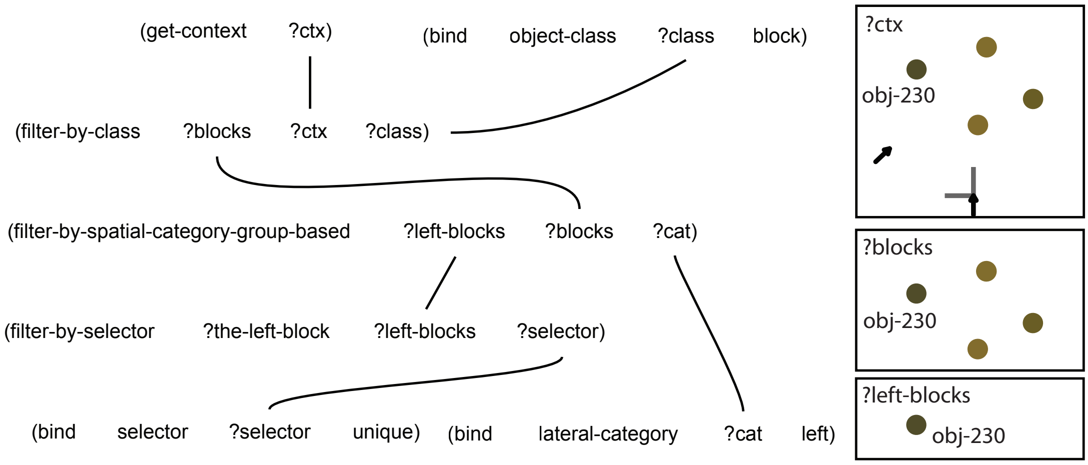

Figure 4.5: On the left side, the IRL-network of the phrase "the left block" with `filter` operations is shown. The images to the right show the progressive filtering of the set of objects in the context.

`filter-by-selector` This operation has as input the set of left blocks. It checks whether the input set contains only a *single* object (`unique`) and returns it if there is only one. This operation implements the discriminating category condition.

4.3 Lenient Semantics

Many scholars propose alternative principles guiding semantic processing. Rather than relying on strict membership and strict discrimination, they require that an object o is the closest object to a category c without further constraining the other objects in the context O and their relationship to the category c. Consequently, other objects in the context O can be strict members of the category c as long as they are not closer to c than o. Psychologists, for instance, have found that in many discrimination tasks the choice of categories seem to be based on the principle of *greatest distance* or *greatest contrast* which only requires the category to establish sufficient difference between the distance of object o and all other objects in the context. These principles are used to explain human behavior in general object discrimination tasks (Hermann and Grabowski, 1976) but have also been applied to spatial language (Herskovits, 1986; Freksa, 1999). Tenbrink (2005), for instance, found that unmodified projective terms are frequently used by participants even though objects were far away from the prototypical axes.

Based on these observations, we propose a novel approach to implementing semantics which we termed *lenient*. Our approach considers similarities to cat-

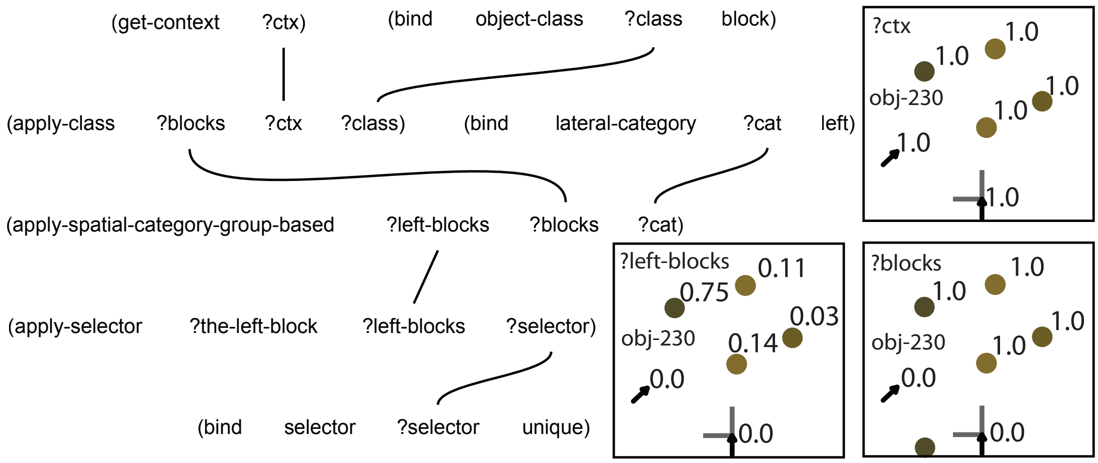

Figure 4.6: On the left side, the IRL-network of the phrase "der linke Block" (the left block) with `apply` operations is shown. The images to the right show the progressive scoring of objects in the context through the operations in the network.

egories without enforcing the strict membership criteria. Figure 4.6 shows the semantic structure for the phrase "the left block" using the lenient approach. The IRL-network is structurally the same. Only the implementation of cognitive operations is changed (signified by the prefix `apply-` instead of `filter-by-`).

`apply-class` Applies an object class by scoring each object using a similarity measure. Here, every object in the context is scored based on its similarity to the object class `block`. The result is available in `?blocks`. (Note that the membership of block is in fact strict: the similarities are either 0 or 1.)

`apply-spatial-category-group-based` Applies the spatial relation `left` by multiplying the similarity of each object with the spatial relation with the similarity to the object class `block`. The result is published in `?left-blocks`.

`apply-selector` This operation then applies the `unique` selector to the objects in `?left-blocks`. Here, this is implemented as choosing the object with the highest similarity score in the input.

The most important thing to note is that no filtering occurs. Rather, objects are scored based on their similarity to concepts and spatial categories. Only at the very end the quantifier picks the referent of the phrase. This sort of processing can deal with the initial problem presented in Figure 4.4. Upon hearing the phrase "the block to the left of the box" (see Figure 4.4), an interpreter is still able to identify `object-1` using the lenient interpretation, because the block `object-1` is the leftmost of all blocks.

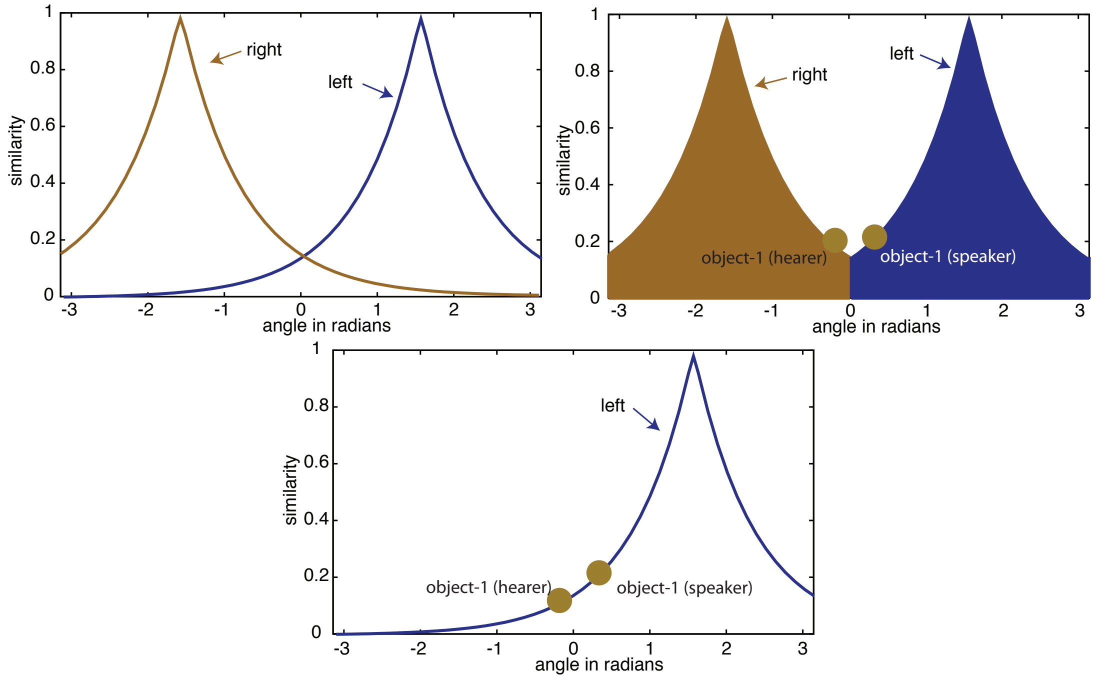

Figure 4.7: Lenient versus strict categorization. Top left figure: similarity functions for `left` and `right` categories over the angle. Top right figure: decomposition of the angular space using the strict approach. The bottom figure shows how the lenient approach uses the similarity function of the spatial category to retrieve the correct object.

Figure 4.7 shows why the lenient approach solves the situation more adequately than the strict approach. The top left figure shows the similarity functions for the spatial categories `left` and `right`. The decomposition of the angular space used in the strict approach is shown in the top right figure. The block `object-1` is categorized by the speaker as being to the left, whereas the same object for the hearer is to the right. When the speaker thus conceptualizes the object as left, the hearer has no chance of retrieving the object using strict interpretation. On the other hand, when applying a lenient discrimination scheme (bottom figure), whether or not the hearer is able to discriminate the correct object depends on whether `object-1` is the most similar object to the category `left` (which is the case, for this example).

4.4 Comparing Strict and Lenient Spatial Semantics

The operationalization of strict and lenient semantics allows us to study the difference between the two approaches systematically. Agents interact in controlled spatial scenes and we measure which of the two approaches performs better in a discrimination task. Here, we concentrate on the semantics only. Therefore, we scaffold syntactic processing and use *direct meaning transfer*. The hearer is passed the IRL-network conceptualized by the speaker without going through production and parsing of syntactic structure. This is equivalent to having a language without uncertainty, ambiguity or loss of information.

Section 4.1.1 describes the interaction script that is the basis of our investigation. Different steps of such an interaction can fail. We consider the following four outcomes of an interaction:

Conceptualization failed (step 2) After the speaker chooses a topic, he has to conceptualize an IRL-network that discriminates the topic. This process fails if the speaker cannot find any IRL-network that allows him to discriminate the object from all other objects in the context.

Interpretation failed (step 3) After the speaker successfully conceptualized a discriminating IRL-network, the hearer interprets this structure by simply evaluating the network. If this evaluation yields no result, the hearer is said to have failed.

Pointing failed (step 4) When the hearer successfully interpreted the semantic structure passed to him by the speaker, he points to the topic he interpreted. The speaker then checks whether the object pointed to is indeed the topic. If this is not the case then pointing failed.

Success If the hearer points to the correct object then the game is a success.

We setup two different populations of agents. In one population, agents are equipped with lenient semantics in the second population all agents are equipped with strict semantics. Both types of agents can handle the same complex spatial semantics such as group-based reference (Tenbrink and Moratz, 2003), landmarks (Mainwaring et al., 2003), frames of reference (Levinson, 1996) and perspectives (Taylor and Tversky, 1996). Agents are given a set of English proximal (near, far) (Kemmerer, 1999) and projective (front, back, left, right) spatial categories (Tenbrink, 2007). The implementation of these complex semantics is part of a larger effort on spatial language (see Spranger, 2011, for an overview).

Performance is tested on different subsets of 897 pre-recorded spatial scenes. We consider two data sets: one containing scenes with *few objects* (on average 4) and the other containing scenes with *many objects* (on average 10).

Figure 4.8 compares the lenient and the strict approach for the two sets of spatial scenes. Clearly, the lenient approach has a communicative advantage over the strict implementation. Success in interaction for the lenient approach is consistently above 85% across the two environmental conditions, whereas the success of strict categorization drops to 22% in the most difficult *many objects* condition. This means that only approx. one in four games is a success using strict interpretation compared to more than 4 out of 5 for the lenient case. Notably, the lenient approach is able to successfully conceptualize the spatial scene for the topic in question in almost all scenes. Only few cases in the *many objects* condition are marked for failure in conceptualization. On the other hand, the strict approach shows enormous problems even conceptualizing for particular objects in particular scenes. Almost all cases of failure are either due to failures of conceptualization or failures of interpretation, where conceptualization takes the major blame for failure. The two conditions show that the more objects there are in a scene the more severely the strict approach is affected.

Apart from the number of objects, the *number of categories* also influences performance. *Failures to conceptualize* are caused entirely by insufficient clustering of the input space. The problem is that there are not enough categories to allow the speaker to discriminate the topic object. On the other hand, *failures to interpret* and *pointing failures* are caused by perceptual deviation. In order to control for lack of categories, we compare four additional conditions: *english*, *double*, *triple* and *quadruple*. The *english* condition is the same as used in the previous results: agents are given sets of English categories. In the *double* condition, the number of categories is doubled. Instead of two lateral categories left and right there are now four. The same holds for frontal and proximal categories. In the triple and quadruple condition, agents are equipped with three and four times as many categories. In each condition the sensorimotor space is equally decomposed by the categories.

Figure 4.9 shows results. The left two groups of bars show the performance of the lenient approach versus the right two groups which show the strict approach. Results reveal not much change for the lenient approach. However, the perfor-

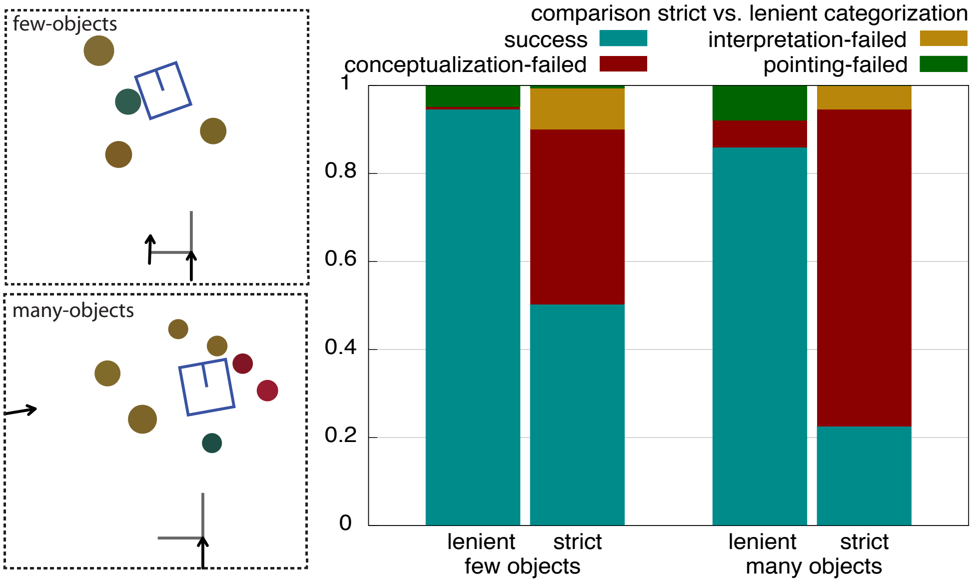

Figure 4.8: Results of comparing strict versus lenient categorization (right image) on different sets of spatial scenes (left images).

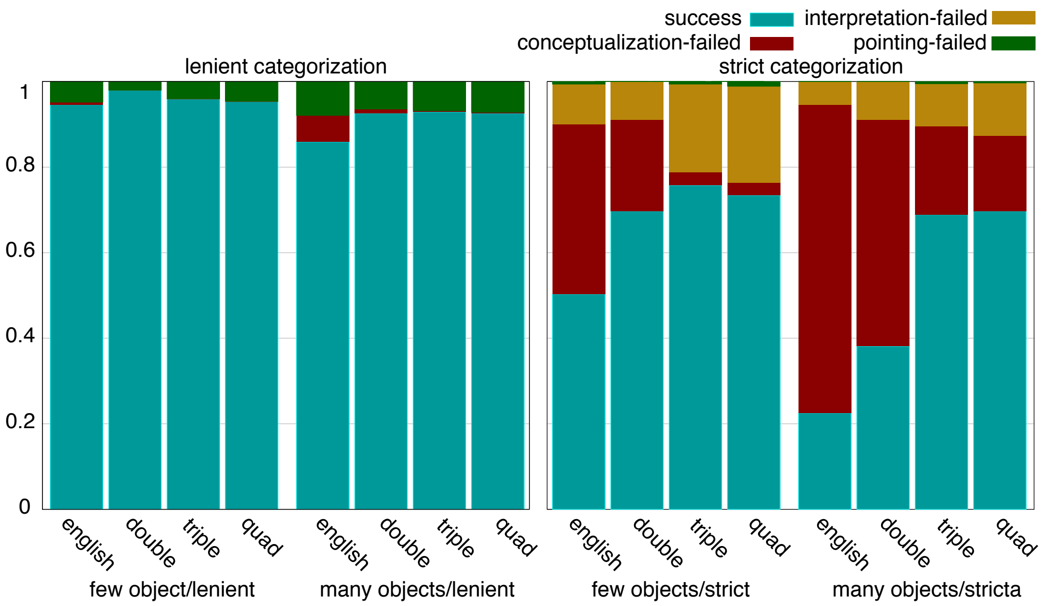

Figure 4.9: Results of comparing different sets of spatial categories and their effect on strict and lenient semantic processing.

mance of the strict approach increases drastically with more categories. But, the graph also shows a saturation effect. Success actually drops again for quadrupled number of categories. Failures of the speaker to conceptualize are replaced by the inability of hearers to interpret and, to a lesser extent, by errors in pointing. This means that the more categories there are available the more impact perceptual deviation has on the strict set approach. The reason is that the more categories, the smaller the area of categories in the sensorimotor space. Consequently it becomes more likely that an object categorized as belonging to a certain category by the speaker will be categorized differently by the hearer.

4.5 Discussion

Up to this point we have given a detailed account of our lenient approach to semantics in search of a solution for the problem of perceptual deviation. However, in the field of linguistic vagueness the aptness of such an approach is highly debated. Although, the problems that are being discussed in this field are of a very different nature than our own, we feel that we can not entirely omit touching upon this discussion.

In linguistics, the discussion of vagueness focusses mainly on gradable adjectives such as "tall]' or 'bold' that have no clear semantic boundaries (see van Rooij, 2011 for an overview). Gradables can be be relative ("tall") or absolute ("flat"). Precise concepts can be made vague, for example by using hedging expressions (Lakoff, 1973) such as "about" and "roughly". And, even seemingly precise concepts are often used in a vague way. For example, a round number such as "twenty" is often used as an approximate (Krifka, 2007).

The need for a model that can deal with vagueness is widely recognized, however the kind of model that should be used is a point of dispute. Traditionally, vagueness focusses on the existence of borderline cases of utterances such as "john is tall". An example is three valued logic where such an utterance can be true, false or undefined. Most modern accounts of vagueness fall under one of two approaches: Degree-based and delineation-based approaches to vagueness (van Rooij, 2011).

Degree-based approaches (Zadeh, 1965; Stechow et al., 1984; Kennedy, 2007) assume that category membership can be expressed in terms of degrees. Such a degree is typically a score between 0 and 1. The model as presented in this chapter is an example of such an approach.

Delineation approaches (Lewis, 1970; Kamp, 1975; Klein, 1980) assume that gradable adjectives are strict predicates, but that the membership of an individual is context dependent. For example, "tall" and "bald" do have cut-off points in every specific context, but the actual cut-off point for all possible contexts is underdetermined. Super-valuation is an example of such an approach.

Clearly, the model we propose is an example of a degree-based approach and is therefore subject to the objections that come with such approaches. Proponents of delineation approaches point out two problems with degree-based analyses. First of all, degree-based approaches fail to preserve necessary logical properties. For example in Fuzzy Semantics $p \wedge \neg p$ is not necessarily false. Secondly, it is not clear what the degrees reflect or where they come from. Are they probabilities? Neuron activation levels?.

So why do we use a degree-based approach in spite of these objections? First of all, most of the problems with degree based approaches are well beyond the scope of the current model. The language games require the agents to discriminate objects to each other, not to establish truth. For this purpose the question if specific logical properties are respected is not of much concern. For example, the inference ”x is taller than y” implies ”y is shorter than x” is not addressed in referential language games. The second more important reason is of a more practical nature: Degrees are the vantage point of our model. The data as described above is continuous. The classification of a perceived object requires some sort of comparison to an internal representation based on similarity measures. A degree-based approach that directly operates on these similarity measures provides therefore a straightforward model of semantic processing. Lastly, delineation-based approaches are not impervious to complications either. Proponents of fuzzy logic (Lakoff, 1973; Wright, 1975; Kamp, 1981a) argue that such accounts are inadequate because they still rely on unnatural borders. It is cognitively implausible that a cut-off point for the word “tall” exists, even for one particular valuation function. And even if it does exist, the ontological status of such a valuation is just as unclear as that of the degrees. So, it is not the case that there is a problem free, ready to use alternative that we are omitting.

4.6 Final Remarks

Obviously, traditional approaches to semantic processing have a lot to offer and made many important contributions, for example, related to reasoning. In our view, the way to combine the two approaches and therefore leverage the great results of logic-based semantic theories is by distinguishing discrimination from description. The notion of truth makes a lot of sense in description tasks, where an accurate description either fits a situation or not. This contrasts with discrimination tasks where truth is not an immediate concern but rather the contrasting of objects from other objects seems dominant. The lenient approach works well in discrimination tasks and can easily be extended to work in description tasks, for instance, by reintroducing acceptability limits. The lenient mechanisms allow agents to track how acceptable a category is for a particular object and, hence, can also be used to make true/false distinction thresholding the similarity landscape and modulating the interpretation of quantifiers in determined noun phrases.

This chapter has argued for a particularly lenient way of grounding meaning in sensorimotor data streams. We have taken the domain of spatial language and illustrated the practical effects and advantages of our model. We compared the performance of the lenient approach to the dominant approach in semantic theory and argued that our approach outperforms traditional semantic processing in discrimination tasks. The experiments show that real world tasks require a rethinking of deep aspects of semantic theory.

Acknowledgements

The research reported here was carried out at the Sony Computer Science Laboratories in Paris and Tokyo. We are greatly indebted to Masahiro Fujita, Hideki Shimomura, and their team for creating the Sony humanoid robots and for making them available for the experiments reported here. This research was funded by the Sony Computer Science Laboratory in Paris with additional funding from the ECAGENTS and ALEAR projects funded by the EU FP6 and FP7 frameworks.