Coverage and Competency in Formal Theories: A Commonsense Theory of Memory

Andrew S. Gordon
USC Institute for Creative Technologies
13274 Fiji Way
Marina del Rey, CA 90292
gordon@ict.usc.edu

Jerry R. Hobbs
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
hobbs@isi.edu

January 21, 2003

Abstract

The utility of formal theories of commonsense reasoning will depend both on their competency in solving problems and on their conceptual coverage. We argue that the problems of coverage and competency can be decoupled and solved with different methods for a given commonsense domain. We describe a methodology for identifying the coverage requirements of theories through the large-scale analysis of planning strategies, with further refinements made by collecting and categorizing instances of natural language expressions pertaining to the domain. We demonstrate the effectiveness of this methodology in identifying the representational coverage requirements of theories of the commonsense psychology of human memory. We then apply traditional methods of formalization to produce a formal first-order theory of commonsense memory with a high degree of competency and coverage.

1 Coverage and Competency

While much research in commonsense reasoning has been directed at describing axiomatic content theories in specific areas, of equal concern are the research methods that are used to develop these content theories. Davis (1998) reflects back on the methodological problems that have hindered progress and recommends a research program based on microworlds. He argues that the goal of commonsense reasoning research is the generation of *competency theories* that can answer commonsense problems that people are able to solve. By emphasizing reasoning competency, Davis makes a strong case for focusing on the function of axiomatic theories rather than their form. However, while the representational form may be indeterminate with respect to its function, representation itself has an even broader role to play across the full spectrum of cognitive behavior, beyond the commonsense reasoning functions of explanation, prediction, planning and design. What is needed of commonsense theories is not only competency, but also enough coverage over the breadth of commonsense concepts to enable use in computational models of memory retrieval, language understanding, perception, similarity, among other cognitive functions. A conservative commonsense reasoning researcher might argue that coverage is an additional constraint on an already difficult task, and is best addressed after suitable competency theories have been put forth. We argue that without addressing the issue of coverage first, competency theories will be intolerant of elaboration and difficult to integrate with each other or within larger cognitive models.

This paper presents a new methodology for authoring formal commonsense theories. The basis of our approach is the tenet that the problems of coverage and competency should be decoupled and addressed by entirely different methods. Our approach begins by outlining the coverage requirements of commonsense theories through the analysis of a corpus of strategies. These requirements are elaborated to handle distinctions made in natural language, as evidenced through the analysis of large English text corpora. We then address the specification of a formal notation (here, first-order predicate calculus) and of a full axiomatic theory. Section 2 of this paper describes the methods used to solve the coverage problem in the domains of commonsense psychology. Section 3 elaborates on the role of natural language in refining these representations, with an example domain of the commonsense psychology of memory. Section 4 presents a formal, axiomatic theory of the com-

monsense psychology of memory aimed at achieving both a high degree of coverage and inferential competency. Section 5 offers our conclusions and considers the challenges of future work in formalizing other domains of commonsense psychology.

2 Commonsense Psychology in Planning Strategies

While the commonsense reasoning research community has long been interested in developing axiomatic theories of the physical world, i.e. naive physics, the field of psychology has become increasingly interested in commonsense reasoning about human psychology itself. Commonsense psychology, also referred to as naive psychology, folk psychology, and the Theory of Mind, concerns the reasoning abilities people use to predict and explain what is going on in their own minds and in the minds of other people. Developmental psychologists have noted that these abilities are strongly age-dependent (Wellman and Lagattuta, 2000; Happe et al., 1998) and have argued that they are central in explaining cognitive deficiencies associated with autism (Baron-Cohen, 2000) and schizophrenia (Corcoran, 2001). Although alternative theories have been proposed (Goldman, 2000), researchers have asserted that our commonsense psychological abilities are facilitated by a tacit representation-level theory of mental attitudes and behavior (Gopnik and Meltzoff, 1997; Nichols and Stich, forthcoming). This hypothesis, referred to as the "Theory Theory", is very much in line with the perspective of the average knowledge representation researcher, whose aim is to describe tacit representational-level theories as explicit, axiomatic theories.

There has been some previous, highly influential, formal knowledge representation research in the area of reasoning about mental states and processes (Cohen and Levesque, 1990; Ortiz, 1998). Interesting beginnings though these theories are, they fall short of what many researchers investigating commonsense psychology feel is required, particularly among those investigating the role that language plays in acquiring mental state concepts (e.g., Dyer et al., 2000).

Gordon (2001a) noted that there is an interesting relationship between concepts that participate in commonsense psychology theories and planning strategies, the abstract patterns of goal-directed behavior that people recognize across analogous planning cases. For example, consider a strategy that was reported by a concert pianist that he used as an aid in memorizing complicated compositions such that they can be executed without referring to sheet music. For particularly challenging passages, the pianist explained that he would focus not on the sensations of his hands hitting the keys during practice, but rather on the visual motions he experienced by watching his hands. He reasoned that his ability to remember complex visual patterns was sometimes more effective than his motor memory, and found that if he again focused his eyes on his hands during a performance, his expectations would guide them to do the right thing. This same strategy may be applicable to workers who operate complex machinery, and even more generally applicable to any performance-based memory task that is directly perceived by the performer. Domain-specific details of any application can be abstracted away so that the description of the strategy does not refer to musical pieces, piano keyboards, or human hands. There are some concepts in this strategy that will remain essential to every instantiation of it in any planning situation. These concepts include the commonsense psychology notions of the *focus of attention*, the *observation of a performance*, the *expected pattern of perception*, and the intention of *memorization*, among others.

Noting the conceptual breadth of planning strategies like this one, Gordon (2001b) devised a methodology for outlining the representational requirements of planning strategies involving the collection and analysis of a large corpus of planning strategies across many different planning domains. 372 strategies were collected in ten different planning domains (e.g., politics, warfare, personal relationships, performance), by referring to domain experts . Pre-formal representations of the strategies were authored to identify the concepts that participated in all instances of each of the strategies, regardless of the specifics of the planning situation. A set of 8,844 concepts were identified as necessary for their representation. This was reduced by combining synonymous terms to a controlled vocabulary of 988 unique concepts.

The ontological scope of these terms was very broad. To better understand this scope, the 988 terms were grouped into 48 representational areas that corresponded to traditional areas of research in knowledge representation or cognitive modeling. Of the 48 representational areas, 18 were closely related to existing areas of knowledge representation work. These 18 included those that focused on fundamental world physics (e.g., time, space, events, sets, values, objects) and the characteristics of people (e.g., organizations, relationships, activities, abilities). The remaining 30 representational areas (outlined in Gordon, 2002) were related to the mental processes of people, broadly speaking. These areas included those

related to planning (e.g., goal management, plan construction, plan adaptation), envisionment (expectation, execution envisionment, other-agent reasoning), and the execution of plans in the world (plan following, execution control, monitoring), among many others. The commonsense psychology concepts from the piano performance strategy above were clustered into the representational areas of Body Interaction (*focus of attention*), Observation of Execution (*observation of a performance*), Managing Expectations (*expected pattern of perception*), and Memory Retrieval (*memorization*).

The conceptual breadth of these 30 commonsense psychology areas is significantly greater than previous work in these areas of formal knowledge representation. However, as this approach is not rooted in inference, no axiomatic theories to drive deductive reasoning about mental states and process are produced by this approach - only an indication of the sorts of concepts that would participate in these axiomatic theories. While it is tempting to simply treat each of these concepts as a predicate in first-order logic, the character of these terms poses a few significant problems. The conceptual specificity of the terms in an area is not uniform. An area such as Goal Management (referring to people's ability to select and prioritize the goals that they will attempt to pursue) includes some very general concepts among the 34 that were identified, such as the mental event of suspending the pursuit of a goal or the mental entity of the currently pursued goal. However, it also calls for more specific terms, such as the mental event of removing an auxiliary goal and the mental event of removing a knowledge goal after it has been achieved.

An even more significant problem exists when the evidence offered by strategy representation provides only a handful of terms to indicate the conceptual breadth of the representational area. This problem is best exemplified by the smallest representational area identified, Memory Retrieval (referring to people's ability to store and retrieve information between the focus of their attention and their memory). Only three memory-related terms occurred in the strategy representations: the mental event of attempting to memorize something so that it could be retrieved from memory at a later time, the mental event of retrieving something from memory into the focus of one's attention, and the mental construct of a memory cue that is the trigger for a memory retrieval event. While it is conceivable that an axiomatic theory could be constructed from predicates based on these three concepts alone, there is a strong sense that our commonsense models of the human memory process are richer than this. In order to solve both the problems of conceptual specificity and sparse concepts, a second phase of refinement is necessary.

3 Commonsense Psychology in Natural Language

The relation between the way people use language in communication and the sorts of formal representations of meaning that are employed in commonsense reasoning theories is very complex, and knowledge representation researchers have generally avoided it as the focus of their efforts. Be we need not be deterred from capitalizing on language as a resource to guide our work. Natural language is our most expressive means of making conceptual distinctions, and the analysis of corpora of written or transcribed natural language can greatly influence the conceptual distinctions we make in our formal commonsense theories where coverage is a major concern.

In this section we describe a method for transitioning from the pre-formal concepts that were identified via the representation of strategies to a set of concepts that will participate in axiomatic theories of the representational area. This method is language based, as it involves the large-scale analysis of natural language text data using tools and techniques borrowed from the field of computational linguistics. The method is labor intensive, requiring expertise outside the field of knowledge representation. In developing this method, we employed three graduate students at the University of Southern California specializing in linguistics or computational linguistics, and several weeks of full time work were required to apply this methodology for each representational area. In describing this methodology, we provide examples from the area of Memory Retrieval.

The first step, *expression elicitation*, is to write down many natural language sentences that include in their meaning concepts from a particular area (e.g., Memory Retrieval). For example, the Memory Retrieval mental event concept of retrieving something from memory into the focus of one's attention is expressible in a wide variety of English expressions:

> He *was reminded of* the time he crashed his car.
>
> The broken headlight *made him think of* when he crashed his car.
>
> He *remembered* the exact location of the car crash.
>
> He *recalled* the name of the street where it took place.

This work was completed largely by collaborative

brainstorming among native English speakers, and typically generated a dozen or so nouns, verbs, adjectives, and idiomatic expressions for each of the concepts in the representational area.

The second step, *lexical expansion*, is to use the initial expressions to seed a more thorough search for related words and expressions in various linguistic reference resources, including traditional dictionaries and thesauri, thematic dictionaries, and phrase dictionaries. The task is simply to look up each of the initially elicited expressions in these resources and to record other associated expressions that are identified. Particularly valuable resources included Addison-Wesley's Longman Language Activator, Collins Cobuild's reference of English Grammar, and Levin's description of English verb classes and alternations (1993). As an example, the initial set of expressions concerning memory was expanded to include verb phrases such as "to know by heart" and "to suppress the memory of", and nouns such as "hint" or "memento". This work was conducted by members of our team with linguistic expertise, and typically generated tens to hundreds of language fragments per representational area.

The third step, *corpus analysis*, is to collect a large database of real examples of the use of language related to the representational area by encoding the relevant vocabulary into finite-state automata that can be applied to very large text corpora. Large scale corpus analysis has become commonplace in modern computational linguistics research, and many research groups have authored software designed to make it easy for researchers to collect instances of particular linguistic patterns by extracting them directly from textual data. Our group utilized the Intex Corpus Processor software (Silberztein 1999a, 1999b), which allowed us to author linguistic patterns as finite-state automata using a graphical user interface. To simplify the specification of patterns, Intex includes a large-coverage English dictionary authored by Blandine Courtois that allows generalization over linguistic variations, such as verb inflections for a given lemma. For example, a pattern for a class of memory expressions can be described generally (here as regular expressions) as "<make> <PRONOUN> think of", and used to index "made him think of" and "makes her think of" among others. Hundreds of generalized linguistic patterns are authored during this step, one for every expression that is identified in the previous step. These are combined into a single finite-state automaton that can be applied to any English text corpus. To collect real examples of the use of these expressions, we applied them to twentieth century fiction and non-fiction works from Project Gutenberg (http://www.gutenberg.net), typically yielding hundreds to thousands of indexes per representational area per averaged-sized book. Sentences containing these indexes were then compiled into a list (concordance) for review.

The fourth step, *model building*, is to review the results of the corpus analysis step to identify the conceptual distinctions made in real language use. The aim of this step is to identify a set of conceptual primitives to be used in an axiomatic theory that is of broad enough coverage to capture the distinctions that are evident in the concordance. This set will serve as a replacement for the initial list identified through strategy representation. The task in this step is to cluster the sentences in the concordance by hand into sets of synonymous uses of the expressions. In our working group, this step was the only step that was not conducted by our linguists and computational linguists graduate students, as it relied more heavily on familiarity with the practices of formalizing knowledge domains. While it is often argued that there are no true sets of synonymous expressions, an effort was made to identify distinctions that will play functional roles in axiomatic theories of reasonable complexity. For example, there are shades of semantic difference between uses of the phrases "repression of memory" and "suppression of memory", but we felt that it was unlikely that axiomatic theories of memory retrieval would be able to define or capitalize on these differences in the near future, so instances of the two uses were judged synonymous to the mental event of causing a concept in memory to become inaccessible.

The model-building step for the area of Memory Retrieval resulted in twelve clusters of synonymous linguistic uses of the expressions, which can be described as follows. People have a *memory ability* (1) that allows them to move *memory items* (2) in and out of the focus of their attention, unless they are *repressed memory items* (3). People have some intentional control over their memory, including the operators of *memory storage* (4) for memorizing things and *memory retrieval* (5) for recalling things into focus. The second operator can fail, however, resulting in a *memory retrieval failure* (6). There are unintentional actions of memory as well, such as making a memory inaccessible through *memory repression* (7). The everyday unintentional function of memory is simply to cause a *reminding* (8), particularly when some other *memory cue* (9) is the focus of attention. This plays a special role in the processes that surround plan execution, where you may *schedule a plan* (10) with the intention of remembering to do it at a certain time in the future, specifically during the event of a *sched-*

uled plan retrieval (11). But sometimes this can fail, yielding a *scheduled plan retrieval failure* (12). In the next section we axiomatize the concepts required to support these twelve clusters.

4 A Formal Commonsense Theory of Memory

Having identified a set of representational constructs that will participate in any commonsense theory of memory of broad coverage, we can now employ more traditional knowledge representation methods of formalization and axiomization. This section presents the results of applying these methods in the form of a competency theory of the commonsense psychology of memory that achieves the identified broad coverage requirements.

4.1 Concepts in Memory

We assume that the mind has (at least) two parts, a focus of attention, or focus, where conscious thought takes place, and a memory, where memories are stored.

$$mind(x,p) \supset (\exists f)[focus(f,p) \wedge part(f,x)]^1$$
$$mind(x,p) \supset (\exists m)memory(m,p) \wedge part(m,x)$$

If x is the mind of a person p, then x has a part f that is p's focus of attention. (We will refer to this as the "focus".) Similarly, memory. The predicate *part* would be explicated in a theory of composite entities (or things made of other things). The second argument of *mind* must be a person or other agent.

$$mind(x,p) \supset agent(p)$$
$$person(p) \supset agent(p)$$
$$person(p) \supset (\exists x)mind(x,p)$$

Persons have minds, but it is not necessarily presumed here that agents other than persons have minds.

We will refer to the entities that occupy these parts of minds as "concepts", without further specifying their nature here except as noted below concerning associations. We assume a theory of the structure of information would further explicate this. A concept can be *in* focus or memory at a particular time, in a mental sense. We will use the predicate *inm* for this.

[1] A note on notation: Conjunction ($\wedge$) takes precedence over implication($\supset$) and equivalence ($\equiv$). Formulas are assumed to be universally quantified on the variables appearing in the antecedent of the highest-level implication.

$$inm(c,x,t) \supset concept(c) \wedge temporal\text{-}entity(t)$$
$$inm(c,x,t) \supset (\exists p,m)[agent(p) \wedge mind(m,p) \wedge part(x,m)]$$

Temporal entities are explicated in the theory of time; they include both instants and intervals. The commonsense theory of thinking or envisionment would include conditions involving concepts being *inm* the focus during the thinking process.

An agent *stores* a concept in memory when there is a change from a state in which the concept is in the person's focus of attention but not in the memory to one in which it is in the memory.

$$agent(p) \wedge focus(f,p) \wedge memory(m,p) \supset (\forall c,t_2)[store(p,c,t_2) \equiv (\exists t_1)change(inm(c,f,t_1) \wedge \neg inm(c,m,t_1), inm(c,m,t_2),t_2)]$$

This axioms is silent about whether the concept is still in focus. The application of the predicate *change* to formulas is a shorthand for a more complex representation involving reifying the *inm* and other relations (Hobbs, 1985). The notion of *change* is explicated in a theory of changes of state, which is related to the theory of time (Hobbs et al., 1987).

Similarly, to *retrieve* a concept from memory is to change from a state in which the concept is in memory and not in focus to one in which it is still in memory but also in focus.

$$agent(p) \wedge focus(f,p) \wedge memory(m,p) \supset (\forall c,t_2)[retrieve(p,c,t_2) \equiv (\exists t_1)change(inm(c,m,t_1) \wedge \neg inm(c,f,t_1), inm(c,f,t_2) \wedge inm(c,m,t_2),t_2)]$$

The predicates *store* and *retrieve* are relations between agents and concepts at particular times.

$$store(p,c,t) \supset agent(p) \wedge concept(c) \wedge temporal\text{-}entity(t)$$
$$retrieve(p,c,t) \supset agent(p) \wedge concept(c) \wedge temporal\text{-}entity(t)$$

The only way for a concept to get into an agent's memory is for it to be stored.

$$inm(c,m,t) \wedge memory(m,p) \supset (\exists t_1)[t_1 < t \wedge store(p,c,t_1)]$$

Note that this rules out pre-existing Platonic ideals in the memory, contra *Meno*. Moreover, thinking about something is a prerequisite for having it in memory.

4.2 Accessibility

Concepts in memory have an "accessibility" which is an element in a partial ordering.

$$memory(m,p) \wedge inm(c,m,t) \supset (\exists a)a = accessibility(c,m,t)$$

accessibility is a function mapping a concept, an agent's memory, and a time into an element of the partial ordering.

$$a = accessibility(c,m,t) \supset concept(c) \wedge temporal\text{-}entity(t)$$
$$a = accessibility(c,m,t) \supset (\exists p)[agent(p) \wedge memory(m,p)]$$

Accessibilities are partially ordered.

$$(\forall s)[(\forall a)[member(a,s) \equiv (\exists c,m,t)a = accessibility(c,m,t)] \supset partially\text{-}ordered(s)]$$

The predicates *member* and *partially-ordered* come from set theory. We will use the symbol $<$ to indicate the partial ordering relation in this set.

There is no assumption that accessibility is comparable across agents.

For any given agent, there is an accessibility value below which concepts are not retrieved from memory.

$$memory(m,p) \supset (\exists a_0)(\forall a_1,c,t)[a_1 = accessibility(c,m,t) \wedge a_1 < a_0 \supset \neg retrieve(p,c,t)]$$

For convenience we can call this the "memory threshold", or *mthreshold*.

$$a_0 = mthreshold(p) \equiv (\exists m)[memory(m,p) \wedge (\forall a_1,c,t)[a_1 = accessibility(c,m,t) \wedge a_1 < a_0 \supset \neg retrieve(p,c,t)]]$$

The theory is silent on how long a particular concept retains a particular accessibility value, but we do explicate some features of the causal structure underlying changes in accessibility.

4.3 Associations and Causing to Remember

One concept can remind an agent of another concept. This occurs when the first concept being in focus causes the second to be remembered.

$$remind(c_1,p,c_2,t) \equiv (\exists f,m)[focus(f,p) \wedge memory(m,p) \wedge cause(inm(c_1,f,t), retrieve(p,c_2,t),t)]$$

The application of the predicate *cause* to formulas is a shorthand for a more complex representation involving reifying the *inm* and *retrieve* relations. The notion of *cause* is explicated in a theory of causality (Hobbs, 2001).

One concept can be associated with another for a given agent.

$$associated(c_1,c_2,p) \supset concept(c_1) \wedge concept(c_2) \wedge agent(p)$$

The specific kinds of association would be partially explicated in a theory of the structure of information. For example, inferentially related concepts are associated for agents that know the inferential relations. In this treatment of memory we will not use any deeper analysis of association; rather we will concern ourselves with the causal consequences of concepts' being associated. We will assume associations are dependent on agents but not on times, although times could easily be incorporated.

$$associated(c_1,c_2,p) \supset (\exists f)[focus(f,p) \wedge (\forall e,t_1,t_2)[change'(e,\neg inm(c_1,f,t_1), inm(c_2,f,t_2),t_2) \supset (\exists a_1,a_2,t_3)cause(e, change(a_1 = accessibility(c_2,p,t_2), a_2 = accessibility(c_2,p,t_3),t_3)] \wedge a_1 < a_2]]$$

That is, if concept c_1 is associated with concept c_2 for agent p, then if e is a change from c_1 not being in p's focus of attention f to c_1 being in f, then e causes an increase in the accessibility of c_2 for p. This accessibility increase may or may not be enough for p to retrieve c_2.

It needs to be part of a theory of thinking or envisionment that agents can cause themselves to have a concept in their focus of attention. Then because of associations among concepts, agents have a causal structure they can manipulate to bring about retrievals from memory. This gives rise to strategies for remembering that involve calling to mind related concepts. For example, we might try to remember someone's name by running through the letters of the alphabet and hoping that the first letter of the name will cause the name to be retrieved.

A theory of goals would have to include an explication of a partial ordering of importance. A concept is more or less important to an agent at a particular time, depending on the relation of the concept to the agent's goals.

$$i = importance(c, p, t)$$
$$\supset\ concept(c) \wedge agent(p)$$
$$\wedge temporal\text{-}entity(t)$$

There is an at least defeasible monotonic relation between importance of a concept and its accessibility.

$$i_1 = importance(c_1, p, t)$$
$$\wedge i_2 = importance(c_2, p, t)$$
$$\wedge a_1 = accessability(c_1, p, t)$$
$$\wedge a_2 = accessability(c_2, p, t) \wedge i_1 < i_2$$
$$\supset\ \neg[a_2 < a_1]$$

This relation is a key part of an explanation for why John's forgetting their anniversary might cause Mary to get angry. If it had been important to him, he would have remembered.

4.4 Ability, Trying, Succeeding, and Failing

A theory of goals and planning would also have to have an explication of the notions of ability, trying, succeeding, and failing. We do not axiomatize them here, but we do sketch what the axiomatization would look like. All of these concepts concern an agent's manipulations of the causal structure of the world to achieve goals.

For an agent to *try* to achieve an eventuality is for the agent to perform actions that tend to cause the eventuality, where that eventuality is a goal of the agent.

$$try(p, e)$$
$$\equiv (\exists s)(\forall e_1)[member(e_1, s)$$
$$\supset tcause(e_1, e) \wedge do(p, e_1)]$$
$$\wedge goal(e, p) \wedge goal(cause(s, e), p)$$

The predicate *tcause*, for "tends to cause", would have to be explicated in a theory of causality. The predicate *do* would be explicated in a theory of events and actions. The last conjunct says that it is one of p's goals that the actions in s cause e. This is necessary because one can imagine situations in which p has a goal e and does actions e_1 that tend to cause e, but does not do these actions with the intention of bringing about e, but rather for some other reason.

To *succeed* in an attempt is to bring about that goal in actuality.

$$succeed(p, e) \equiv try(p, e) \wedge occurs(e)$$

To *fail* in an attempt is to attempt but not succeed.

$$fail(p, e) \equiv try(p, e) \wedge \neg occurs(e)$$

The predicate *occurs* would come from a theory of events and actions. Note that we are assuming that eventualities are individuated finely enough, e.g., by time of occurrence, that if someone is trying to climb over a wall and fails the first five times and then succeeds, this counts as six different eventualities, the first five of which don't occur.

Ability is much more difficult to characterize. We first begin with the notion of possibility. An eventuality is *possible* with respect to a set of constraints if the constraints do not cause or entail the eventuality's not obtaining.

Planning is a matter of exploiting the causal structure of the world in order to achieve goals. These plans will typically require, in addition to actions on the part of the agent, that certain conditions be true in the world that are beyond the control of the agent. For example, we can drive a car to work if the streets aren't flooded. An agent is able to achieve some effect if and only if whenever the agent has the goal of achieving that effect and the required world conditions beyond the agent's control are right, it is possible for the agent to achieve the effect. In other words, an agent is able to do something if that something is possible with respect to a set of constraints that includes the agent's desire to do it and the right world conditions being true.

Because people can cause concepts to be in their focus of attention and this may cause them to remember other concepts, people have an ability to remember things. It is also possible for people to *try* to remember things, and thus to succeed or fail to remember things.

If the accessibility of a concept is above the threshold, it is possible for the agent to retrieve it, relative to some set of constraints s.

$$mind(m, p)$$
$$\wedge mthreshold(p) < accessibility(c, m, t)$$
$$\supset (\exists s) possible(retrieve(p, c, t), s)$$

4.5 The Meanings of "Remember" and "Forget"

The English word "remember" can refer to a range of notions within this complex. At the simplest level, it can mean that the agent has the concept in memory and that it is accessible, but not necessarily in focus. In this sense, you remembered twenty minutes before encountering this sentence that Columbus discovered America in 1492.

$$memory(m, p) \wedge inm(c, m, t)$$
$$\wedge mthreshold(p) < accessibility(c, m, t)$$
$$\supset remember(p, c, t)$$

Even though the fact was not in focus, it was accessible in memory.

A somewhat stronger notion of remembering is when there has actually been a retrieval from memory. For a concept to be retrieved is for it to be remembered.

$$retrieve(p, c, t) \supset remember(p, c, t)$$

This rule was deliberately glossed in the passive. There is no notion of p's agency in the retrieval of a concept as we have defined *retrieve*. Thus, this notion of remembering covers cases where a fact simply pops into an agent's head, with no prior effort or intention.

A stronger sense of "remember" is one in which the agent plays a causal role in the remembering.

$$cause(p, retrieve(p, c, t), t) \supset remember(p, c, t)$$

This happens when we are told to remember who the president of the United States is, and somehow immediately we do. This sense of "remember" is what is conveyed in imperatives like

Remember that bears are unpredictable.

(Often such sentences are used to invite the hearer to draw an inference rather than retrieve something from memory, but in these cases it is an implicature that it was already in memory.)

The above rule is silent about whether the causality is immediate or there are intermediate actions on p's part designed to jog his memory. A stronger notion of remembering involves the latter. There is a distinct attempt to retrieve something from memory, and it succeeds. Since *succeed* as defined above entails trying, we can state this simply as follows:

$$succeed(p, retrieve(p, c, t)) \supset remember(p, c, t)$$

Again, this notation is a shorthand for a more complex one involving reification.

There are at least two levels of forgetting. In the simplest, the accessibility of a concept in memory has fallen below the memory threshold. To forget a concept is for the accessibility of the concept to change from being above the memory threshold to being below it.

$$\begin{aligned} &forget(p, c, t) \\ &\equiv (\exists\, m, t_1, a_1, a) \\ &\quad [a < mthreshold(p) < a_1 \\ &\qquad \wedge change(a_1 = accessibility(p, c, t_1), \\ &\qquad\qquad a = accessibility(p, c, t), t)] \end{aligned}$$

It is a theorem that if an agent p forgets something at a particular time, p does not remember it at that time, under any notion of remembering.

$$forget(p, c, t) \supset \neg remember(p, c, t)$$

One might argue that another sense of "forget" occurs when something is not remembered at the appropriate time, even though it was accessible. For example, someone dashes into the surf, is pulled out to sea, is rescued, and says, "I forgot about the undertow." One could say the concept was accessible; it just wasn't accessed. But it is probably cleaner to say that its accessibility changed, since there will be many factors that induce changes in accessibility, and to leave "forget" defined as above.

This section illustrates the relation between core theories and the lexicon. The core theory of some domain is constructed in a careful, coherent way, and the predicates explicated in the core theory can then be used to characterize the various uses of the relevant lexical items.

4.6 Remembering To Do

Our plans for achieving goals spread across time. For example, the goal of eating dinner tonight might involve stopping at the grocery store on the way home. The timely performance of an action requires us to be consciously aware of the need to perform the action at the time of its performance. Since things cannot be retained continuously in the focus of attention, it is necessary to remember to do actions before doing them. Thus, as a precondition for doing an action, remembering to do it must also be a part of the plan of which the action is a part.

$$agent\text{-}of(p, e) \wedge at\text{-}time(e, t) \wedge focus(f, p) \supset enable(inm(e, f, t), e, t)$$

That is, if p is the agent in an event e that takes place at time t, then e is enabled by being in p's focus of attention at time t. Thus, remembering to do something can become part of a plan, and hence an intention. As with all actions, a person can succeed or fail at remembering to do something. The predicate *agent-of* would come from a theory of the structure of events; the predicate *at-time* comes from that part of the theory of time that relates events to times (Hobbs, 2002); the predicate *enable* comes from the theory of causality.

It may be useful to have an explicit predicate for the important notion of remembering to do an action. An agent remembers to do something if that action was part of a plan of the agent's, and hence a goal, and if this action's being remembered at some time

prior to (or not after) the time of the action causes the action to be in focus at the designated time of the action.

$$(\forall p, f)[focus(f, p) \\ \supset (\forall e, t)[remember\text{-}to(p, e, t) \\ \equiv (\exists t_1)[goal(do(p, e, t), p) \\ \wedge cause(retrieve(p, e, t_1), inm(e, f, t), t_1)]]]$$

This definition does not presume that the action was completed. Factors other than forgetting could have prevented it.

4.7 Repressing

At least since the time of Freud, memories can be repressed. The passive predicate *repressed* requires less in the way of ontology than the active action of "repressing", so we will consider that first.

If a concept c is repressed at time t for an agent p, then c is in p's memory but its accessibility is less than p's memory threshold.

$$repressed(c, p, t) \wedge memory(m, p) \\ \supset accessibility(c, m, t) < mthreshold(p)$$

If a concept is repressed, it is unpleasant to the agent.

$$repressed(c, p, t) \supset unpleasant(c, p)$$

The predicate *unpleasant* would be explicated in the theory of emotions and the theory of goals.

Moreover, the unpleasantness of the concept plays a causal role in the concept's being repressed.

$$repressed'(e, c, p, t) \\ \supset partially\text{-}cause(unpleasant(c, p), e, t)$$

That is, if e is the condition of c's being repressed in p at time t, then c's unpleasantness to p partially causes this repression. The predicate *partially-cause* is explicated in the theory of causality; essentially it means that the first argument is a subaggregate of an aggregate of conditions and events that together bring about the second argument.

The conjunction of the consequents of these three rules may constitute a sufficient condition for repression.

It is problematic to say that an agent represses a memory. We may want to say in a theory of envisionment, or thinking, or consciousness, that agents are aware of what they are doing. But to store something in memory in a way that it can't be accessed is as contradictory as being told not to think of an elephant. There are two ways around this problem. The first is to say that there are some actions that an agent may do without being conscious of them. The second, the Freudian approach, is to say that agents have within them subagents that can perform actions the superagent is not aware of. These two approaches are probably equivalent.

5 Conclusion

In this paper we have demonstrated a new methodology for constructing formal theories in commonsense konwledge domains. We have shown how a close examination of a very general task—strategic planning—leads to a catalog of the concepts and facts that must be encoded for general commonsense reasoning. These can be sorted into a manageable number of coherent domains, one of which is memory. We can then mine textual corpora for further concepts and facts that must be characterized. A formal theory of these concepts is then constructed, but in a way that is constrained by the need to capture the full range of relevant facts.

Memory is a particularly interesting domain to illustrate this with, because although human memory has been extensively studied in psychology, there have been few attempts at formal axiomatizations. (Davis (1994) axiomatized some features of memory.) Moreover, although it has rarely been pointed out in the planning literature, remembering is a crucial element in plans aimed at achieving goals over long periods of time.

This work is part of a larger enterprise directed toward the axiomatization of the most fundamental areas of knowledge required in planning and language, especially commonsense psychology.

Acknowledgements

Abe Kazemzadeh, Milena Petrova, and Anish Nair participated in the language analysis portion of this research. This paper was developed in part with funds from the U.S. Army Research Institute for the Behavioral and Social Sciences under ARO contract number DAAD 19-99-D-0046, and in part with funds from the Advanced Research and Development Agency (ARDA). Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Department of the Army or of ARDA.

References

[1] Baron-Cohen, S. (2000) Theory of mind and autism: a fifteen year review. In S. Baron-Cohen,

H. Tager-Flusberg, and D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience*, second edition. Oxford, UK: Oxford University Press.

[2] Cohen, P. and Levesque, H. (1990) Intention is Choice with Commitment. *Artificial Intelligence* 42, 213-261.

[3] Corcoran, R (2001) Theory of Mind in Schizophrenia. In: D. Penn and P. Corrigan (Eds.) *Social Cognition in Schizophrenia*. APA.

[4] Davis, E. (1994) Knowledge preconditions for plans. *Journal of Logic and Computation*, 4(5), 253-305.

[5] Davis, E. (1998) The Naive Physics Perplex, *AI Magazine*, Winter 1998.

[6] Dyer, J., Shatz, M., and Wellman, H. (2000) Young children's storybooks as a source of mental state information. *Cognitive Development* 15, 17-37.

[7] Goldman, A. (2000) Folk Psychology and Mental Concepts. *Protosociology* 14, 4-25.

[8] Gopnik, A. and Meltzoff, A. (1997). *Words, thoughts, and theories*. Cambridge, Mass.: Bradford, MIT Press.

[9] Gordon, A. (2001a) Strategies in Analogous Planning Cases. In J. Moore and K. Stenning (eds.) *Proceedings*, 23rd Annual Conference of the Cognitive Science Society, Hillsdale, NJ: Lawrence Erlbaum Associates.

[10] Gordon, A. (2001b) The Representational Requirements of Strategic Planning. Fifth symposium on Logical Formalizations of Commonsense Reasoning. (http://www.cs.nyu.edu/faculty/davise/commonsense01/).

[11] Gordon, A. (2002) The Theory of Mind in Strategy Representations. *Proceedings*, Twenty-fourth Annual Meeting of the Cognitive Science Society (CogSci-2002), George Mason University, Aug 7-10. Mahwah, NJ: Lawrence Erlbaum Associates.

[12] Happe, F., Brownell, H., and Winner, E. (1998) The getting of wisdom: Theory of mind in old age. *Developmental Psychology*, 34 (2), 358-362.

[13] Hobbs, J. (1985) Ontological Promiscuity. *Proceedings*, 23rd Annual meeting of the Association for Computational Linguistics, pp. 61-69. Chicago, IL. July 1985.

[14] Hobbs, J. (2001) Causality. *Proceedings*, Commonsense 2001, Fifth Symposium on Logical Formalizations of Commonsense Reasoning, pp. 145-155. New York University, New York, NY. May 2001.

[15] Hobbs, J. (2002) Toward an Ontology of Time for the Semantic Web. *Proceedings*, Workshop on Annotation Standards for Temporal Information in Natural Language, Third International Conference on Language Resources and Evaluation, Las Palmas, Spain. May 2002.

[16] Hobbs, J., Croft, W., Davies, T., Edwards, D. and Laws, K., 1987. "Commonsense Metaphysics and Lexical Semantics", *Computational Linguistics*, vol. 13, nos. 3-4, July-December 1987, pp. 241-250.

[17] Levin, B. (1993) *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

[18] Nichols, S. and Stich, S. (forthcoming) How to Read Your Own Mind: A Cognitive Theory of Self-Consciousness. In Q. Smith and A. Jokic (Eds.) *Consciousness: New Philosophical Essays*, Oxford University Press.

[19] Ortiz, C. (1999) Introspective and elaborative processes in rational agents. *Annals of Mathematics and Artificial Intelligence* 25, 1-34.

[20] Silberztein, M. (1999a). Text Indexing with INTEX. *Computers and the Humanities* 33(3), Kluwer Academic Publishers.

[21] Silberztein, M. (1999b). INTEX: a Finite State Transducer toolbox. *Theoretical Computer Science* 231(1), Elsevier Science.

[22] Wellman, H.M., and Lagattuta, K. H. (2000). Developing understandings of mind. In S. Baron-Cohen, H. Tager-Flusberg, and D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience*, second edition. Oxford, UK: Oxford University Press.