Cognitive Science 32 (2008) 1304–1322

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1080/03640210802455175

Ability, Breadth, and Parsimony in Computational Models of Higher-Order Cognition

Nicholas L. Cassimatis[a], Paul Bello[b], Pat Langley[c]

[a]*Department of Cognitive Science, Rensselaer Polytechnic Institute*
[b]*Office of Naval Research*
[c]*School of Computing and Informatics, Arizona State University*

Received 20 July 2007; received in revised form 2 September 2008; accepted 5 September 2008

Abstract

Computational models will play an important role in our understanding of human higher-order cognition. How can a model's contribution to this goal be evaluated? This article argues that three important aspects of a model of higher-order cognition to evaluate are (a) its ability to reason, solve problems, converse, and learn as well as people do; (b) the breadth of situations in which it can do so; and (c) the parsimony of the mechanisms it posits. This article argues that fits of models to quantitative experimental data, although valuable for other reasons, do not address these criteria. Further, using analogies with other sciences, the history of cognitive science, and examples from modern-day research programs, this article identifies five activities that have been demonstrated to play an important role in our understanding of human higher-order cognition. These include modeling within a cognitive architecture, conducting artificial intelligence research, measuring and expanding a model's ability, finding mappings between the structure of different domains, and attempting to explain multiple phenomena within a single model.

Keywords: Higher-order cognition; Human-level intelligence; Cognitive models

1. Understanding higher-order cognition

Computational modeling is a particularly important part of understanding higher-order cognition. One reason for this is that precise models can help clarify or obviate often troublesome theoretical constructs, such as "representation" and "concept." A second is that the characteristics of human intelligence appear to be so different from other topics of scientific research as to call into question whether a mechanistic account of human intelligence is possible. Being instantiated in a computational model would resolve doubts about whether a theory

Correspondence should be sent to Nicholas L. Cassimatis, Rensselaer Polytechnic Institute, 108 Carnegie, Troy, NY 12180. E-mail: cassin@rpi.edu

was implicitly presupposing an intelligent "homunculus" and would make the possibility of intelligence resulting from natural phenomena more plausible.

In this article, "higher-order cognition" refers to inference, problem solving, and language use that goes beyond immediate sensations and memories. Such cognition is often studied in subfields such as reasoning, problem solving, analogy, syntax, semantics, and pragmatics. We are particularly concerned with a challenging puzzle: How do mechanisms that do not exhibit higher-order cognition (such as retrieval from long-term memory and access to short-term memory buffers) combine to produce higher-order cognition? How do mechanisms that do not reason, use language, or have goals combine into a system that does? No known modeling approach—including production systems, dynamical systems, or neural networks—currently exhibits the full range of higher-order human cognition. Our aim is to eliminate the gap between current approaches to cognitive modeling and the abilities of human intelligence.

Two desired traits of cognitive models are *ability* and *fidelity*; that is, we attempt to identify mechanisms that have the power and flexibility of human intelligence. We also wish to confirm that these mechanisms are at least similar to those that underlie human cognition. How should we evaluate a cognitive model's progress toward these goals? Like Newell (1973), we argue for the need to move beyond models of isolated cognitive phenomena. We claim that, although fitting model behavior to human data can be an important activity, it is but one method for evaluating cognitive models. We then stress the importance of ability, breadth, and parsimony in model evaluation. Finally, we use ongoing research programs to illustrate how cognitive models can be motivated and evaluated using these criteria.

There are many discussions that touch on ability, breadth, or parsimony. For example, Anderson and Lebiere (2003) used Newell's (1990) criteria for cognitive models to compare the ACT-R and connectionist approaches to modeling. Many of these criteria relate primarily to ability, although the range of abilities is broad and parsimony is briefly mentioned. This article differs from previous discussions in relating ability to model fitting, giving examples of concrete and precise measures of these criteria and by motivating the discussion explicitly from the goal of understanding higher-order cognition.

2. The model fit bias

The observe–hypothesize–test view of the scientific method is often applied to evaluating cognitive models thus: First, one collects quantitative human data in an experimental setting, then one develops a cognitive model that reproduces this behavior and predicts unobserved behavior, after which one conducts further experiments to confirm these predictions. Although such work can play an important role in evaluating models, and we do not claim it is unnecessary or somehow undesirable, we argue that over-emphasizing model fits relative to other criteria does not address all the goals of cognitive science and can impede progress toward models that provide general accounts of higher-order cognition.

2.1. Modeling the fundamental data

People can solve problems in new situations, carry out complex chains of reasoning, interpret what they see and perceive, engage in extended conversations about a wide range of

topics, and learn complex structures that support these abilities. There are no existing cognitive models that can reason, solve problems, converse, or learn as well as people. Although there are models that predict reaction times or error rates in specific situations, they do so only on one or a few tasks and, thus, do not generally account for any of the aforementioned human abilities.

The fact that models with exemplary, near-perfect fits to quantitative data are possible to construct without making significant progress toward expanding the situations and domains over which cognitive models can deal with illustrates that the model fits alone are not sufficient for evaluating models of higher-order cognition. For example, consider two hypothetical sentence processing models. Model A fits eye movement and reaction time data but makes no inferences about the sentence meaning. Model B makes many correct inferences about sentences (measured, for example, by answering questions about them) at a level that far exceeds the state of the art in computational linguistics or cognitive modeling. Model B does not predict reaction times or eye movements. Model B would clearly be an important contribution because it advances our ability to provide computational accounts of the inferential power evident in human language use. However, if we primarily emphasize model fits, then Model B does not count as much of a contribution, and Model A must be favored. Our claim is not that Model B is better than Model A. Model A is very likely to account for processes that B does not. Rather, our claim is that over-emphasizing model fits can under-emphasize certain kinds of modeling efforts that make progress toward solving some very difficult computational questions about higher-order cognition.

This point can also be made in terms often used to motivate the importance of model fits. The role of models (and scientific theories generally) is to explain and predict observations. The number and range of observations a model explains and predicts is often used to test how accurately the model characterizes reality. These observations can include data carefully collected in a laboratory as well as readily observable facts, such as the daily rising of the sun and the solubility of salt in water. One of the most uniquely observable facts about humans is their ability to reason, solve problems, converse, and learn. Thus, when evaluating the cognitive plausibility of a model, among the observations that should be considered are those that pertain to ability. In short, ability is part of the data on human cognition and the extent to which a model has this ability is an important part of evaluating its plausibility as a model of human cognition.

2.2. Enabling ability before fitting models

In order fit a model to data about performance in a task, one must first have a model that can perform the task. With higher-order cognition, however, it is often the case that no computational methods are known that can exhibit the ability to perform many tasks. Discovering computational methods with this ability is, therefore, important and has many characteristics that distinguish it from model fitting work.

In cognitive modeling research, there is often more than one mechanism that produces a particular kind of behavior. For example, there are both neural network (McClelland & Patterson, 2002) and rule-based (Pinker & Ullman, 2002) accounts of past-tense morphological processing, and there are both mental model (Johnson-Laird, 1983) and mental logic (Braine

& O'Brien, 1998; Rips, 1994) accounts of behavior in many reasoning tasks. In these cases, research often attempts to determine which mechanisms are actually involved in these tasks by observing behavior corresponding to the differing predictions each model makes.

In much of higher-order cognition, however, there are no known mechanisms that exhibit the kind of behavior we seek to understand. We know of no computational processes that can learn, use language, reason, or solve problems in as wide and complex a set of situations as people. In such cases, it is impossible to fit models because there are no candidate models to fit.

Because finding mechanisms that enable such models is such a difficult problem, it cannot merely be treated as a preliminary stage of model fitting research. Finding computational methods with human-level cognitive abilities will likely involve several steps of progress along the way, each of which will need to be evaluated in some manner. To the extent that quantitative model fits are not well-suited to measuring ability, additional criteria will be required. Because finding computational methods with human-level ability is a large task, it will require many projects evaluated primarily according to those criteria. The remainder of this article proposes some such criteria and illustrates their use.

3. Evaluating the ability, breadth, and parsimony of a cognitive model

Although we have argued that quantitative model fitting, as typically defined, is not alone sufficient for evaluating accounts of higher-order cognition, the field still requires guidelines for measuring progress. In this section, we propose some additional criteria for evaluating cognitive models. For each criterion, we discuss its analogues in other fields and the ways in which it has played an important role in the history of cognitive science. Unfortunately, just as there is no all-encompassing, precise definition or procedure for model fitting, we must content ourselves for now with general descriptions of these criteria and specific methods of using them in certain situations. In subsequent sections, we will provide examples of such methods.

3.1. Ability

We have argued that one of the most interesting and important facts about cognition is that people can solve novel problems, make nontrivial inferences, converse in many domains, and acquire complex knowledge structures, and that they can do so at a level beyond the reach of currently known computational methods. Thus, it is important to ask how much a model advances the ability of computational methods to explain and produce higher-order cognitive phenomena.

Many important contributions to cognitive science have involved ability. Chomsky's (1959) arguments against associationist models of language relied on the claim that they did not have the ability to model the hierarchical and recursive nature of human language syntax. His argument focused on linguistic competence rather than the details of performance. Similarly, Newell, Shaw, and Simon's (1958b) Logic Theorist was an advance because it demonstrated that a computational mechanism, search through a problem space, could prove the same logic

theorems that humans could. The degree of match to human data, quantitative or otherwise, was far less important than the demonstration that a certain class of mechanism could explain some kinds of human problem solving. Finally, back propagation in neural networks (Rumelhart, Hinton, & Williams, 1986; Werbos, 1974) was viewed as a significant achievement not because it fit human data on learning patterns like the exclusive-or function, but because, counter to impressions generated by Minsky and Papert (1969), it demonstrated their ability to learn some of these functions.

The aforementioned efforts each caused a genuine revolution within cognitive science because they helped advance the ability of formal or computational methods to explain human behavior. Although none of these efforts endeavored to fit detailed data initially, each resulted in a framework that enabled more precise accounts of specific behavior and ultimately led to empirically rich bodies of research. These efforts suggest that, when faced with a choice between increasing a modeling framework's cognitive abilities and improving fits against already-modeled phenomena, there are clear benefits if some researchers choose to work on ability without immediate concern for quantitative model fits.

As a cautionary tale regarding the dangers of narrowly characterizing ability, we consider the history of computer chess playing. Researchers as early as Turing (1946) believed that chess playing would be a good demonstration of the power of computational accounts of human intelligence. Early efforts at successfully programming computers to play chess aimed to give them the ability to win games and were not concerned with fitting detailed human performance data. Simon's boast (recounted in Crevier, 1993) about progress in chess was that "a computer would be chess champion of the world within ten years" (p. 109) and not that, say, high-quality eye-movement predictions would be soon possible. The reason for this was that a major question at the time was how any mechanical process could produce intelligent behavior, such as chess playing. The first approach to produce good results was the heuristic search (Newell, Shaw, & Simon, 1958a). This constituted progress because it showed that computational processes could lead to some forms of intelligent behavior (and thus an advance according to the ability criterion) and because, in fact, people actually carry out some lookahead search when playing chess.

Subsequent computational investigations into chess focused almost exclusively on improving the chess rating of computer chess programs (one measure of ability). The general approach was to exploit programming improvements and growing computational power to increase the numbers of possible moves a computer could explore. The result was that computer chess programs matched and, in many cases, exceeded human ability but did so by performing inhuman amounts of lookahead.

It is common to conclude from the history of chess research that ability is a flawed criterion for cognitive models. There are, however, three problems with this conclusion. First, even when most aspects of a model are implausible, some of them may be similar in some way to actual cognitive mechanisms. For example, although humans do not perform "brute-force" search, there is significant evidence (Dingeman, 1978; e.g., from verbal protocols) that they do perform some search. Further, much work into human chess playing has investigated specific heuristics people use to perform search.

This history is consistent with the sequence, discussed in the last section, from modeling ability to modeling specific mechanistic details. Although chess programs have from the

beginning generally used implausible amounts of search, they did reflect the fact that humans used some search and have established the framework within which many aspects of human chess playing have been understood.

Second, models can help explain human intelligence, even when they do not use *any* mechanisms that it is implausible to believe humans have. Some cognitive modeling research efforts do not specifically address mechanisms at all. For example, creators of many Bayesian cognitive models use stochastic simulation mechanisms, such as Gibbs Sampling (Geman & Geman, 1984), which require implausibly large numbers of simulations of an event, because they are not attempting to model the mechanisms of human cognition, but instead to identify the constraints or knowledge those mechanisms are using. In human chess playing research, for example, it is common to investigate how humans formulate aspects of a game (Dingeman, 1978). One could imagine using search mechanisms quite different from human search to show that certain representations of chess playing yield to human-like playing patterns and use this as evidence that humans use those representations. These examples illustrate how models whose mechanisms are not faithful to human cognition can nevertheless help explain aspects of it.

A third problem with using chess research as an argument against ability concerns the sense of ability being considered. Human beings are not only able to play chess, but they are also able to learn the rules of chess, speak about their chess playing strategies, adapt to changes in rules, and play many other games. The algorithms used in computer chess today do not have any of these capabilities and, thus, when ability is construed to include them, they are not only failures under the fidelity criterion, but also according to the ability criterion. Thus, as we discuss in the next section, when breadth and not just level of ability is considered, existing chess-playing systems are not a good counter-example to the importance of ability as a criterion because they are lacking in (some important aspects of) ability.

3.2. Breadth and parsimony through unifications

As just mentioned, one of the characteristics of human intelligence we wish to explain is how it is able to succeed in such a broad array of situations. We further wish to do so with a single theory that posits as few mechanisms as possible. There are several advantages to explaining a set of phenomena with one theory rather than many. First, so long as the single account is internally consistent, one can be confident that its explanation of those phenomena is consistent, whereas explaining them with many theories may rely on inconsistent assumptions. Thus, when Newton provided one theory that explained the phenomena covered by Galileo's law of uniform acceleration and Kepler's laws of planetary motion, he demonstrated that those accounts were consistent with one another.

Theory unifications are important scientific contributions because they serve Occam's razor by increasing the ratio of phenomena explained to theoretical principles posited. Several important achievements in the history of science have involved unifications. Newton's three laws and one gravitational force subsumed Kepler's laws of planetary motion and Galileo's mechanics, providing a unified account for a variety of phenomena. Others immediately recognized it as an important achievement because of the unification itself rather than any new empirical results.

Unification is particularly important in cognitive science for several reasons. Newell (1990) listed several. Pertaining specifically to higher-order cognition is the fact that much of the progress of science has been to provide naturalistic explanations of phenomena previously explained by vital or goal-directed forces. For example, the theory of evolution and the economic theory of markets both show how globally purposeful behavior can emerge from local interactions. Biology has shown how much of what was once attributed to vital forces can be explained through physical and chemical interactions. To the extent that cognitive modelers are successful, cognition would be accounted for using principles that are as "un-goal" directed as, for example, gravitational or electromagnetic forces and thus unified with the rest of our understanding of the natural world. In the case of higher-order cognition, which superficially seems to be governed by principles so different from those regulating such physical forces, the unification would be especially dramatic.

Further, although human cognitive mechanisms are likely to be to some extent heterogonous, the generality of cognition implies that that there must be common elements (and corresponding theoretical unifications) in many forms of cognition. Because entities such as automobiles, parliaments, and interest rates did not exist when humans cognition evolved, the mechanisms used to reason about them must be the same as those used to reason about aspects of the world that humans did evolve to deal with; for example, physical and social events and relations. Thus, many of the same principles must govern cognition in all of these domains.

The history of cognitive science confirms the importance of unifications. Several achievements in cognitive science have been significant primarily because they unified previous results rather than because they predicted or explained new phenomena. Examples include Soar's (Laird, Newell, & Rosenbloom, 1987) account of many different weak methods in terms of problem-space search and impasses, the REM model (Shiffrin & Steyvers, 1997) of multiple memory phenomena in terms of the probability of storage errors, and Chomsky's (1981) unification of constraints in transformational grammar under a small set of principles and parameters.

Cognitive architectures (Newell, 1990) have been a particularly important contribution to breadth and parsimony in theories of higher-order cognition. Key abilities that support higher-order cognition include carrying out multi-step reasoning, solving novel problems, conversing about many topics, and learning complex structures. A cognitive architecture provides a theory of the memories, representational formalisms, and mental processes that are invariant in human cognition across domains and that enable adaptation to new domains. Thus, cognitive architectures are particularly well-suited for implementing theories about the generality and adaptability of human cognition.

Cognitive architectures are further important because, like other kinds of theory unification, they can increase the impact of individual pieces of work. The broader a theory's coverage of a phenomena, the greater the number of potential ramifications for individual efforts that elaborate on it. For instance, Hamilton's (1833) elaboration of Newtonian mechanics had a wider impact than if it had been merely an elaboration of Galileo or Kepler's theory. In cognitive science, when the mechanisms of an architecture account for multiple phenomena, revisions in those mechanisms can have an impact on theories of each of those phenomena. The more phenomena modeled within an architecture, the broader the potential explanatory value of revisions or extensions to it. For example, because ACT-R's production system and

theory of memory are used in models of analogy, sentence processing, and problem solving, revisions of those systems will alter our understanding of all of those processes and will also be constrained by what is already known about them.

These reflections on breadth and parsimony suggest that, although there are benefits to refining models of specific phenomena, there is also great value in creating computational frameworks that provide unified accounts for a wide array of cognitive activities.

To summarize, the history of cognitive science and other disciplines suggests there are many ways to evaluate the field's progress toward explanations of higher-order cognition. Hypothesis testing and model fits are appropriate in some cases for gauging how accurately a model approximates particular aspects of cognition, but they do not measure whether a model or architecture has the power to explain the breadth and complexity of the human mind. Quantitative model fits measure success on only one criterion: how faithful a model is to reality. Other important issues concern whether a theoretical framework has the basic ability to predict observed phenomena and the range of phenomena it can cover with a small number of principles. Criteria such as ability, breadth, and parsimony therefore have a crucial role to play in evaluating candidate theories of higher-order cognition.

4. A modern example of model comparison

We have shown how efforts to increase the ability, breadth, and parsimony of computational approaches to modeling human higher-order cognition have played seminal roles in the development of cognitive science. In this section, we provide examples of model comparison in the domain of language. Rather than selecting a "winner," our goal in making these comparisons is to demonstrate how these criteria can play an important role in modern cognitive science, how progress on them can be measured quantitatively, and how they can be used to evaluate cognitive models.

4.1. Evaluating models of language use

Language use is a quintessential example of higher-order cognition. It has been studied in many subfields of cognitive science (most notably in artificial intelligence, cognitive psychology and, of course, linguistics) with very different methodologies. We will use three specific research efforts based on different methodologies to illustrate how the criteria of ability, breadth, and parsimony can be used to compare and evaluate cognitive models. We will first briefly describe each model and discuss its contribution along these dimensions. We will then discuss the trade-offs among the approaches and use them to illustrate the relationship between artificial intelligence and cognitive science.

4.1.1. The limited capacity model

We first consider Lewis and Vasishth's (2005) sentence processing model, which we refer to as the "limited capacity model." It attempts to explain how humans can quickly infer the complex grammatical structure of sentences despite verbal memory, which does not include information about the order of words and a severely limited focus of attention. Because order

is an essential part of a sentence's syntactic structure, it is surprising that there is little evidence for explicit encoding of word order in short-term memory.

The limited capacity model accounts for the fact that human sentence processing conforms to order constraints in syntax (e.g., that subjects normally precede verbs in English) without any explicit representation of order. These can be illustrated using the sentence, "Mary knew John appreciated her gift." When "appreciated" is focused on, the model attempts to retrieve a noun phrase subject. Because there is no explicit order information in memory, "Mary" is not immediately ruled out for retrieval, and the subject can predict that there will be subsequent noun phrases (i.e., "the gift) in the sentence. However, three factors favor "John" being retrieved from among the other noun phrases in the sentence. First, "John" is remembered when "appreciated" and "her gift" are perceived. Thus, the fact that memories are always in the past implicitly leads to order constraints in processing. Second, Mary is already encoded as being the subject by "knew" and, thus, cannot also be the subject of "appreciated" (because, in part, it was the only previous noun phrase when "knew" was perceived). Finally, memory decays through time so that, other things being equal, more recent items (i.e., "John") are more likely to be retrieved than less recent items ("Mary"). Factors such as these explain how sentence processing conforms to syntactic ordering constraints without explicitly representing order in memory.

Lewis and Vasishth (2005) evaluated their model with fits to reading times. Although these fits are impressive and many aspects of the model are original, their dependent measures and methods of fitting reading data are standard. We now consider how the model contributes to ability, breadth, and parsimony.

By explaining sentence processing effects in terms of memory and attention mechanisms with no explicit syntactic representations, the limited capacity model extends the verbal memory and attention literature's breadth of impact. It also contributes to parsimony of cognitive theory by reducing the need to posit separate mechanisms for syntactic and other forms of processing. Also, because the model is created within a cognitive architecture, ACT-R, it expands the explanatory scope of that architecture and, thus, brings more unity to the field insofar as ACT-R can explain a wide variety of phenomena. These contributions would all have been significant with less accurate model fits or even just qualitative predictions.

Regarding ability, Lewis and Vasishth (2005) conceded that there are several kinds of syntactic constructions, such as "self-embeddings," that their model does not parse (e.g., "The rat the cat ate died."). This is consistent with the fact that people often find such sentences difficult to parse, but it does not reflect the fact that people nevertheless often can overcome this difficulty and parse these sentences. More generally, modern, wide-coverage, syntactic theories in formal linguistics involve formal structures and operations (such as type hierarchies with default inheritance, Pollard & Sag, 1994; and empty categories, Radford, 1997) that are not included in the limited capacity model. Without such mechanisms or structures, it remains to be seen whether the assumptions of the model are able to achieve broad grammatical coverage.

Although these remarks do not make use of any mathematical or statistical methods, they nevertheless illustrate how ability, breadth, and parsimony can be discussed with some precision. The list of grammatical constructions the limited capacity model explains (a measure of breadth) can be specifically enumerated, and the precise number of mechanisms the model

shares with other ACT–R models (a measure of parsimony) is easy to determine by inspection. This is an example of how the precision generated by computationally instantiating a theory of cognitive architecture makes work within that framework easier to discuss and evaluate.

4.1.2. The corpus-driven approach

The corpus-driven approaches into language that we discuss here would not typically be considered cognitive modeling, although we argue that, on the basis of our criteria, it makes a significant contribution to our understanding of language use. A central goal of this work is to avoid the difficulties of handcrafting a broad-coverage human language grammar by automatically learning grammatical rules from an annotated corpus. Handcrafting grammars has been difficult because of the large number of grammatical constructions that must be covered and the many exceptions to these constructions.

This work relies heavily on annotated corpora. The existence of corpora such as the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1994), which include the syntactic structure for thousands of sentences, has led to much activity (summarized in Lease, Charniak, Johnson, & McClosky, 2006) in artificial intelligence and computational linguistics. The Penn Treebank includes sentences from sources (such as *The Wall Street Journal*) paired with their grammatical structures. Given this corpus, researchers design algorithms that attempt to infer the probabilistic context-free grammar that generated the sentences. A probabilistic context-free grammar rule is a context-free rule associated with a conditional probability. For example, S $\rightarrow$ (.3) NP VP is interpreted as asserting that when a sentence (S) occurs, it is in 30% of cases generated by a noun phrase (NP) followed by a verb phrase (VP). By combining the rules that generated a parse of a sentence, one can compute the probability of that parse having been generated. Parsers for probabilistic context-free grammars output the most likely parse. Learning methods are often evaluated by the *precision* and *recall* of the grammars they produce. Precision is the proportion of phrases generated by the learned grammar that exist in the corpus, and recall is the proportion of phrases in the corpus that the learned grammar generates. Precision and recall rates in the mid-90% range have been reported.

These results have had a significant impact on the field. This can be illustrated with the "prepositional phrase attachment problem." For example, the sentence, "I saw the bird with the telescope," has multiple possible syntactic structures. In one, "with the telescope" is a prepositional phrase that modifies "the bird," implying that the bird has the telescope. In another, "with the telescope" modifies "see," implying that the telescope was used to see the bird. Resolving these "attachment ambiguities" seems to require reasoning about the world (e.g., that birds do not typically use telescopes) in addition to purely syntactic processing. That the corpus-driven approach resolves a surprisingly high proportion of (although by no means all) attachment ambiguities, together with other successes using corpora, has been counted as evidence that statistical or "distributional" information is much more potent than has been implied by those arguing (Chomsky, 1976) that the language children hear is not sufficient for them to learn grammatical structure and, thus, significant aspects of syntax must be innate.

This work contributes to the breadth and parsimony of cognitive theory in several ways. The corpus-driven approach can potentially increase the parsimony of theories of language development and use because it reduces (although not necessarily eliminates) the need to posit learning and sentence processing mechanisms beyond those involved in processing

statistical information. Although most learning algorithms operate over all sentences at once (as opposed to incrementally, as in human language learning), they do demonstrate the power of the statistical information latent in human language use. This work also contributes to breadth because of the wide range of sentence types and corpora upon which it has shown success. Sentence processing models in psycholinguistics and syntactic theories from formal syntax cannot claim to correctly identify large fractions of phrases in large-scale corpora. Finally, this work also advances the ability of computational models of language use because there were previously no known computational methods (in psycholinguistics or computational linguistics) for inducing such wide-coverage grammars.

Because precision and recall are measures of ability and because the number of sentences and corpora used reflects breadth, this work illustrates that there are contexts where breadth and ability can be precisely quantified and enable price model comparisons. A review (Lease et al., 2006) of some past research in this field lists several approaches that make clear claims about their superiority over previous methods based on the improvements in precision and recall they generate.

All these contributions have been made despite the fact that most of the algorithms used in this research would fail entirely on many of the eye-tracking or reading time tests often used to evaluate theories of sentence processing. More out of convenience than necessity, these algorithms parse sentences in a parallel, bottom-up manner (i.e., by considering all words simultaneously and computing the phrases they can form) rather than in the incremental order in which humans process them.

Two criticisms of this approach concern ability. First, most current corpora, including the Penn Treebank, are based on context-free, or even less powerful, grammars. Although the relative simplicity of these grammars makes them more amenable to statistical methods, there are many regularities (for example, gender agreement and head/argument ordering uniformities) in language that context-free grammars do not naturally capture. Second, there are many aspects of sentence processing (and language use in general) that cannot be studied in corpora as they are currently constituted. For example, consider the following two sentences: "The couple went for a walk. He held her hand." "He" and "she" clearly refer to the male and female members of the couple, respectively. But existing corpora only encode co-reference between words in the corpus. Thus, because the actual antecedents of "he" and "she" are not written or spoken and only inferred, statistical inferences made on purely what is written or said cannot determine these co-reference relationships. Further, ellipsis provides an example in which the corpus includes the antecedent, but the reference to it is not actually spoken. For example, in "Mary likes ice cream and so does John," it is clear that John likes ice cream. However, because John liking ice cream is not specifically mentioned in the corpus or its annotation, we cannot test whether a parser associates the appropriate action with "so does John." This makes ellipsis very difficult to study within a corpus. Of course, these are problems with the corpus-based approach as it is today. They do not preclude richer corpora from being developed to study such phenomena.

4.1.3. Mapping syntactic structure onto physical structure

The final research effort we study attempts to show how syntactic structure can be mapped onto the structure of physical reasoning problems so that a physical reasoning model can

thereby process sentences (Cassimatis, 2004; Murugesan & Cassimatis, 2006). There are several reasons to attempt to map syntactic structures to physical structures. The concepts involved in syntax (e.g., anaphora, bindings, gaps) superficially appear much different from the concepts used in the physical reasoning. Like a similar mapping found between social and physical cognition (Bello, Bignoli, & Cassimatis, 2007), finding a mapping between these two domains would make it more plausible that other surprising mappings between domains exist and that architectural approaches to cognitive modeling based on small numbers of principles and mechanisms can have wide explanatory breadth. Also, as we describe later, these mappings have a direct impact on views about the modularity, learnability, and innateness of language.

The mapping between language and physical structures is based on the fact that verbal utterances[1] are actions taken by people. Like physical actions, verbal utterances occur over temporal intervals, belong to categories, combine into larger actions, and are normally taken to achieve some goal. This mapping enables the most powerful grammatical theories to be reformulated using representations for physical actions and events. To formally confirm that such mappings are possible, Murugesan and Cassimatis (2006) used Head-Driven Phrase Structure Grammar (HPSG; Pollard & Sag, 1994). HPSG was used because its coverage is competitive with other major theories while being very amenable to computational implementation. The following examples illustrate how structures used by HPSG can be represented using physical structures.

4.1.3.1. Constituency and linear order as parthood and time: Almost all grammars contain some rules, such as $S \rightarrow NP + VP$. In physical terms, we represent this using part, temporal, and category relationships. For example, this rule says that an action of category noun phrase utterance followed immediately by an action of category verb phrase utterance can combine to form an action of category sentence utterance.

4.1.3.2. Feature unification and identity: HPSG and many other grammars use feature information, for example, to say that the gender of a verb and its subject should agree. Features of physical objects must also "agree." For example, a gun should be loaded with cartridges of the same caliber, and the octane of gas put into the car should be the same as the octane of gas that the car requires. Agreement in both the physical and verbal world can be thought of as an identity relation between features.

4.1.3.3. Type hierarchies: Many grammars rely (heavily in the case of HPSG) on hierarchies of categories. These also exist in the physical world (e.g., iron objects are metal objects), and which physical laws apply to an object depends on its category.

4.1.3.4. Co-reference: The fact that two phrases share the same reference (e.g., as with "Mary" and "herself" in "Mary likes herself") can be encoded as an identity relation. In this case, the reference (R1) of one phrase is said to be identical to the reference (R2) of another phrase: $R1 = R2$. Identity relations and the need to resolve identity ambiguity are an important aspect of physical reasoning, as when one must infer whether an object that has emerged from an occluder is the same object one saw go behind the occluder or merely a different object with the same appearance.

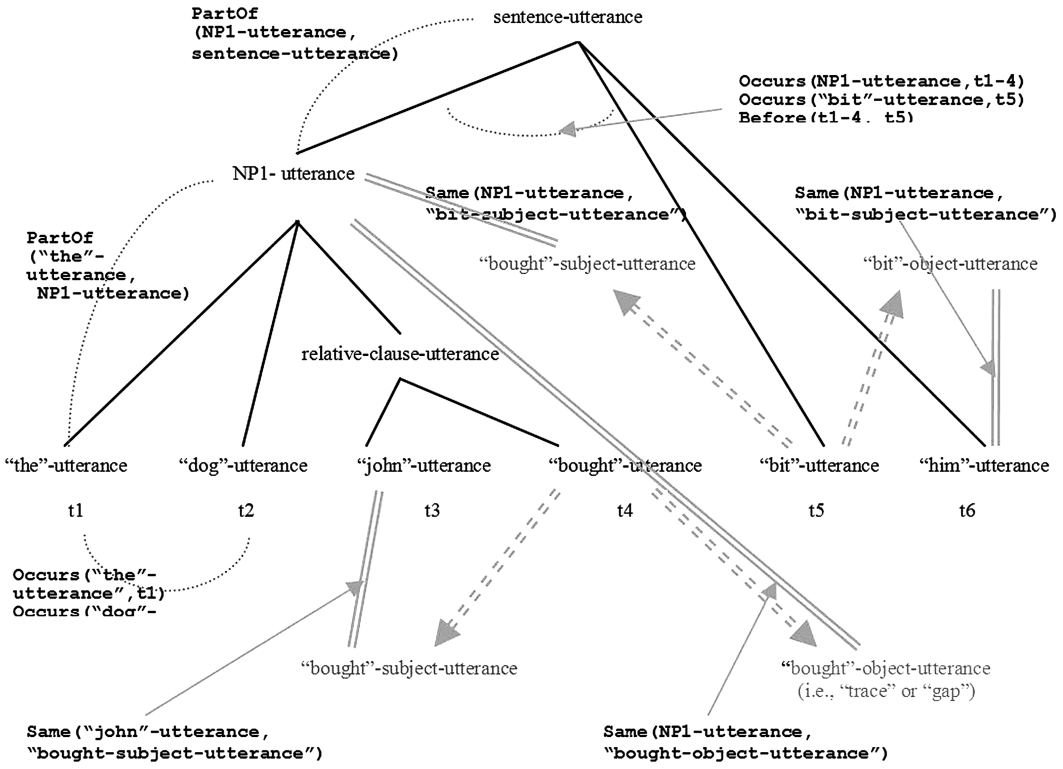

Fig. 1. The syntactic structure of the sentence represented using relations from physical reasoning. *Note:* NP = noun phrase.

Fig. 1 illustrates this mapping. It depicts a parse tree for a sentence where each element of the parse is encoded using physical relation.

Even seemingly obscure and language-specific constraints can be easily reformulated using this mapping between syntactic and physical structures. For example, Radford (1997) formulated "the c-command condition on binding" thus: A bound constituent must be c-commanded by an appropriate antecedent. He defined c-command by stating that a node X c-commands Y if the mother of X dominates Y, $X \neq Y$ and neither dominates the other. Because c-command is a constituency relation, it can be reformulated using the parthood relation thus:

> X **c-commands** Y if PartOf(X,Z) and there is no X' such that PartOf(X,X') and PartOf(X',Z) (i.e.,"Z is the mother of X"); PartOf(Y,Z) ("Z dominates Y"); Same(X,Y) is false ("X≠Y"); and PartOf(X,Y) and PartOf(Y,X) are both false ("neither dominates the other").

The mapping we have been describing has several consequences. First, because (as mentioned earlier) human syntax and the physical relations used in the mapping are so superficially different, the possibility of finding other mappings between dissimilar domains increases. Second, language is often neglected (sometimes explicitly; Newell, 1990) in cognitive architecture research and thought by many (e.g., Chomsky, 1976) to be a module "encapsulated" from the rest of cognition. The syntax mapping raises the possibility that human language processing

involves the same mechanisms studied in other domains. At a minimum, the mapping demonstrates that domain-general mechanisms can generate and process language whose structure is as rich as human syntax. The mapping therefore reduces the need to posit a special language faculty or module to explain human verbal behavior. Finally, many arguments (e.g., Laurence & Margolis, 2001; Lidz & Waxman, 2004) for the innateness of language are based on the presumption (Chomsky, 1980; called the "poverty of the stimulus") that children have too little verbal input to infer the grammatical structure of English on their own. However, if language processing shares many or all of the mechanisms of physical reasoning, then the amount of experience children can bring to bear on language development includes their physical interaction with the world and is therefore much larger, thus reducing the potency of poverty of the stimulus arguments.

This mapping primarily makes contributions to ability and unity. First, it enables a model of syntactic parsing to be built using the same mechanisms as an updated version of a model of physical reasoning, the mapping clearly demonstrates that physical reasoning mechanisms (with only the addition of a representation of the sound of words) are sufficient to explain syntactic parsing. Second, because computational methods in AI and cognitive modeling for processing syntax and those for processing other forms of information have often been so different and difficult to integrate, the mapping increases the ability of computational methods to integrate pragmatic, semantic, and syntactic information during parsing. However, the mapping and the model based on it do not yet have the ability to learn a grammar, even though it potentially has a role in explaining human language acquisition.

4.2. Trade-offs and model evaluation methods

Having illustrated how the criteria of ability, breadth, and parsimony can be used to evaluate the contribution of individual cognitive models, we now use these criteria to compare these models. Specifically, we will show that even though the evaluation methods their creators used lead to certain trade-offs and different research directions, their results can nevertheless inform each other's work. Many of the trade-offs we discuss are not logically entailed, but because there are only finite resources for any research project, they often do become practical necessities in the short term.

Not being bound by the necessity to fit quantitative data on sentence processing has enabled the mapping model to create more grammatical coverage and has allowed the corpus-driven model to provide insight into language learning over a very wide range of corpora. The mapping approach presupposed a temporal reasoning capacity, which is not as well-characterized psychometrically as the aspects of verbal memory and attention that the limited capacity model assumes. To have first collected the relevant psychometric data would have required a large effort that would have made the project infeasible. Likewise, although there is a considerable body of work on language acquisition, the measures, methods, subject pools, and often even the results vary so dramatically that it would have been impractical to attempt to regularize all the data and then relate it to the corpora used to test grammar learning algorithms. Of course, one consequence of accepting these trade-offs is that, unlike the limited capacity model, the mapping model and corpus-driven approach do not make detailed claims about the specific mechanisms involved in sentence processing.

The mapping model and the corpus-driven approach illustrate a trade-off between different kinds of abilities. Because the mapping model involves a more expressive grammatical formalism than the grammars used in most corpus-driven work, it can capture more of the deep, cross-linguistic regularities in grammar. Further, the mapping from grammatical to physical structure enables reasoning about nonlinguistic items to constrain sentence processing, whereas context-free grammars (which capture only category, order, and parthood relations) are ill-suited for such nonlinguistic reasoning (which involves many other kinds of relations). Statistical learning methods for more complex grammatical formalisms are much less-well understood and more difficult to implement. Further, a complex grammar would require better-trained human corpus annotators and much more of their time. In the short term, it would therefore be difficult (or at least prohibitively expensive) for the mapping approach to achieve the breadth, precision, and recall rates of the corpus-driven approach. There is a thus a trade-off between the corpus-driven approach's ability to learn and the mapping model's ability to explain deep linguistic regularities and the integration of syntactic and other forms of information in processing.

That choosing a certain methodology leads these efforts to certain trade-offs in the short term does not preclude each research program from benefiting considerably from the others. For example, ACT-R's memory retrieval system operates, in part, on the basis of statistical information from past experience. If the operation of this subsystem could be meaningfully related to the statistical methods used to induce probabilistic context-free grammars in the corpus-driven work, then the ACT-R model would gain significant language learning abilities, whereas the corpus learning approach would be constrained by and perhaps exploit insights from what is known about human learning and memory. The mapping model suggests directions for both research programs. For example, because the mapping relies heavily on temporal and identity relations, work incorporating methods for learning these relations could potentially increase the ability of the other two approaches to deal with more complex grammars. Such work would, in turn, add a learning ability now lacking in the mapping model. We believe that these potential synergies between research efforts would happen, not in spite of the fact that each research program does not subject itself to the others' evaluation regimes, but because ignoring certain constraints in the short term frees each approach to develop certain ideas in sufficient depth to make them broadly relevant.

4.3. Cognitive modeling and artificial intelligence

Our example model comparisons illustrate that, with regard to ability, parsimony, and breadth, artificial intelligence and cognitive modeling are highly interrelated endeavors. It is clear that advances in computational models of higher-order cognition can be significant contributions to artificial intelligence research. As mentioned earlier, one of the greatest challenges to creating such models is that there are no known computational methods that exhibit many aspects of human higher-order intelligence. Finding such methods is one of the goals of research in artificial intelligence. An advance in the abilities of cognitive models of higher-order intelligence is thus also a contribution to artificial intelligence.

Further, work that increases the power of known computational methods can also be of interest to cognitive modeling, even when it does not aim to do so. For example, in the early

literature on the corpus-based approaches (typically associated with artificial intelligence or computational linguistics rather than cognitive modeling), there was little discussion of the actual mechanisms used by people in language understanding. However, as outlined in the last section, this work had a significant impact in our understanding of and research into human language use and development because it demonstrated that statistical information had more grammatical information latent within it than many had previously suspected. As another example, even though the early work into chess-playing machines made few, if any, attempts to relate the mechanisms they used to actual human mechanisms, our discussion of chess showed that this work nevertheless did contribute meaningfully, in part, by establishing a framework within which to ask research questions in subsequent studies of human chess playing.

Examples such as these illustrate that when model evaluation is broadened beyond model fits to include ability, breadth, and parsimony, results in artificial intelligence (operationally defined, for example, by the journals and conferences to which it is disseminated) are often in fact a contribution to our understanding of human cognition. It also illustrates that, as in the early chess and corpus-driven language research, systems that use mechanisms that differ extensively from the mechanisms of human cognition and that do not precisely fit or predict empirical data (for example, about reaction times and error rates) can make a significant contribution to our understanding of human cognition.

5. Conclusion

Computational models play several roles in explaining higher-order cognition in humans and, consequently, there are several different ways to evaluate them. Researchers would like to know whether a model posits mechanisms that (a) at least approximate those that implement human cognition and (b) are capable enough to explain the broad range of human cognitive abilities. They should also prefer alternatives that are parsimonious and that provide a unified account of these phenomena. Although ability, parsimony, and breadth are in general very difficult to define formally and measure precisely, we have demonstrated that, in specific contexts, it is possible to precisely characterize the contribution of a research result to achieving these ends.

We have argued that quantitative model fits are only one of many activities that can contribute to a computational understanding of higher-order cognition. Our discussion about ability, parsimony, and breadth, together with the example model comparisons, suggest a number of other approaches to evaluating and driving progress in cognitive models for higher-order cognition.

5.1. Breadth and cognitive architectures

Cognitive architectures are theories of structures and processes that are invariant across much or all of human cognition. When an architecture is used to model cognition in a new domain, the breadth and parsimony of cognitive theory is extended because more phenomena are explained using the same set of mechanisms. As the next point amplifies, these benefits

are increased to the extent that multiple models within an architectural framework make the same assumptions.

5.2. Parsimony through a single model

One aspect of higher-order cognition that is especially difficult to replicate computationally is people's ability to function in a wide variety of situations. Most computational frameworks become intractable as the amount of world knowledge becomes larger. The common practice of modeling behavior only on one specific task does not demonstrate that an architecture scales as needed. By evaluating a framework's ability to support a single model that operates across a broad range of tasks, we can assure that research within that framework does not evade the difficult computational challenges of higher-order cognition. This research agenda would move the field toward theories that reproduce the broad functionality and adaptability observed in humans.

5.3. Increasing ability

A model that is unable to reason, solve problems, converse, and learn in situations where humans clearly are able to do so is an as-of-yet incomplete theory of cognition. Improving a model's cognitive abilities is thus an important step toward developing it into a comprehensive and unified account of higher-order cognition. There is no single approach to evaluating ability, but once it has been recognized as an important goal, it is often straightforward to measure. Examples of precise characterizations of ability mentioned in this article include Chomsky's language hierarchy, chess ratings, precision and recall rates in corpus parsing, and formal demonstrations that one set of concepts (e.g., those involved in physical reasoning) are sufficient to characterize the structure of other domains (e.g., the syntax of human language).

5.4. Unity through mappings

Theories of the cognitive architecture posit that certain basic mechanisms are involved throughout the range of human cognition and thus, at least implicitly, assume that there are many common underlying structures to domains in which people operate. Finding mappings between the structure of domains supports claims for parsimony and breadth made by a particular architecture, but can also play an important architecture-independent role. For example, as more domains whose structures are mapped to those described earlier (e.g., time, identity, parts, and categories), any cognitive model that explains reasoning with those structures gains in explanatory power.

5.5. Artificial intelligence

It is common to characterize models that do not quantitatively fit data as being "AI" and not "cognitive science." We have argued that this distinction is historically inaccurate (Langley, 2006) and that both fields today, especially insofar as ability is a priority, have overlapping goals. Cognitive modelers need to develop computational methods with more ability, which is

also the goal of artificial intelligence. Further, our discussion of the contributions of the corpus-driven approach show that, even when work in artificial intelligence is conducted without attention to psychological constraints, it can still significantly advance our understanding of human cognition.

Both the history of cognitive science and ongoing research demonstrate that concentrating on ability, breadth, and parsimony can generate results that have significant implications for the field. These include issues such as how language interacts with the rest of cognition and how people learn grammar. Many more such outstanding questions regard how various forms of learning integrate with each other and with reasoning, how people are often able to retrieve relevant facts and knowledge from a very large memory store in less than a second, and how emotion interacts with reasoning and problem solving. We believe that the results to date demonstrate that the approaches to evaluating computational models described herein can help drive progress toward theoretical frameworks with greater ability, parsimony, and breadth and therefore lead to significant progress on many challenging and important questions in cognitive science.

Note

1. We consider spoken utterance here. Text and other forms of non-spoken language can be mapped onto hypothetical spoken utterances.

References

Anderson, J., & Lebiere, C. (2003). The Newell test for a theory of cognition. *Behavioral and Brain Sciences, 5,* 587–601.

Bello, P., Bignoli, P., & Cassimatis, N. L. (2007, July). *Attention and association explain the emergence of reasoning about false belief in young children.* Paper presented at the 8th International Conference on Cognitive Modeling, Ann Arbor, MI.

Braine, M. D. S., & O'Brien, D. P. (1998). *Mental logic*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cassimatis, N. L. (2004, August). *Grammatical processing using the mechanisms of physical inferences.* Paper presented at the 26th annual conference of the Cognitive Science Society, Chicago, IL.

Chomsky, N. (1956). Three models for the description of languge. *IRE Transactions on Information Theory, 2,* 113–124.

Chomsky, N. (1959). A review of Skinner's "Verbal Behavior." *Language, 35*, 26–58.

Chomsky, N. (1976). On the nature of language. In S. R. Harnad, H. D. Steklis, & J. Lancaster (Eds.), *Annals of the New York Academy of Sciences* (Vol. 280, pp. 46–57). New York: New York Academy of Science.

Chomsky, N. (1980). *Rules and representations*. Oxford, England: Basil Blackwell.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht, The Netherlands: Foris.

Crevier, D. (1993). *AI: The tumultuous history of the search for artificial intelligence.* New York: Basic Books.

Dingeman, A. (1978). *Thought and choice in chess*. Berlin: de Gruyter.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 7*, 721–741.

Hamilton, W. R. (1833). Introductory lecture on astronomy. *Dublin University Review and Quarterly Magazine, 1*.

Johnson-Laird, P. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence, 33*, 1–64.

Langley, P. (2006). *Intelligent behavior in humans and machines*. Stanford CA: Stanford University Press, Computational Learning Laboratory, CSLI.

Laurence, S., & Margolis, E. (2001). The poverty of the stimulus argument. *The British Journal for the Philosophy of Science, 52*, 217–276.

Lease, M., Charniak, E., Johnson, M., & McClosky, D. (2006, July). *A look at parsing and its applications*. Paper presented at the American Association for Artificial Intelligence.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*, 375–419.

Lidz, J., & Waxman, S. (2004). Reaffirming the poverty of the stimulus argument: A reply to the replies. *Cognition, 93*, 157–165.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*, 313–330.

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences, 6*, 465–472.

Minsky, M. L., & Pappert, S. A. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Murugesan, A., & Cassimatis, N. L. (2006, July). *A model of syntactic parsing based on domain-general cognitive mechanisms*. Paper presented at the 8th annual conference of the Cognitive Science Society, Vancouver, Canada.

Newell, A. (1973). You can't play *20 questions* with nature and win. In W. G. Chase (Ed.), *Visual information processing* (pp. 285–305). New York: Academic.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A., Shaw, J. C., & Simon, H. A. (1958a). Chess-playing programs and the problem of complexity. *IBM Journal of Research and Development, 2*, 320–335.

Newell, A., Shaw, J. C., & Simon, H. A. (1958b). Elements of a theory of human problem solving. *Psychological Review, 65*, 151–166.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences, 6*, 456–463.

Pollard, C., & Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Radford, A. (1997). *Syntactic theory and the structure of English: A minimalist approach*. Cambridge, England: Cambridge University Press.

Rips, L. J. (1994). *The psychology of proof: Deduction in human thinking*. Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533–536.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review, 4*, 145–166.

Turing, A. M. (1946). *Proposed electronic calculator*. Unpublished report for the National Physical Laboratory. Published in A.M. Turing's ACE Report of 1946 and other papers (eds. B. E. Carpenter and R. W. Doran, Cambridge, MA: MIT Press, 1986).

Werbos, P. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. Unpublished doctoral dissertation, Harvard University, Cambridge, MA.