CHAPTER 10

Towards interpersonal perception

> *"What is this thing, anyway?" said the Dean, inspecting the implement in his hands. "It's called a shovel," said the Senior Wrangler. "I've seen the gardeners use them. You stick the sharp end in the ground. Then it gets a bit technical."* (Pratchett, 1991a)

Harnad (2002) argues that creatures can learn about categories of objects or other entities either through toil or theft. Sensorimotor toil is his term for "trial-and-error learning, guided by corrective feedback from the consequences of miscategorisation." This is inherently expensive in terms of time and risk. Linguistic 'theft', on the other hand, allows categories to be passed on from other individuals, at a much reduced cost to the recipient. Harnad is mostly concerned with arguing that it can't be theft all the way down, that there must be some grounding in toil. In this thesis, the robot has so far been doing a lot of toil, so it would be interesting to see if it could start doing some theft.

The goal of this chapter is to build the tools necessary for the robot to familiarize itself with novel activities. Since automatic action understanding is currently very limited, social interaction is used as scaffolding for this learning process, to 'steal' structural information about activities from a cooperative human. Learning about activities is important because they provide tools for exploring the environment. Chapter 3 showed that, with a built-in poking activity, the robot could reliably segment objects from the background (even if it is similar in appearance) by poking them. It could determine the shape of an object boundary in this special situation, even though it cannot do this normally. This is the desirable feature of activities for learning – they provide special enabling contexts. In fact, they are key to creating an open-ended developmental cycle (see Figure 10-1). Particular, familiar situations allow the robot to perceive something about objects or object properties that could not be perceived outside those situations. These objects or properties can be tracked into other, less familiar situations, which can then be characterized and used for further discovery. Just as the segmented views provided by poking of objects and actors by poking can be collected and clustered as discussed in Chapter 5, provided precisely what was needed to train up an object detection and recognition system, tracking those objects provides exactly what is needed to learn about other activities, which in turn can be used for further learning.

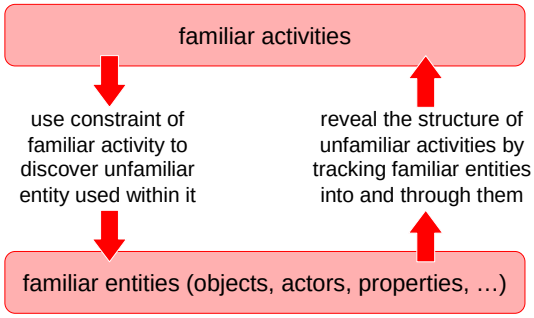

Figure 10-1: If the robot is engaged in a known activity (top), there may be sufficient constraint to identify novel elements within that activity (bottom). Similarly, if known elements take part in some unfamiliar activity, tracking those can help characterize that activity. Potentially, development is an open-ended loop of such discoveries. Familiar activities can be used to learn about components within those activities (for example, the object struck during poking) and then tracked out into novel activities; then when the robot is familiar with those activities it can turn around and use them for learning also.

10.1 Learning through activity

The intersection of communication, perception and development encompasses some well-established fields of research – for example, language acquisition. It has been observed that language acquisition involves a search through a large search space of models guided by relatively sparse feedback and few examples. This so-called "poverty of the stimulus" relative to the complexity of the models being acquired is taken to imply that infants must have a good search strategy, with biases well matched to the nature of appropriate solution. This is a claim of innate constraints, and is historically controversial. Examples stressing under-determination in language learning include Quine's "Gavagai" example (Quine, 1960), where Quine invites us to imagine ourselves walking with a native guide in a foreign country, and seeing a rabbit pass just as the guide says "gavagai" – and then consider all the possible meanings this utterance might have.

Pragmatic constraints offer one way out of this sea of ambiguity. For example, Markman (1989) proposes a set of particular constraints infants might use to map words on to meanings. These constraints are along the style of the following (with many variations, elaborations and caveats) :-

- ▷ Whole-object assumption. If an adult labels something, assume they are referring to the whole object and not a part of it.
- ▷ Taxonomic assumption. Organize meanings by "natural categories" as opposed to thematic relationships. For example when a child is asked to find a "dog", he/she may fetch the cat, but won't fetch dog-food.
- ▷ Mutual exclusivity. Assume objects have only one label. So look for an unnamed object to which a new label can be applied.

These constraints are intended to explain a spurt in vocabulary acquisition where infants begin to acquire words from one or a few examples – so-called fast-mapping. They are advanced not as absolute rules, but as biases on search.

Tomasello raises several objections to the constraint-based approach represented by Markman Tomasello (1997). Tomasello favors a "social-pragmatic" model of language acquisition that places language in the context of other joint referential activity, such as shared attention. He rejects the "word to meaning mapping" formulation of language acquisition. Rather, Tomasello proposes that

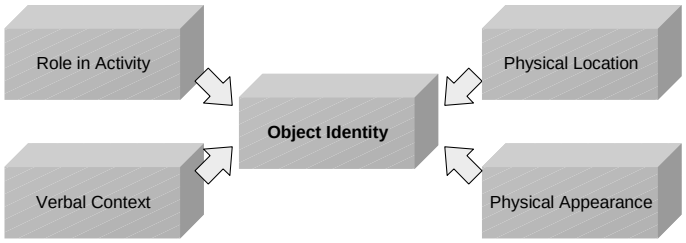

Figure 10-2: Perceptual judgements are fundamentally about identity: what is the same, what is different. Identity judgements should depend (at least) on activity, location, appearance, and verbal context. These in turn can be influenced by a teacher.

language is used to invite others to experience the world in a particular way. From Tomasello (1997) :-

> The social-pragmatic approach to the problem of referential indeterminacy ... begins by rejecting truth conditional semantics in the form of the mapping metaphor (the child maps word onto world), adopting instead an experientialist and conceptualist view of language in which linguistic symbols are used by human beings to invite others to experience situations in particular ways. Thus, attempting to map word to world will not help in situations in which the very same piece of real estate may be called: "the shore" (by a sailor), "the coast" (by a hiker), "the ground" (by a skydiver), and "the beach" (by a sunbather).

Regardless of the utility of Tomasello's theory for its proper domain, language acquisition in infants, it seems a useful mindset for tackling interpersonal perception, which is in essence all about inviting the robot to view the world in a particular way.

Tomasello and his collaborators developed a series of experiments designed to systematically undermine the constraints approach to learning as typified by Markman and others. The experiments investigate word learning among children in the context of various games. The experiments are instructive in showing a range of situations in which simple rules based directly on gaze or affect would fail in at least one case or other. The experiments all avoid giving children (18-24 months old) ostentive naming contexts, and rather requiring them to pull out meanings from the "flow of interaction".

For example, in one experiment, an adult makes eye-contact with a child subject and says "Let's go find the toma," where toma is a nonsense word the child has never heard before. They then go to a row of buckets, each if which contains an object with which the child is not familiar. One of these objects is randomly designated the "toma". If the session is a control, the adult goes directly to the bucket containing the toma, finds it excitedly and hands it to the child. Otherwise, the adult first goes to two other buckets in sequence, each time taking out the object, scowling at it, and replacing it, before "finding" the toma. Later, the child is tested for for the ability to comprehend and produce the new word appropriately. The results show equally good performance in the test and control scenarios. Tomasello argues that this situation counts against children using simple word learning rules such as "the object the adult is looking at while saying the novel word," "the first new object the adult looks at after saying the novel word," "the first new object the infant sees after hearing the novel word," or such variants.

Tomasello's theories and experiments are provocative, and suggest an approach quite different from the simple associative learning that is most often seen in robotics. Work on interpersonal perception on Cog draws heavily on (a grossly simplified caricature of) these ideas. The basic idea

for interpersonal perception drawn from Tomasello's work is that information about the identity of an object needs to be easily transferred between perception of activity, location, speech, and appearance (Figure 10-2). Without this flexibility, it is hard to imagine how scenarios such as the experiment described above or others proposed (Tomasello, 1997) could be dealt with.

10.2 Places, objects, and words

It is currently unreasonable to expect a robot to understand a "flow of interaction" without help. Unaided segmentation of activity is a very challenging problem (see Goldberg and Mataric (1999) for one effort in the robotic domain). The human interacting with the robot can greatly simplify the task by making the structure of the activity unambiguous. Two mechanisms for this are particularly easy to deal with: vocalizations and location. If places and words are used consistently in an activity, then it is straightforward to model the basic "flow of interaction" they define.

The robot was given a verbal, 'chatting' behavior to augment the object-directed poking behavior developed in 3. This used the vocabulary extension mechanism developed in Chapter 9, the egocentric map developed in Chapter 8, and the ability to recognize poked objects developed in Chapter 5. If the robot hears a word while fixating a particular object, and that word has not been heard in other context, then the word is associated with the object. If this happens several (three) times, the association is made permanent for the session. Invocation of an object by name triggers the egocentric map to drive the eyes to the last known location of the object, and the foveal object recognition module to search for the object visually (see Figure 10-3).

This simple naming capability serves as a baseline for the rest of this chapter, which will show how the robot can learn new opportunities for associating names and objects in situations without ostentive showing.

10.3 Learning the structure of a novel activity

Before the robot can learn through an activity, it must be able to learn about that activity. If it cannot do this, then it will remain restricted to the set of built-in activities provided by the programmer. Ideally, it should be possible to demonstrate a task to a robot and have it learn to do it. As already mentioned, unaided segmentation of activity is a very challenging problem in machine perception. It is perhaps more productive to see activity segmentation as something that is explicitly communicated to the robot, rather than something it learns autonomously.

While individual parts of a task may be difficult to describe formally, its abstract structure or control flow will often be amenable to simple description. For example, the overall branch-and-loop flow of a sorting task is easily expressed, but the actual sorting criterion may depend on differentiating two classes of objects based on a small difference in their appearance that would be easier to demonstrate than to describe. If we go ahead and communicate the task structure to the robot, it can be used to guide interpretation of the less easily expressed components. Figure 10-4 shows a schematic for how this may be achieved. The basic idea is for the robot to interact with the instructor vocally to acquire a "sequencing model" of that task, and then to ground that model based on a demonstration of the task. The demonstration is annotated by the instructor both in terms of the sequencing model and in terms of previously grounded elements.

As the human tutor demonstrates a task, they are expected to verbalize their activity. Initially the robot cannot make much of the physical demonstration, but it can process the speech stream, and attempt to recover the structure of the task from that. In particular, the robot will attempt to determine "states" of the task – points at which the demonstration returns to what is effectively the

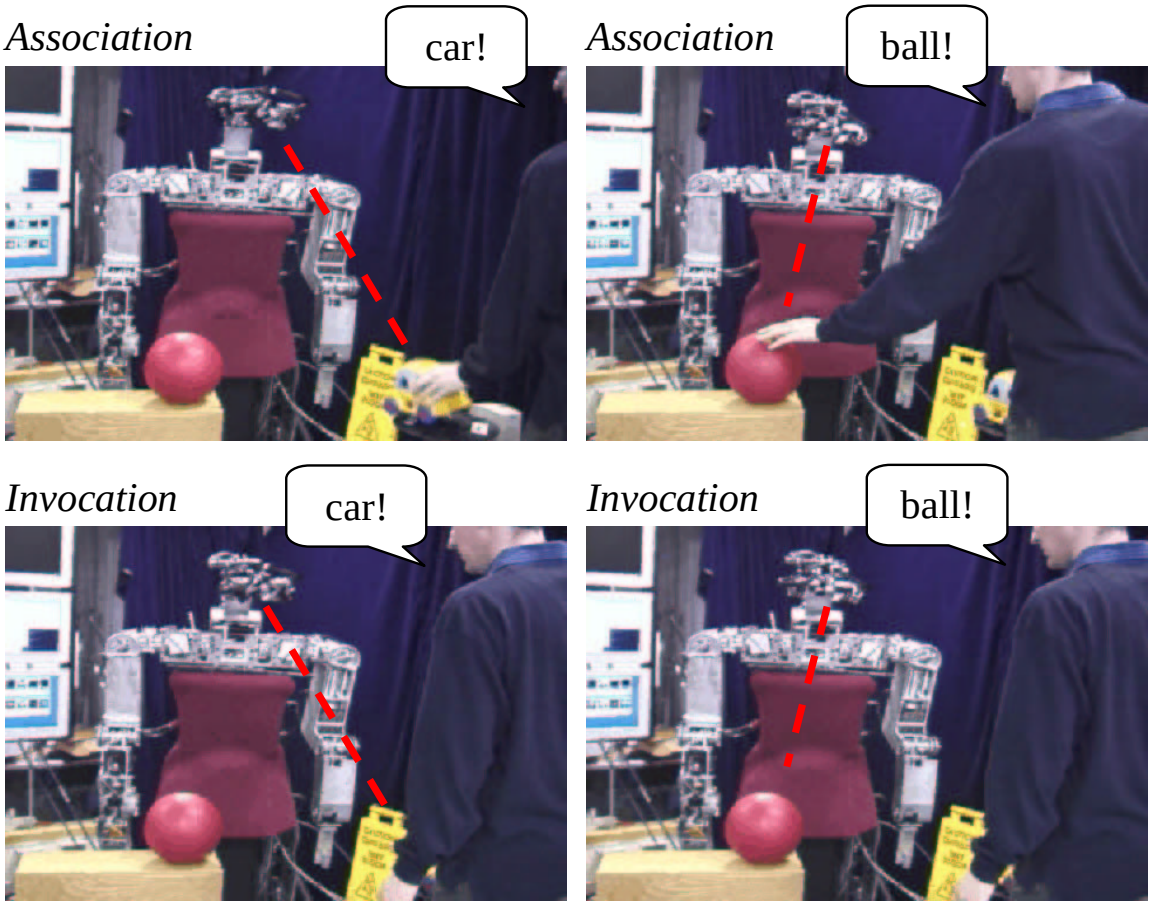

Figure 10-3: Association and invocation via the egocentric map. When the robot looks at an object and recognizes it, its head rolls into an inquisitive look. If a word is spoken at this point (e.g. "car!" or "ball!" in top two frames – note that the human is bringing the robot's attention to an object with his hand) then that word is associated with the object the robot is viewing. If that word is spoken again later (as in the lower frames – note that the human is standing back, only interacting through speech), then the robot queries the egocentric map for the last known location of the associated object, turns there, and looks for the object.

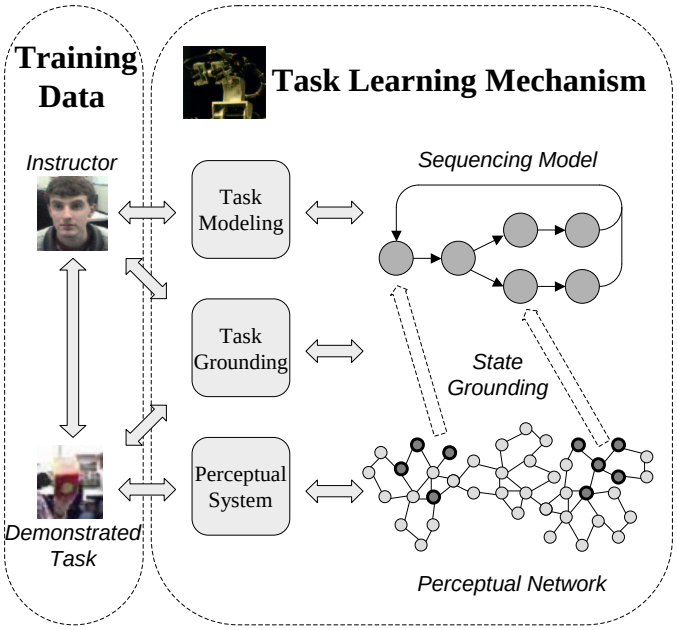

Figure 10-4: A model of task segmentation. The instructor demonstrates the task while providing verbal annotation. The vocal stream is used to construct a model of the task. Generic machine learning methods are then used to ground this model in the robot's perceptual network, guided by feature selection input from the human. The idea is to avoid ever presenting the robot with a hard learning problem; the learning algorithms are intended to be "decoders" allowing the human to communicate changes in representation, rather than to learn in the conventional sense.

same mode. This is straightforward to do if the human tutor speaks simple vocal labels corresponding to actions, configurations, objects, or whatever the tutor finds mnemonic. The type of labels used does not need to be pre-specified, and could vary from word to word.

The method used to recover the task structure is based on n-gram modeling procedures developed for speech recognition – although there are many other methods (Murphy, 1996), this one was chosen for its simple, easily predicted behavior. Here, we estimate the probability of event sequences from models trained on sequence frequency counts from a corpus. Models vary in the amount of history they incorporate – bigram models, trigram models etc. Low order models are limited in the dependencies they can capture, but can be trained up with relatively little data. High order models are more expressive but harder to train. Best results are achieved when n-gram models of many different orders are used, and interpolated based on the amount of training data available for each context (see Figure 10-5 for a simulated example). Once the robot has a model for the task structure, the goal is to relate that to the actual physical demonstration the human tutor is making. Machine perception is necessarily noisy and full of ambiguity. The degree to which this is so for a given task will fundamentally limit the complexity of any modeling the robot can do, if we permit uncertainty to compound on uncertainty. By first establishing a task model through a relatively noise-free protocol for which we can depend on error-correcting feedback from the human tutor, we limit the impact that uncertainty in grounding one element of the model will have on all the others.

Figure 10-6 shows an example of this for a sorting activity, implemented on the robot Kismet. Note that words are used here without the robot needing to know their meanings – it is sufficient that they be used consistently enough for the structure of the task to be made obvious.

1 2 1 4 1 2 1 3 1 2 . . .

. . . 1 3 1 2 1 4 1 2 1 4 . . .

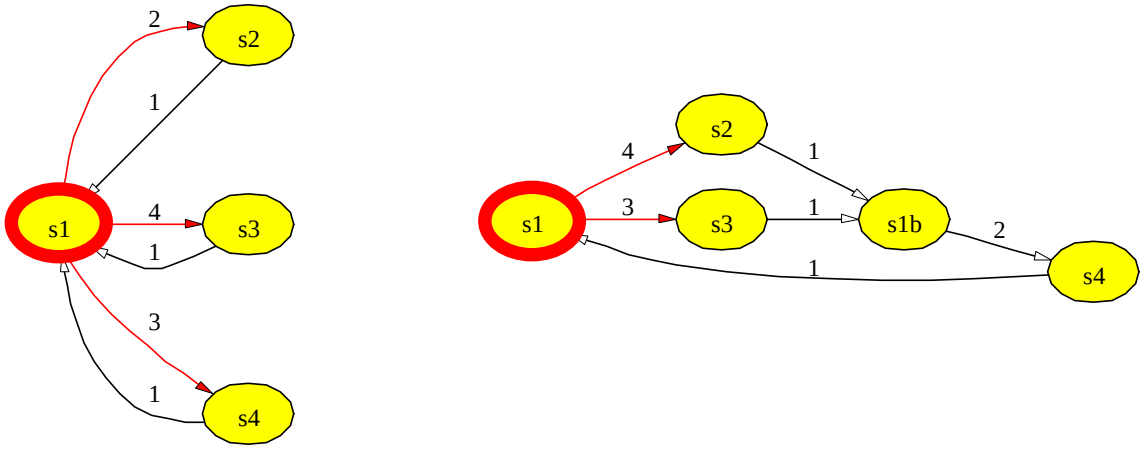

Figure 10-5: The top row of this figure shows a simple artificial sequence presented to the task segmentation system. For the purposes of visualization, the current estimate of the task structure is converted to a state machine and rendered automatically using AT&T Graphviz (Gansner and North, 2000). Initially, the sequence is interpreted as alternating between 1 and any of 2, 3, or 4 (left). Once more data is available, the model expands to incorporate the fact that there is a second-order alternation between 2 and either 3 or 4 (right).

Human speech	Human action	Robot speech	Robot action
. . .	. . .	. . .	. . .
say yellow	[shows yellow toy]	yen	[looks at toy]
say yellow		yelo	
yes! say green		green	
. . .	. . .	. . .	. . .
yellow	[shows yellow toy]	yelo	[looks at toy]
left	[moves toy left]	left	[tracks toy]
green	[shows green toy]	green	[looks at toy]
right	[moves toy right]	right	[tracks toy]
yellow	[shows yellow toy]	yelo	[looks at toy]
left	[moves toy left]	left	[tracks toy]
. . .	. . .	. . .	. . .
yellow	[shows yellow toy]	yelo	[looks at toy]
next?		ahm left?	[nods left]
yes!	[moves toy left]	left!	
green	[shows green toy]	green	[looks at toy]
next?		ahm right?	[nods right]
yes!	[moves toy right]	right!	
. . .		. . .	

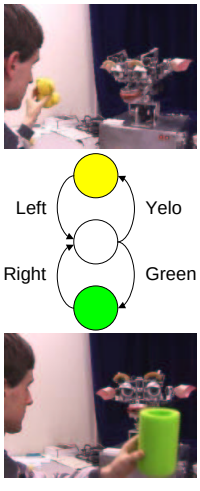

Figure 10-6: Extracts from a dialogue with Kismet. The first extract (say yellow...) illustrates how the robot's active vocabulary was extended. On Cog, this is replaced with the automatic mechanism described in Chapter 9. The second extract shows how a simple sorting activity was annotated for the robot. The final extract shows the robot being tested on its understanding of the form of the activity. The robot's utterances were transcribed phonetically, but are written in a simple form here for clarity. To the right is shown the simple state machine model of the activity deduced by the robot (graph drawn by hand).

10.4 Learning the rules of a novel activity

The structure learning mechanism in the previous section is useful, but if a person interacting with the robot is cooperative in their choice of vocabulary, it is in fact overkill. It also does not deal well with nested activities, where one activity is suspended temporarily to deal with another – for example, if during the sorting behavior the person wants to check if the robot knows the name of an object. When transferring the mechanism from Kismet to Cog, the emphasis was shifted from learning global task structure to learning local rules (which could grow to support long-distance interactions). On Cog, seeing objects or hearing words are treated as the basic events in the system. The robot continually searches for useful new ways to describe events, where being 'useful' means having predictive power. The events it considers are :-

- **Conjunctions:** if two events are noted to occur frequently together, and rarely occur without each other, an event called their conjunction is formed. This event is defined to occur whenever the two events do in fact occur together. The formation of an event simply means that statistics related to it are tracked. Once an event is formed, it doesn't matter if the conditions for its creation cease to hold.
- **Disjunctions:** if two events are noted to occur frequently together, but also occur independently in other situations, an event called their disjunction is formed. This event is defined to occur whenever one or both of the two original events occur.
- **Implications:** Causal versions of the above events also exist, which are sensitive to event order and timing.

These composite events are intended to allow the robot to make meaningful generalizations, by allowing the same physical event to be viewed in ways that are sensitive to past history. Figure 10-7 demonstrates the use of such generalizations to solve one of Tomasello's experiments – linking an object with its name through an extended search activity. Searches are presented to the robot as following a fairly strict script: first the word 'find' is uttered, then the name of the object to search for is mentioned. Then a series of objects are fixated. The word 'no' is uttered if the object is not the target of the search. The word 'yes' indicates that the search has succeeded, and the object currently fixated is the target of the search. These facts can be discovered using event generalization. The word spoken after 'find' gets a special implication event associated with it, let us call it `word-after-find` (of course, no such symbols are used internally, and the word 'find' initially has no special significance – it could be replaced with any other word, such as 'seek,' 'cherchez,' or 'fizzle-tizzle'). If the search is for an object whose name the robot knows (through a pre-established disjunction) that is also noted as a simultaneous event with `word-after-find`. The object seen when 'yes' (`object-with-yes`) is said matches this and an implication is formed between the two. This implication is sufficient to link an *unknown* word following 'find' with the object seen when 'yes' is said, via the `word-after-find` and `object-with-yes` generalizations (again, the choice of the word 'yes' has no special significance, and could be replaced with 'frob').

When the rule-learning mechanism above was integrated with the full Cog code-base, then the search activity became much simpler to learn, requiring less generalization. This is because for the 'chatting' behavior, the robot has internal state to track when it hears a name of an object it is not looking at (so that it can then direct its gaze towards that object). With this structure built in, the robot simply has to map the search activity on to it, which it can do with just two observations:

- 'Find' is followed by mention of an absent object
- 'Yes' is said when a previously absent object is in view

The generalizations already built in to the robot's state are sufficient to achieve the first Tomasello scenario.

Human speech	Human action	Robot speech	Robot action
…	…	…	…
say	[shows ball]	say	[looks at ball]
beh		ball	
say	[shows car]	say	[looks at car]
keh		car	
say	[shows cube]	say	[looks at cube]
keh		cube	
say		say	
	[waits]	cube	
	[shows ball]		[looks at ball]
say		say	
	[waits]	ball	
…	…	…	…
	[attracts attention]		[looks at person]
find		find	
ball		ball	
no	[shows cube]	no	[looks at cube]
no	[shows car]	no	[looks at car]
yes	[shows ball]	yes	[looks at ball]
…	…	…	…
	[attracts attention]		[looks at person]
find		find	
toma		toma	
no	[shows ball]	no	[looks at ball]
no	[shows cube]	no	[looks at cube]
yes	[shows bottle]	yes	[looks at bottle]
say	[shows cube]	say	[looks at cube]
		cube	
say	[shows bottle]	say	[looks at bottle]
		toma	
…	…	…	…

Figure 10-7: Extracts from a dialogue with Cog. First, the robot is taught to name the object it is looking at when the word ‘say’ is spoken. This is done by speaking the word, then prompting the robot with a short utterance (beh and keh in this example). Short utterances prompt the robot to take responsibility for saying what it sees. A link is formed between ‘say’ and prompting so that ‘say’ becomes an alternate way to prompt the robot. Then the robot is shown instances of searching for an object whose name it knows (in the one example given here, the ball is the target). Finally, the robot is shown an instance of searching where an unfamiliar object name is mentioned (‘toma’). This allows it to demonstrate that it has learned the structure of the search task, by correctly linking the unfamiliar name (‘toma’) with the target of search (a bottle). Ideally, to match Tomasello’s experiment, all the objects in this search should be unfamiliar, but this was not done. In the infant case, this would leave open the possibility that the infant associated the unfamiliar word with the first unfamiliar object it saw. In the robot case, we have access to the internal operations, and know that this is not the cue being used.

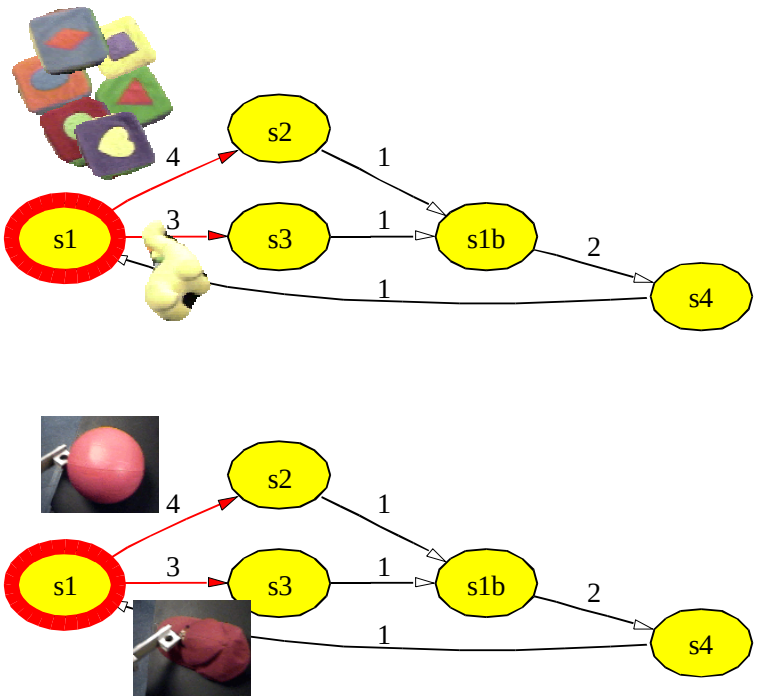

Figure 10-8: Task structure should allow the robot's perceptual biases to be overruled. for example, objects are differentiated by the robot purely based on color histogram. This could cause problems for an object which looks very different from different sides (such as a toy cube, top). These views could be united within a task where they are all treated the same way (for example, by performing one action on the cube and another on another toy). If two distinct objects are treated as the same by the robot because of color similarity (such as a ball and baseball cap, bottom), then their difference can be exaggerated by using them differently within a task.

10.5 Limitations and extensions

There are many limitations to the activity learning described in this chapter, including :-

- The cues the robot is sensitive to are very impoverished, relative to what a human infant can perceive. For example, there is no direct representation of the teacher, and no perception of prosody or non-verbal cues.
- If multiple activities share similar vocabularies, their is the potential for interference between them. The issue of capturing the overall activity context has not been addressed.
- The basic events used are word and object occurrences, which do not begin to capture the kind of real world events that are possible. So the robot could not respond to non-speech sounds, or changes in distance, or any of the infinite possible events that are not simply word/object appearances.

To begin to deal with this last point, a simple mechanism was developed to get the robot's attention to an unnamed feature or feature combination (as opposed to simply an object) using periodicity detection. All perceptual features on Cog are monitored over a sixty second time window to detect the occurrence of periodicity. Hence if it is desired that the robot attend to the color of objects as opposed to their identity or size, for example, then objects of contrasting colors can simply be

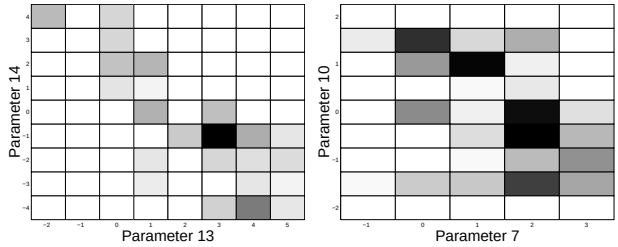

Figure 10-9: Searching for correlations in the robot's perceptual features. The joint probability distribution of pairs of features is estimated, and compared with the product of their independent distributions. This figure shows the top two correlations in an experiment where the robot applied one of four actions to an object (tapping it from the side, slapping it away, etc.). The highest ranking correlation, left, captures a physical constraint on the angle of approach at the moment of impact. The next correlation (right) captures the gross displacement of the object away from the robot as a function of the type of gesture (the correlation is noisy because of erratic motor control). With verbal annotation of the actions, this correlation could be enhanced (by tagging failed actions for removal) and selected out for use.

shown to the robot. The periodic signal oscillation increased the salience of a channel in a manner similar to the behavioral influences used on Kismet (Breazeal and Scassellati, 1999). But at the time of writing, this was not strongly integrated with the activity learning mechanisms.

The idea of influencing the robot's perception through shared activity also could and should be developed further. Perception is not a completely objective process; there are choices to be made. For example, whether two objects are judged to be the same depends on which of their many features are considered essential and which are considered incidental. For a robot to be useful, it should draw the same distinctions a human would for a given task. To achieve this, there must be mechanisms that allow the robot's perceptual judgements to be channeled and molded by a teacher. This would also useful in situations where the robot's own abilities are simply not up to the challenge, and need a helping hand. Once the structure of tasks can be communicated to the robot, it should be possible to use that high-level structure to modify the robot's perception. It is easiest to see this in the case of modifying biases in pre-existing abilities of the robot. For example, we could emphasize the difference between objects the robot sees as identical, or draw connections between different views of the same object that the robot sees as distinct (see Figure 10-8).

More generally, we can use the task structure to initialize a set of focused problems in machine learning, where divergent paths in the task are treated as labels for the (initially unknown) physical features that cause that divergence. By correlating features across the robot's perceptual space with these labels, we can select those that might contribute to the decision, and then train up a classifier. Figure 10-9 shows an example of searching for such correlations in the robot's perception of its own poking behavior.

10.6 Summary

This chapter has shown that a virtuous circle of development is possible (see Figure 10-1. If we want robots to be able to cope with novel tasks, they will need a deep understanding of the activities around them. We can treat the range of naming situations a robot can deal with as a test of the depth of that understanding. Consider search: if the robot understands the purpose of searches, how they succeed and fail, then that naturally extends the range of naming situations it can deal with beyond

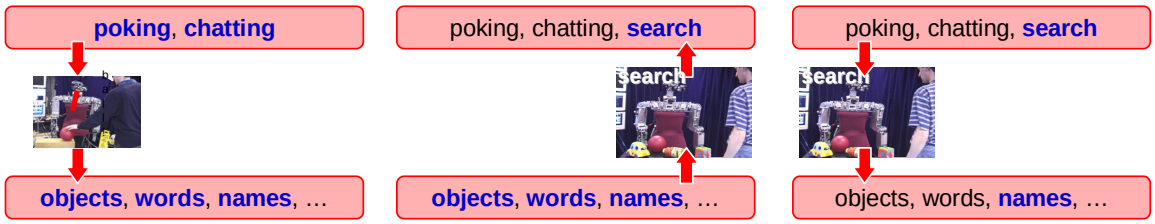

Figure 10-10: The chapter developed a specific example of the virtuous circle of development. First, poking allows the robot to explore objects, and then chatting allows names to be associated with those objects. Then the robot tracks those named objects as a human demonstrates a search task, learning about the structure of search from these examples. Finally, the robot uses this knowledge as a new way to learn names for objects without having to see an object and hear its name simultaneously, as is the case for chatting.

simple ostentive associations, as this chapter showed. In the infant development literature, considerable emphasis is placed on the child's ability to interpret the behavior of others in terms of intent using a "theory of mind". Such an ability is very powerful, but so also is the more computational viewpoint of processes as branches, loops and sequences. This is an alternative route to establish an initial shared perspective between human and robot, and could potentially complement a theory of mind approach (for example, the work of Scassellati (2001) on Cog).

In the activity-learning system described here, the meanings of words grow out of their role within an activity. In the search activity, 'yes' and 'no' come to denote the presence or absence (respectively) of the search target. In another scenario, they may denote something entirely different. For example, it would be possible to train the robot to keep holding out its arm while 'yes' is said, and to drop it upon hearing 'no' – in which case the words now denote action continuance or termination respectively. This plasticity means that we can avoid the problem of trying to form a global theory of all the meanings a word can take on.

Richer meanings are possible when multiple-word utterances are permitted (Roy et al., 2002, 2003), rather than the isolated words dealt with in this chapter. An interesting direction for future research would be to derive grammatical forms as a compression of the structure of an extended activity into a single sentence.