Chapter 5
Sketches Toward a Theory of Complex Action

5.1. Introduction

It is often neglected that the reason for having a brain is in order to be able to act. Cognition is not an end in itself. Knowing things only does the organism good if this knowledge can be put to use in formulating courses of action. And there is more to action than motor control. Here I want to examine two simple, even banal actions that we take absolutely for granted: shaking hands and making coffee. I will use these examples as a vehicle to explore some of the complexity in the systems of knowledge and processing that underlie action.

One goal of this exploration is to help set the stage for studying social cognition in the sense of the previous chapter: the system of largely unconscious knowledge that we bring to bear on our interactions with other people. As stressed there, an essential part of this knowledge concerns social norms and moral judgment. The literature often speaks of norms and morality in terms of the ability to make judgments of others' or perhaps one's own actions. But this is not enough: normative judgments of one's own past and potential future actions must play a role in choosing what to do; and normative judgments of others' actions must play a role in choosing how to interact with them. However, when I began to think about these problems, I soon discovered that there is no hope of getting all the way to morality very soon. Even the simplest actions, with little social significance, reveal a daunting complexity that has to be understood first. For the present, the main connection is that morality and social convention are rules about actions, so we would do well to understand something about what actions are.

I have another motive in undertaking this exploratory work. One of the important questions in the linguistic theory of the last 40 years has been how much of the human language faculty is a cognitive specialization and how much has been "borrowed" from other faculties, either in the contemporary brain or in the course of evolutionary development from primate ancestors. I have often argued (e.g. chapter 1) that this question cannot be answered without comparable theories of other capacities. Just looking at brains or at behavior is not enough: to compare another capacity to language, we need something comparable to the detailed account of linguistic structure that has been built up over the years. I believe that the only domain for which we have such an account is music, in the work I've done with Fred Lerdahl (Lerdahl and Jackendoff 1983). But music, like language, is a very special human capacity. Action is another promising candidate: it too is a temporally structured domain, one that may well be shared to some degree with other species. So it would be of interest to compare it to language, if only we can develop a theory of its structure. And of course the outcome is informative no matter how much of it turns out to overlap with language.

I have come across nothing in the psychological literature that speaks to the structure of complex actions from the point of view of the performer of the action. Zacks and Tversky 2001 is a good survey of research on event perception, for which the state of the art is fairly
rudimentary compared to object perception. Nevertheless, there are important connections between event perception and the formulation of action that will be addressed as we go along. More relevant is research in AI that has addressed some of these issues over the years, in literature on planning and on robotics. Many of these proposals find close parallels in the realm of the present more psychological approach.  

My methodology here will be to muster a lot of common-sense observations and to attempt to give them some theoretical structure, showing where the inquiry intersects with problems that have been addressed in the literature. Clearly the story is very preliminary. To flesh it out, one would want to work out many more such examples, and, more importantly, work out research methodologies for examining the hypotheses experimentally.

I see this enterprise as involving a number of leading questions:

- What is the repertoire of structures that can be composed to form complex actions? This question is parallel to the question in linguistic theory of the grammar used to compose sentences.
- In any particular action, what are the stored parts out of which it is built? This corresponds to the issue of the lexicon in linguistic theory (the AI literature has used the terms library (Fikes and Nilsson 1971, Sacerdoli 1977) and Actionary (Badler et. al. 1999)).

These two questions concern the form of the cognitive structures underlying complex actions. I find it plausible that such cognitive structures are involved in the answers to the following four questions as well.

- How are complex actions constructed on line in preparation for execution? This is the problem of planning.
- How are complex actions executed on line? This plus the previous question correspond to the issue of language production.
- How are complex actions by others recognized? This corresponds to the issue of language perception and recognition.
- How are complex actions specified by linguistic descriptions? For instance, how does one formulate instructions to tell someone else how to perform a complex action?

Finally, as in linguistics, one can ask the following learning-theoretic questions:

- How are new stored complex actions acquired? How are the means of composing complex actions online acquired?
- Is there an innate basis that provides the overall form in terms of which complex actions are built and learned?

¹For some reason, I don’t find myself using the term goal with the same freedom as these other literatures. I’m not sure why.

5-2
I won’t deal with these questions systematically, but they will be constantly addressed in different ways throughout the discussion. Another question, which I will not address, is how all this is neurally instantiated, not just the motor program but all the cognitive apparatus that leads up to it.

5.2. Shaking hands

The simple act of shaking hands reveals several silent features of actions. First, it has structure in two separate cognitive planes, the physical and the social. Second, the structure in each of the planes and the interaction between them exhibits interesting complexity. Third, shaking hands is a cooperative action, requiring coordination among actors.

Shaking hands is of course a social convention. Little kids don’t seem to “get it”; it’s something they have to learn to do. But just to say “shaking hands is a social convention” doesn’t explain anything. At the very bottom, we must store something in memory that might be characterized as our “knowledge of shaking hands”. This has to have at least two parts: how to do it and when to do it. Both of these have to be learned, of course. In addition, going back to the issues of chapter 4, there is the more basic question of what kind of knowledge a social convention is.

One’s stored knowledge of shaking hands surely does not provide a full set of instructions for all possible situations, which may vary considerably in both physical and social dimensions. So the knowledge must be stored in some schematic form, the “shaking-hands action type”, which is modulated or adapted or adjusted to suit particular token circumstances. In this respect, action categorization is much like perceptual categorization: not all dogs or tables or bicycles look exactly alike, and one’s stored encoding of the category must be modulated or adapted or adjusted to fit particular tokens one may encounter. No surprises here.

On the other hand, since an actor ultimately has to use the stored schema to produce a token action, the theory of action has to include principles of composition that allow the actor to modulate, adapt, or adjust the schema to suit the particular circumstances and to realize it as a complex of motor instructions. Speech production is a familiar case: the motor production of a stored word is modulated by the intonation contour, the overall speech rate, the emotional tone, and possibly the fact that the speaker is chewing gum or holding a cigarette in his mouth.2

2The necessity of adapting a schema to current action obviously has a counterpart in robotics. One approach (Badler et al. 1999, Binfield et al. 1999, Kipper and Palmer 1999) encodes schemas in memory as Parameterized Action Representations. These are filled in to suit current circumstances with parameters such as objects to be manipulated, manner and duration of action, trajectory of motion, and necessary preparatory actions. The approach is not too distant from what is developed here.

5-3
5.2.1. The social plane of shaking hands. As discussed in the last chapter, humans conceptualize people, actions, and artifacts in terms of two independent but interacting planes, the physical and the social. The physical plane involves concepts of physical objects moving in space and exerting forces on each other. The social plane involves concepts of persons: individuals with whom we can have social relations. (This is a narrower class than animates, which also include mosquitoes and worms; however, pets may be counted as "honorific persons.") People and their social actions are typically encoded on both planes simultaneously; the division corresponds rather nicely to the folk-psychological division between body and soul.

One’s knowledge of shaking hands of course involves the physical plane. One has to know how to actually perform the action: extend one’s hand appropriately, grasp the other’s hand, shake, and let go. I’ll discuss this in a moment. First let’s think about the social significance of shaking hands: it is an assertion or confirmation of social connection or solidarity. There are at least five circumstances where one may shake hands: greeting, taking leave, congratulating, closing a deal, and introducing oneself? All but one of these uses are symmetrical, in that either participant can initiate the action. But congratulation requires that the person doing the congratulating initiate the action, not the person being congratulated.

In order to know when it is appropriate to assert solidarity or social connection, the actors must be able to identify the larger social frame in which an action of greeting, taking leave, etc. can be identified. You don’t understand shaking hands if you are just constantly shaking hands inappropriately, or don’t know to do it when it’s called for. More generally, we might ask why we greet and congratulate each other, anyway. And why is a physical gesture of social connection so necessary as part of such actions? And what is the nature of the “social connection” that shaking hands symbolizes? I leave these larger questions aside as parts of the general theory of social cognition.

There are also choices of physical gesture to be produced, which depend on at least (a) culture, (b) gender, (c) formality of the occasion, (d) degree of intimacy between the individuals. In some cultures, mutual bowing takes the place of shaking hands; in others, high-fiving. In some cultures on some occasions, the appropriate gesture between a man and a woman is the man kissing the woman’s hand. In our culture, a certain degree of intimacy permits (or even demands) the substitution of a hug and/or a kiss, the choice depends in part on the gender and sexual orientation of the participants (heterosexual men probably require the greatest intimacy in order to permit this substitution). The European version has kisses on both cheeks (or three iterations in Holland). Most of this is so familiar as to be nearly transparent. Nevertheless, choosing the right physical action to “express the social meaning” requires complicated knowledge of the status of the participants and their cultural expectations.

Notice that being introduced sometimes coincides with greeting, but not always. You’re casually conversing with a stranger on an airplane, and at some point you somehow mutually decide it’s appropriate to tell each other your names. Suddenly you can’t resist shaking hands.
Modulations of the physical action can have social meaning too. Most blatantly, declining an offered handshake is a blatant snub. So is offering a more formal action when a more intimate action is appropriate. Conversely, offering too intimate an alternative can be seen as social aggression. The intensity of the handshake and the character of the accompanying eye contact, whether consciously produced or not, also carry social meaning such as assertions of dominance or of disengagement.

How much of this is stored as part of the “knowledge of shaking hands”? Some pairs might come from elsewhere, simplifying the “shaking hands” schema in memory. First, one must independently judge degree of intimacy for other purposes in social interaction, for example in the choice of formal vs. informal second person pronouns (you vs. tu in French) or honorifics in Japanese, as well as overall choice of linguistic register (formal vs. casual). Second, since hugging is primarily a sign of affection, the choice between shaking hands and hugging depends in pairs on whether overt displays of affection are independently judged appropriate. For instance, hugging is less appropriate in the context of a formal ceremony than when welcoming a visitor into one’s home. Third, the modulation of the intensity of the handshake and accompanying eye contact probably falls out of more general principles for signaling approach, avoidance, and like and dislike. But if we leave these matters out of the handshake schema in memory, then we put the burden on the combinatorial system, which must integrate judgments of intimacy, formality, gender, and attitude into a script for performance.

On the other hand, the need for some of these judgments can be shortcut by storing in memory how one interacts with particular individuals (my brother doesn’t like to hug, my sister-in-law greets me with a kiss on the mouth etc.). I take such particular memories to be analogous to one’s memory for particular objects: one not only has a schema for tables in general but also one for the table in one’s kitchen (an issue to which I return in a while).6

5.2.2. The physical plane of shaking hands. For a first approximation, the physical action can easily be described in terms of a sequence of five subactions:

1. reach hand to other person > grasp other hand > shake > release grasp > withdraw hand

My main task here is to show that there is more structure to shaking hands than this sequence. But before going on, let’s talk about the theoretical status of (1). For the moment assume that it is stored in long-term memory as the schema for the physical aspect of shaking hands. Suppose that at some point it is called up into working memory as a possible action; the instantiation, with

6 The relation between stored schemata and stored special cases of the schema appears in the linguistic literature under the rubric of “inheritance hierarchies” (Pollard and Sag 1987). Similar notions of inheritance are invoked in the AI literature on actions, for instance Kautz 1990, Pollack 1990, Kipper and Palmer 19xx.

5-5
its modulations specified in working memory, has the status of a plan. Suppose that the actor commits to the plan; he now gains the status of an intention. Finally, to carry out the schema, it has to be instantiated in actual motor instructions which are read off left to right; it thereby becomes a voluntary action. In other words, our intuitive ontology of plan, intentions, and voluntary actions corresponds nicely to the various roles that an action structure can play in cognition and execution. (We return to planning and intending in chapter 6.)

Now let us consider the sequence in (1) in more detail. It makes sense to structure (1) into constituents, where each constituent has a subaction or a subconstituent as "head." The head of the entire sequence in (1) is the actual shaking of hands: that's the point of the sequence, the subaction that the whole thing is for. Reaching and grasping then constitute a preparation for shaking; and releasing and withdrawing constitute the "finishing-off" or "coda." This organisation is reflected in an interesting asymmetry in the way we describe the subactions. One reaches and grasps in order to shake; but one certainly does not shake in order to release and withdraw. Rather, one shakes in order to affirm social connection; and one releases and withdraws in order to return to the original state, to "put things back in order."

Is there any organization within the preparation and the coda? I think so. It seems reasonable to say that one reaches in order to grasp, that is, reaching is a preparation for grasping. And it seems reasonable to say that one releases grasp in order to withdraw. The overall structure, then, can be mapped as a tree structure like (2).\(^6\)

````
(2) shaking hands
    preparation
      reach
      grasp
    head
      shake
      ungrasp
    coda
      prep
      head
````

The shake constituent also has more structure. It is a sequence of up and down motions. We'd want to check this empirically, but my sense is that they're oscillations above and below a neutral position, starting by moving upward to high position, and ending by moving upward from low position to the neutral height. In any event, this sequence has no fixed length, and it contains no single action that serves as head. So we might notate this in a tree structure as (3), using H, L,

\(^{1}\)A linguistic fine point pointed out by Bratman 1990: I mean here a "plan for shaking hands", i.e. an action under consideration. I plan to shake hands expresses a commitment or intention, the next step toward execution.

\(^{6}\)Trees along these lines appear in the AI literature, for instance Lisan and Allen 1990.

5-6
and N for high, low, and neutral position respectively, and using the star as a sign of indefinite repetition (following the custom of the "Kleeene star" in formal languages). I'll treat the return to neutral position as a coda of the shake constituent.

![Diagram](image)

Much of this structure requires coordination with the other participant in the course of execution. Your reach is modulated by perceived height and distance of the other participant. You don't begin shaking hands by sticking your hand in the other person's face: you are aiming for a point roughly midway between you - higher if the person is taller, lower if the person is shorter (or seated). The strength of your grasp is calibrated to match the other person's (well, most people do this, anyway); the tempo and amplitude of the shake must be coordinated; the ungrasp must be pretty close to simultaneous. The modulation of the reach is based on visual perception; from the grasp on, everything is calibrated tactilly and proprioceptively.

The coordinated timing of these actions between individuals cannot be the result of a miracle. Someone must initiate the action and the other must pick it up, very quickly (250 ms. or less?), and often not consciously. I think that the only time I'm conscious of the case is when they misfire, for example if I loosen my grip and the other person keeps shaking. The social context likely primes the action, so one's reaction to someone else initiating a handshake is probably faster in an appropriate context than in an inappropriate one. And there is probably a conventional approximate default length for the shaking stage, which primes the appropriate point to let go.

These actions may be concurrent with others. Suppose you happen to see an acquaintance walking toward you on the street. The scenario might go like this: You might extend your hand in advance, while still walking, before you've actually converged to within shaking distance. As you shake hands, you are also talking. Without releasing your grasp, you may turn around each other, so as to continue in the direction you've been going, and then turn away from each other concurrently with ungrasping and withdrawing your hands. And the withdrawal may merge fluidly into waving to each other. So the total movement script in this case superimposes the handshaking schema on a number of other motion patterns. Like talking while chewing gum, this requires a theory of action composition.

What aspects of the physical pattern are stored in memory as part of the handshaking schema? It's plausible that it's just the "shake" constituent (3). This presupposes that the hands are grasping each other at an intermediate position between the two participants. In order to execute the "shake" constituent, each participant would have to get to the proper position, which
would involve constructing an action plan for the preparation constituent in (2). And in order not to be stuck together holding hands, each participant would have to construct an action plan for the coda constituent in (2). In other words, the “shake” constituent might be provided with just enough structure on its own for the rest to be constructed online. Again, for this to be feasible, it is necessary to provide the theory of action with the potential for action composition, i.e. putting together a complex action from stored parts.

Let’s be a little clearer about what constitutes “enough structure”. The “shake” constituent cannot now just say “move your hand up and down some number of times”. Rather, it has to say something like

Grasping the other participant’s right hand with your right hand in a mutually comfortable midline position.

Move your right hand up and down some number of times.

All the information in the first clause was previously encoded in the endpoint of the “reach” and “grasp” constituents; it is now encoded as the “neutral” position in the “shake” constituent (which required it anyway).

How might a tree work all this information in? We want to say there is a basic position that is modulated by up and down movements. (4) is an attempt. The Head is the neutral position; this is modulated by the up and down motion, which is notated as a constituent called Modulation. The Modulation is connected to the larger action by a dotted line, which is meant to indicate that it is concurrent with the Head rather than successive to it. (If we had three directions at our disposal, we could notate this constituent branching off at right angles to the page.) Such a tree configuration is also useful for characterizing motions like waving (“hold your hand up in the air and wiggle it back and forth”); it also bears some similarity to the sorts of structures posited for hand position and movement in sign language (see various models in Fischer and Siple 1990).

Just for fun, I’ve also notated the possibility of eye contact as an additional concurrent action, connected to the complex again with a dotted line. Eye contact is a concurrent action, i.e. a sort of co-head, rather than a modulation, because it can take place independently of hand grasping. By contrast, the modulation of waving grasped hands up and down cannot take place without hand grasping. The English description of positions in (4) is of course proxy for an encoding of body position in an appropriate spatial/proprioceptive format. I’m more interested here in the way all these pieces have to be connected together.

5-8
To sum up at this point: the hypothesis is that a structure on the order of (4) is stored in memory as the physical plane of shaking hands, and that the preparation and code of (2) are simply constructed online in working memory to provide appropriate transitions into and out of the basic position.

In turn, (4) is linked in both long-term and working memory with the structure of handshaking in the social plane. In the social plane, the structure is "confirming social connection in the context of greeting, etc." The same social structure is also linked in long-term memory to alternative physical realizations such as hugging, high-fiving, kissing the hand, and so on. It's interesting that all of these have structure rather like (4); they can be characterized as adoption of a position for a brief period of time, sometimes with modulation, sometimes not. And each of these positions require strategies of approach and withdrawal, which it might be reasonable to suppose are constructed online. And they require cues for coordination (Which comes first? Does the lady extend her hand, or does the gentleman start lowering his head and pursing his lips while reaching for her hand? It depends.)

5.2.2. Variable instantiation, including by self. I've slipped something into (4) that makes it more abstract than the previous stories. The Position constituents speak of two actors grasping hands and making eye contact, not of "self and other actor". Now in any event we need a perceptual schema that allows us to recognize two other people shaking hands. And maybe (4) is that schema. But notice what it would take to use (4) that way. In any particular situation, the people involved are particular people, not generalized people, and we're not just recognizing two people shaking hands, we're recognizing, say, Roosevelt and deGaulle shaking hands. In order to achieve such recognition, it is necessary to treat actor1 and actor2 as variables, which are instantiated on particular occasions by different individuals. In other words, recognition of others' actions as actions requires online variable instantiation, of the sort familiar from language processing. Here we connect with the event perception literature (Zacks and Tversky 2001). In particular, Cavanaugh et al. 2001 suggest that familiar complex motion patterns such as walking

5-9
are stored as "high-level animation" or "sprites", ready to help identify perceived patterns. Shaking hands might be one of those.

Now, what would happen if by chance one of the variables in (4) were filled by self (or ego)! The result would be the action of "me and other actor shaking hands." Now there is a basic asymmetry about the world: I cannot, simply by intending it, move your body, but I can move my own. That is, there is an eccentric connection between the conceptual formulation of action and the motor system: the motor system can be engaged to formulate a motor program if and only if the actor in a conceptualized action is self. If this is the case, (4) could serve to schematize handshake at the most general level, applying both to the perception and the production of handshakes. The latter could occur only if one of the actors is self.

On the other hand, there is often considerable disparity between one's ability to perceive an action and to perform it. Lots of people can tell a good dancer or basketball player from a bad one, but can't produce the actions with any competence at all; my Uncle Bernie was an avid and sophisticated operagoer, but could barely carry a tune. Presumably, in order to use one of these schemas for performing an action, one must engage linked mental structures specialized for proprioception and motor control (speaking as usual is a good example). Practicing an action refines these proprioceptive and motor structures; watching someone else do it does not. The disparity between perception and production also shows up in the many actions that one can recognize that one has never performed. For example, I don't believe I've ever kissed a lady's hand, and I've certainly never been a lady having her hand kissed. But one needs some mental encoding of such actions, so one can recognize them and judge if they are correctly performed in an appropriate context.

The idea of stored action schemas that serve both perception and production may seem a little outrageous until we remember that words have a similar status. Recall from chapter 3 that these are stored linkages of mental structures, and of course they are used both in language perception and language production. As with actions, one may have perceptual command of a word — be able to interpret it in context — without necessarily having productive control — being able to use it fluently and appropriately. So again the parallel with language proves useful in understanding what we're doing here.

The overall point, though, is that a structure like (4) is on the cusp between action perception and action production — a crucial juncture in mental structure that is not in my experience very well addressed in the literature. Two intriguing issues suggest themselves,

2This turns up in chapter 6 as a formal condition on "actional attitudes" such as intending: the Actor of the action must be bound to the intender.

3A schema like (4) resembles in spirit the Parameterized Action Representations of Badier et al. 1999, where the goal is to animate a virtual robot on the basis of natural language instructions. However, if the present interpretation of (4) is correct, it goes beyond
although we hardly know enough to explore them yet. First, one might wonder whether structures like (4) are what mirror neurons are - neurons that respond both to seeing someone else perform an action and to performing the action oneself. Second, action schemas like (4) would be a suitable mental vehicle for learning by imitation: one might code the action by perceiving someone else doing it, replace the particular actor by a variable, then instantiate the variable as self and let the motor system cut in and execute the action. The going wisdom these days is that other primates are considerably less proficient at imitation than humans (Donald 1998, Hauser 2000). It would be interesting to ask where in the chain of mental structures the difference lies (evidently not in the known mirror neurons, since they were first found in monkeys).  

5.2.4. Joint action. A final point about shaking hands is that it is conceptualized as a joint action. It is not just me taking your hand and shaking it, nor is it me taking your hand plus you taking my hand. It is us doing this together. This crucial difference between simultaneous action and truly joint action has been stressed by Sheeke 1995 and Clark 1996, among others (see chapter 5). Joint action need not be social in intent, for instance lifting a heavy object together. But it does nevertheless create a sense of social connection, through "you and me acting as one".

Cark points out that a joint action usually requires an "offer" by one participant and an "uptake" by the other. The offer is what has been termed here initiation; the uptake is simply following suit in taking part. It should be clear from what we've done here that practically every phase of shaking hands requires initiation and response.

I'll leave shaking hands here. Let's do something else.

5.3. Making coffee

Making coffee lacks a social component but compensates with a variety of other complexities. This case is representative of a whole repertoire of "practical knowledge" that involves using objects. Think of the whole sequence of routines you go through on getting up in the morning - taking a shower, getting dressed, personal grooming, feeding the cats, clearing up the kitchen, reading the paper, preparing and eating breakfast, and so on. There is a tremendous amount of organization, complexity, and flexibility here. I want to give detailed attention to just one of these tasks to illustrate what the human mind must be capable of in formulating and carrying out these sorts of actions.

Parameterized Action Representations in that it can be deployed for action recognition as well as for action production.

*Notice, by the way, that neither of these suggestions requires Theory of Mind - all they require is observation of external action.
The first thing we might notice is that the action of making coffee varies depending on what kind of coffeemaker you use. One’s knowledge might extend to certain kinds of coffeemakers and not others. I don’t have a clue how to use an espresso machine; the old percolator in my family’s summer home requires a different technique from the automatic filter coffeemaker in my kitchen. So even in this restricted domain, one’s knowledge of action is highly tool-specific. For the analysis here, I’ll take the coffeemaker in my kitchen as the operative example.

5.3.1. Basic structure. Again beginning at the grossest level, the steps can be described as in (5).

\[
(5) \quad \text{put in coffee} \quad \text{put in water} \quad \text{> turn on coffeemaker} \quad \text{> wait until coffee is done}
\]

As in shaking hands, there is more structure beyond the temporal sequence. Everything here is preparation for the coffeemaker performing its function while the actor waits. That is, the actor actually doesn’t perform the head of the action. In order to encode the structure of this action, then, we have to deal with the roles of both the actor and the machine. The actor has to know what the machine is supposed to do.

One feature we haven’t seen before is two unordered subactions. It doesn’t matter which order you put in coffee and put in water; as long as you do both before you turn on the coffeemaker. Nevertheless, you probably have a default (or habitual) order in which you do these subactions. I typically do the water first. (Similarly at larger scales of action: it doesn’t logically matter which order I make the coffee, take a shower, and feed the cats in the morning, but they are all necessary subtasks of the “getting started in the morning” action, and I have a default order.)

In order to start turning (5) into a tree structure, we have to ask how we want to build such “habitual but not necessary” temporal order into the grammar of action and encode it in notation. Our cases so far have involved necessary temporal order: e.g. you can’t shake hands till you’ve grasped hands. These show up in the tree as attachment dependencies: preparation to head, or head to code. By contrast, putting in water and putting in coffee are logically independent of each other. For the moment, I’ll just notate them as two independent Head branches in the Preparation.

Another question is whether these two steps are preparation for turning on the machine, or for a larger constituent that consists of turning on the machine and letting it work. (6a) shows the first possibility and (6b) the second. I kind of favor the latter, though without yet having a notion what would count as evidence. (The evidence would parallel constituency tests in linguistics, and might involve either intuitive or experimental procedures.)
5.3.2. Complexity in the subactions. Let’s look more closely at “actor puts water in machine.” The way I do it is to take the pot out of the coffeemaker, put the right amount of water in it from the faucet, pour the water into the back of the coffeemaker, and put the pot back in its proper place. In a tree structure this might look like (7).

There is more detail, for example in the constituent “fill pot to proper level from faucet”. One problem is what “proper level” is. This is correlated with the amount of coffee one wants to make, which has to be a free parameter in the specification of the task, with perhaps a default setting (for me, six cups). This parameter will show up again in “put coffee in machine”, when one has to measure out the coffee grounds.

A more immediate problem arises in how to lay out the tree. The actor puts the pot under the faucet and turns the water on (not necessarily in that order). Then two things happen at once:

5-13
water runs into the pot, and the actor monitors the water level (which may include some extra steps of checking, such as lifting the pot to eye level). When the level reaches criterion, the actor terminates the process by turning off the water and/or removing the pot from under the faucet. Here the process of the water running into the pot seems like the Head, but the actor’s removing the pot from under the water doesn’t seem like a Coda in the sense we’ve been using the term so far. So let me call these two constituents Process and Termination, more or less following practice in the linguistic literature on event structure and aspect.

We also need a way to note the actor’s monitoring the rising water level. Our little grammar of actions needs to be supplemented to incorporate “checking steps.” Toward that end, let’s introduce the notation ?x? “check to see if x is the case”, where x is a state of affairs — here, whether the water level is at criterion. The action ?x? has two possible continuations, depending on the answer. With all this in place, we might elaborate “fill pot to proper level from faucet” as (8).

(8) fill pot to proper level from faucet

Preparation | Head | Head | Process | Termination
---|---|---|---|---
put pot under faucet | turn on faucet | water runs into pot | y | Head
| enough water? | continue | remove pot | turn off faucet | Head
| n3 | keep checking | from under faucet | |

How much of this actually has to be stored as part of the knowledge of making coffee? Filling things to criterion under a faucet is a more general task, and turning the faucet on and off to get water for whatever purpose is a still more general task. So perhaps the knowledge of making coffee only specifies the endstate of the process, namely that the pot has the right amount of water in it, and the rest is constructed on the fly.

Left out of this are even finer details: You have to grasp the pot, disengage it from the machine, and walk with it to the sink; after filling it, you have to walk back to the machine. While the pot is filling, you have to maintain it in a fixed position, which requires applying more force to hold it up as the water adds to its weight. Turning the faucet on and off calls for mechanical

---

9 The resemblance to the “test” step in the old Miller, Galanter, and Pribram (1969) TOTE units (“test-operate-test-exit”) does not go unnoted. Tests are inevitable as a part of any sort of flexible behavior. On the other hand, the particular test in (8) is not a discrete step: it goes on continuously throughout the process until the condition is satisfied.

5-14
knowledge of the faucet: where to apply force, and how much. How the pot disengages from the machine might be part of the knowledge of making coffee with this machine (what else would you use this knowledge for?), but grasping the pot, holding it and walking surely are not. Your mechanical knowledge of the faucet is not part of the coffeemaking schema, rather it is its own little schema, full of details of your kitchen faucet, other faucet types you know, perhaps related to a more generic schema that helps you deal with faucets you have never encountered before. All the subactions that are not part of the coffeemaking schema need to be recruited on line and integrated into the structure of ongoing action.

More interesting is the possibility that the pot needs to be cleaned before being filled. We'll come back to this shortly.

The "put coffee in machine" constituent in (6) offers a few new sorts of details. My coffeemaker has a permanent filter that has to be taken out and cleaned; most people use paper filters instead. In either case, the filter has to be dealt with, the coffee container has to be located (in my case, in the freezer) and opened, coffee has to be measured into the filter, the coffee container has to be closed and replaced in its proper location. The structure, partly expanded, looks about like (9).

\[
\begin{align*}
\text{Preparation} & \quad \text{Head} & \quad \text{Coda} \\
\text{prepare filter} & \quad \text{prepare coffee} & \quad \text{close up filter part of machine} \\
\text{get out coffee} & \quad \text{measure coffee} & \quad \text{put away coffee} \\
\text{Prep Head Coda} & \quad \text{Prep Head Coda} & \quad \text{Prep Head Coda} \\
\text{open take out shut} & \quad \text{move measure replace} & \quad \text{open put in shut} \\
\text{freezer coffee can freezer} & \text{top of coffee top of coffee can} & \text{can}
\end{align*}
\]

I've left out the cleaning and replacing of the filter (more use of the sink), all of which is subordinate to "prepare filter"; and I haven't touched on the further structure of the lower constituent labeled "measure coffee", where you have to (a) coordinate the amount of coffee with the amount of water and (b) use the whole motor process of repeatedly scooping coffee out of the can and into the filter, while monitoring the total amount scooped — a more elaborate version of the measuring routine in (8). The scooping, of course, presents more fascinating problems for motor control.

The main thing elaborated in (9) is this sequence of openings and closings. The Codas of closing the freezer and putting the top on the coffee can are not necessary for the task. Rather, as in the Coda for shaking hands, they're necessary for restoring the status quo ante. It seems to be
a general problem in socializing one's children that they leave the coffee out of tasks. That's why we're always cleaning up after them and turning off the lights in rooms they left hours ago.

But back to measuring coffee. Most of the structure in (9) is doubtless not part of the coffee-making routine. Rather, my guess is that the coffee-making routine includes the measuring of coffee into the filter plus knowledge of where you keep the coffee; but all the fetching, opening, and closing is constructed online. Nevertheless, all these subtractions must be performed, so they must be integrated into the sequence of motor instructions. The snapshot is that a performed action has a deeply embedded structure—in this case, (7) and (9) embedded in (6b), and (8) further embedded in (7), plus all the pieces I haven't bothered to expand into their full structure. This tree is complex enough that I will spare us all the difficulty of writing it all down in one place.

I'll just mention three other very general complications that need to be fitted in. First, what happens when you're in the middle of making coffee and the phone rings? You have to find a point to break off (I'll bet it's likely to be a constituent boundary), leave a pointer to where you are in the task, answer the phone, then after the phone call return to where you were in making coffee. It's a little like the situation in language when you need to interject a comment in the middle of a discourse. And, just as when your comment gets so long that you lose track of the original thought, you may forget to go back to making coffee after a long phone conversation.

On the other hand, you may interleave the two activities, talking on the phone while continuing to work on the coffee. So we need also the possibility of multi-tasking, guided by shifts of attention between the two.

Finally, making coffee doesn't require the same actor throughout. My wife may start the coffee and I may finish it up. Or she may measure the water while I measure the coffee. And we both keep track of which steps have been completed. Perhaps this is the result of composing a very general joint-activity routine with the coffee-making routine.

5.4. Building structure

5.4.1. Parallels to sentence production. In analyzing each of these actions, I've constantly asked what parts of the structure are stored as part of this action per se and what parts are built online by adjoining other stored actions and by modulating the pattern in various ways. A picture is beginning to emerge. The stored schema that pertains specifically to shaking hands encodes the actual shaking action and its social meaning. The stored schema for making coffee is a structure of semi-connected vignettes: measuring a predetermined amount of water from the faucet into the pot, pouring the water from the pot into the coffeemaker, measuring a commensurate amount of coffee from the can, turning the machine on, and letting it do its work. In both cases, all the "connective tissue" is missing. Where does it come from, and how does it come to be integrated?

We can't take the position that it all comes for free. As people in robotics point out, we can't just leave it all out and expect our robot to perform, using its nonexistent common sense.

5-16
On the other hand, often it seems that these subsections are recruited from more general processes. One walks from one place to another all the time, not just from the coffee maker to the sink and back. One reaches for all kinds of things in all kinds of positions, not just for people’s hands; one uses the sink for all kinds of things besides making coffee. One opens the freezer for all manner of reasons — and getting things in and out of the freezer is just a special case of getting things in and out of cabinets and cupboards. However, some of these actions may have specialized parts, such as how to use your very own kitchen faucet.

Moreover, the construction of actions has to allow a certain degree of flexibility and creativity. Suppose one is going to measure water into the pot, and the pot turns out to contain some dregs of coffee. Then one doesn’t want to go straight into filling it with water; one wants to clean the pot first. Similarly, consider what happens if you go to measure coffee and find there isn’t any. One possible response is to abandon the task altogether: no coffee for you this morning. But if the need is desperate — suppose you’re expecting guests for whom you have to run the “hospitality” routine — you might go out and buy some coffee. In this case the composed structure of (9) contains, instead of “take coffee can out of freezer”, a huge Preparation constituent, which includes driving and navigating and buying routines. And the driving routine may in turn require as Preparation a search for your car keys. But all these routines must be bound to the need for coffee, so that you go to the right place and buy the right thing (if your short-term memory holds out). So the composed structure of an action can become deeply embedded, in unpredictable ways.

This is beginning to look to me a bit like the construction of sentences. You can’t store in your head all the sentences you can speak: there are just too many of them. Rather, you store bits of language in memory and combine them in real time to create a sentence that suits your current needs. Moreover, just as you don’t usually construct a whole sentence in your head before starting to say it, you probably don’t construct a whole complex action in your head before your start performing it. Rather, you add on pieces as you need them (or as you anticipate needing them). Still, in both cases, pieces must be attached in such a way that the output is coherent. In the case of language, coherence amounts to some approximation of grammaticality, not always achieved. We’re trying to figure out here what the notion of coherence of an action might mean.

Standard views in linguistics have it that the stored pieces are some tens of thousands of words plus some unknown but relatively small number of combinatorial rules that combine the words into structures. However, according to a view developing independently in a number of different quarters, there is no strong distinction between words and rules — rather there is a continuum running from simple words such as dog, through lexms such as kick the bucket, which

---

have syntactic structure, through regular morphemes such as the English regular plural -s, through idiosyncratic constructions such as give more you read, the less you understand, all the way to very general principles of combination such as VP → V- NP. The basic principle for building sentences is called unification: the idea is that one clips together stored pieces any way possible, consistent with the constraints each of them imposes, such that every bit of the composed structure is accounted for from one stored piece or another (Shieber 1986). Such an approach seems appropriate for the construction of actions from multiple schemas.

Like the “semi-connected vignettes” I’ve posited for some action schemas, some of the stored pieces of language are discontinuous. For instance, the idiom take... for granted is certainly a stored unit, but it cannot be used in a sentence unless something fills in the gap between its two parts. Another familiar example is the French split negation ne...pas. On the other hand, what binds such items together is that their discontinuous parts are connected in the tree structure. This seems also true of the relevant parts of the coffeemaking schema.

The parallels go even further. Among the principles involved in sentence composition are principles of “referential binding.” A particularly apt case might be relative clauses: a relative clause has to contain a pronoun element of some sort that co-refers with the noun the clause is attached to. Consider the sentence There’s the coffee that I made last night. The relative clause is understood as I made (some) coffee last night, not, say, I made a cake last night. This constraint is reminiscent of situations we have encountered here. For instance, going to the store and buying cheese is an “ill-formed” preparation for baking coffee, and it doesn’t make sense to wash a spoon when what needs to be clean is the pot. In other words, the formal device of variable binding that connects a relative clause to its head noun is also about right for connecting objects in subordinate actions to objects in the main actions they prepare. (And if the relative clause gets real long and you lose track, you may be the equivalent of getting to the store and forgetting what you were supposed to buy.)

Another such parallel arises in the necessity of correlating the amount of coffee you measure out with the amount of water you measure out. Such a correlation is directly expressed in a sentence like For each cup of water, put one scoop of coffee - that is, a sentence involving quantification. (It’s unclear to me whether one could have such correlations in action structures without having linguistic quantification to support it).

5.4.2. Filling in the blanks: If actions are stored in memory in highly schematic form, how are all the other pieces recruited and integrated into the main schema one is executing?

A key emerges, I think, from the problem posed a moment back: suppose when you go to fill the coffeepot you discover it needs to be cleaned. The cleaning routine itself wouldn’t be a part of coffeemaking: one cleans all kinds of things and the motor control involved is

Chomsky’s (1995) “Merge” operation, supposedly maximally simple and general, is one case of unification but not the only one.

5-18
Phenomenally complicated - try to pay attention to your hands when you’re washing dishes.

But when it is necessary to clean the pot, the cleaning routine has to be inserted as a Preparation for (8), perhaps as in (10).

(10) \[ \text{fill pot to proper level} \]
    \[ \text{Preparation} \quad \text{Head} \]
    \[ \text{clean pot} \quad \text{(8)} \]

One way to make this happen would be to add a checking step in the stored structure, perhaps like (11).

(11) \[ \text{fill pot to proper level} \]
    \[ \text{Preparation} \quad \text{Head} \]
    \[ \text{if} \]
    \[ \text{?pot clean? - continue} \]
    \[ \text{..} \]
    \[ \text{clean pot} \]

I'm not fond of this solution. It misses the point that every time we start to use a utensil we want it to be clean. Moreover, if we put a checking step in the structure at every juncture where something might be needed or something might go wrong, we risk the computational explosion of the dreaded Frame Problem (McCarthy and Hayes 1969). E.g. when walking from the coffee maker to the sink, do you have to constantly check to make sure the floor is still there and your feet are still connected to your legs? There are too many ways things can go wrong and too many possible repairs to build them all into particular routines.

If an action schema does not stipulate all the necessary preparatory routines, what motivates adding them to the composed structure? Let's go back to shaking hands. The shaking constituent stipulates the hand configuration between the two participants. But at the point where you decide to shake hands, your hands aren't at the right place. The insertion of reaching and grasping is motivated by this discrepancy. But how should we flesh out this term "motivates", short of putting the reaching and grasping back into the handtake script?

I'm going to again invoke the model of language. The linguistic parallel to stored action schemas is stored words and idioms, which can be integrated into sentence structures in an unlimited number of different contexts. What "motivates" their being activated is that the speaker has a piece of meaning in mind. The psycholinguistic evidence (e.g. Dell, Burger, and Svec 1997, Levelt 1999) shows that the "call" to the lexicon is very general, along the lines of "does anyone
in there mean *this*?* We can think of lexical items actively "volunteering" in response to the need for a particular meaning to be expressed, and competing actively with each other for expression; a process of selection picks out the one that is actually uttered. (We get speech errors if the competition isn't properly resolved, e.g. *problem* for *trouble plus problem.*)

A similar conception seems appropriate for complex actions. When one decides it's appropriate to shake hands, there is a discrepency between the position one's hand happens to be in and the position it needs to be in, i.e. grasping the other person's hand. The result is a "call" to the "action lexicon" for an action to get one from there to there. The "reach for x and grasp x" routine stored in memory "volunteers" and is attached as a Preparation to "shake!" But as part of being attached, its variables are instantiated by the relevant objects in the current situation (the way a verb's variables are instantiated by its subject and object), so that one reaches to the right place and grasps the right thing. Similar mismatches of position are going to get you to walk from the coffeemaker to the sink and back, and to drive to the store to buy more coffee and return (which in turn requires you to walk to your car). That is, the discrepency "I'm here and I need to get there" can call up a variety of routines depending on the context. This intuitive account is fleshed out in the AI literature on planning (Pollack 1990) and robotics (Badler et al. 1999). In the latter of these, a standard example is, while the virtual robot is sitting down, giving it a command to walk somewhere. In order for the command to be executable, first a Preparation of standing up must be adjoined to the structure; the program does this automatically.

Another very general sort of Preparation based on discrepency involves the use of an object that one is not already holding — for instance "fill pot to proper level from faucet." As Preparation, one must pick up the pot. But in Preparation for that, the pot must be located — so I have to know where to find it. My knowledge of my coffeepot is that it sits in the coffeemaker (so I know to know where that is) — and this was built into structure (7). But suppose the coffeepot isn't there. I have some backup knowledge: it might be in the dishdrain or the dishwasher, so I'll look there. Failing that, I have to institute a general search. The point is that the discrepency of needing a particular object triggers a search for it, and the search can be driven by knowledge very specific to the object. The object belongs here, but it might also be in such-and-such other places. Multiply that over all the objects in your house and the knowledge is phenomenally rich. Where do you keep the cake-pans? the spare key? Aunt Betty's pewter candle snuffer? And it's not just in your house: Where in the supermarket is the pasta? the beer? and so on.

A general routine along the lines of "find object x" can fill in many of the complicated details of opening and closing in the "measure out coffee" routine in (9). This simplifies the stored schema, but of course then all the extra pieces have to be added online. If the coffee's not found in the expected place, the backup location is the store, triggering the construction of the whole trip to the store routine.

Not every case of discrepency involves a mismatch of position. Remember the dirty coffeepot. When it comes time to measure water into the pot, "wash x" is called up to rectify the
discrepancy between its present state and the desired state of a clean pot. If the pot is already clean, no discrepancy is detected, so nothing needs to be added into the water-measuring routine. Where does the presumption of a clean pot come from? At worst, it could be part of the routine, like the proper starting position for shaking hands. But more likely, there's a general constraint on use of utensils that is unified with any particular food-preparation routine — more composition.

Finally, I've kept talking about Coda as returning to the status quo ante. Here the discrepancy is between the state of affairs at the end of the Head of the action and the state obtaining before the action began. In the case of shaking hands, this Coda is motivated by necessity: you can't be attached to the other participant forever. In the case of the various steps of making coffee, the Coda is motivated more by attention and foresight, the desire to keep things in order. So this sort of Coda is more dispensable, especially for children, whose attention and foresight is not up to their parents' standards.

Speaking of foresight, a special higher-order sort of discrepancy is the sort that leads you to notice you're running low on coffee and thereby write it on the shopping list — preparing a Preparation for an action that isn't even planned yet. It's not at all clear how this might be incorporated into tree structures for complex actions.

What I find interesting about this approach to action composition is the hypothesis that the main routine does not need to be responsible for calling all the subsections. "Making coffee" does not have to be full of checkpoints like "is the pot clean?" "is there coffee in the freezer?" "is the top off the coffee can already?" "am I at the sink yet?" and "is the sink still there?" Rather, subactions are called by the integrative process that adapts the schema to a particular action suited to the present context. The integrative process is sensitive to discrepancies between the present situation and the requirements of the routine in progress. Moreover, having detected a discrepancy, the integrative process does not specify a particular subaction, it simply calls for any action that can rectify the discrepancy, and like words, suitable subactions "volunteer."

A more difficult case is when no suitable actions volunteer. This is perhaps the point where conscious planning has to step in, searching for a way to get to the desired point in a series of steps, each of which is already a stored schema. I'll not follow this line up, but here is where we make contact with other traditions in the planning literature, such as the old Newell and Simon (1972) General Problem Solver.

5.4.3 Choosing among alternative actions. There is a further parallel to lexical selection. Recall that when speech production calls the lexicon, all remotely appropriate words get activated, and then comes a process of selecting one as the one to be integrated into the present utterance. Parallel situations arise in constructing actions, often consciously (as word choice is, very occasionally). Consider a constituent of "making coffee" we haven't looked at yet: "Turn on
machine." Suppose I press the switch, but then the machine doesn’t go on. I may notice this because the little red light doesn’t go on, or because ten seconds later the water doesn’t start hissing, or because two minutes later there still isn’t coffee in the pot. For any of these cues to alert me, I need to know something about what to expect from the machine – particular knowledge about how the device works – and my attention needs to be drawn to the discrepancy.

Suppose I detect the discrepancy. What happens next? I might see if maybe the coffeemaker is unplugged, and if so plug it in and then look again to see if the red light is on. Again, I don’t think we want to say that checking the plug is part of my knowledge of making coffee. Rather, this would seem to be the first thing that comes to mind in a general collection of strategies for dealing with any electrical appliance that doesn’t operate as expected. My wife says her next strategy is “Call Ray,” a very general strategy for attempting to repair anything. Then there are other electrical appliance strategies such as checking whether a fuse has blown (in which case you continue by replacing it), checking whether the power in the house is out (in which case you continue by checking the neighborhood, calling the power company, etc.), and maybe even taking the appliance apart.¹⁴

One option for every repair routine is “abandon task.” If this option is chosen, abandonment percolates up the structure of the action to everything for which this subaction is necessary. So if you abandon trying to get the coffeemaker to turn on, you also abandon making coffee, and therefore (under the usual scenario) you also abandon drinking coffee. On the other hand, alternative courses of action can be pursued further up the tree. If the coffeemaker doesn’t work, so making coffee fails, one can, if the need is sufficient, go out to a café (if one is open).

There are also cases where your knowledge of a device includes a specific repair strategy. For instance, when I turn on the front burner of my stove, sometimes it doesn’t light; and sometimes then I can get it to light by also turning on the rear burner. Or maybe you have to hit your car radio a special way to get it to turn on, or you have to jiggle the key in your front door a certain way to get it to turn. We carry around dozens if not hundreds of these tricks in memory.

¹³More particular knowledge of objects: You have to know where the switch is on the machine, and how to press it. Remember that these are not always obvious with a machine you’ve never encountered before – my favorites are the lamps in hotels.

¹⁴In turn, this invokes one’s knowledge of the causal structure of the device and how the various attempted remedies may bring about their effect. For some people and some repairs this sort of causal structure is absent; for instance, most of us don’t know much about how a radio works. In such a case any repair routines we might know are simply operational (say to bang on the top of it), not based on any especially deep understanding of the device. Much of the practice of medicine, especially before the 20th century, has been of this nature. On the other hand, when one does understand how a device works, one’s responses to malfunction are probably different and more flexible.

5-22
The upset is that the discrepancy of a device not working as it should is sometimes repaired by a very general action, and sometimes by a very specific action.

An old-fashioned programming approach to this plethora of strategies would be to have an ordered checklist. And sometimes we do use such a structured meta-strategy. But I don’t think that’s a sufficiently general solution — there’s no reason always to try everything in the same order. Rather, there are situations in which no mere list of possible repairs will do the trick. Suppose you and an acquaintance spot each other at a party, and it’s appropriate to shake hands. But her hands are full of food, or you’ve been eating chicken wings so your hands are full of grease. Or she’s way across the room talking to someone else. On such an occasion you may improvise a symbolic handshake, say reaching in her direction and shaking your hand rigidity but without grasping hers. You may have never done this before; it’s a new action, created on the spot. So it can’t be on a list. (On the other hand, you may then store it away in a list for future emergencies.)

In the spirit of the approach to the free composition of actions we’re trying out here, let’s consider another possibility. Suppose that, like multiple words “volunteer” to fit a desired meaning, multiple actions can “volunteer” to fill a gap in action — in fact, all reasonably appropriate actions do so (including “perverted” or “coerced” versions of the normal routines such as the symbolic handshake). Returning to the coffee example, in response to detection of the discrepancy “switch pressed but coffeemaker hasn’t gone on,” a lot of different actions with different levels of generality are activated.

Which option is selected? Following the same intuitions that motivate the literature on rational choice, optimization, and heuristics, let’s guess that the primary criterion for selection is cost or effort. Checking the plug doesn’t cost very much, checking the fuse costs a good deal more, asking the coffeemaker apart a whole lot more. For Hildy, calling me (if I’m in earshot) costs less than checking the fuse.15 So the order in which you try different repair strategies depends on their relative cost. Notice that relative cost has to be computed in a situation-specific fashion. For instance, the cost of abandoning attempts to get the coffeemaker to work depends on how much you want coffee (or think you’ll want it soon). If making coffee is a Preparation for the “hospitality” routine, the cost is higher than if the coffee is just for yourself.

All of this presumes that you can estimate an action’s cost in advance of performing it. My impression from a passing acquaintance with the literature is that the computations rapidly get fairly horrendous (though perhaps more so than in lexical selection). Here at last is a place where morality will impinge on the theory of action: moral judgment places biases on the value or cost of certain actions, and therefore affects the choice among alternatives.

15For the term “coerced”, see section 6.x.

16The cost to me is likely to play a role in her decision too: she’s nicer than some other people I’ve known.
5.5. Summing up

What have we got out of this exercise?

- Even the simplest and most routine actions reveal a complex hierarchical structure, some of which is stored in memory as action schemas and some of which is the result of composing stored action schemas online.
- The structure of an action can involve both social and physical planes. The physical plane can involve both functional description (what is to be accomplished) and more strictly physical description (what motions accomplish it). The latter is linked to the actual motor script that realizes the action as muscle activations guided by proprioception.
- The structural relations underlying the composition of actions include
  (a) the combination of actions as Head, Preparation, and Course of a larger action;
  (b) concurrent Modulation of a Head;
  (c) Modulation of a Process by checking, which can terminate the Process;
  (d) temporally unordered Heads;
  and probably other possibilities.
- These structural options can be thought of as constituting parts of the “grammar” of actions.
- Some aspects of a stored action schema can be applied either to perceiving others performing the action or to executing one’s own action.
- One stores vast amounts of information about how various devices work, potentially at every level of specificity from very particular (the faucet in my kitchen) to very general (electrical appliances). Perhaps “naive physics” is one of the most general schemas in this class — “how physical objects and substances work.”
- One also stores vast amounts of information about the canonical location of particular objects.
- Much of the online composition of action is not driven by explicit choices in stored routines: making all the choice points explicit would lead to computational explosion (the frame problem). Rather, stored actions are to some degree skeletal and schematic. The full complexity of executed actions is a consequence of
  (a) composing multiple stored actions;
  (b) instantiating and binding variables in the schemas to suit the current context, including composition with other actions.
- The variables may include locations to be moved to and speed of motion.
- Composition is often motivated by a discrepancy between the current situation and the situation required to initiate an intended action. The discrepancy can be a matter of physical necessity (getting your hand to the right place or handling an object), a matter requiring perceptual attention (the dirty coffeepot), or a matter of foresight (running low on coffee).
- A discrepancy triggers a call to the “action lexicon”, which is answered promiscuously by all reasonably appropriate stored actions. The action actually executed is selected according to minimum cost, where cost is at this point a highly context-dependent wild card.
- The question of how to select among alternative actions is a crucial problem. Here is where the calculation of cost and benefit comes in; here is where heuristic reasoning comes in.
Some of the discussion here is reminiscent of Schank and Abelson's (1977) "scripts" and Minsky's (1975) "frames". These approaches proposed that one stores structured knowledge of complex actions, so that one doesn't have to build them up from primitives every time (using a General Problem Solver or the like). One important difference is that they were mostly couched in terms of what you need to know about actions in order to understand stories -- for instance, what you need to know about restaurants so you know there's an implicit waiter in the narrative "Bill ordered a hamburger and it was burned so he didn't leave a big tip." Here, though, we're interested in what you need to know about actions in order to do them. On the other hand, it's plausible that it amounts to the same kind of knowledge in the end. For example, our question of what you have to store in the coffeemaking routine might translate into asking exactly what verbal and pictorial information you have to put on the instruction sheet that comes with the coffeemaker. (And I believe these earlier approaches recognized the potential convergence.)

My impression is that the script/frame idea floundered for a number of reasons: first, there were just too many contingencies that couldn't be mentioned explicitly in the script; second, it proved difficult to characterize situations that involved mixed scripts (e.g. a birthday party at a restaurant); third, there was no notion of inheritance hierarchy that permitted a smooth transition from very explicit to very general scripts. The approach here attempts to avoid these problems by (a) invoking compositionality, in particular compositionality that is driven not by explicit instructions in the schemas themselves, and (b) organizing action schemas in terms of inheritance hierarchies of generality. Here we have drawn on analogies to lexical access and composition in speech production, an area of cognitive science that has blossomed in the past 25 years. (The closest analogy in the AI of the '70s was perhaps the notion of "demons" (Lindsay and Norman 1977); Minsky's (1986) "Society of Mind" is along similar lines.)

What I haven't dealt with at all is the process of consciously constructing complex plans, where you're starting more or less from scratch, and many many steps have to be put together for an action to work -- say, inventing a coffeemaker. Here I've been concerned with the more or less automatic construction of actions that we perform all the time in our daily lives. The point is to see how rich even these dumb unnoticed actions are. Of course, the more of these schemata you have in memory, the smoother your action becomes: a call to the action lexicon produces lots of useful results among which one can select, rather than no useful results, so that one must rely on conscious ingenuity.

As in language, there is the question of learning: how do you get all these schematic actions in your head, so you have this repertoire? Some of them might be explicitly taught, but others aren't. Where do such new items come from? They have to be constructed from pieces assembled online. Where do those pieces come from? As in the case of language, it might make sense to look for a primitive basis of features in terms of which actions can be assembled. This has to include not just basic actions but also the structural principles for connecting them, such as
the notions of decomposition into Head, Preparation, and Coda. So we are looking at the prospect of a sort of "universal grammar" that is the initial state for acquiring an action lexicon.

Let me conclude by summing up the parallels to the theory of language. Action structures, like linguistic structures, can be full of embedded constituent structure. Like linguistic structures, they seem to be determined by some sort of a grammar that specifies the structural options. Like the linguistic lexicon, there appears to be an enormous action lexicon, partly indexed by the objects involved in the actions. Like the linguistic lexicon, the action lexicon is structured in terms of inheritance hierarchies, which relate very general to very specific schemes with all degrees of generality in between. Like the composition of linguistic structures, the composition of action structures involves instantiation and binding of variables, and possibly even quantification. We have been able to make some plausible claims about the online construction of actions by taking seriously the analogy with language production. Moreover, like linguistic structures, action structures can be used not only to produce actions of one's own but also to understand the actions of others.

Those who wish to deny language much in the way of special character might therefore be tempted to say that we have shown language to be simply an outgrowth of action in general. I think such a conclusion would be misguided. First of all, the theory of action at this point is little more than armchair speculation, backed up by some research in planning and robotics but no depth of coverage and no psychological experimentation. So it hardly can serve as a serious justification for repudiating decades of linguistic research.

In addition, even at the primitive stage achieved here, it is possible to recognize important differences. To be sure, language partakes of many general principles of structural organization, memory, and processing that are shared by other systems of mind. That should be no surprise. However, one thing that makes language special is its role in the architecture of the mind, as a bidirectional conduit between the structure of thought and overt communicative expression. Action has quite a different role. Another thing that makes language special is the particular forms of structure that the general principles apply to. The tree structures for actions have different categories and arrangements of constituents from the tree structures for language, especially those for phonology and syntax. If there is any structural common ground to be found between action and language, it is in conceptual structure, which presumably can encode aspects of complex actions. But conceptual structure is the organization of thought in general, and is not particular to language.

So the conclusion is mixed: the wondrous recursive creativity of language is not as special as it is often claimed to be. Nevertheless, language is a special system because of what it does and the particular structural materials it uses to do it.