Abstract

The main claim of this paper is that connectionism offers cognitive science a number of excellent opportunities for turning methodological, theoretical, and meta-theoretical schisms into powerful integrations—opportunities for forging constructive synergy out of the destructive interference which plagues the field. The paper begins with an analysis of the rifts in the field and what it would take to overcome them. We argue that while connectionism has often contributed to the deepening of these schisms, it is nonetheless possible to turn this trend around—possible for connectionism to play a central role in a unification of cognitive science. Essential to this process is the development of strong theoretical principles founded (in part) on connectionist computation; a main goal of this paper is to demonstrate that such principles are indeed within the reach of a connectionist-grounded theory of cognition. The enterprise rests on a willingness to entertain, analyze, and extend characterizations of cognitive problems, and hypothesized solutions, which are deliberately overly simple and general—in order to discover the insights they can offer through mathematical analyses which this simplicity and generality are makes possible.

In this paper, seven interrelated principles are articulated, analyzed, and applied. The three areas of application concern (1) computational, (2) linguistic, and (3) philosophical problems in cognitive science: (1) the integration of connectionist and symbolic computation; (2) the syntactic, semantic and phonological components of grammar; and (3) the explanation of the systematicity and productivity of higher cognition.

With respect to (1), principles integrating connectionist and symbolic computation are developed by establishing mathematical relationships between two levels of description of a single computational system: at the lower level, the system is formally described in terms of highly distributed patterns of activity over connectionist units, and the dynamics of these units; at the higher level, the same system is formally described by symbolic structures and symbol manipulation. Specific treatment of recursion is developed, and the principles are shown to be strong enough to allow the specification of arbitrary formal languages (and hence Turing machine computation).

Applied to natural language, application (2), the computational principles entail that a central organizing principle of grammar should be *optimality*: a grammar is a means of determining which of any set of structural analyses of an input is the most well-formed. Such a grammar consists of a set of "soft" rules or constraints, each of which is in principle violable in the appropriate context. This constitutes a novel framework for formal grammar which emerges from the connectionist computational substrate. It is shown how such soft rules allow for precise treatment of the complex interaction of semantic and syntactic factors in a linguistic problem that has been the subject of considerable research in recent years, that of split intransitivity. Work showing that soft constraints embodying optimality principles allow significant progress in the development of a theory of universal phonology is briefly summarized.

Finally, the principles are shown to successfully meet the foundational challenge (3) posed by (Fodor and Pylyshyn, 1988): to use connectionism to explain the systematicity and productivity of language and thought, without merely implementing the traditional symbolic explanation. On the new account described here, symbols and rules have a novel status in cognitive theory: they play essential roles in mathematical proofs which explain these crucial properties of higher cognition, but they do not play a role in algorithms which causally generate this behavior.

The paper concludes with an assessment of how well the principles meet the goals set for them.

Principles for an Integrated Connectionist/Symbolic Theory of Higher Cognition

Paul Smolensky[1]
Department of Computer Science &
Institute of Cognitive Science
University of Colorado
Boulder, Colorado 80309-0430
`smolensky@cs.colorado.edu`

Géraldine Legendre[2]
Department of Linguistics &
Institute of Cognitive Science
University of Colorado
Boulder, Colorado 80309-0295
`legendre@tramp.colorado.edu`

Yoshiro Miyata[3]
School of Computer & Cognitive Sciences
Chukyo University
101 Tokodate, Kaizu-cho
Toyota, 470-03 Japan
`miyata@sccs.chukyo-u.ac.jp`

Report CU-CS-600-92, Computer Science Department, University of Colorado at Boulder
Report 92-08, Institute of Cognitive Science, University of Colorado at Boulder
Report 92-1-02, School of Computer & Cognitive Sciences, Chukyo University

July 2, 1992

[1]This research has been supported by NSF grants IRI-8609599, ECE-8617947, and IST-8609599 and BNS-9016806; by the Sloan Foundation's computational neuroscience program; by the Optical Connectionist Machine Program of the University of Colorado Center for Optoelectronic Computing Systems and by the University of Colorado at Boulder Council on Research and Creative Work.

[2]The research presented here was partly supported by the University of Colorado at Boulder Council on Research and Creative Work.

[3]The research reported here supported by the Optical Connectionist Machine Program of the University of Colorado Center for Optoelectronic Computing Systems (sponsored in part by NSF/ERC grant CDR-8622236 and by the State of Colorado Advanced Technology Institute).

Contents

1 Goals and Overview

In this paper we will argue that connectionism offers cognitive science a number of excellent opportunities for turning methodological, theoretical, and meta-theoretical schisms into powerful integrations—opportunities for converting much of the destructive interference which has historically plagued the field into constructive synergy. The source from which these opportunities radiate is the unification of connectionist and symbolic computational principles; the importance of such unification seems to be increasingly recognized (e.g. (Estes, 1988; Hendler, 1989; Hinton, 1990; Touretzky, 1991)). The general approach to unification we follow here was articulated in (Smolensky, 1988).

In this first section, we articulate three goals for cognitive science and argue that, despite their importance, these goals have by and large been seriously compromised in cognitive research. In the remainder of the paper we argue that, despite the difficulty of pursuing these three goals simultaneously, meaningful progress can in fact be made on all of them. As evidence we present a research program, reviewing past results, presenting many new results, and suggesting a few future directions. By presenting an extended example, we hope to raise the plausibility of the proposition that within reach of available research techniques is a cognitive science that is genuinely a coherent discipline—rather than merely a cover term for an incoherent collection of related, but largely divergent, sister disciplines.

1.1 Model- and Principle-Centered Cognitive Science

At the meta-theoretical level, our goal is to show that connectionism can make a different kind of contribution to cognitive science than has been previously recognized. This difference can be appreciated by dividing research in cognitive science very broadly into what we'll call the "model-centered" and "principle-centered" approaches. In both approaches, the research involves both more specific analyses and more general cognitive claims; the difference lies in the focus of attention. [1]

Model-centered research focusses on specific analyses—usually in the form of implemented computational models—of particular behavioral phenomena—typically, effects measured in psychological experiments, or particular computational tasks which abstract from cognitive behavior. In model-centered research, the model is usually offered in the scope of more general theoretical claims, but there is a marked asymmetry between these claims and the model. The general theoretical claims are clearly in the background, and there is rarely a major effort directed toward articulating them with great care or in analyzing them in depth. As a result, at least in the form offered, they are usually too imprecise to allow any strong conclusions to follow. Instead, it is the model that is assigned the job of demonstrating the desired behavior; this is then taken as evidence in favor of the general claims.

In principle-centered research, the relative roles of specific accounts and general theoretical claims is reversed. Considerable effort is invested in clearly formulating the general claims, to make them worthy of the appellation *principles.* While general in scope, they must be sufficiently precise that one can actually deduce (general) consequences from them. Furthermore, it must be possible to instantiate the general principles in specific accounts, in order to demonstrate their adequacy for explaining particular empirical phenomena. In the model-centered approach, the

[1] We do not pretend to have a thorough and careful analysis of the the model- vs. principle-centered distinction and all its connections to cognitive science; if we did, it would surely require a paper of its own. The following remarks, obviously oversimplified, are intended only to partially clarify the distinction.

general background claims serve to enhance the appeal of the particular model in the spotlight; in the principle-centered approach, the specific accounts serve as a background to demonstrate the empirical adequacy of the general principles.

The model- vs. principle-centered distinction is rather strongly correlated with others in cognitive science, most notably, the subdisciplines. In neuroscience, cognitive psychology, and much of AI, the model-centered approach is well intrenched; in philosophy of mind, linguistics, and the remainder of AI, the principle-centered approach tends to dominate. In philosophy of mind, of course, general theoretical claims are by and large the entire point of the enterprise, and specific analyses are clearly relegated to the background (when they figure at all). In linguistics, the rise of principle-centered research over earlier styles of research which emphasized descriptive grammars of particular languages is taken by generative linguistics to be perhaps its foremost achievement; to the extent that the generative approach is dominant within but not coextensive with modern linguistics, the principle-centered approach dominates but does not entirely characterize linguistics. AI, reflecting in part the distinction between theory and practice in computer science, is more evenly split by the principle- and model-centered divide, which correlates with what has long been known as the "neat vs. scruffy" dichotomy; the "scruffies' credo" that "the program **is** the theory" is an extreme formulation of the model-centered approach. In cognitive psychology, theory in general remains largely subordinated to empirical investigation, and it is hardly surprising therefore that as computational cognitive modeling has developed, it has focussed largely on providing specific accounts of specific empirical findings. This discipline as a whole appears rather dubious about the possibility of general theoretical principles.[2] Thus the model-centered approach has a strong hold in psychology, although some research (e.g., on cognitive architecture) illustrates that the principle-centered approach is not entirely unrepresented. And in neuroscience, the dominance of the empirical over the theoretical is even more extreme; neuroscientists, even more than psychologists, like to view their object of study as far too complicated to submit to general theoretical principles.

Indeed, complexity is at the heart of the model- vs. principle-centered distinction. In order for a modeling problem to be well-posed, the data to be fit should be sufficiently rich and complicated to highly constrain the possible models. Certainly the data of neuroscience, cognitive psychology, and linguistics provide a virtually limitless source of empirical complexity to serve this function. In contrast to the model-centered approach, the principle-centered approach demands a willingness to treat this complexity incrementally—starting with characterizations of the cognitive phenomena to be understood which are deliberately, and carefully, 'over'-simplified and 'over'-generalized. Simple and general principles which provide this understanding then serve as the foundation for tackling more complex characterizations of the phenomena. Focusing on fairly simple principles runs counter to a strong computational current in cognitive science best represented by AI, where the entire research project is often equated with the development of ever more complicated and powerful computational systems. This is especially true in connectionist research, were the need for greater computational power is particularly pressing, and ever greater network complexity is often taken to be the only possible way of proceeding. The computational principles to be presented here will involve connectionist networks that are simple enough to appear almost antique to readers familiar with state-of-the-art networks. Yet the reward for greater simplicity is greater

[2]This may in part come from failing to clearly distinguish such principles from empirical generalizations of broad scope, such as the power law of practice. To the extent that such generalizations exist, we would claim that they play the role of *empirical facts to be explained* by theoretical principles; they are not the principles themselves. General theoretical principles may well exist even in the absence of broad-scope empirical generalizations.

analyzability. For all the simplicity of the fundamental principles they rest on, the results developed here constitute significant progress not just in our *understanding* but also in the *power* of connectionist computation, as it pertains to higher cognition. There are merits to studying computational systems at multiple points in the tradeoff between complexity and analyzability.

The pre-eminent goal of this paper is to show that, contrary to the currently prevailing view, connectionism can serve to heal the meta-theoretic split between model- and principle-centered cognitive science. That connectionism has proved itself an extremely fruitful vehicle within the model-centered approach is a well-attested (if not uncontroversial) belief. Not surprisingly, examples are most common within the subdisciplines where this approach dominates: connectionism has been a major impetus in promoting the success of theory (viz. modeling) within neuroscience; connectionist modeling has rapidly become a central part of theory (viz. modeling) within cognitive psychology; and connectionist (or "neural network") techniques in AI and in applications have enjoyed extraordinary success within a remarkably short time.

By contrast, with respect to *principle-centered* research, connectionism's contribution has been *strikingly* smaller. Inadequacies of connectionist research construed as principle- rather than model-centered have been pointed out by a number of critiques, some early language-oriented examples being (Lachter and Bever, 1988; Pinker and Prince, 1988) and a recent psychology-oriented example being (McCloskey, 1992).

The fact—call it "F"—that connectionism is not effective as a vehicle for principle- as opposed to model-centered research has served as yet another source of divisiveness within cognitive science. Proponents of the principle-centered approach can use F to dismiss or deride connectionism. Proponents of connectionist modeling can recruit F for a new argument against the validity of a principle-centered approach.[3] Agreeing on the validity of F, both connectionist modelers and anti-connectionists favoring principle-centered research can use it to their drive a wedge into the rift between them.[4]

We aim to dissolve this conflict by arguing that F is in fact false, that connectionism *can* make important contributions to principle-centered research. Rather than further widening the schism between model- and principle-centered cognitive science, connectionism can actually help to *integrate* the two approaches, by providing *both* a powerful platform for modeling *and* new general principles that increase the scope of principle-centered research within cognitive science. That connectionist methods can contribute to principle-centered research has long been argued by Feldman (e.g., (Feldman, 1981; Feldman, 1985; Shastri and Feldman, 1985) and Grossberg (Grossberg, 1982; Grossberg, 1988)), and more recently by McClelland (McClelland, 1991). The case we make here for principle-centered connectionist research is based, however, on a set of principles whose content defines a theory quite different from those proposed by these authors.

We are convinced that in order to achieve this, connectionism's contributions to the development of cognitive principles must come in the form of integration with—not replacement of—symbolic principles. This takes us from the issue of meta-theoretic integration of the model-

[3] E.g.: "Connectionist models have proven themselves as models of human behavior, and as our current best description of neural computation, so they are the best models of the mind/brain we have; but, clearly, connectionist models, in all their essential complexity and mystery, are *just too complicated* to admit the sort of precise, powerful, general principles presupposed by the principle-centered approach—therefore, such principles don't exist."

[4] This situation illustrates an argument structure which underlies several of the other schisms in cognitive science discussed below, and which is known in philosophy as "one person's *modus ponens* is another's *modus tolens*." Since F is essentially $p \rightarrow$ not-q where p = "connectionism is valid" and q = "the principle-centered approach is valid," the pro-connectionist uses F to conclude not-q (by assuming p) while the anti-connectionist uses F to conclude not-p (by assuming q).

centered and principle-centered approaches to the issue of theoretical integration of the connectionist and symbolic approaches, and to related issues of integrating the methodologies of the cognitive subdisciplines.

1.2 Three Goals for Cognitive Science

To further focus this discussion, and to serve as a reference point for the remainder of this paper, we articulate as goals for cognitive science the development of a body of theory characterized by the following three attributes:

(1) **Theoretically Integrated.** A theory:

- a. that addresses the full challenges of higher-level cognition,
- b. while achieving both "horizontal" and "vertical" integration.

(2) **Methodologically Integrated.** A theory:

- a. that integrates the theoretical insights into the various cognitive domains contributed by the numerous and varied relevant disciplines;
- b. whose development effectively exploits the diverse methodologies of these disciplines; and
- c. whose content constructively feeds into these disciplines and furthers their own particular goals.

(3) **Meta-Theoretically Integrated.** A theory that supports both:

- a. **model-centered research** by providing the technical resources for building detailed accounts of empirical cognitive data; and
- b. **principle-centered research** by providing cognitive principles of considerable generality which are precise enough to support explanations and which are grounded in mathematically sound and powerful foundations.

In (1a), we use "higher-level cognition" to refer to those rather abstract domains—e.g., language, problem solving, reasoning, abstract planning—where nearly all existing cognitive theory relies heavily on some sort of symbolic computational model. These domains are to be contrasted with lower-level ones—e.g., pattern recognition, perception, motor control, memory—where symbolic computation plays a significantly less dominating role in current theory.

By a "horizontally integrated" theory in (1b) we mean one that provides a coherent account of the interrelation of cognitive processes across the wide range of domains (exemplified in the preceding paragraph) that span both lower- and higher-level cognition. "Vertical integration" requires a coherent theory of the interrelation of the multiple levels of organization spanning from the neural level up to the highest mental levels (e.g., that of the universal theory of grammar). While horizontal integration has been a focus of cognitive architecture research from its early days to the present (Newell, 1990; Van Lehn, 1991), the importance of the goal of vertical integration for cognitive science has only begun to be broadly recognized more recently (e.g., (NSF Planning Workshop Report, 1991)). [5]

[5] It might be argued—as indeed it is in (NSF Planning Workshop Report, 1991)–that the full notion of vertical integration ought to extend further "upward" to the social level. The absence of this level is an acknowledged weakness of this paper.

We will now briefly argue that all of the Three Goals are extremely important for cognitive science, and that most research programs have failed to really face the challenge they *together* present. We will deliberately oversimplify to bring out the main points in a minimum of space.

At the heart of cognitive science lies a profound paradox. Formal theories of logical reasoning, grammar, and other higher mental faculties compel us to think of the mind as a machine for rule-based manipulation of highly structured arrays of symbols. What we know of the brain compels us to think of human information processing in terms of manipulation of a large unstructured set of numbers, the activity levels of interconnected neurons. Finally, the full richness of human behavior, both in everyday environments and in the controlled environments of the psychological laboratory, seems to defy rule-based description, displaying strong sensitivity to subtle statistical factors in experience, as well as to structural properties of information. To solve the *Central Paradox of Cognition* is to resolve these contradictions with a unified theory of the organization of the mind, of the brain, of behavior, and of the environment. Such a theory would make cognitive science truly *a discipline*, rather than a convenient cover term for the panoply of divergent disciplines that each represent one perspective on cognition, but which each avoid facing the Central Paradox.

Research in cognitive science, like that within the more specialize allied disciplines, has also exploited many devices to avoid facing the Paradox head on. The Three Goals have been designed to block them all, as we now see.

One kind of research that would appear to come close to simultaneously addressing all the Three Goals is work in which principles of neural or connectionist computation are formulated and applied to some cognitive domain. In such research, however, the domains studyied tend to be limited to lower-level cognition; a typical "cognitive" problem addressed in such work is phoneme classification, often treated in fact as an engineering pattern recognition task. Such research makes valuable contributions to cognitive science; but it falls seriously short on the goal of Theoretical Integration, and completely avoids facing the Central Paradox by simply ignoring the half of the paradox involving symbolic computation and higher-level cognition. A large fraction of connectionist research falls into this category. The approach to the Paradox taken by such research is, either implicitly or explicitly, one of *denial*: the existence of the symbolic half of the Paradox is simply denied. Indeed, connectionist or neural modelers of the eliminativist school make a major virtue out of this denial.

This type of denial is the underside of a coin that has been valued currency in cognitive science since its beginning. The top side of the coin is functionalism, a doctrine that has been used to conclude that mental computation is utterly sealed off from neural computation. This doctrine has for decades made a virtue of ignoring neuroscience, and, again, of approaching the Paradox by denial.

With connectionism's rise to prominence over the past half-decade, many admirers of functionalism and its relatives adopted the implementationalist position that the proper role of connectionism in cognitive science is merely to implement existing symbolic theory. This constitutes one kind of violation of Methodological Integration—part (2c)—because literal implementation cannot inform the higher-level account; connectionism would thus provide nothing whatever new to the theory of higher-level cognition. By contrast, the work described in this paper—addressing grammar (cf. (5d) and Section 3) and the explanation of the productivity of cognition (cf. (5e) and Section 4)—exploits vertical integration mediated by connectionism to make major innovations within higher cognitive theory.

But the more compelling motivation behind the goal of Methodological Integration is the historically dominant relationship between the subdisciplines of cognitive science: sectarian warfare, thinly veiled by politically correct endorsement of multidisciplinarity. Those neuroscientists aware of research on higher cognition have dismissed most of it as, at worst, addressing phenomena about as real as ESP, and, at best, premature by at least a century. Cognitive scientists studying higher cognition have happily agreed to wait 100 years before paying any serious attention to neuroscience. Psychologists have dismissed the central data of linguistics as meaningless artifacts of linguists' own intuitions, and have dismissed linguistic theory as a grossly over-indulgent abuse of abstraction. Linguists have returned the favor by dismissing empirical psychological work on language as irrelevant to any interesting issue and psychological theory of language as hopelessly empiricist. Neuroscientists, psychologists and linguists have managed to get together, however, in order to dismiss AI as largely irrelevant for human cognition due to an almost complete lack of any appropriate empirical constraints. And AI researchers have joined with all their critics for the purpose of dismissing philosophy of mind as contributing nothing to cognitive science but worthless verbiage.

This picture is an apt caricature of the true attitudes prevalent among cognitive scientists, on our experience; the features may be exaggerated, but they are indeed the essential ones. A reasonable case could be made that this factional fighting has produced so much destructive interference between the subdisciplines that mainstream cognitive science has in the past been a whole that is much *less* than the sum of its parts.

But cognitive science is changing. The goal of Methodological Integration articulates the challenge of putting aside the historical feuds, a challenge to which increasingly many cognitive scientists are rising.

The final goal of Meta-Theoretic Integration has already been motivated in Section 1.1. For the purposes of this paper, we will take it as established that both connectionist and symbolic computation have proved themselves capable of supporting model-centered research (although not often with respect to the same cognitive phenomena). We will therefore put (3a) aside and focus our attention on (3b), which we will refer to simply as the goal of *Principle-Centering*. This goal is satisfied by much cognitive research based on symbolic computation, but becomes an important constraint on a theory once connectionist computation starts to play a major role. The power and solidity of the foundations of connectionist computation are in much need of development, especially for the purposes of the theory of higher cognition, emphasized in the Theoretical Integration goal (1a), but neglected in most connectionist work. And it is important for cognitive theory employing connectionism to complement the development of particular network models with the simultaneous development of general cognitive principles, in order to constructively engage with other cognitive theory.

Having highlighted some limitations of existing research in cognitive science, we want to emphasize that our purpose is *not* to dismiss this research; on the contrary, it is to argue for the importance of *all* of the Three Goals, and to demonstrate the need for research programs that tackle the difficult challenge of integrating widely disparate approaches to cognitive research. The main burden of this paper is to argue that, by letting the Three Goals fashion all facets of a research program, it is in fact possible to simultaneously advance on all of them—and possible to make significant strides towards a satisfactory resolution of the Central Paradox of Cognition.

1.3 Role of Connectionism

The research program we present here, which we'll call the *Sub-Symbolic Paradigm* or simply *SSP*, develops an integration of symbolic and connectionist computation to pursue the Three Goals. While our motivation for such integration has much in common with the increasingly popular movement towards hybrid systems (e.g., (Hendler, 1989)), we use the term "integrated" rather than "hybrid" to emphasize that in SSP, the connectionist and the symbolic are two perspectives on a single computational component, rather than two computationally separate components of a larger system.

Choosing to incorporate connectionism into an approach to the Three Goalsentails both exciting opportunities and tremendous difficulties. Such a mixture is the most one can reasonably expect in any approach that squarely faces the Central Paradox.

A number of the opportunities afforded by connectionism were discussed at some length in the paper (Smolensky, 1988) that introduced SSP. It was argued that connectionism can provide a powerful means of achieving vertical integration by adopting a level of description intermediate between those of neurons and of symbols, and by exploiting this intermediate ("subsymbolic") level as a bridge for bringing into contact theories of neural and mental computation. Furthermore, it was argued, connectionism provides a computational account of a unified cognitive architecture from which can emerge quite varied processes or *virtual machines* that serve the varying needs of diverse cognitive domains; this is a powerful means of achieving horizontal integration. Thus the same fundamental connectionist computational mechanisms can be seen to underly perceptual processes, memory, and certain higher-level processes; different principles of organization emerge as higher-level descriptions of different kinds of connectionist networks operating in different kinds of information-processing environments.

Connectionism, however, makes each of the Three Goals difficult to achieve. Let's consider them in turn.

Given the difficulty of relating connectionist computation to the symbolic computation dominating the theory of higher-level cognition, connectionist research has tended to focus heavily on lower-level processes and neglect much of the challenge of higher-level cognition—which has been the heart of cognitive science since its inception. Thus connectionism has in practice seriously compromised the important aspect (1a) of Theoretical Integration emphasizing higher cognition.

The small proportion of connectionist research devoted to higher cognition (and the perhaps larger quantity of rhetorical debate surrounding it) has tended towards one or the other extreme of eliminativism or implementationalism. Eliminativist connectionist models are used to argue that some concept from symbolic theory is misguided or superfluous; implementationalist connectionist models are used to more-or-less directly implement symbolic concepts. Our assessment of the impact within cognitive science of both kinds of research is a kind of polarization that is completely antithetical to Methodological Integration, goal (2). Theorists of higher-level cognition overwhelmingly evaluate the eliminativist research as seriously missing the mark by failing to address the target symbolic concepts in anything like the adequacy required to justify the conclusion that these concepts can be eliminated. Neither are they generally impressed by the implementationalist research, since the implementations tend to neither do justice to the true symbolic concepts nor to look very computationally enlightening or neurally plausible. We believe the net effect is that most cognitive scientists studying higher-level cognition *at best* are becoming confirmed in their initial suspicion that connectionism has little to offer them in the pursuit of their scientific goals—at worst, they are concluding from strong eliminativist claims that connectionists are trying to seriously roll back the clock in our understanding of higher-level cognition.

Finally, like the two integrative goals, Principle-Centering, (3b), is a major challenge for a connectionist approach. Connectionist cognitive research has typically been heavily model-centered, rather than focussing on precise, powerful, and formally well-founded cognitive principles, which may be illustrated or embodied in particular models. Connectionist models are sufficiently difficult to understand that it is often quite mysterious why they fail, and—perhaps even more disturbing—just as mysterious why they succeed when they do. Such mysteries must give way to principled theory if connectionism is to achieve Principle-Centering—but this is a serious technical challenge.

1.4 Illustrative Problem

To illustrate the intent of the Three Goals, and to introduce some of the ways that research in SSP bears on these goals, we now briefly consider one of the general research questions on which SSP has made significant progress in recent years:

(4) How can principles of neural computation inform those of mental computation to the point of strengthening the universal theory of grammar?

Theoretical Integration manifests itself in question (4) in two ways: (a) in the choice as the target cognitive domain one of the most abstract and central areas of higher-level cognition, the organizing principles of human language; and (b) in looking to bottom-up input from neural computation as a source of new insight—a hallmark of vertical integration. Methodological Integration is pursued by taking the characterization of the problem to be solved—development of a universal theory of grammar—from the discipline most centrally concerned with the organization of human language: theoretical linguistics. Methodological Integration manifests itself further in the way SSP addresses question (4): by incorporating a number of the insights (e.g., principles of syllable structure) and methodologies (e.g., the study of well-formedness) of theoretical linguistics. Answers to question (4) serve the goals of theoretical linguistics at the same time as advancing cognitive science. Finally, Principle-Centering shapes question (4) by identifying the goal with general principles, as opposed to specific models of particular examples of linguistic behavior. The particular means SSP uses to address the question involves incorporating into linguistic theory a powerful new mathematical principle central to the connectionist computational framework: optimization of well-formedness or *Harmony*.

Some of the results achieved by SSP concerning question (4) are now briefly summarized, along with other results; Section 3 contains the more extended discussion.

1.5 Summary of Results

Those contributions of SSP which are reported in this paper can be summarized as follows:

(5) **Summary of results of SSP**

a. **Representation.** SSP develops a mathematical formalism showing precisely how a mental representation can be *simultaneously* a fully distributed pattern of numerical activities at one level of analysis *and* the functional equivalent of a (possibly recursive) symbolic structure when analyzed at a higher level.

b. **Processing.** SSP shows in mathematical detail, illustrated by computer simulations, how mental processing can be simultaneously a massively parallel process of spreading activation at one level of analysis and, at a higher level, a kind of parallel holistic manipulation of symbolic structures—even those containing recursive embedding. In some cases, the functions computed by this symbol manipulation can be formally specified by symbolic programs.

c. **Well-formedness.** SSP and related connectionist research demonstrate that the overall effects of spreading activation can sometimes be analyzed at a higher level as a process of *optimization*, in which a representation is constructed that maximizes a connectionist measure of well-formedness we call *Harmony*.

d. **Grammar.** SSP shows how to combine the three preceding results (a-c) to define a new formalism for grammar, a formalism which can be used for the characterization of formal languages and which has been successfully employed to address long-standing problems in natural language phonology and syntax to which solutions within purely symbolic theory have been problematic. This formalism constitutes a novel integration of connectionist and symbolic computation, and rests on both symbolic and connectionist technical advances.

e. **Productivity.** SSP combines (a-c) to shed new light on a central problem in the foundations of cognitive science: the explanation of how higher cognition can achieve, with finite and fixed resources, competence that is highly systematic, coherent, compositional, and productive.

Results (5a–c) are discussed in Section 2; (5d), in Section 3; and (5e) in Section 4.

These results (5) constitute direct progress in the achievement of the Three Goals. The cognitive problems of grammar and productivity addressed in (5d,e) are among the most central in higher cognition, falling squarely under Theoretical Integration (1a). Indeed, results (5d) on language bear directly on question (4); the universal theory of grammar is surely one of the bastions of higher cognitive theory. And—as emphasized in (Fodor and Pylyshyn, 1988), a highly influential critique of connectionist theory—the problem of productivity addressed in (5e) has also been recognized as one central to the understanding of higher-level cognition. Results (5a–c) on representation, processing, and well-formedness provide the supporting pillars of a vertically integrated theory, built from general notions of connectionist computation that cut across many cognitive domains and simultaneously provide horizontal integration. Thus the results (5) contribute significantly towards the achievement of Theoretical Integration.

Goal (2), Methodological Integration, is clearly impacted as well. Results (5d,e) on grammar and productivity target problems that have been of central interest in the disciplines of theoretical linguistics and philosophy of mind, and the formulation of these problems taken up by SSP are those that have been developed by the practitioners of those fields; the data in question are those recognized in the related disciplines; the theoretical constructs of the other disciplines play a major role; and the established methodologies of those disciplines have been incorporated into the emerging multidisciplinary methodology of SSP.

Finally, Principle-Centering, (3b), is served by all the results (5). In each case, the results take the form of powerful and general cognitive principles, centered on novel concepts for understanding cognition that arise from viewing a lower-level connectionist computational model and a higher-level symbolic computational model as two descriptions, at different levels of analysis, of one and the same computational system. The connectionist technical innovations of SSP—including, with others, the technical components of all the results (5)—have contributed substantially to the soundness and power of the mathematical framework supporting a connectionist theory of higher cognition. Original symbolic technical contributions have also provided innovations in grammar formalism. Key have been mathematical bridges between the continuous, numerical model of computation underlying connectionism and the discrete, symbolic computation of virtual machines that emerge naturally as higher-level approximate descriptions of appropriately designed connectionist systems. These mathematical techniques provide the kind of technical leverage needed to make progress on the Central Paradox of Cognition.

1.6 Presentation of Results

The remainder of the paper provides more detailed discussion of (5) and the technical contributions of SSP: Section 2 concerns the computational issues of (5a–c); Section 3, the linguistic questions of (5d); and Section 4, the foundational issues of (5e). For each of the three research projects presented in these three sections, we sketch the general cognitive and computational principles providing the core, and indicate how progress has resulted from newly integrated theory and research methodology. Aside from half of each of Sections 2.2.1 and 3.1.5, the results discussed are reported here for the first time.

Given the ambition of the enterprise and the few person-years so far invested in it, some rather obvious caveats are in order. First, while considerable effort has been put into the formulation and analysis of the principles we present, it is clear that these are working hypotheses that will undergo major refinement, revision, and possibly rejection in the course of subsequent research. While rather precise formulation of the principles is necessary for analyzing them, this level of precision should not be mistaken as implying that we believe they are now ready to be etched in stone. More important than the exact formulation of the principles is the general theory of higher cognition they embody and *what kind of research they make it possible to do.* We intend the principles given here to be taken as representing a class of principles, largely unexplored, which embody the same general theory and which allow the same kind of research. Each of the principles to be presented in its current version performs a certain function in the theory; in the future, each is likely to be replaced by a stronger, perhaps more complex, version which performs much the same function in the overall theory but which does so more effectively.

Secondly, the computational principles presented are quite general, and are therefore *potentially* applicable to a wide variety of problems in higher cognition. But of course the extent of their actual validity in various cognitive domains remains for the most part to be put to the empirical test. The adequacy of the principles as judged by certain computational, linguistic, and foundational criteria are respectively addressed in Sections (2)–(4), but this of course constitutes a tiny fraction of the tests to which one needs to put such general principles. In particular, crucial directions for future work are investigation of the neural underpinnings of these principles, and their embodiment in detailed psychological models. A few opportunities for pursuing the latter direction are pointed out in Sections 2.4 and 3.3.[6] The absence of empirical test of the principles discussed below against the kind of data central to cognitive psychology and neuroscience is an acknowledged deficiency in the current state of the research program. As explained in Section 1.1, the *initial* emphasis of this research on issues central to the theory of computation, linguistics, and philosophy of mind, rather those of neuroscience or cognitive psychology, is a consequence of differences among these fields in the prominence of principle-centered research, along with our high-level goal of demonstrating that connectionism can contribute to principle-centered and not just model-centered research in cognitive science.

[6] It is less clear at this point how best to pursue the question of whether the principles have adequate support in empirical neuroscience. The right kind of question, of course, is *not* "how much detailed neural data can we explain with Harmony maximization," but rather, "is there a higher level of analysis of neural computation where Harmony is maximized, to a reasonable approximation." It is worth noting that while the motivations for introducing Harmony were psychological and computational rather than neural (Smolensky, 1983; Smolensky, 1984b; Smolensky, 1984a; Smolensky, 1986), neural motivation was explicitly cited for the introduction of very closely related Lyapunov functions in (Cohen and Grossberg, 1983; Hopfield, 1982; Hopfield, 1984). Thus, while the real question remains to be addressed, it is nonetheless historically accurate to say that principles virtually identical to that of Harmony maximization are neurally motivated.

2 Integration of Connectionist and Symbolic Computation

This section concerns the computational backbone supporting the SSP research program: research studying how symbolic computation can arise naturally as a higher-level virtual machine realized in appropriately designed lower-level connectionist networks.

2.1 The Principles

We now present the current formulation of the fundamental computational principles, briefly discussing each in turn before we proceed in subsequent sections to provide the details. These principles are hypothesized to operate in higher cognitive domains, where cognitive theory has posited symbolic representations that play a central role.

(6) **Integrated Representation.**

a. When analyzed at the lower level, mental representations are distributed patterns of connectionist activity; when analyzed at a higher level, these same representations constitute symbolic structures.

b. Such a symbolic structure $\mathtt{s}$ is a set of *filler/role bindings* $\{\mathtt{f}_i/r_i\}$, defined by a collection of structural roles $\{r_i\}$ each of which may be occupied by a filler $\mathtt{f}_i$—a constituent symbolic structure.

c. The corresponding lower-level description is an activity vector

(7) $\mathbf{s} = \sum_i \mathbf{f}_i \otimes \mathbf{r}_i$

which is the sum of vectors representing the filler/role bindings. In these *tensor product representations*, the pattern for the whole is the superposition of patterns for all the constituents. The pattern for a constituent is the tensor product of a pattern for the filler and a pattern for the structural role it occupies.[7]

d. In certain cognitive domains such as language and reasoning, the representations are *recursive*: fillers which are themselves complex structures are represented by vectors which in turn are recursively defined as tensor product representations. The set of fillers is then the same as the set of whole structures.

Tensor product representations (Dolan, 1989; Dolan and Dyer, 1987; Smolensky, 1987b; Smolensky, 1990) generalize many of the traditional kinds of connectionist representations, and while the generic tensor product representations are fully distributed, special cases reduce to fully or semilocal representations.[8] The units in tensor product representations can sometimes be interpreted

[7] The tensor product $\otimes$ generalizes the outer product of matrix algebra. If $\mathbf{u} = (u_1, u_2); \mathbf{v} = (v_1, v_2)$ then their tensor product $\mathbf{u}\otimes\mathbf{v}$ is a *second-rank tensor*, a vector whose elements are normally labelled with *two* subscripts: $(\mathbf{u}\otimes\mathbf{v})_{ij} \equiv u_i v_j$; the elements of $\mathbf{u}\otimes\mathbf{v}$ are thus $(u_1 v_1, u_1 v_2, u_2 v_1, u_2 v_2)$. In general, a tensor of rank n has elements labelled with n subscripts, and the tensor product extends in the obvious way to tensors of arbitrary rank; e.g., the tensor product of a rank-2 tensor $\mathbf{S}$ and a rank-3 tensor $\mathbf{T}$ is a rank-5 tensor $\mathbf{R} = \mathbf{S}\otimes\mathbf{T}$ with elements $R_{ijklm} \equiv S_{ij}T_{klm}$. The recursive construction of tensor product representations requires the use of tensors of rank higher than two, which is why simpler matrix algebra does not suffice.

[8] Thus even those who believe with (Feldman, 1989) that neural representations are not highly distributed should not see this as any objection to the use of tensor product representations as a low-level model of mental representations. If desired, special cases of the tensor product representation can be designed with any desired degree of locality, up to and including representations which dedicate a single node to an entire structure (e.g, proposition).

as the conjunction of a feature of a filler and a feature of it role; in these cases, the approach is a formalization of the idea of *conjunctive coding* (Hinton et al., 1986; McClelland and Kawamoto, 1986; Rumelhart and McClelland, 1986) Psychological models of human memory have also employed tensor product or closed related representations; see Section 2.4. Tensor calculus (unlike matrix algebra) allows such representations to be defined recursively; in this paper, we will considerably extend the analysis of the recursive capabilities of tensor product representations.

(8) **Integrated Processing.**

a. When analyzed at the lower level, mental processes consist in massively parallel spreading of numerical activation; when analyzed at a higher level, these same processes constitute a form of symbol manipulation in which entire structures are manipulated in parallel.

b. Certain of these processes can be described precisely in terms of higher level programs. Like traditional computer programs, these programs describe complex functions by sequentially combining primitive symbolic operations. Such programs specify the *input/output function that is computed*, but the complex sequences of primitive operations do *not* constitute procedures by which these functions are actually computed.

c. These processes are capable of fully productive recursive structure processing.

Structure-sensitive symbolic processing of tensor product representations is achieved by means of operations from tensor calculus which check conditions on constituents and which use linear transformations to move constituents in given structural roles to new ones, or to modify the fillers in given roles. Such operations are naturally embodied in connectionist networks (Dolan, 1989; Dolan and Dyer, 1987; Dolan and Smolensky, 1989; Legendre et al., 1991; Smolensky, 1987b). The new extensions of this principle (8b,c) are developed here, primarily in Sections 2.2.1–2.3.

(9) **Harmonic Principle—Connectionist formulation.**

a. The lower-level descriptions of the activation spreading processes constituting mental representation satisfy certain mathematical properties which entail that these processes can be analyzed as constructing the representation including the given input structure which *maximizes Harmony.*

b. At the lower level, the Harmony is computed as a particular mathematical function of the numbers comprising the activation pattern and the connection weights. In many important cases, the core of the Harmony function can be written at the lower level as the simple quadratic form:

(10) $H = \mathbf{a}^{\mathbf{T}}\mathbf{W}\mathbf{a} = \sum_{i,j} \mathbf{a}_i \mathbf{W}_{ij} \mathbf{a}_j$

Here **a** is the network's activation vector and **W** its connection weight matrix.

c. In recursive domains, a simple condition on the weight matrix **W**, analogous to translation-invariance in many visual processing systems, ensures that the Harmony function is *embedding invariant.*

Principle (9) combines the tensor analysis techniques underlying (6) and (8) with another technique from mathematical physics: Lyapunov functions for analysis of the convergence of dynamical systems. (The Harmony function is such a Lyapunov function.) (9a,b) have been developed by many people, including (Cohen and Grossberg, 1983; Golden, 1986; Golden, 1988; Hinton and Sejnowski, 1983; Hinton and Sejnowski, 1986; Hopfield, 1982; Hopfield, 1984; Hopfield, 1987; Smolensky, 1983; Smolensky, 1986). The new extensions concerning recursion (9c) are developed in Section 3.1.3.

(11) **Harmonic Principle—Symbolic formulation.**

a. At the higher level, mental processes can be analyzed as constructing the maximum-Harmony symbolic structure which includes the given input.

b. The Harmony is computed at the higher level as a function of the symbolic constituents comprising the structure:

(12) $H = \sum_{c_1,c_2} H_{c_1;c_2}$

Each $H_{c_1;c_2}$ is the Harmony of having the two symbolic constituents c_1 and c_2 in the same structure (c_1 and c_2 are constituents *in particular structural roles*; they may be the same).

c. In recursive domains, the mutual Harmony $H_{c_1;c_2}$ depends only on the relative structural positions of the two constituents c_1 and c_2, and not on the absolute positions at which they are embedded.

The most recent extensions of this principle to recursive domains (11c) is developed in Sections 3.1.2–3.1.3; the basic results (11a,b) were derived in (Legendre et al., 1990a). (12) is a straightforward mathematical consequence of the linear character of tensor product representations (7) and the bilinear nature of H (10); together, these imply that the Harmony can be computed not only at the lower, connectionist level in terms of the activation values of units, but also at the higher, symbolic level in terms of the constituents comprising a structure. For according to (10), $H = \mathbf{a}^{\mathbf{T}}\mathbf{W}\mathbf{a}$, and according to (7), if $\mathbf{a}$ represents a structure, then

$$\mathbf{a} = \sum_i \mathbf{f}_i \otimes \mathbf{r}_i = \sum_i \mathbf{c}_i$$

where the constituent i is represented by the vector $\mathbf{c}_i = \mathbf{f}_i \otimes \mathbf{r}_i$. Putting these together, we have

$$H = \mathbf{a}^{\mathbf{T}}\mathbf{W}\mathbf{a} = (\sum_i \mathbf{c}_i)^{\mathbf{T}}\mathbf{W}(\sum_j \mathbf{c}_j) = \sum_{i,j} \mathbf{c}_i^{\mathbf{T}}\mathbf{W}\mathbf{c}_j = \sum_{i,j} H_{c_i;c_j}$$

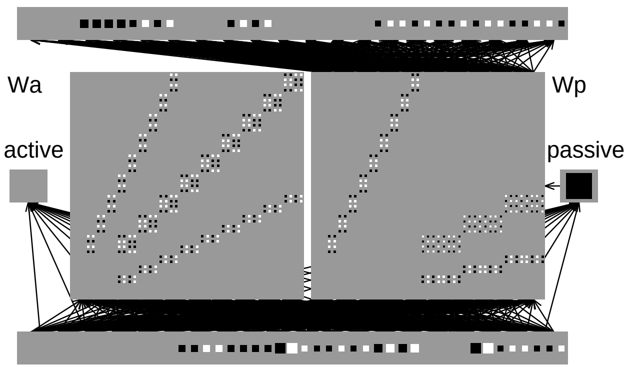
which is (12). The definition that has been used here,

(13) $H_{c_i;c_j} \equiv \mathbf{c}_i^{\mathbf{T}}\mathbf{W}\mathbf{c}_j$

provides the link between the lower and higher levels. At the higher level, $H_{c_i;c_j}$ is a simple number, the Harmony resulting from having the constituents c_i and c_j in the same structure (remembering that each such constituent c_k is a particular filler $\mathtt{f}_k$ in a particular structural role r_k). This higher-level quantity is actually computable via (13) from the lower level quantities, the numbers comprising the distributed activity vectors $\mathbf{c}_k = \mathbf{f}_k \otimes \mathbf{r}_k$ (for $k = i,\ j$) and the numbers comprising the weight matrix $\mathbf{W}$.

The computational research within SSP has developed the concrete mathematical techniques needed to perform computations using principles (6)–(11), and has realized these computations in computer simulations. One simple simulation, which we return to several times below, was designed purely to demonstrate the formal capabilities of the technique; it takes as input a distributed

Output = cons(V,cons(C,cons(A,B)))



Input = cons(cons(A,B),cons(cons(Aux,V),cons(by,C)))

Figure 1: Active/PassiveNET processing a passive sentence

pattern of activity representing the tree structure underlying an English sentence, determines by inspecting the structure whether the form is that of an active or passive sentence, and, accordingly, produces as output a distributed representation of a tree structure encoding a predicate-calculus form of the semantic interpretation of the input sentence (Legendre et al., 1991). The network, "ACTIVE/PASSIVENET", performs all the required symbol manipulation in parallel, and handles entire embedded sub-trees (e.g., complex NPs) as readily as it does simple symbols.

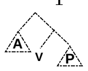
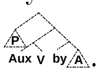
ACTIVE/PASSIVENET processes sentences with two possible syntactic structures: simple active sentences of the form  and passive sentences of the form . Each is transformed into a tree representing V(A, P), namely . Here, the agent  and patient  of the verb V are both arbitrarily complex noun phrase trees. (The network could actually handle arbitrarily complex V's as well. Aux is taken as a marker of passive, e.g., *are* in *are admired.*)

Figure 1 shows the network processing a passive sentence `((A.B).((Aux.V).(by.C)))` as in *Few connectionists are admired by Jerry* and generating `(V.(C.(A.B)))` as output.

2.2 An Example: Binary Trees

While the tensor product technique is general enough to apply to virtually any kind of symbolic structure, here we will consider only the special case of binary trees, the basic data structure of LISP.[9]

2.2.1 A Partially Distributed, Stratified Representation

The work reported in this section extends earlier results presented in (Legendre et al., 1991; Smolensky, 1990).

[9]This special case is also of particular interest to syntax. With respect to formal languages, it suffices for Context-Free Languages (using Chomsky Normal Form); with respect to natural languages, the Government-Binding Theory (Chomsky, 1981) now includes tree binarity as one of its basic principles (Kayne, 1984).

A binary tree may be viewed as having a large number of positions with various locations relative to the root: we can adopt *positional roles* r_x labelled by binary strings (or bit vectors) such as $x = 0101$ which is the position in a tree accessed by the LISP function `cadadr` .[10] Decomposing the tree using these structural roles (positions), each constituent of a tree is an atom (the filler) bound to some role r_x specifying its location. A tree $\mathtt{s}$ with a set of atoms $\{\mathtt{f}_i\}$ at respective locations $\{x_i\}$ has the tensor product representation $\mathbf{s} = \sum_i \mathbf{f}_i \otimes \mathbf{r}_{\mathbf{x}_i}$.

A more recursive view of a binary tree sees it as having only *two* constituents: the atoms or subtrees which are the left and right children of the root. In this *recursive role decomposition*, fillers may either be atoms or trees: the set of possible fillers is the same as the original set of structures S.

A *recursive* representation is one obeying, $\forall \mathtt{s}, \mathtt{p}, \mathtt{q} \in S$:

(14) $\mathtt{s} = \mathtt{cons}(\mathtt{p}, \mathtt{q}) \Rightarrow \mathbf{s} = \mathbf{p} \otimes \mathbf{r_0} + \mathbf{q} \otimes \mathbf{r_1}$

Here, $\mathtt{s} = \mathtt{cons}(\mathtt{p}, \mathtt{q})$ is the tree with left subtree $\mathtt{p}$ and right subtree $\mathtt{q}$, while $\mathbf{s}, \mathbf{p}$ and $\mathbf{q}$ are the vectors representing $\mathtt{s}, \mathtt{p}$ and $\mathtt{q}$. The only two roles in this recursive decomposition are r_0, r_1: the left and right children of root. These roles are represented by two vectors $\mathbf{r_0}$ and $\mathbf{r_1}$.

A recursive representation obeying (14) can actually be constructed from the positional representation, by assuming that the (many) positional role vectors are constructed recursively from the (two) recursive role vectors according to:[11]

(15)

$$\mathbf{r_{x0}} = \mathbf{r_x} \otimes \mathbf{r_0} \quad \mathbf{r_{x1}} = \mathbf{r_x} \otimes \mathbf{r_1}.$$

For example, $\mathbf{r_{0101}} = \mathbf{r_0} \otimes \mathbf{r_1} \otimes \mathbf{r_0} \otimes \mathbf{r_1}$. The vectors representing positions at depth d in the tree are tensors of rank d (taking the root to be depth 0). Thus the tree $\mathtt{s} = \mathtt{cons}(\mathtt{A}, \mathtt{cons}(\mathtt{B}, \mathtt{C})) = \mathtt{cons}(\mathtt{p}, \mathtt{q})$, where $\mathtt{p} = \mathtt{A}$ and $\mathtt{q} = \mathtt{cons}(\mathtt{B}, \mathtt{C})$, is represented by:

(16)

$$\begin{aligned} \mathbf{s} &= \mathbf{A} \otimes \mathbf{r_0} + \mathbf{B} \otimes \mathbf{r_{01}} + \mathbf{C} \otimes \mathbf{r_{11}} = \mathbf{A} \otimes \mathbf{r_0} + \mathbf{B} \otimes \mathbf{r_0} \otimes \mathbf{r_1} + \mathbf{C} \otimes \mathbf{r_1} \otimes \mathbf{r_1} \\ &= \mathbf{A} \otimes \mathbf{r_0} + (\mathbf{B} \otimes \mathbf{r_0} + \mathbf{C} \otimes \mathbf{r_1}) \otimes \mathbf{r_1} = \mathbf{p} \otimes \mathbf{r_0} + \mathbf{q} \otimes \mathbf{r_1}, \end{aligned}$$

in accordance with (14). a Since we are now adding tensors of different rank, the vector space V of these recursive representations is somewhat more complicated than the cases considered in (Smolensky, 1990): V is the direct sum of the spaces of tensors of different rank. An element of V can be viewed as a vector whose elements are tensors of different ranks, or, as one long vector, gotten by concatenating all the numbers comprising those tensors. In the first notation, the complete vector $\mathbf{p}$ is written

$$\mathbf{s} = \{S^{(0)}, S^{(1)}, S^{(2)}, \ldots\};$$

in the second notation,

$$\mathbf{s} = \{S^{(0)}_{\varphi}, S^{(1)}_{\varphi\rho_1}, S^{(2).}_{\varphi\rho_2\rho_1}, \ldots\}$$

[10] `cadadr(s) = car(cdr(car(cdr(s))))` , that is, the left child (0; `car`) of the right child (1; `cdr`) of the left child of the right child of the root of the tree `s` .

[11] By adopting this definition (15) of $\mathbf{r_x}$, we are essentially taking the recursive structure that is implicit in the subscripts x labelling the positional role vectors and mapping it into the structure of the vectors themselves, in such a way that this structure can be manipulated by simple connectionist processing mechanisms.

Each $S^{(d)}$ is a tensor of rank $d+1$ representing depth d in the tree; e.g., $S^{(2)}$ is a rank-3 tensor built up by adding together tensor products of the form $\mathbf{f}\otimes\mathbf{r_{x_2}}\otimes\mathbf{r_{x_1}}$ for atoms `f` at depth-2 positions $x = x_2x_1$. For example, consider **s** in (16); here,

$$S^{(2)} = \mathbf{B}\otimes\mathbf{r_0}\otimes\mathbf{r_1} + \mathbf{C}\otimes\mathbf{r_1}\otimes\mathbf{r_1}$$

The numbers comprising $S^{(2)}$ always have the form $S^{(2)}_{\varphi\rho_2\rho_1}$; in (16) these numbers are:

$$S^{(2)}_{\varphi\rho_2\rho_1} = \mathbf{B}_\varphi\mathbf{r_0}_{\rho_2}\mathbf{r_1}_{\rho_1} + \mathbf{C}_\varphi\mathbf{r_1}_{\rho_2}\mathbf{r_1}_{\rho_1}$$

In (16), there is only one atom `A` at depth 1, so $S^{(1)}$ is simply:

$$S^{(1)} = \mathbf{A}\otimes\mathbf{r_0}$$

which is comprised of the numbers

$$S^{(1)}_{\varphi\rho_1} = \mathbf{A}_\varphi\mathbf{r_0}_{\rho_1}$$

In (16), there are no atoms at any depths other than 1 and 2, so all the other tensors $S^{(d)}$ for $d \neq 1, 2$ are zero.

In general, the tensor $S^{(d)}$ for depth d has one subscript φ for the filler vectors (e.g., **B** and **C**), and d subscripts $\rho_d\rho_{d-1}\cdots\rho_2\rho_1$ for the depth-d role vectors

$$\mathbf{r_x} = \mathbf{r_{x_d x_{d-1} \cdots x_2 x_1}} = \mathbf{r_{x_d}}\otimes\mathbf{r_{x_{d-1}}}\otimes\cdots\otimes\mathbf{r_{x_2}}\otimes\mathbf{r_{x_1}}$$

Thus $S^{(d)}$ is rank-1 with respect to fillers, rank-d with respect to roles, and rank-$(d+1)$ overall.

The vector operation **cons** for building the representation of a tree from that of its two subtrees is given by (14). As an operation on V this can be written:

$$\begin{aligned}&\mathbf{cons} : (\{P^{(0)}_\varphi, P^{(1)}_{\varphi\rho_1}, P^{(2)}_{\varphi\rho_2\rho_1}, \ldots\}, \{Q^{(0)}_\varphi, Q^{(1)}_{\varphi\rho_1}, Q^{(2)}_{\varphi\rho_2\rho_1}, \ldots\}) \mapsto \\ &\quad \{\mathbf{0}, P^{(0)}_\varphi\mathbf{r_0}_{\rho_1}, P^{(1)}_{\varphi\rho_2}\mathbf{r_0}_{\rho_1}, \ldots\} + \{\mathbf{0}, Q^{(0)}_\varphi\mathbf{r_1}_{\rho_1}, Q^{(1)}_{\varphi\rho_2}\mathbf{r_1}_{\rho_1}, \ldots\},\end{aligned}$$

or, more compactly:

$$\begin{aligned}&\mathbf{cons} : (\{P^{(0)}, P^{(1)}, P^{(2)}, \ldots\}, \{Q^{(0)}, Q^{(1)}, Q^{(2)}, \ldots\}) \mapsto \\ &\quad \{\mathbf{0}, P^{(0)}\otimes\mathbf{r_0}, P^{(1)}\otimes\mathbf{r_0}, \ldots\} + \{\mathbf{0}, Q^{(0)}\otimes\mathbf{r_1}, Q^{(1)}\otimes\mathbf{r_1}, \ldots\}.\end{aligned}$$

(Here, **0** denotes the zero vector in the space representing atoms.) Using matrix multiplication in V, this can simply be written:

(17) $$\mathbf{cons}(\mathbf{p}, \mathbf{q}) = \mathbf{W_{cons0}}\ \mathbf{p} + \mathbf{W_{cons1}}\ \mathbf{q}$$

(parallel to (14)), where the non-zero elements of the matrix $\mathbf{W_{cons0}}$ are given by:

$$\mathbf{W_{cons0}}^{(d+1,d)}_{\varphi\rho_{d+1}\rho_d\cdots\rho_2\rho_1,\varphi\rho_{d+1}\rho_d\cdots\rho_2} = \mathbf{r_0}_{\rho_1}$$

and $\mathbf{W_{cons1}}$ obeys the same equation, with $\mathbf{r_1}$ replacing $\mathbf{r_0}$. This is the block of the full matrix $\mathbf{W_{cons0}}$which maps depth d trees to depth $d+1$ trees.

For later purposes it is worthwhile to work this equation into another, eventually simpler, form. We start by rewriting it as:

$$\mathbf{W_{cons0}}^{(d+1,d)}_{\varphi\rho_{d+1}\rho_d\cdots\rho_2\rho_1,\varphi'\rho'_d\rho'_{d-1}\cdots\rho'_2\rho'_1} = \delta_{\varphi,\varphi'}\delta_{\rho_{d+1},\rho'_d}\delta_{\rho_d,\rho'_{d-1}}\cdots\delta_{\rho_2,\rho'_1}\mathbf{r_0}_{\rho_1}$$

where

(18) $$\delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

are the elements of the identity matrix $\mathcal{I}$. Using the matrix $\mathcal{I}$ we can write it more succinctly as:

$$\mathbf{W_{cons0}}^{(d+1,d)} = \mathcal{I} \otimes \mathcal{I} \otimes \cdots \mathcal{I} \otimes \mathbf{r_0}$$

where there are d factors of $\mathcal{I}$. The full matrix $\mathbf{W_{cons0}}$ has one such block for each depth d:

$$\mathbf{W_{cons0}} = \mathbf{r_0} + \mathcal{I} \otimes \mathbf{r_0} + \mathcal{I} \otimes \mathcal{I} \otimes \mathbf{r_0} + \mathcal{I} \otimes \mathcal{I} \otimes \mathcal{I} \otimes \mathbf{r_0} + \ldots$$

If we now define the "recursion matrix"

(19) $$\mathcal{R} \equiv 1 + \mathcal{I} + \mathcal{I} \otimes \mathcal{I} + \mathcal{I} \otimes \mathcal{I} \otimes \mathcal{I} + \cdots$$

then we get a very simple expression for $\mathbf{W_{cons0}}$:

(20) $$\mathbf{W_{cons0}} = \mathcal{R} \otimes \mathbf{r_0}$$

We will later interpret this equation as simply saying that $\mathbf{W_{cons0}}$ is the recursive version of the operation of multiplying by $\mathbf{r_0}$; this weight matrix "pushes" or embeds a tensor representing a tree into the role of left child of the root of a new tree; combining this with $\mathbf{W_{cons1}}$ gives the **cons** operation (17).

Taking the `car` or `cdr` of a tree `s` —extracting its left or right subtree—is equivalent to "unbinding" r_0 or r_1; assuming the role vectors to be linearly independent, these unbinding operations can be performed accurately, via linear operations **car** and **cdr** (Smolensky, 1990, Section 3.1): the generalized inner product (tensor contraction) of $\mathbf{s}$ with an "unbinding vector" $\mathbf{u_0}$ or $\mathbf{u_1}$. Unbinding vectors form the dual basis to the role vectors: they comprise the inverse matrix to the matrix of all role vectors. **car** can be realized as multiplication in V by the matrix $\mathbf{W_{car}}$ with non-zero elements:

$$\mathbf{W_{car}}^{(d,d+1)}_{\varphi\rho_{d+1}\rho_d\cdots\rho_2,\varphi\rho_{d+1}\rho_d\cdots\rho_2\rho_1} = \mathbf{u_0}_{\rho_1}$$

and **cdr** by the corresponding matrix $\mathbf{W_{cdr}}$ with $\mathbf{u_0}$ replaced by $\mathbf{u_1}$. This matrix block maps tree depth $d+1$ to depth d. Using the recursion matrix $\mathcal{R}$, the full matrix $\mathbf{W_{car}}$ for all depths can be written simply as:

(21) $$\mathbf{W_{car}} = \mathcal{R} \otimes \mathbf{u_0}^{\mathbf{T}}$$

This connectionist representation of trees enables massively parallel processing. Whereas in the traditional sequential implementation of LISP, symbol processing consists of a long sequence of `car`, `cdr`, and `cons` operations, here we can compose together the corresponding sequence of $\mathbf{W_{car}}$, $\mathbf{W_{cdr}}$, $\mathbf{W_{cons0}}$ and $\mathbf{W_{cons1}}$ operations into a single matrix operation. Adding some minimal nonlinearity allows us to compose more complex operations incorporating the equivalent of conditional branching. For example, in ACTIVE/PASSIVENET, the parse tree `s` of a passive sentence is transformed to `cons(cdadr(s), cons(cdddr(s), car(s)))` by the matrix $\mathbf{W_p} = \mathbf{W_{cons0}}\mathbf{W_{cdr}}\mathbf{W_{car}}\mathbf{W_{cdr}} + \mathbf{W_{cons1}}(\mathbf{W_{cons0}}\mathbf{W_{cdr}}\mathbf{W_{cdr}}\mathbf{W_{cdr}} + \mathbf{W_{cons1}}\mathbf{W_{car}})$. In the terminology of production systems, this "action matrix" $\mathbf{W_p}$ is gated by a "condition unit" which determines whether the input pattern $\mathbf{s}$ represents the parse tree `s` of a passive sentence, by checking whether `caddr(s)` = `Aux`. This condition unit can be a linear threshold element whose activity is 1 when its net input $I \geq 0$, and 0 otherwise; its net input $I = -\|\mathbf{W_{car}}\mathbf{W_{cdr}}\mathbf{W_{cdr}}\mathbf{s} - \mathbf{Aux}\|^2$ is always negative, except that $I = 0$ when the desired condition `caddr(s)` = `Aux` is satisfied.

2.2.2 Fully Distributed Recursive Representations

The recursive representation of binary trees described in the previous subsection might be called a *stratified* representation: different levels of the tree are represented separately (over different connectionist units)—they are essentially *concatenated* by the direct sum in the construction of the vector space V, rather than truly superimposed, as in fully distributed representations. Among other things, this means that a finite-sized network will have a sharp cutoff in the depth of trees it can represent, rather than displaying *graceful saturation*—where the accuracy with which information in a tree is represented gracefully degrades beyond a certain depth. We will now see how to overcome this limitation, without sacrificing the recursive character of the representation. The representation space is actually somewhat simpler than in the stratified representation, and the equations are basically the same, but the operators involved are somewhat more complex. The new technique, as before, can be studied by a combination of mathematical analysis and computer simulation.

The crux of the idea is to add to the fundamental role vectors $\{\mathbf{r_0}, \mathbf{r_1}\}$ of the stratified representation a third vector $\mathbf{v}$ which serves basically as a place holder, like the digit 0. Instead of representing an atom `B` at position r_{01} by $\mathbf{B}\otimes\mathbf{r_0}\otimes\mathbf{r_1}$, we use $\mathbf{B}\otimes\mathbf{v}\otimes\mathbf{v}\otimes\ldots\otimes\mathbf{v}\otimes\mathbf{r_0}\otimes\mathbf{r_1}$, using as many $\mathbf{v}$s as necessary to pad the total tensor product out to produce a tensor of some rank $D+1$. Now, atoms at all depths are represented by tensors of the same rank; the new vector space of representations of binary trees is just a space V' of tensors of rank $D+1$, and the representations of all atoms can fully superimpose: this representation is *fully distributed.*

Trees up to depth D can now be represented with complete accuracy (assuming the three vectors $\{\mathbf{r_0}, \mathbf{r_1}, \mathbf{v}\}$ are linearly independent). The stratified representation of Section 2.2 can be straightforwardly embedded as a special case of this new fully distributed representation by mapping $\mathbf{r_0} \to (\mathbf{r_0}, 0)$, $\mathbf{r_1} \to (\mathbf{r_1}, 0)$ and by setting $\mathbf{v} \equiv (\mathbf{0}, 1)$ where $\mathbf{0}$ is the zero vector with the same dimensionality as $\mathbf{r_0}$ and $\mathbf{r_1}$. That is, the special case in which $\{\mathbf{r_0}, \mathbf{r_1}, \mathbf{v}\}$ is decomposed as the direct sum of the 2-dimensional space spanned by the old vectors $\{\mathbf{r_0}, \mathbf{r_1}\}$ and the one-dimensional space spanned by $\mathbf{v}$ reduces the new fully distributed representation to the direct-sum-across-depths stratified representation. But the general case is gotten by choosing $\{\mathbf{r_0}, \mathbf{r_1}, \mathbf{v}\}$ to be, say, three linearly independent vectors in three-space, each with non-zero components along all three coordinate axes; in this case, every unit in the connectionist network will take part in the representation of every atom, regardless of its depth in the tree. As before, for extracting tree elements, we need the unbinding vectors $\{\mathbf{u_0}, \mathbf{u_1}, \mathbf{u}\}$ that form the basis dual to $\{\mathbf{r_0}, \mathbf{r_1}, \mathbf{v}\}$.

The recursive characterization of this new representation requires the following new operations:

$$\mathbf{w} \odot \mathbf{T} \equiv \sum_{\rho_D} \mathbf{w}_{\rho_D} \mathbf{T}_{\varphi\rho_D\cdots\rho_2\rho_1} \qquad \mathbf{T} \odot \mathbf{w} \equiv \sum_{\rho_1} \mathbf{T}_{\varphi\rho_D\cdots\rho_2\rho_1} \mathbf{w}_{\rho_1}$$
$$\mathbf{T} \circ \mathbf{w} \equiv \mathbf{v}\otimes(\mathbf{T} \odot \mathbf{w}) \qquad \mathbf{T} \star \mathbf{w} \equiv (\mathbf{u} \odot \mathbf{T})\otimes\mathbf{w}$$

$\mathbf{w} \odot \mathbf{T}$ is an inner product of $\mathbf{w}$ with the left-most ("deepest") role-index of $\mathbf{T}$, and $\mathbf{T} \odot \mathbf{w}$, with the right-most ("highest"); the resulting tensor has rank one less than that of $\mathbf{T}$. $\mathbf{T} \circ \mathbf{w}$ is a rank-preserving inner product, a version of $\mathbf{T} \odot \mathbf{w}$ in which a right inner product with $\mathbf{w}$ is taken, and an extra $\mathbf{v}$ is added to pad the new tensor back up to full rank. $\mathbf{T} \star \mathbf{w}$ is a rank-preserving outer product, a version of $\mathbf{T}\otimes\mathbf{w}$ in which the tensor product with $\mathbf{w}$ is taken only after "unpadding" $\mathbf{T}$ via an inner product with $\mathbf{u}$, the vector dual to $\mathbf{v}$.

Now the new equations generalizing those of Section 2.2 are:

(22)

$$\mathtt{s} = \mathtt{cons}(\mathtt{p},\mathtt{q}) \;\Rightarrow\; \mathtt{s} = \mathbf{p} \star \mathbf{r_0} + \mathbf{q} \star \mathbf{r_1}$$
$$\mathtt{p} = \mathtt{car}(\mathtt{s}) \Rightarrow \mathbf{p} = \mathtt{s} \circ \mathbf{u_0} \qquad \mathtt{q} = \mathtt{cdr}(\mathtt{s}) \Rightarrow \mathbf{q} = \mathtt{s} \circ \mathbf{u_1}$$

The equations, like (14), characterizing the recursive properties of the stratified representation still hold, except that the rank-altering inner- and outer-product operations $\odot$ and $\otimes$ of Section 2.2 are now replaced by their rank-preserving counterparts $\circ$ and $\star$. As before, all operations can be readily realized as matrix operations on the representational vector space, now V'. ACTIVE/PASSIVENET has been reimplemented with this new technique, using a representation in which all connectionist units participate in the representation of all tree depths. As before, the network behaves perfectly, as indeed it provably must.

Having produced a fully distributed recursive representation, it remains to develop specific ways to exploit the full advantages of distributed representation. One such direction is discussed in Section 3.1.6; here, we consider only the issue of graceful saturation with depth. One general technique is to take a high (perhaps infinite) dimensional, fully accurate representational space and to project onto a lower dimensional subspace. Here we want to pick such a subspace so that less accuracy is associated with greater depth. A promising idea for how to achieve this comes from revisiting the means described above for specializing the fully distributed representation to the stratified representation. Strict depth stratification arises from choosing $\mathbf{v} = (\mathbf{0}, 1)$ because $\mathbf{v}$ is orthogonal to the subspace spanned by $\{\mathbf{r_0}, \mathbf{r_1}\}$. If instead we choose $\mathbf{v}$ to have small but non-zero projection onto this subspace, each depth will have small but non-zero representation on subspaces that are primarily dedicated to the representation of lesser depths. We can therefore set up a fully distributed representation with some large depth limit D (possibly infinite), and then project onto a lower-dimensional subspace, achieving a "soft depth limit" beyond which the representation saturates gracefully.

2.3 TPPL

We have begun to develop an explicit symbolic formalism—TPPL for "Tensor Product Programming Language"—which enables high-level formal characterization of the computations performed by connectionist networks processing tensor product representations. TPPL contains analogues of simple programming language control structures (like `if-then-else`) and basic symbolic computation operators (like `car`, `cdr`, `cons`) which are formally defined using elements of the tensor calculus. TPPL enables both a formal higher-level symbolic description of the lower-level networks and a calculus for proving their correctness.

To illustrate the idea, ACTIVE/PASSIVENET is described in TPPL by the program:

$$\begin{aligned}
AP(s) &\equiv \mathtt{if}\ PassiveP(s)\ \mathtt{then}\ PassiveF(s)\ \mathtt{else}\ ActiveF(s) \\
PassiveP(s) &\equiv [PassiveMarkerF(s) = \mathtt{Aux}] \\
PassiveMarkerF(s) &\equiv \mathtt{car}(\mathtt{cdr}(\mathtt{cdr}(s))) \\
PassiveF(s) &\equiv \mathtt{cons}(\mathtt{cdr}(\mathtt{car}(\mathtt{cdr}(s))), \mathtt{cons}(\mathtt{cdr}(\mathtt{cdr}(\mathtt{cdr}(s))), \mathtt{car}(s))) \\
ActiveF(s) &\equiv \mathtt{cons}(\mathtt{car}(\mathtt{cdr}(s)), \mathtt{cons}(\mathtt{car}(s), \mathtt{cdr}(\mathtt{cdr}(s))))
\end{aligned}$$

The primitive operations `car, cdr, cons` are defined using the inner- and outer-product operations of tensor calculus, as already explained for the stratified representation in Section 2.2.1 and as extended to the fully distributed representation in Section 2.2.2. As illustrated for $PassiveF$ in Section 2.2.1, the matrix realizations of these operations are straightforwardly combined to produce a single matrix that performs the entire function $PassiveF$, which can then be implemented in one layer of connection weights. The same applies to the other functions, $ActiveF$ and $PassiveMarkerF$.

Predicates like $PassiveP$ of the general form[12] $[s = r]$ are defined by the function $\theta(\|\mathbf{s} - \mathbf{r}\|)$, where $\theta(x)$ is 1 if $x = 0$ and 0 otherwise (or some smoothed version such as the Gaussian $\theta(x) = e^{-\beta x^2}$). This can then be conveniently implemented in the connectionist network by a linear threshold unit, or, perhaps more naturally, by a radial basis function unit.

Finally, the construct `if` P `then` F_1 `else` F_2 is defined as the function $P \cdot \mathbf{F}_1 + (1 - P) \cdot \mathbf{F}_2 = P \cdot (\mathbf{F}_1 - \mathbf{F}_2) + \mathbf{F}_2$. This is in turn implemented either (i) by having the output of the unit u implementing P gate (through multiplicative connections) the connections implementing $\mathbf{F}_1$, while $1 - u$ gates the connections implementing $\mathbf{F}_2$; or (ii) by having u gate connections implementing $\mathbf{F}_1 - \mathbf{F}_2$ while connections implementing $\mathbf{F}_2$ are ungated.

TPPL provides a third level for formally describing networks like ACTIVE/PASSIVENET for processing recursive tensor product representations. The lowest level is that of the individual units and connections of a connectionist network which employs fairly traditional machinery such as radial basis function units and multiplicative connections to perform fully distributed, massively parallel computation. The next higher level uses tensor calculus and θ functions to concisely but precisely describe the numerical vector processing which the net implements. Finally, TPPL provides the highest-level, but still formally precise, description of the system as performing *symbolic structure processing.* These TPPL programs describe the symbolic function being computed using expressions like "`cons(cdr(car(cdr(`s`))), cons(cdr(cdr(cdr(`s`))), car(`s`)))`", but the actual computation does *not* involve sequential application of `car`, `cdr`, and `cons`. It is possible to define each level of description in terms of the one below it with sufficient rigor to allow proofs of correctness which ensure that the network computes the symbolic mapping defined by the TPPL program.

2.4 Contracting Representations and Human Memory Models

Tensor product representations allow the exact representation of complex structures, but at the cost of large connectionist networks when deep structure is involved. Tensor calculus provides techniques of *tensor contraction* which reduce the dimensionality of tensors by dropping and summing elements.[13] One direction of future research is investigation of the consequences for representation and processing of reducing the size of tensor product representations via contraction. In addition to computational motivations, this investigation has a psychological one as well. Evidence for the psychological reality of tensor product representations comes from a line of psychological research into human memory (Anderson et al., 1984; Humphreys et al., 1989; Metcalfe-Eich, 1982; Murdock, 1982; Pike, 1984; Wiles et al., 1990); this research shows that memory traces can often be well modelled by what have been called "matrix," "correlation," or "holographic" representations of stimuli. These representations, several of which have been recently studied from a more computational perspective in (Plate, 1991), can be analyzed as either simple cases of tensor product representations, or contracted versions of them; these reduced tensor product representations

[12] Note that this allows conditions such as `caddr(`s`)` = `cadr(`r`)` in which neither element of the comparison is a constant.

[13] Contraction can be viewed as a generalization to arbitrary rank-tensors of the *trace* operation in matrix algebra, where the matrix elements $\{T_{ij}\}$ are collapsed to the single number $\sum_k T_{kk}$ by summing the diagonal elements and ignoring the others.

appear to be adequate, at least for stimuli with only a small degree of structure.

2.5 Further Computational Principles

Following the goal of Principle-Centering (3b), SSP requires a solid mathematical framework of connectionist computation within which to develop its cognitive principles; while this framework has made major progress, there remains a long way to go. The research discussed in this paper exploits principles of representation and processing (6)–(11) that grow out of only two techniques, tensor algebra and Lyapunov functions. While we are arguing that the research described here shows the value of these techniques for addressing central issues in higher cognition, it is obvious that we also need many other principles that derive from the wide variety of mathematical approaches to the study of connectionist and symbolic computation. [14]

One important problem for which principles are badly needed is the characterization of the representations and knowledge developed in connectionist networks with hidden units. Two examples of techniques for addressing this problem which are among those that may lead eventually to solid principles, and which are referred to below, are *skeletonization* (Mozer and Smolensky, 1989b)—a technique of incrementally and automatically eliminating from a network the least important connections or units—and *contribution analysis*, a rather complex statistical technique for automatically determining the responsibilities of hidden units (Sanger, 1989; Sanger, 1990).

2.6 Methodology

This project is made possible by methodological innovations intimately involving all the Three Goals. The work requires identifying what is central in symbolic computation for theories of higher-level cognition, and how that can be formalized within discrete mathematics; what is central in the higher-level characterization of connectionist computation, and how that can be formalized within continuous mathematics; and then how these elements of continuous and discrete mathematics can be unified in terms of general principles, concrete means of formal calculation, and computer simulation.

3 Optimization in Grammar

Language has always played a special role in cognitive science, and poses a special challenge for a connectionist-grounded theory. The theory of symbolic computation is intimately bound with the theory of formal languages; language-like representations have long been central in cognitive theory; Chomsky's work on the grammars of formal languages and on the formal grammars of natural languages has always defined a central place in the field; and Chomsky's arguments for the combinatorial, recursive character of the knowledge of language have been taken to generalize to much of higher cognition. For all these reasons and more it is important for any theory of higher cognition, especially one attributing a significant role to connectionist computation, to address the

[14] The approaches to mathematical analysis of connectionist computation represented in (Smolensky et al., tion), for example, range from computability with and computational complexity of networks of threshold elements, to numerical analysis, the theory of analog computation, dynamical systems theory, control theory, information theory, statistical mechanics, and statistical estimation theory. All these mathematical perspectives on connectionist computation, and others as well, are likely to contribute principles which are at least as important for higher cognition as those discussed in this paper.

problem of language, and of grammar in particular. In this section we take up the SSP research on grammar.

It will be useful to distinguish two ways of viewing the grammar of a language: descriptively, as a function which identifies, via an abstract specification, the correct linguistic structure to output for each given input, and algorithmically, as a device for actually constructing this output.[15]

In descriptive grammar, as studied particularly within theoretical linguistics, an extremely powerful methodology has developed in which a central role is played by the study of the *well-formedness* of various structures. This notion applies not only to traditional linguistic problems such as those of phonology and syntax, but also to problems of central interest to natural language processing and computational linguistics, such as semantic interpretation: purely syntactic structures such as parse trees are not the only structures that can be separated by a grammar into those which are well- and ill-formed—the same is true of structures that combine both syntactic and semantic information. Thus, e.g., in a number of unification-based approaches to syntax and semantics, the "correct" semantic interpretation of an input sentence is analyzed as the semantic part of the well-formed structure which contains the input, together with associated syntactic and semantic information (e.g., (Shieber, 1986)).

Thus a powerful concept around which to build a connectionist-grounded theory of grammar is that of linguistic well-formedness. And the principles (6)–(11) turn out to be exactly what we need to do just that. The further fundamental principles underlying current SSP work on grammar (Legendre et al., 1990a), and building directly on (11), begin with:

(23) **Fundamental principle of Harmonic Grammar—General formulation**

a. The well-formedness of a linguistic structure is measured by the Harmony of that structure.

b. Descriptively, the grammar assigns to an input that linguistic structure which is most well-formed, i.e., has maximal Harmony. The descriptive grammar can therefore be specified by the Harmony function itself, which measures the well-formedness of all possible linguistic representations that could be assigned to an input.

c. Algorithmically, the grammar is a Harmony-maximizing connectionist network, the Harmony function of which specifies the descriptive grammar.

This and subsequent principles constitute a formal development of conceptual ideas linking Harmony to linguistics which were first proposed in Lakoff's *cognitive phonology* (Lakoff, 1988; Lakoff, 1989) and Goldsmith's *harmonic phonology* (Goldsmith, 1990; Goldsmith, In press b).

The theory now proceeds along two somewhat different paths, which further specify (23) in different ways. The first is a numerical formulation, the second a non-numerical formulation using a kind of abstract albegra to replace numerical calculation.

---

[15]For example, the phonological component of a grammar might receive as its input a string of phonemes constructed in the morphological component by concatenating the phonemes of a verb stem with the phonemes of a verb ending. The phonological component's job might then be to output a structure in which (a) some phonemes may have been altered to meet various phonological constraints in the language, and (b) hierarchical structure has been added which groups phonemes into syllables, syllables into metrical feet, etc., and (c) prominence structure has been added, marking varying degrees of stress on syllables, etc. Another example would be the syntactic/semantic component of a grammar, which might receive as input a string of word tokens, and might produce as output a structure in which (a) the words are tagged as to lexical class; (b) the string is parsed into phrases; and (c) semantic structures capturing various aspects of the string's meaning are included.

3.1 Numerical Theory

The numerical theory is the futher specification of (23) which follows one most directly from the preceding principles.

3.1.1 Principles

The fundamental principle of the numerical theory is a direct consequence of (9) (Legendre et al., 1990a):

(24) **Fundamental principle of Harmonic Grammar—Numerical formulation**

a. The explicit form of the Harmony function can be computed to be a sum of terms each of which measures the well-formedness arising from the coexistence, within a single structure, of a pair of constituents in their particular structural roles.

b. The descriptive grammar can thus be identified as a set of *soft rules* each of the form:[16]

> If a linguistic structure S simultaneously contains constituent c_1 in structural role r_1 and constituent c_2 in structural role r_2, then add to $H(S)$, the Harmony value of S, the quantity $H_{c_1,r_1;c_2,r_2}$ (which may be positive or negative).

A set of such soft rules or constraints defines a *Harmonic Grammar.*

c. The constituents in the soft rules include both those that are given in the input and the "hidden" constituents that are assigned to the input by the grammar. The problem for the algorithmic grammar is to construct that structure S, containing both input and "hidden" constituents, with the highest overall Harmony $H(S)$.

The distinction between well- and ill-formed inputs, according to this theory, is a numerically graded one: the higher the value of $H(S)$ for the structure S assigned by the grammar to an input, the more well-formed is that input. The soft rules in a Harmonic Grammar can potentially interact very strongly; the Harmony-maximizing structure, and its degree of well-formedness, can be highly sensitive to combinations of factors in the input.

[16] We have yet to work out in any detail the relation between the numerical Harmony values or 'strengths' attached to soft rules and the probability values attached to rules in various stochastic grammar models, e.g., variable rules (Labov, 1969; Cedergren and Sankoff, 1974) or probabilistic formal grammars (see, e.g., (Resnick, 1992)). The Harmony value appearing in a soft rule is not itself a probability, and not directly a measure of the variability with which the rule is followed. On the other hand, as developed in detail in the original Harmony Theory (Smolensky, 1983; Smolensky, 1986), there is a direct mathematical relation between Harmony values and probabilities; very crudely, the Harmony values are logarithms of probabilities. If the principles of Harmony Theory relating to stochastic inference were to be added to the theory presented in this paper, then explicit connections would appear between probabilities of constraint violation—one kind of variability—and the Harmony values appearing in the soft rules. Two important caveats are immediately required, however. First, the conceptual and technical issues surrounding probabilistic grammars or variable rules are much too complex (e.g., (Kay and McDaniel, 1979)) to prejudge the outcome of such a development. Secondly, and relatedly, the correct relation between probabilities and Harmony values is quite indirect, and *it would be* **completely** *mistaken to believe that the Harmony values in soft rules correlate directly with the* **frequency** *with which they are violated.* What this error most crucially ignores is the highly varying degree to which rules come into conflict with each other. When two rules conflict, the strengths of the rules govern which prevails; but in the absence of detailed knowledge of how frequently a rule comes in conflict with stronger rules, we have no way of determining the frequency of its violation.

Following goal (2), Methodological Integration, in applying Harmonic Grammar to a particular linguistic phenomenon in natural language, we take as a starting point the working hypothesis that the particular kinds of constituent structures posited by the best current linguistic theories of that phenomenon are indeed valid higher level descriptions of the relevant mental representations. We then study patterns of well-formedness judgements to identify candidate constituent interactions; this gives us a set of candidate soft rules (24b), in which the numerical constants $H_{c_1,r_1;c_2,r_2}$ are unknown. *These numerical values can then be automatically determined* from the well-formedness judgements elicited from native speakers by an appropriately generalized version of the connectionist learning algorithm, back-propagation (Rumelhart et al., 1986). But before describing an application of Harmonic Grammar to natural language syntax/semantics, we will turn to some issues in the syntax of formal languages.

3.1.2 Context-Free Harmonic Grammars

One means for assessing the expressive power of Harmonic Grammar (HG) is to apply it to the specification of formal languages. Can, e.g., any Context-Free Language (CFL) L be specified by an HG? Can a set of soft rules of the form (24b) be given so that a string $s \in L$ iff the maximum-Harmony tree with s as terminals has, say, $H \geq 0$? A crucial limitation of these soft rules is that each may only refer to a *pair* of constituents: in this sense, they are only *second order.* (It simplifies the exposition to describe as "pairs" those in which both constituents are the same; these actually correspond to first order soft rules, which also exist in HG.)

For a CFL, a tree is well-formed iff all of its *local trees* are—where a local tree is just some node and all its children. Thus the HG rules need only refer to pairs of nodes which fall in a single local tree, i.e., parent-child pairs and/or sibling pairs. The H value of the entire tree is just the sum of all the numbers for each such pair of nodes given by the soft rules defining the HG.

It is clear that for a general context-free grammar (CFG), pairwise evaluation doesn't suffice. Consider, e.g., the following CFG fragment, $G_0 : A \rightarrow B\ C,\ A \rightarrow D\ E,\ F \rightarrow B\ E$, and the ill-formed local tree $(A\ ;\ (B\ E))$ (here, A is the parent, B and E the two children). Pairwise well-formedness checks fail to detect the ill-formedness, since the first rule says B can be a left child of A, the second that E can be a right child of A, and the third that B can be a left sibling of E. The ill-formedness can be detected only by examining *all three* nodes simultaneously, and seeing that this triple is not licensed by any single rule.

One possible approach would be to extend HG to rules higher than second order, involving more than two constituents; this corresponds to H functions of degree higher than 2. Such H functions go beyond standard connectionist networks with pairwise connectivity, requiring networks defined over hypergraphs rather than ordinary graphs. There is a natural alternative, however, that requires no change at all in HG, but instead adopts a special kind of grammar for the CFL. The basic trick is a modification of an idea taken from Generalized Phrase Structure Grammar (Gazdar et al., 1985), a theory that adapts CFGs to the study of natural languages.

It is useful to introduce a new normal form for CFGs, *Harmonic Normal Form* (HNF). In HNF, all rules of are three types: $A[i] \rightarrow B\ C,\ A \rightarrow a$, and $A \rightarrow A[i]$; and there is the further requirement that there can be only one branching rule with a given left hand side—the *unique branching condition.* Here we use lowercase letters to denote terminal symbols, and have two sorts of nonterminals: general symbols like A and *subcategorized* symbols like $A[1], A[2], ..., A[i]$. To see that every CFL L does indeed have an HNF grammar, it suffices to first take a CFG for L in Chomsky Normal Form, and, for each (necessarily binary) branching rule $A \rightarrow B\ C$, (i) replace the symbol A on the left hand side with $A[i]$, using a different value of i for each branching rule with a given left hand side, and (ii) add the rule $A \rightarrow A[i]$.

Subcategorizing the general category A, which may have several legal branching expansions, into the specialized subcategories $A[i]$, each of which has only one legal branching expansion, makes it possible to determine the well-formedness of an entire tree simply by examining each parent/child pair separately: an entire tree is well-formed iff every parent/child pair is. The unique branching condition enables us to evaluate the Harmony of a tree simply by adding up a collection of numbers (specified by the soft rules of an HG), one for each node and one for each link of the tree. Now, any CFL L can be specified by a Harmonic Grammar. First, find an HNF grammar G_{HNF} for L; from it, generate a set of soft rules defining a Harmonic Grammar G_H via the correspondences:

G_{HNF}	G_H
a	R_a: If a is at any node, add -1 to H
A	R_A: If A is at any node, add -2 to H
$A[i]$	$R_{A[i]}$: If $A[i]$ is at any node, add -3 to H
start symbol S	R_{root}: If S is at the root, add $+1$ to H
$A \rightarrow \alpha$ ($\alpha = a$ or $A[i]$)	If α is a left child of A, add $+2$ to H
$A[i] \rightarrow B\ C$	If B is a left child of $A[i]$, add $+2$ to H
	If C is a right child of $A[i]$, add $+2$ to H

The soft rules R_a, R_A, $R_{A[i]}$ and R_{root} are first-order and evaluate tree nodes; the remaining second-order soft rules are *legal domination* rules evaluating tree links.

This HG assigns $H = 0$ to any legal parse tree (with S at the root), and $H < 0$ for any other tree; thus $s \in L$ iff the maximal-Harmony completion of s to a tree has $H \geq 0$.

Proof. We evaluate the Harmony of any tree by conceptually breaking up its nodes and links into pieces each of which contributes either $+1$ or -1 to H. In legal trees, there will be complete cancellation of the positive and negative contributions; illegal trees will have uncancelled -1s leading to a total $H < 0$.

The decomposition of nodes and links proceeds as follows. Replace each (undirected) link in the tree with a pair of directed links, one pointing up to the parent, the other down to the child. If the link joins a legal parent/child pair, the corresponding legal domination rule will contribute $+2$ to H; break this $+2$ into two contributions of $+1$, one for each of the directed links. We similarly break up the non-terminal nodes into sub-nodes. A non-terminal node labelled $A[i]$ has two children in legal trees, and we break such a node into three sub-nodes, one corresponding to each downward link to a child and one corresponding to the upward link to the parent of $A[i]$. According to soft rule $R_{A[i]}$, the contribution of this node $A[i]$ to H is -3; this is distributed as three contributions of -1, one for each sub-node. Similarly, a non-terminal node labelled A has only one child in a legal tree, so we break it into two sub-nodes, one for the downward link to the only child, one for the upward link to the parent of A. The contribution of -2 dictated by soft rule R_A is similarly decomposed into two contributions of -1, one for each sub-node. There is no need to break up terminal nodes, which in legal trees have only one outgoing link, upward to the parent; the contribution from R_a is already just -1.

We can evaluate the Harmony of any tree by examining each node, now decomposed into a set of sub-nodes, and determining the contribution to H made by the node and its *outgoing* directed links. We will not double-count link contributions this way; half the contribution of each original undirected link is counted at each of the nodes it connects.

Consider first a non-terminal node n labelled by $A[i]$; if it has a legal parent, it will have an upward link to the parent that contributes $+1$, which cancels the -1 contributed by n's corresponding sub-node. If n has a legal left child, the downward link to it will contribute $+1$, cancelling the -1 contributed by n's corresponding sub-node. Similarly for the right child. Thus the total contribution of this node will be 0 if it has a legal parent and two legal children. For each *missing* legal child or parent, the node contributes an uncancelled -1, so the contribution of this node n in the general case is:

(25) $H_n = -$(the number of missing legal children and parents of node n)

The same result (25) holds of the non-branching non-terminals labelled A; the only difference is that now the only child that could be missing is a legal left child. If A happens to be a legal start symbol in root position, then the -1 of the sub-node corresponding to the upward link to a parent is cancelled not by a legal parent, as usual, but rather by the $+1$ of the soft rule R_{root}. The result (25) still holds even in this case, if we simply agree to count the root position itself as a legal parent for start symbols. And finally, (25) holds of a terminal node n labelled a; such a node can have no missing child, but might have a missing legal parent.

Thus the total Harmony of a tree is $H = \sum_n H_n$, with H_n given by (25). That is, H is the *minus* the total number of missing legal children and parents for all nodes in the tree. Thus, $H = 0$ if each node has a legal parent and all its required legal children, otherwise $H \leq 0$. Because the grammar is in Harmonic Normal Form, a parse tree is legal iff every every node has a legal parent and its required number of legal children, where "legal" parent/child dominations are defined only pairwise, in terms of the parent and one child, blind to any other children that might be present or absent. Thus we have established the desired result, that the maximum-Harmony parse of a string s has $H \geq 0$ iff $s \in L$.

We can also now see how to understand the soft rules of G_H, and how to generalize beyond Context-Free Languages. The soft rules say that each node makes a negative contribution equal to its valence, while each link makes a positive contribution equal to its valence (2); where the "valence" of a node (or link) is just the number of links (or nodes) it is attached to in a legal tree. The negative contributions of the nodes are made any time the node is present; these are cancelled by positive contributions from the links only when the link constitutes a legal domination, sanctioned by the grammar.

So in order to apply the same strategy to unrestricted grammars, we will simply set the magnitude of the (negative) contributions of nodes equal to their valence, as determined by the grammar.

We can illustrate the technique by showing how HNF solves the problem with the simple three-rule grammar fragment G_0 introduced early in this section. The corresponding HNF grammar fragment G_{HNF} given by the above construction is $A[1] \rightarrow B\ C$, $A \rightarrow A[1]$, $A[2] \rightarrow D\ E$, $A \rightarrow A[2]$,

$F[1]{\rightarrow}B\ E$, $F{\rightarrow}F[1]$. To avoid extraneous complications from adding a start node above and terminal nodes below, suppose that both A and F are valid start symbols and that B, C, D, E are terminal nodes. Then the corresponding HG G_H assigns to the ill-formed tree $(A\ ;\ (B\ E))$ the Harmony -4, since, according to G_{HNF}, B and E are both missing a legal parent and A is missing two legal children. Introducing a now-necessary subcategorized version of A helps, but not enough: $(A\ ;\ (A[1]\ ;\ (B\ E)))$ and $(A\ ;\ (A[2]\ ;\ (B\ E)))$ both have $H = -2$ since in each, one leaf node is missing a legal parent (E and B, respectively), and the $A[i]$ node is missing the corresponding legal child. But the correct parse of the string $B\ E$, $(F\ ;\ (F[1]\ ;\ (B\ E)))$, has $H = 0$.

This technique can be generalized from context-free to unrestricted (type 0) formal languages, which are equivalent to Turing Machines in the languages they generate (e.g., (Hopcroft and Ullman, 1979)). The ith production rule in an unrestricted grammar, $R_i:\ \alpha_1\alpha_2\cdots\alpha_{n_i} \rightarrow \beta_1\beta_2\cdots\beta_{m_i}$ is replaced by the two rules: $R_i':\alpha_1\alpha_2\cdots\alpha_{n_i} \rightarrow \Gamma[i]$ and $R_i'':\Gamma[i] \rightarrow \beta_1\beta_2\cdots\beta_{m_i}$, introducing new non-terminal symbols $\Gamma[i]$. The corresponding soft rules in the Harmonic Grammar are then: "If the kth parent of $\Gamma[i]$ is α_k, add $+2$ to H" and "If β_k is the kth child of $\Gamma[i]$, add $+2$ to H"; there is also the rule $R_{\Gamma[i]}$: "If $\Gamma[i]$ is at any node, add $-n_i - m_i$ to H." Finally, there are the soft rules R_a, R_A, and R_{root}, defined as in the context-free case.[17]

3.1.3 Embedding-Invariant Grammars

The recursive structure of parse trees is reflected in the recursive notion of grammatical well-formedness. CFGs show this in a pure form: the well-formedness of a local tree is independent of the location where it happens to be embedded in the overall tree; "$A[i]{\rightarrow}B\ C$" means that the local tree $(A[i]\ ;\ (B\ C))$ is well-formed, *regardless of where it is embedded.* This is reflected in the soft rules of the corresponding HG, which take the form $R_{B,A[i]}$, "If B is a left child of $A[i]$ then add $+1$ to H": this rule applies regardless of the embedding location of $A[i]$. What property of the lower-level connectionist network will entail this *embedding invariance*? In this section we answer this question assuming the stratified recursive tensor product representation of Section 2.2.1; extending it to the fully distributed recursive representations of Section 2.2.2 is a subject of current research.

Soft rules like $R_{B,A[i]}$ are English paraphrases of terms in the Harmony function (24b) of the form: $H_{c_1,r_1;c_2,r_2}$. Here the relevant constituents are $c_1 = B$ and $c_2 = A[i]$, and the roles are such that B is a left child of $A[i]$: if $r_2 = r_x$ (the tree position indexed by bit-string x), then $r_1 = r_{0x}$ (the left child of position x). Construed as applying at all possible embedding positions x, then, $R_{B,A[i]}$ encapsulates a large number of terms in the Harmony function, all of the form $H_{B,r_{0x};\,A[i],r_x}$ where x varies over all possible tree positions. These Harmony values must be the same for all possible positions x to achieve embedding invariance. These Harmony values are in turn determined from the lower-level connection weights $\mathbf{W}$ and lower-level distributed patterns representing the constituents in their respective roles (13):

(26) $H_{B,r_{0x};\,A[i],r_x} = (\mathbf{B}{\otimes}\mathbf{r_0}{\otimes}\mathbf{r_x})^{\mathbf{T}}\,\mathbf{W}\,(\mathbf{A[i]}{\otimes}\mathbf{r_x})$

What condition on the lower-level weight matrix $\mathbf{W}$ will make this independent of x? Some tensor

[17] Unlike Context-Free Grammars, the derivations generated by unrestricted grammars cannot be represented as static data structures using simple trees; thus the data structure over which the preceding Harmonic Grammar is defined is a more general kind of graph. Development of tensor product representations for the appropriate graph structures with the appropriate recursive properties has yet to be explicitly worked out, but the methods described here for the case of binary trees should extend naturally to the more general case.

algebra which we exhibit shortly shows that $\mathbf{W}$ should satisfy the following recursion formula:

(27) $\mathbf{W} = \mathbf{W_{root}} + \mathcal{I} \otimes \mathbf{W}$

where $\mathcal{I}$ is the identity matrix for the role vector space containing $\{\mathbf{r_0}, \mathbf{r_1}\}$. $\mathbf{W_{root}}$ includes only the weights corresponding to constituent pairs in root position, where x = the empty string (labelling the root). $\mathbf{W_{root}}$ is a weight matrix giving the well-formedness of local trees in root position, while $\mathbf{W}$ is the embedding-invariant well-formedness measure on full trees that is generated by $\mathbf{W_{root}}$. This recursion relation guarantees embedding invariance quite generally, and is capable of going beyond CFGs to encode well-formedness dependencies that are longer-distance than those between a child and its parent. When $\mathbf{W}$ obeys the condition (27), we will say that the weight matrix or the network is *recursive*.

Because the property of recursiveness (27) plays a significant role in the foundational arguments of Section 4, it is worth discussing in a little more detail. First, let's verify that (27) does entail the desired embedding invariance, in which the Harmony values of (26) are independent of embedding position x. Recall that $x = x_d \cdots x_2 x_1$, is a bit-string of length d for a node of depth d; when x is the empty string, the embedding position is the root of the tree. If x is non-empty, let the last $d-1$ bits of x be denoted $y = x_{d-1} \cdots x_2 x_1$, so that $x = x_d y$; y is the parent of x in the tree. Recall also that $\mathbf{r_x} = \mathbf{r_{x_d}} \otimes \cdots \otimes \mathbf{r_{x_2}} \otimes \mathbf{r_{x_1}} = \mathbf{r_{x_d}} \otimes \mathbf{r_y}$. Now suppose that $\mathbf{W}$ is recursive (27) and evaluate (26):

$$\begin{aligned} H_{B,r_{0x};\,A[i],r_x} &= (\mathbf{B} \otimes \mathbf{r_0} \otimes \mathbf{r_x})^{\mathbf{T}}\, \mathbf{W}\, (\mathbf{A[i]} \otimes \mathbf{r_x}) \\ &= (\mathbf{B} \otimes \mathbf{r_0} \otimes \mathbf{r_x})^{\mathbf{T}}\, [\mathbf{W_{root}} + \mathcal{I} \otimes \mathbf{W}]\, (\mathbf{A[i]} \otimes \mathbf{r_x}) \\ &= (\mathbf{B} \otimes \mathbf{r_0} \otimes \mathbf{r_x})^{\mathbf{T}}\, \mathbf{W_{root}}\, (\mathbf{A[i]} \otimes \mathbf{r_x}) + (\mathbf{B} \otimes \mathbf{r_0} \otimes \mathbf{r_{x_d}} \otimes \mathbf{r_y})^{\mathbf{T}}\, \mathcal{I} \otimes \mathbf{W}\, (\mathbf{A[i]} \otimes \mathbf{r_{x_d}} \otimes \mathbf{r_y}) \end{aligned}$$

By definition, $\mathbf{W_{root}}$ is zero except for structure embedded at the root, so the first term will be zero unless x is empty and the embedding position is the root. In this case, the first term is the Harmony of the structure in root position, and the second term is zero, since $\mathbf{W} \otimes \mathcal{I}$ has no non-zero elements at level zero. On the other hand, if x is non-empty and the embedding position is not the root, then the first term vanishes by definition of $\mathbf{W_{root}}$, and we are left with the second term:

$$\begin{aligned} H_{B,r_{0x};\,A[i],r_x} &= (\mathbf{B} \otimes \mathbf{r_0} \otimes \mathbf{r_{x_d}} \otimes \mathbf{r_y})^{\mathbf{T}}\, \mathcal{I} \otimes \mathbf{W}\, (\mathbf{A[i]} \otimes \mathbf{r_{x_d}} \otimes \mathbf{r_y}) \\ &= \mathbf{r_{x_d}}^{\mathbf{T}} \mathcal{I} \mathbf{r_{x_d}}\, (\mathbf{B} \otimes \mathbf{r_{0y}})^{\mathbf{T}}\, \mathbf{W}\, (\mathbf{A[i]} \otimes \mathbf{r_y}) \\ &= \|\mathbf{r_{x_d}}\|^2\, H_{B,r_{0y};\,A[i],r_y} \end{aligned}$$

Assuming that the fundamental role vectors $\mathbf{r_0}$ and $\mathbf{r_1}$ are normalized to unit length, then $\|\mathbf{r_{x_d}}\|^2 = 1$, and we have the conclusion that the Harmony at embedding position x is the same as that at x's parent, y. Applying this recursively, we conclude that the Harmony at embedding position x is the same as that at its parent's parent, …, until the recursion grounds out at the root position, the case considered earlier. Thus (27) entails that (26) is independent of x, as required.

Solving the recursion equation (27) is easy; just put the equation into itself repeatedly:

$$\begin{aligned} \mathbf{W} &= \mathbf{W_{root}} + \mathcal{I} \otimes \mathbf{W} \\ &= \mathbf{W_{root}} + \mathcal{I} \otimes [\mathbf{W_{root}} + \mathcal{I} \otimes \mathbf{W}] \\ &= \mathbf{W_{root}} + \mathcal{I} \otimes \mathbf{W_{root}} + \mathcal{I} \otimes \mathcal{I} \otimes [\mathbf{W_{root}} + \mathcal{I} \otimes \mathbf{W}] \\ &= \cdots \end{aligned}$$

$$
\begin{aligned}
&= \mathbf{W_{root}} + \mathcal{I}\otimes\mathbf{W_{root}} + \mathcal{I}\otimes\mathcal{I}\otimes\mathbf{W_{root}} + \cdots \\
&= (1 + \mathcal{I} + \mathcal{I}\otimes\mathcal{I} + \cdots)\otimes\mathbf{W_{root}} \\
&= \mathcal{R}\otimes\mathbf{W_{root}}
\end{aligned}
$$

where $\mathcal{R}$ is the "recursion matrix" defined above in (19). Indeed in Section 2.2.1 we have already seen instances of just this characterization of the recursive matrices $\mathbf{W_{cons0}}$ (20) and $\mathbf{W_{car}}$ (21): in the former case, $\mathbf{W_{root}} = \mathbf{r_0}$ and in the latter case $\mathbf{W_{root}} = \mathbf{u_0}^{\mathbf{T}}$. So, in conclusion, as an alternative to (27), we can characterize a weight matrix $\mathbf{W}$ as recursive if it has the form

(28) $\quad \mathbf{W} = \mathcal{R}\otimes\mathbf{W_{root}}$

This equation is rather easy to interpret: it says that to create a recursive weight matrix $\mathbf{W}$, take the relatively small number of basic weights in $\mathbf{W_{root}}$ which determine the well-formedness of structures in root position, and copy each of them to all the homologous locations throughout the network which correspond to all possible embedding positions. Thus this equation gives rise to embedding invariance in a way analogous to how translation invariance is often achieved in connectionist networks for vision: there is a relatively small number of weights determining the processing of the patch of image at the origin, and each of these is copied to all the homologous locations in the network which correspond to translating the patch away from the origin (e.g., (Fukushima, 1980); also, (Minsky and Papert, 1969)). (28) characterizes a low-level invariance structure imposed on the network, in which individual connection weights in homologous network locations are required to be equal. It has the high-level consequence that the well-formedness of pairs of symbolic constituents $H_{c_1;c_2}$ depends only on their relative, not their absolute, positions in the tree. And this in turn is just what we needed for H to specify a recursive measure of well-formedness, such as that required to characterize Context-Free Languages.

3.1.4 Extension: Unification

Building on the analysis of embedding-invariant well-formedness and of CFLs, a natural next step is to extend the theory to incorporate *unification*. Unification is of particular interest for at least two reasons. First, it is a very general and powerful operation at the core of symbolic computation (e.g., (Robinson, 1965; Shieber, 1986)) which is fundamentally compatible with the connectionist computational notion of simultaneous satisfaction of multiple constraints, e.g., as defined by Harmony maximization (11) (for a "structured connectionist network" approach to unification as constraint satisfaction, see (Stolcke, 1989)). Secondly, unification is the central computational notion of a class of symbolic approaches to natural language syntax and semantics (e.g., (Pollard and Sag, 1987; Pollard and Sag, 1991; Sag et al., 1986)). It seems likely that the principles (6)–(11) can be further developed, to the point where the maximum-Harmony structure constructed by the network (11) can be characterized as the one that best unifies the input with the structures embodied in the network's connections—its knowledge base.

3.1.5 An Application to Natural Language Syntax/Semantics

We now move on from formal to natural languages. We have applied Harmonic Grammar to the study of a particular phenomenon in the syntax and semantics of natural language: *unaccusativity* (Legendre et al., 1990b). Before introducing this phenomenon and our research concerning it, we want to emphasize our goal in this work: to assess the strengths and weaknesses of HG as a means of integrating connectionist and symbolic computation for furthering the theory of higher cognition; to promote HG's development through contact with real linguistic phenomena; to lead us to alternative ways of pursuing the Three Goals. Our purpose in this paper is not to sell our analysis of unaccusativity—readers interested in that phenomenon in its own right are referred to (Legendre, 1992; Legendre et al., tion; Legendre et al., In press a; Legendre et al., In press b). Rather, we hope that even a rather brief description of a concrete application of HG will provide support for our claim that this research project is a fruitful way of integrating connectionism into existing symbolic research in linguistics, as a way of furthering the Three Goals. If the integration of connectionist and symbolic methodologies being developed in this project can become a useful tool in theoretical linguistics and/or psycholinguistics, it will have made its intended contribution, even if future work shows the approach to unaccusativity to need modification. Indeed, one direction of further development of the HG methodology has already emerged: the algebraic approach which we discuss next. This newer approach has already yielded a range of results in phonology; it may also do so in syntax/semantics.

The phenomenon of unaccusativity[18] or *split intransitivity* can be introduced to those unfamiliar with it as follows. Of central importance in the interaction of syntax and semantics is the *argument structure* of verbs, which, on one view, relates the syntactic roles of a verb's arguments to the semantic roles of the interpretation of those arguments as participants in the described event.[19] Perhaps not surprisingly, given its central role in language, argument structure has turned out to be quite a challenging problem, and a number of linguists have focussed attention on the simplest case: events with one participant described by intransitive verbs.

In practically every language in which intransitive verbs have been carefully examined, it appears that they split into two classes, depending on whether their single argument behaves like the subject or direct object of a transitive verb in that language. In some languages, the distinction is reflected in the morphological marking of the verb: in Lakhota (a Siouan language), the two arguments of a transitive verb are directly encoded into the verb via a morpheme X corresponding to the subject and a morpheme Y corresponding to the direct object. Intransitive verbs contain a single morpheme corresponding to their single argument, which in some cases is X (typically agentive verbs, but not always), and in others is Y (typically non-agentive verbs, but not always)

[18] We chose unaccusativity phenomena as our first testbed because it rated well according to eight criteria: (a) it is one part of a central issue in syntax/semantics (argument structure); (b) it is a phenomenon well-studied by linguists, ... (c) yet it is not completely understood, and offers the potential for new contributions; (d) it sometimes exhibits strong, complex interactions of multiple syntactic and semantic factors (as shown in our research in French), and so lends itself to one of the strengths of HG; (e) it involves some syntactic structure (according to many, but not all, researchers), ... (f) but it can be approached without dealing explicitly with issues of embedding (which the technical core of HG was not yet well equipped to handle when we began this work), ... (g) yet it can serve as a stepping stone for subsequent research on related phenomena that do crucially involve embedding; and, finally, (h) it is a domain in which Legendre has been working for some years, using traditional symbolic methods (Legendre, 1989b). It is a pragmatic entailment of (a-b) that unaccusativity is a controversial phenomenon, and we assert (c-g) not because these claims are undisputed, but because we feel we have solid arguments to defend them—in the particular case of French, where our HG work to date has focussed. As suggested in (d-g), unaccusativity is useful in the early stages of development of HG because it allows us to test whether HG can deal with the complexity of *interactions* that involve syntactic structure, without requiring that we deal explicitly with complexity in the syntactic structure itself; yet it gets a foot in the door leading to such complexity as well. Concerning (h), we note that the Harmonic Grammar work is fully integrated into an ongoing in-depth study of unaccusativity in French which includes more traditional linguistic research into both the syntactic and semantic approaches to analyzing the phenomenon, and which is part of a larger cross-linguistic study of unaccusativity in a wide range of languages (Legendre, tion; Legendre, 1992).

[19] For example, the argument structure of the passive verb *was kissed* tells us that in the sentence *John was kissed by Mary*, the argument of *was kissed* filling the syntactic role of *subject*—*John*—plays the semantic role of *patient* in the correct semantic interpretation: the sentence describes an event in which John *receives* a kiss.

(Legendre and Rood, 1992; Williamson, 1979). In languages with impoverished morphology such as English and French, there is no immediately visible distinction among intransitive verbs; yet a distinction can nearly always be observed in certain syntactic phenomena. For example, French has a construction in which the main verb *croire* 'believe' occurs with a participial complement, corresponding to English *I believe John gone*. In the French construction, only the direct object of transitive verbs or the argument of *some* intransitive verbs can appear as the object of *croire*; the subject of transitive verbs and the arguments of the remaining intransitive verbs cannot.[20] The same distinction can be observed in a half-dozen other syntactic constructions in French (Legendre, 1989b; Legendre, 1992).

A variety of theoretical approaches have shed light on unaccusativity phenomena. Some analysts have emphasized the parallel between the split among intransitives in certain syntactic contexts and the corresponding split in behavior of the subject and direct object of transitives, as exemplified in the previous paragraph for the *croire* construction. They advocate a "deep" syntactic distinction, claiming in essence that the argument structure of some intransitive verbs dubbed "unaccusative" calls for a deep direct object, rather than a deep subject, as called for with "unergative" intransitives (Burzio, 1986; Perlmutter, 1978; Perlmutter, 1989). In such treatments, a deep direct object is often claimed to be a necessary condition for an intransitive verb to appear in a given (e.g., *croire*) construction—but such a condition is usually not sufficient, as the acceptability of the resulting sentence is sometimes sensitive to semantic and aspectual properties of both the intransitive verb and its argument (e.g., telicity, stativity, volitionality, animacy, indefiniteness (Legendre, 1989b; Levin and Rappaport, 1989)). This is one reason that different constructions do not split the intransitive verbs in exactly the same way: the problem of *unaccusativity mismatches*. The deep syntactic approach must incorporate an explicit account of the interacting semantic and aspectual factors if it is to give a complete account of the phenomena, including the complex pattern of mismatches.

Another line of work attempts to establish that such semantic and aspectual factors are themselves sufficient to provide a complete account, and that a deep syntactic distinction is unnecessary (e.g. (Dowty, 1991; Van Valin, 1990)). This controversy cannot be separated from such important and extremely controversial issues as mono- vs. multi-level syntax. But the level of complexity in the interaction between semantic and syntactic factors that underlie split intransitivity is often relegated to a secondary place in the discussions: broad tendencies are emphasized, while the extent of their validity, and the factors leading to their violation, are typically neglected.

As part of the research on unaccusativity in French including our HG work (Legendre, 1992), have been carrying out detailed studies based on a data base of acceptability judgements which we have assembled: 8393 sentences involving 183 intransitive and 225 transitive verbs in 11 syntactic environments. Studies of the 3608 sentences in this data base involving intransitive verbs have corroborated some of the claimed universal semantic and aspectual tendencies, not corroborated others, and identified new regularities. We have tested a number of the claims in (Van Valin, 1990), and have determined, for example, that his account is silent concerning about 70% of our data, since it falls outside his verb classification based on the Aktionsart classes introduced by (Vendler, 1967) and further developed by (Dowty, 1979); that the putatively universal tendency of Activity verbs to behave unergatively is indeed fairly well respected in French, but that the other Aktionsart

[20] *Je croyais le pont détruit*, 'I believe the bridge [to be] destroyed'; *Je croyais le pont effondré*, 'I believe the bridge [to have] collapsed'; but **Je croyais la bombe détruit le pont*, 'I believe the bomb [to have] destroyed the bridge'; **Je croyais le pont explosé*, 'I believe the bridge [to have] exploded'. (*détruire* is strictly transitive; *s'effondrer* and *exploser* are strictly intransitive.)

classes show only weak tendencies; and that the claimed tendencies for the aspectual property telicity[21] are only observed moderately, in one direction. Testing a recent proposal in (Dowty, 1991), we have discovered, for example, that the claim that atelicity and agentivity together "definitely" entail unergative behavior is fairly well corroborated, but the claim that telicity and non-agentivity together entail unaccusative behavior is not well respected. An adequate treatment of these proposals, and others such as (Zaenen, 1989), is well beyond the scope of this paper, especially since our results show that the phenomena are much more complex than has often been suggested in the literature (Legendre, 1992).

Our conclusion at this point in our study is that, in French, (a) semantic and aspectual factors play a major role in unaccusativity phenomena; (b) their role is more complex than has been previously proposed; (c) they are not individually or conjunctively sufficient to provide a complete account; (d) a major role is also played by a deep syntactic distinction; and (e) the syntactic and semantic factors interact strongly.

Our HG account has co-evolved with and contributed to our study of syntactic and semantic accounts; in fact, it builds on and integrates these accounts. The HG account involves representations of deep grammatical functions (DGFs) subject and direct object, and so incorporates the syntactic approach. Unlike symbolic theories, however, a given intransitive verb does not *require* its argument to have one or the other DGF—instead, it has a *preference* for one over the other; any linguistic structure in which this preference is violated has its well-formedness (Harmony) reduced by a particular amount which characterizes the *strength* of that verb's preference. This preference is encoded in one of the lexical soft rules in the HG account. This rule interacts strongly with (a) other syntactic soft rules, (b) a set of semantic soft rules, and (c) soft rules concerning syntactic/semantic correspondences. An example of type (a) is a rule that says the well-formedness of a *croire* construction is diminished if the target NP is not a deep direct object; these rules embody the syntactic constraints on the relevant constructions that derive from the syntactic approach. In this syntactically simplified account, the only syntactic factors appearing in these rules is the construction in which the intransitive is embedded, and a non-surface ("hidden") variable which we interpret as the DGF of the intransitive's argument. An example of a semantic rule of type (b) is one asserting that the well-formedness of *croire* constructions is reduced if the target NP is volitional; these rules capture the semantic and aspectual tendencies of the semantic approach. The properties figuring in these rules are telicity, progressivizability[22], volitionality, animacy, and definiteness; our study of the French data has shown acceptability judgements to be sensitive to all these factors. Finally, an example of a linking rule of type (c) is one stating that the well-formedness of a structure is increased if the argument of a progressivizable predicate is assigned the DGF direct object.

The HG framework integrating these different types of constraint on well-formedness allows them all to interact strongly enough to account for the French data. Appendix A explains how this account handles several interesting interactions. Based on an earlier data base which we assembled, it correctly accounts for the acceptability judgements of all but 3 of the 885 sentences involving 143 intransitive verbs in five syntactic constructions (Legendre et al., In press a).

Extensions and Relations to Other Research. It is interesting to note that HG involves a kind of numerical counting that Dowty also invokes in his account (Dowty, 1991). His framework

[21]A verb V is *telic* if it is "inherently bounded," i.e., it is semantically impossible to say in English *He V-ed for hours*; otherwise it is *atelic*.

[22]I.e., ability of a verb V to appear in the progressive; in English: *He is V-ing*

centers on prototypes for Agent and Patient which are each characterized by a list of properties or implications: a participant in an event is assigned the role of Agent or Patient (an important element of his account of unaccusativity, as well as many other phenomena) in such a way as to *maximize the number of correct implications*. This is analogous to the way an argument of a verb is assigned a role in our HG account—indeed, HG allows us to formalize Dowty's account as a collection of soft rules like "If a volitional argument is assigned the role Agent, add 1 to the well-formedness of the structure." Generally, in HG, the number appearing in such a soft rule need not be "1," but rather a quantity specifying the strength of the implication "Agent ⇒ volitional" relative to other interacting soft rules. Dowty in fact explicitly suggests that such relative weighting might be desired within his framework (Dowty, 1991, page 574); but without a formalism such as HG that systematically integrates numerical weighting with structural rules, it would not be possible to pursue Dowty's suggestion except in an *ad hoc* manner. Thus the form of computation employed in our current HG account of unaccusativity, and even the content of many of the rules, constitutes in many respects a systematic formalization of Dowty's account.[23]

Our account of unaccusativity in French copes with a very large and complex set of data, but at a price: the account itself is very difficult to understand (see Appendix A). This is a problem endemic to connectionism, and while our HG account seems quite complex when compared to other linguistic accounts of unaccusativity, it seems quite simple and comprehensible compared to most connectionist models. One approach for dealing with the complexity of the account would be to systematically simplify it (or, equivalently, the connectionist network that embodies it) using a variety of techniques developed in recent years, e.g. skeletonization (Mozer and Smolensky, 1989b; Mozer and Smolensky, 1989a), weight elimination (Weigend et al., 1990), and a number of other regularization techniques from neural network research and statistics. This would provide a nested *series* of grammatical accounts of the phenomena which explicitly trade off coverage of data for complexity of the account. Thus the simplest, readily comprehensible, accounts will cover what the grammar itself identifies as the *core* data, and from these understanding of the phenomena could be built up through increasingly complex accounts that cover increasing portions of the data.

A more comprehensive HG approach to French unaccusativity is described in (Legendre et al., In press b). This new approach addresses the complete data base of 8393 sentences mentioned earlier, and avoids some of the inherent ambiguities concerning our interpretation of the "hidden variable" of the earlier account as the DGF of the argument: the new account handles embedded *transitive* verbs as well as intransitives, and explicitly treats the argument of an intransitive like either the subject or direct object of a transitive—whichever produces maximal Harmony in a given syntactic context. Furthermore, rather than allowing each intransitive predicate to have an arbitrary preference for the DGF of its argument, as in the earlier account, connectionist learning is used to automatically extract from the data a small number of predicate "features" which—like telicity and progressivizability in the earlier account—serve to determine each verb's preference for DGF (and type of argument). A preliminary version of this new account (using six learned predicate features) correctly accounts for the acceptability judgements of all but 104 of the 8393 sentences. With intransitives, the focus of the study, all but 14 of 3608 sentences are correctly accounted for. Statistical cluster analysis of the learned features reveals some provocative connections to aspectual properties, but the results are too preliminary to take seriously.

[23] Dowty, however, intends "argument selection" as generalizations concerning possible argument structures. The soft rules that we take to define the well-formedness of structural descriptions of sentences would have to be reinterpreted as defining the well-formedness of argument structures in the lexicon in order to reflect this aspect of Dowty's analysis.

Since much of the interest in split intransitivity concerns universal properties or tendencies, an important future direction is to investigate what insights HG can bring to bear in the cross-linguistic study of the phenomenon. One goal would be to develop a HG framework that covers a wide variety of languages within a common set of soft rules, and which will allow us to determine what is universal in the patterns of rule strengths cross-linguistically. A possible outcome mirroring HG results in phonology discussed in Section 3.2.1 is that a single core set of soft constraints operate in all languages; what varies across languages is the *relative strength* of the constraints. A formal framework embodying this conception of the relation between universal grammar and the grammars of particular languages is a unique contribution of Harmonic Grammar.

Moving beyond split intransitivity, another important direction for extending this work involves the study of syntactic structures crucially involving embedding; these have only recently become accessible within HG thanks to recent advances in the technical base discussed in Section 2.

3.1.6 Extension: Distributed Representation of Syntactic/Semantic Roles

A final direction for future HG work is to explore the underlying connectionist distributed represenetations. A central connectionist principle asserts that distributed representations encode the feature- or similarity-structure of information, and a main motivation for distributed tensor product representations is to allow the application of this principle to the *roles* of symbolic structures. Here, natural languages provide a distinctly better domain of study than purely formal languages, because of all the meaningful information that is encoded through real linguistic structure. In a purely formal binary tree, the two recursive roles *left child* and *right child* are just two primitive, distinct roles; in our simulations we represent them as two arbitrary distributed vectors. But linguistic theories provide a rich set of features (explicitly or implicitly) for describing the roles in syntactic and semantic structure. These include hierarchically structured grammatical functions (subject, direct object, ...), thematic roles (agent, patient, ...), X-bar syntactic configurational roles (head, specifier, argument, adjunct ...), syntactic and semantic feature structural roles (number, gender, ...), and many more. Tensor product representations make it possible to study the consequences—for representation, processing, learning, and grammatical description—of directly encoding such information via distributed role vectors.

3.2 Algebraic theory

In applying Harmonic Grammar to phonology, Prince and Smolensky discovered that in a wide variety of phonological problems, the numerical strengths of soft rules arrange themselves so that the rules form *strict dominance hierarchies*. In these hierarchies, the soft constraints can be ordered from weakest to strongest in such a way that each constraint is stronger than *all* the weaker constraints *combined*; thus a given constraint must be satisfied (if possible), regardless of whether that entails violation of any number of weaker constraints—unless satisfying the constraint requires violating still stronger constraints that can otherwise be satisfied. In such situations, all the information carried by the numerical strengths of the soft rules can be re-expressed non-numerically as the ranking of the rules in the dominance hierarchy. In this special case, principle (24) can be reformulated in non-numerical terms: (Prince and Smolensky, tion; Prince and Smolensky, 1991):

(29) **Fundamental principle of Harmonic Grammar—Algebraic formulation**

a. A descriptive grammar is an axiomatically defined algebraic prefence relation $\succ$ among linguistic structures; $S_1 \succ S_2$ is interpreted as "S_1 is more Harmonic (well-formed) than S_2."

b. Given an input I, such a grammar assigns as output that linguistic structure S containing the input which is maximally Harmonic; i.e., $S \succ S'$ for all other structures S' containing I.

c. The well-formedness relation $\succ$ among linguistic structures is defined compositionally from well-formedness relations among the substructures from which the linguistic structures are built.

d. Most of the basic well-formedness relations and means of combination needed for the grammars of individual languages are universal: they appear in the grammars of all languages. What primarily distinguishes the grammars of individual languages is *the particular ways the universal well-formedness rules are combined* (e.g., the particular ranking of constraints in dominance hierarchies).

This principle can be viewed as one means of formalizing the linguistic notion of *markedness* introduced in the 1920s and 1930s by the Prague School (e.g., (Jakobson, 1962; Jakobson, 1971)) and employed to some extent also in generative grammar (e.g., (Chomsky and Halle, 1991, Chapter 9)) Despite the long history of the idea, and its widespread application in virtually all branches of linguistics, "markedness has so far resisted a satisfying treatment, and no clearly defined theory of markedness has emerged." (Battistella, 1990, p. 5). The formalization provided by (29) is new, however, and consitutes an original *symbolic* contribution of SSP; it makes it possible for a certain conception of markedness to play a full and formal role in grammatical theory. Principle (29) can also be viewed as a formalization of certain notions of *optimality* or *economy* employed in linguistic theory (e.g., recently, (Chomsky, 1991; Chomsky, 1992)).

3.2.1 Applications to Phonology

The algebraic formulation of Harmonic Grammar has been applied by Prince and Smolensky to a variety of problems in phonology.[24] Examples of problems which have been treated to date include: (a) the universal typology of basic syllable structure; (b) a detailed analysis of the unusual Berber syllabification system; (c) classic interactions of various phonological processes such as those exhibited in Lardil and Yawelmani; and (d) the universal typology of stress systems. In (a), for example, the universal typology arises simply by considering all possible dominance rankings of the following universal well-formedness relations[25]:

(30) **Universal Harmony conditions for syllable structure:**

a. A syllable is *more* Harmonic if it contains an onset position.

b. A syllable is *less* Harmonic if it contains a coda position.

c. A syllable is less Harmonic if it contains unrealized (deleted) phonemes.

d. A syllable is less Harmonic if it contains epenthetic (inserted) segments.

Basic syllable structure typology.

[24] Many of these were presented in July 1991 in a four-week course "Connectionism, Harmony Theory, and Linguistics" which Prince and Smolensky co-taught at the Linguistic Institute sponsored by the Linguistics Society of America at the University of California, Santa Cruz (Prince and Smolensky, 1991).

[25] For present purposes, we assume every syllable to have a nucleus (typically a vowel); any preceding consonants in the syllable fill the *onset* position, and any following consonants in the same syllable occupy the *coda* position.

> The possible basic syllable structures in the world's languages are just those that arise from all possible dominance hierarchies formed from conditions a–d.

To illustrate how this typology works out, suppose a language contains the morpheme /ti/ to which it adds the affix /a/ forming /ti+a/. How is this syllabified? That is, if we give the syllable-structure component of the grammar this input, what is its output? What is the maximally Harmonic syllable structure? One possible structure is a bisyllabic parse, the first syllable being *ti* and the second *a*, *.ti.a.* Since the second syllable *.a.* lacks an onset (it has no pre-vocalic consonant), according to (30a), this syllabification is less Harmonic than alternatives with no missing onsets. One such alternative output would be *.ti.ya.*; this can be analyzed as the pronunciation of the syllabification *.ti.□a.*, where the second syllable is a tree structure which has an onset node that is not filled by any phoneme in the input—this empty onset node □ is then "filled in" with a phonetically suitable epenthetic segment such as *y*. Since the syllable structure *.ti.□a.* contains an onset for each syllable, it is more Harmonic than the first syllabification *.ti.a.* according to (30a)—but it is *less* Harmonic according to (30d) since it contains the epenthetic segment *y* or equivalently the empty syllable position node □.

How is this stalemate resolved? By the dominance hierarchy of a given language. Universal grammar provides the Harmony conditions (30) but it is up to individual languages to determine the relative strength of these conditions, which are locked in eternal conflict. If a language ranks (30a) above (30d) in its dominance hierarchy, then *.ti.□a.* is overall more Harmonic than *.ti.a.*, since the condition that favors the former, (30a), strictly dominates the condition that favors the latter, (30d). In a different language in which (30d) dominates (30a), *.ti.a.* would be the more Harmonic.

This example suggests how the typology of syllable structures arises. In a language that ranks (30a) high in its dominance hierarchy, the most Harmonic syllable structures will always be those in which every syllable contains an onset position (even if an empty one, which must be filled in with an epenthetic segment). In such a language, *all syllables must have onsets*; that is, for any input string of phonemes, the output (the maximally Harmonic syllabification) will consist solely in syllables with onsets. On the other hand, languages that rank (30d) high in their dominance hierarchies will be languages that *forbid epenthesis.* Languages that rank (30c) high will *forbid deletion* while those placing (30b) high will *forbid codas.* Considering all possible dominance hierarchies, we get a spectrum of languages, some which require onsets, others which forbid codas, or deletions or insertions. On the other hand, we get *no* languages which *forbid* onsets, or *require* codas, or which insert or delete segments except when this is necessary to provide an onset or to avoid a coda. We explain the typological variation found in the syllable structures of natural languages as a logical consequence of the general formal principle (29) and the simple substantive universal Harmony conditions (30).

The new formulation of phonology based on principle (29) makes a number of contributions to linguistic theory, of which a few are:

(31) **Contributions to phonological theory**

a. A precise formal framework is provided for powerful kinds of constraint-based reasoning, some of which are new, and some of which have previously been available, but only informally. As examples of the latter kind, from the dominance hierarchy of well-formedness conditions, one can deduce that (i) certain structures X will appear (in the maximally Harmonic structure) *unless* that would violate certain other (higher ranked) constraints; while (ii) certain other structures Y will *not* appear *except* in order to satisfy certain other (higher ranked) constraints. This straightforwardly achieves the evasive result of formalizing grammatical reasoning involving "do X unless ..." and "do Y only ..." rules.

b. Accounts of the interaction of various phonological processes based on the ordering of rules in a sequential derivation are replaced by a declarative characterization of phonological well-formedness in terms of the relative strengths in a given language of mostly universal and some language-particular constraints. In place of language-specific derivational processes, we have the universal process of Harmony maximization (a process which in fact underlies the connectionist account of many other cognitive domains besides language, including lower-level processes such as perception and memory retrieval—note the implications for horizontal Theoretical Integration, (1b)).

c. Formal means are provided for deriving universal typologies from universal (soft) constraints, and for situating language-particular systems within a theory of universal phonology. For example, what previously appeared to be a rather singular, bizarre syllabification system in Berber can now be formally analyzed as a natural, albeit extreme, special case of the universal theory of syllabification[26].

These results constitute significant progress in the development of a formal declarative theory of universal phonology, a theory which has been rather elusive despite much effort directed towards it. Indeed, prominent generative phonologists have even tried to argue that such a theory is nonexistent (Bromberger and Halle, 1989). These results support the claim of Section 1.4 that SSP has made significant progress on exploiting insights derived from principles of neural computation to further the universal theory of grammar: the challenge laid down in question (4).

3.3 Extensions: Connectionist Computational Substrate

The Algebraic Theory within HG (Section 3.2) has had striking descriptive successes, but much remains to be understood about its algorithmic underpinnings. This theory *can* be implemented in the Numerical Theory: as the dominance hierarchy of constraints is mounted, their numerical strength grows exponentially [as in (Minsky and Papert, 1969, Chapter 7)]. Local connectionist networks can be used to compute with such constraint hierarchies, and we have done so for our account of syllabification in Berber (31c). It is unclear whether there are lower-level distributed representations that *naturally* cause exponential weightings to emerge at the higher level.

The connectionist network computing Berber syllabifications turns out to display a surprising property: except in one particular situation, a greedy algorithm for H maximization always finds the *global* H maximum, despite the fact that such an algorithm is only guaranteed to find a *local* maximum. D. E. Rumelhart (personal communication) has suggested that this startling success is a consequence of the exponential soft rule strengths required to numerically implement the algebraic theory. If confirmed in future research, this hypothesis would have a surprising and important consequence for the overall theory. On the face of it, the exponential growth of rule strengths seems like an inelegant hack required to get a connectionist net to display the unnatural behavior

[26]In most languages, a vowel (consonant) must be parsed as the nucleus (onset or coda) of a syllable, so the possible parses of a phoneme string into syllables is fairly constrained. However, in the Tahlhiyt dialect of Berber, as spoken in the Imdlawn valley of the Western Higher Atlas (Dell and Elmedlaoui, 1985), every phoneme (except /a/) can be parsed into any syllable position, greatly increasing the complexity of the syllabification process, descriptively as well as algorithmically.

of strict dominance. But Rumelhart's hypothesis allows us to see strict dominance as a *natural* consequence of a connectionist implementation, *under the additional well-motivated assumption that a greedy algorithm should in general tend to find global H maxima.*

The principles governing the use of connectionist networks to compute in the Numerical Theory have been developed, but there has so far been little study of these principles in actual practice. Is the effectiveness of networks performing greedy Harmony maximization that we observed in Berber syllabification typical? Or is something like the simulated annealing of the original Harmony Theory (Smolensky, 1983; Smolensky, 1986) (also (Hinton and Sejnowski, 1983; Kirkpatrick et al., 1983)) generally needed to avoid local H maxima? In either case, what can be said of the time course of parsing in these networks? How does this compare to the real time syntactic and semantic processing observed in human subjects, and to models of these processes based on symbolic computation? These questions remain to be addressed.

Learning Harmonic Grammars, including embedding-invariant ones discussed in Section 3.1.3, is obviously a major issue for future research. A general technique for connectionist learning of rule systems recently developed by Mozer, McMillan and Smolensky (McMillan et al., 1991a; McMillan et al., 1991b; McMillan et al., 1992) may be applicable. In this technique, during learning, the weights are driven by a variety of means towards distinguished regions of weight-space which correspond to legal rules; these means include hard and soft constraints on the weights during gradient error descent, and periodic projection of the weights into the nearest point in a distinguished region. In the examples studied in (McMillan et al., 1992), such networks were able to learn, from a small sample of only positive examples, exactly the sets of symbolic permutation rules from which the data were generated—even inducing the categories of symbols which condition the applicability of the rules. The technique was applied with good results to the syntactic- to semantic-role assignment task in English studied in (McClelland and Kawamoto, 1986; Miikkulainen and Dyer, 1988; Miikkulainen and Dyer, 1991). These networks had *a priori* knowledge of the general syntactic form of the rules to be learned, but not of the content (e.g. categories) of those rules.

3.4 Methodology

The research described in this section is made possible by integrating several methodologies: elicitation and analysis of well-formedness judgements by native speakers, theoretical analysis of the structure of linguistic representations, development of novel symbolic formalizations of optimization, mathematical analysis of connectionist computation, and the design of specialized connectionist processing architectures, learning algorithms, and network analysis techniques.

It is worth contrasting the methodology developed here with the two most prevalent methodologies currently practiced for relating connectionism and language, which illustrate a number of the general points made in Section 1.1 concerning model- vs. principle-centering and in Section 1.3 concerning eliminativism and implementationalism. Our strategy, embodied in principles (23) through (29), may be summarized as follows:

(32) **Harmonic Grammar methodology**

a. abstract from particular connectionist models to (relatively simple) general connectionist principles;

b. use these principles to derive a general grammar formalism;

c. use this formalism to develop specific analyses of particular linguistic data (which can be viewed both at a lower level as a connectionist net and at a higher level as a set of soft rules);

d. generalize from these analyses to the universal properties of the grammars of human languages.

This strategy embodies the principle-centered approach, avoiding both eliminativism and implementationalism.

By contrast, the most typical approach to applying connectionism to language (Berg, 1991; Jain, 1991; Miikkulainen and Dyer, 1988; Miikkulainen and Dyer, 1991; McClelland and Kawamoto, 1986; Mozer, 1990; Pollack, 1988; Pollack, 1990; Rumelhart and McClelland, 1986; Servan-Schreiber et al., 1991; St. John and McClelland, 1990) might be summarized as follows:

(33) **Traditional methodology for connectionist language research**

a. identify some linguistic phenomenon of interest;

a. construct specific data sets that exhibit this phenomenon;

a. train and test some (relatively complex) particular connectionist network on these data;

a. try to draw more general linguistic conclusions that go beyond these particular data.

The traditional strategy is model-centered, uses quite complex networks, and is often fairly eliminativist; it attempts to connect linguistics to connectionism by encoding particular data into particular networks, while the SSP strategy is to connect simplified general high-level principles of connectionist computation with general linguistic principles: a direct manifestation of (3b), Principle-Centering.

Our point is not that the traditional approach does not yield interesting and important experimental results about what connectionist networks can learn, represent, and compute—it certainly does. On the contrary, our claim is the traditional and SSP approaches can offer complementary kinds of contributions to cognitive theory, and that together they allow connectionism to promote Meta-Theoretic Integration within cognitive science.

Other examples of more principle-centered research striving to explicitly integrate connectionist and linguistic principles include (Elman, 1991; Goldsmith, In press a; Goldsmith and Larson, In press; Hare, 1990; Larson, In press). This work has brought to light some valuable relationships between connectionist computation and linguistics, implicitly if not explicitly investigating the applicability to linguists of connectionist computational principles such as those describing activation and inhibition of adjacent prominence values in a linear sequence, or those governing similarity and continuity in temporally adjacent output feature vectors. The SSP work pushes further towards principle-centering, and moves to principles such as Harmony maximization which are higher-level and computationally stronger.[27]

Another alternative strategy (Rager and Berg, 1990; Touretzky, 1989; Touretzky and Wheeler, 1991; Wheeler and Touretzky, In press) could aptly be termed "implementationalist," in that it

[27]The Goldsmith-Larson approach (Goldsmith and Larson, In press), based on activation and inhibition principles, yields some suggestive computer simulation examples of Berber syllabification. It is interesting to compare this to the extensive analyses of this problem that have been developed using both the Algebraic and Numerical SSP approaches. The latter includes a connectionist implementation which bears some resemblance to the Goldsmith-Larson network—but because our network is mathematically derived to maximize a specially-designed Harmony function, precise theorems concerning the correctness of its competence can be proved. Furthermore, it is integrated into the broad grammatical framework of HG, and a universal theory of syllabification (31c) (Prince and Smolensky, 1991). The Goldsmith-Larson model has been solved in closed form and extensively studied analytically in (Prince, 1992), in which are proved a number of general properties, some linguistically promising, others pathological.

uses connectionist mechanisms to directly implement (often quite serial) symbolic rule application. Advocates of this methodology argue that it results in major revision of symbolic theory, but it can be argued that the implementational relationship it enforces between connectionist and symbolic computation calls on the weaknesses, rather than the strengths, of both: the kind of connectionist network used for this sort of implementation typically fails to exploit the power of connectionist computation resulting from learning algorithms, distributed representation, mutual constraint satisfaction, and optimization; and the kind of symbolic representations and operations that get implemented in these kinds of networks are typically rather impoverished. Such a computational compromise would appear to impose severe limits on a vehicle for advancing linguistic theory, although advocates of the approach argue that such constraints are actually a virtue.

4 Explaining the Productivity of Cognition

In the founding days of cognitive science, Chomsky used the productivity of language—native speakers' competence to understand or generate a potentially infinite number of sentences—to argue that the mind operates on combinatorial principles: that mental representations, like the sentences of a language, consist of arrays of symbols that can be combined in a potentially infinite variety of ways; that mental processes operating on recursive principles can appropriately handle this potentially infinite combinatorial variation.

This combinatorial principle has become the basis for nearly all research in higher cognition, extending well beyond the scope of Chomskyan linguistics: it has provided the fundamental strategy for explaining—both in general, conceptual terms, and in the form of elaborate, explicit theories and AI programs—how various mental faculties can embody the competence to handle an enormous and rich variety of problems on the basis of a finite store of knowledge.

Thus it is of both foundational and practical interest to know: Does a connectionist approach to cognitive science employ the combinatorial strategy? If so, how? Connectionist representations and connectionist knowledge are apparently large, unstructured (activation) vectors and (weight) matrices—mere collections of numbers—and these do not appear to have the recursive combinatorial structure that is the minimal prerequisite to even admit the *possibility* of a combinatorial strategy. On the other hand, if connectionist-grounded cognitive science rejects the combinatorial strategy, what is offered in its place? Does the striking productivity of mental faculties now go unexplained? Or is this productivity simply denied?

This issue was raised in Fodor and Pylyshyn's influential critique of connectionism (Fodor and Pylyshyn, 1988), which claimed that connectionism's only two choices were to either leave productivity unexplained, or to implement the symbolic explanation. Through an ongoing debate in the literature (Fodor, 1991; Fodor and McLaughlin, 1990; Fodor and Pylyshyn, 1988; Smolensky, 1987a; Smolensky, 1988; Smolensky, 1991; Smolensky, tion), and through continued development of the SSP theory, Smolensky has been concerned to show how a theory based on distributed connectionist representations can employ the combinatorial strategy *without* "merely implementing" a symbolic instantiation of that strategy, e.g., Fodor's "Language of Thought" (Fodor, 1975; Fodor, 1987).

To clarify this position, it is useful to adopt Marr's (Marr, 1982) distinction between what he called the computational, algorithmic, and implementation levels of analysis, and to consider the combinatorial strategy at each of these levels. Since the computational level concerns itself with the abstract input/output function that is computed, the combinatorial strategy at this level consists in analyzing the inputs and outputs of cognitive processes (e.g., the components of a grammar) as combinatorial structures, and in analyzing the function that maps the inputs to outputs in terms of this combinatorial structure. The implementation level involves physical properties such as physical location, duration, and causation, and the combinatorial strategy at this level involves characterizing in combinatorial terms the physical properties of inputs and outputs and the causal processes that mediate between them. The algorithmic level can be viewed as involving abstract, non-physical notions of "locations" (in data structures), "times" (during the execution of a procedure), and "causation" (state changes effected by primitive operations). Thus to adopt the combinatorial strategy at the algorithmic level is to ascribe to the constituents in combinatorial structures roles not just in the inputs and outputs (as at the computational level), but also in internal data structures which endow them with abstract internal locations at abstract intermediate times, and in algorithms which impute to them abstract causal properties in the chain of abstract events that lead from input to output.

Adopting the combinatorial strategy at Marr's computational level is clearly weaker than adopting it at the algorithmic or implementation levels. To adopt what Fodor and Pylyshyn call the "Classical" explanation of productivity is to adopt the combinatorial strategy at the algorithmic level, and possibly also at the implementation level. Yet the results presented above show how productivity can be explained by adopting the combinatorial strategy only at the computational level, and not at lower levels. Such an explanation, then, is non-Classical; it does not presume that the algorithms that generate the behavior can be stated over symbolic constituents, endowing them with (abstract) causal power, and that connectionism comes in at the implementation level to realize these Classical algorithms. Yet it does just the same show how connectionist algorithms can realize higher cognitive functions that are fully productive.

Note that the sense in which the SSP explanation is non-Classical is relevant not just to foundational issues, but also to major enterprises such as AI, symbolic cognitive modeling, natural language processing, and computational linguistics: in each case, the major research activity is the search for *algorithms* operating over symbol structures. According to SSP, symbol structures play a crucial role in the analysis of the functions computed by cognitive processes, but not generally in the algorithms that causally generate these processes.

4.1 Structure of the New Explanation

> "We conclude that assuming that mental representations are activation vectors does not allow Smolensky to endorse the Classical solution of the systematicity problem. And, indeed, we think Smolensky would grant this since he admits up front that mental processes will not be causally sensitive to the strong compositional structure of mental representation. That is, he acknowledges that the constituents of complex mental representations play no causal role in determining what happens when the representations get tokened. '. . . Causal efficacy was not my goal in developing the tensor product representation . . . ' ([Smolensky] 1988b; p. 21). What are causally efficacious according to connectionists are the activation values of individual units; the dynamical equations that govern the evolution of the system will be defined over these. It would thus appear that Smolensky must have some *non*-Classical solution to the systematicity problem up his sleeve; some solution that does *not* depend on assuming mental processes that are causally sensitive to constituent structure. So, then, after all this, what *is* Smolensky's solution to the systematicity problem? Remarkably enough, *Smolensky doesn't say.*" (Fodor and McLaughlin, 1990, pp. 200–201; emphasis original—*of course*).

In an attempt to maximize clarity in presenting this explanation, we begin by outlining the structure of the argument in a slightly formal manner: [28].

(34) a. Mental representations in higher cognitive domains satisfy a property P_{reps} in which constituents play an essential role.

b. Certain higher mental processes obey a property P_{proc} in which constituents do *not* figure.

c. P_{reps} logically entails that higher mental representation is systematic.

d. P_{reps} and P_{proc} logically entail that certain higher mental processes are productive.

e. The role of constituents in P_{reps} and in the explananda—systematicity, productivity—endows them with *explanatory* roles.

f. The lack of role of constituents in P_{proc} means they are non-causal.

The kernel of the argument, in other words, is this:

(35) The constituents, which are defined only at the higher level:

a. figure in *proofs* concerning the behavior—they are explanatory;

b. do **not** figure in *algorithms* that generate this behavior—they are non-causal.

This is no mystery: there *are* (connectionist) algorithms generating the behavior, but they are defined only at the lower level; however, they obey a property P_{proc} which has provable mathematical consequences, in conjunction with P_{reps}, for the higher level behavior; these consequences involve the constituents because P_{reps} does.

4.2 Non-causally Explanatory Constituents

The properties P_{reps} of the representations (34a) and P_{proc} of the processing (34b) have already been articulated among the fundamental principles of SSP, but we reformulate them explicitly here:

(36) **Property** P_{reps} **:** Mental representations in higher cognitive domains are tensor product representations (6). For some domains, e.g. the representations of propositions and linguistic expressions, the tensor product representation is recursive (14) (e.g., the binary tree representation of Section 2.2). [Recapitulation of (6).]

That is: to say that a system of mental representations $\mathcal{M}$ can represent the proposition $P = R(A, B)$ is to say that there is some recursive tensor product representation, say that of binary trees in Section 2.2, with respect to which $\mathcal{M}$'s representational vector space contains the vector

$$\mathbf{P} = \mathbf{R}\otimes\mathbf{r_0} + \mathbf{A}\otimes\mathbf{r_0}\otimes\mathbf{r_1} + \mathbf{B}\otimes\mathbf{r_1}\otimes\mathbf{r_1}$$

(here, just as in Section 2.1, we use a LISP-like binary tree `cons(R, cons(A, B))` for "$R(A, B)$"; see (16)).

The relevant property P_{proc} of mental processes is:

[28] The argumentation involved in this research is being developed with the help a computer tool, EUCLID, which Smolensky and colleagues (principally Bernard Bernstein) have built with the support of NSF grant IST-8609599, with additional support from Symbolics and Apple (Bernstein, 1992; Bernstein et al., 1989; Smolensky et al., 1987; Smolensky et al., 1988). EUCLID is a hypertext system for supporting the construction, evaluation, and communication of complex arguments.

(37) **Property** P_{proc} **:** In the cognitive domains of (36) for which the tensor product representation is recursive, the weights are also recursive; they obey the property (27) (equivalently (28)) of Section 3.1.3. [Recapitulation of part c, (9) and (11).]

Now the first logical consequence follows immediately from P_{reps}:

(38) **Systematicity.** If $\mathcal{M}$ can represent $P = R(A, B)$, then it can also represent $P' = R(B, A)$.

For by (36), if the tensor product representational vector space of $\mathcal{M}$ contains the vector $\mathbf{P}$ representing P, then it also contains the vector representing P', namely:

$$\mathbf{P}' = \mathbf{R} \otimes \mathbf{r_0} + \mathbf{B} \otimes \mathbf{r_0} \otimes \mathbf{r_1} + \mathbf{A} \otimes \mathbf{r_1} \otimes \mathbf{r_1}$$

This establishes (34c).

The second logical consequence follows from P_{reps} and P_{proc} together, for domains in which the representations and weights are both recursive. We saw in Section 3.1.3 that the lower-level recursive properties P_{reps} (36) of the activity patterns and P_{proc} (37) of the weights together entail that at the higher level, the Harmony of having two constituents in the same structure is independent of their level of embedding, and depends only on their relative positions. Furthermore, in Section 3.1.2 we saw that this embedding invariance generates Harmony functions which can be used to express any Context Free Language; that is, the network can distinguish ill- and well-formed strings from such a language. Indeed we saw how this extends to arbitrary formal languages. Thus:

(39) **Productivity.** Recursive networks possess the unbounded productive competence to distinguish well-formed sentences of a formal language from ill-formed strings.

This establishes (34d) (and recapitulates (8c)).

Now we see that the principles of SSP do indeed provide a non-Classical explanation for the systematicity and productivity of higher cognition. To recapitulate, less formally: the patterns of activity which are mental representations have a combinatorial (tensor-product) structure which mental processes are sensitive to; the constituents in these representations figure crucially in the statement of certain high-level regularities (e.g. systematicity and productivity) in behavior; the combinatorial structure of the representations figures centrally in the explanation of this behavior (via mathematical deduction); but the constituents do not have causal power in the sense of figuring in mental algorithms for generating behavior: these causal algorithms can *only* be stated at a level lower than that of mental constituents, the level of individual connectionist units.

The work on TPPL and HG shows that the techniques from which this novel explanation of systematicity and productivity follow are not just philosophical curiosities: TPPL provides an explicit formal programming language for structure processing with massive parallelism (Section 2.3); HG provides a novel means of specifying formal languages (Section 3.1.2) and thus capturing Chomsky's classic idealized conception of the productivity of human linguistic competence; and HG goes further and provides new means of advancing the descriptive (Section 3.1.5) and explanatory (Section 3.2.1) adequacy of the theory of actual human languages.

HG thus bears on one of the central issues in the foundations of cognitive science: the psychological reality of the rules of higher cognition, such as those of grammar. Implementationalism imputes *full* psychological reality to grammatical rules, assuming them to be part of processing algorithms; eliminativism imputes *no* psychological reality to such rules, assuming them to have no role in the description of mental representations or processes. HG reifies the intermediate hypothesis that such rules play a crucial *explanatory* role with respect to systematic, productive behavior but do *not* play any *causal* role in the computational system that generates this behavior.[29]

4.3 Methodology

The work described in this section is entirely dependent on a unified methodology, embracing mathematical analysis, traditional philosophical analysis, novel neural network design, and computer simulation. The technical work is driven in large part by problems in the foundations of cognitive science, and there is simply no distinction between pursuing these problems and developing new and more powerful techniques for integrating connectionist and symbolic cognitive principles.

5 Summary and Conclusion

In this paper we have argued for the following claims.

(40) **Summary of main claims**

a. Connectionism can provide not only cognitive models but also cognitive *principles*, thereby fostering the Meta-Theoretic Integration (3) of model- and principle-centered research.

b. These principles promote Theoretical Integration (1) as follows: they achieve horizontal integration by embodying general computational mechanisms that apply widely across cognitive domains; they achieve vertical integration by unifying lower-level connectionist computation with higher-level symbolic computation. These principles therefore address central problems in higher cognition while offering meaningful progress towards resolving the Central Paradox of Cognition, the tension between viewing the mind/brain as a structure processing engine and as a numerical processing engine.

c. The research based on these principles represents significant Methodological Integration (2), bringing together problems, data, theoretical constructs, and research techniques from both connectionist and symbolic computation, from linguistics, and from philosophy of mind; the connectionist component also constitutes important contact with psychological and neural modeling, and while such opportunities have not yet been pursued in the research projects reported here, they have been pursued vigorously by many other researchers.

To close the paper we summarize and assess the arguments we have presented here. To the extent that these arguments provide support for the claims (40), we hope they also argue the utility for cognitive science of taking seriously the Three Goals, the importance and feasibility of facing the Central Paradox of Cognition head on, and the value of exploiting connectionism for principle-centered as well as model-centered research.

[29] With respect to the role of combinatorial structure in *learning*, we add in passing that previous research has shown that the structure of simple combinatorial environments, as encoded through tensor product representations, can also be *learned* by connectionist algorithms, and that, once such structure has been learned, it permits rapid learning of new information sharing in this structure (Brousse, 1991; Brousse and Smolensky, 1989).

5.1 Meta-Theoretic Integration

In this paper we have presented seven principles: (6), (8), (9), (11), (23), (24), and (29). These principles concern the formal characterization—at two levels of description—of mental representations and processes, and of the well-formedness of representations, especially linguistic ones. Taking into consideration all of the Three Goals, these principles should be assessed on the basis of how well they meet the following eight desiderata:

(41) **Desired attributes for cognitive principles:**

a. sufficiently general to have wide applicability for the study of lower- and higher cognition;

b. sufficiently precise that we can reason directly from and about them;

c. sufficiently powerful that the conclusions they entail make novel contributions to the theory of higher cognition;

d. sufficiently closely tied to existing cognitive principles that strong theoretical and methodological connections can be made;

e. sufficiently expressive to enable their instantiation in specific accounts of particular cognitive phenomena ...

f. ...which can be supported empirically by the relevant data from neuroscience, experimental psychology, linguistics, and philosophy;

g. capable of being given an adequate computational implementation; and

h. grounded in sound mathematical foundations;

With respect to the criterion of generality (41a), while it was noted in Section 1.6 that the extent to which the principles discussed here can be empirically validated (41f) is very much an open question, it should nonetheless be clear that the *potential* applicability of the principles is very broad, given their computational generality. Principles (6) and (8) concerning the integration of symbolic and connectionist representations and processes is potentially applicable to virtually all cognitive domains in which higher-level symbolic representations have been theoretically attested; although the power of the connectionist substrate to support more powerful symbolic operations will need to be continually expanded for some time yet, for example, along the lines of Section 3.1.4. As for the principle (9) of Harmony maximization, this too has great generality; in fact, it applies to lower-level cognition as well. If we ignore the assumption of higher-level symbolic representations, and stick purely to a lower-level description, mechanisms for simultaneously satisfying multiple soft contraints have figured prominently in connectionist and related models of perception and memory (e.g., (Bechtel and Abrahamsen, 1991; McClelland et al., 1986), and important subclasses of these have the formal properties required to be characterized as Harmony-maximizing.

Concerning the criterion of sufficient precision to support logical inference (41b), the following summarizes the logical relations presented above:

(42) **Logical structure of the principles:**

a. Integrated Processing (8) is a direct mathematical extension of Integrated Representation (6).

b. The Symbolic Harmonic Principle (11) is a mathematical consequence of Integrated Representation (6) and the Connectionist Harmonic Principle (9).

c. The General (23) and Numerical (24) Formulations of the Fundamental Principle of Harmonic Grammar follows from the Symbolic Harmonic Principle (11).

d. The Algebraic Formulation of the Fundamental Principle of Harmonic Grammar (29) is a direct outgrowth of the Numerical Formulation (24) in the special case of strict dominance.

e. The new explanation of cognitive productivity (34)–(35) is a direct consequence of the principles themselves, rather than extrapolation from particular models of specific phenomena.

As for the power of the principles to inform the theory of higher cognition (41c), the research on grammar (Section 3) and explaining the productivity of higher cognition (Section 4) both involve significant conceptual and theoretical innovations resulting directly from the principles on problems of central interest in linguistics and the foundations of cognitive science.

Regarding degree of theoretical and methodological contact with existing cognitive research (41d), the fact that the linguistic principles (23)–(29) directly address well-formedness puts them immediately in contact with the central theory and methodology of linguistics; and the fact that all the principles directly concern the integration of symbolic and connectionist representations and processes puts them in direct contact with foundational research on the psychological reality of symbols and rules, and the relation of connectionism to the combinatorial strategy for explaining higher cognition.

The expressive power (41e) of the principles is sufficient to enable their instantiation in specific accounts of traditional symbol manipulation (Section 2) and particular linguistic phenomena in formal language theory (Section 3.1.2) and the syntax/semantics (Section 3.1.5) and phonology (Section 3.2.1) of natural languages. However it remains a major project for future research to further test and extend this expressiveness (e.g., along lines suggested in Sections 3.1.4–3.1.6).

As for empirical validation of the specific accounts (41f), we have claimed such with respect to the data of linguistics (Sections 3.1.5 and 3.2.1) and philosophy of mind (Section 4), We have indicated the urgent need for such validation with respect to psychological and neural data (Section 1.5), and have suggested a few possible directions for pursuing the former (Sections 2.4 and 3.3).

Regarding computational implementation (41g), the principles derive from analysis of the higher-level properties of certain kinds of connectionist networks, and so there is an explicit route to connectionist implementation of particular accounts based on such principles. Five such implementations were discussed in Sections 2, 3.1.5, and 3.3. Much more research is however required to study the effectiveness of such networks (see Section 3.3), to control their size (see Section 2.4), and to extend them to serve more powerful principles (e.g., along the lines suggested in Sections 2.5 and 3.1.4).

With respect to the final criterion, the soundness of the mathematical foundations on which the principles rest (41h), we have claimed that the technical results embodied in principles (6)–(11), including the specific treatment of tree structures and formal languages (Sections 2.2–2.3 and 3.1.2–3.1.4) constitute significant progress in laying the foundations of integrated connectionist/symbolic computation. At the same time, we have also emphasized how much farther we need to go on this crucial issue (Section 2.5).

5.2 Methodological and Theoretical Integration

The methodologies which work together synergistically to make possible the research described here come from many corners of cognitive science: mathematical analysis of connectionist computation; numerical computer simulation; design of specialized connectionist processing architectures, learning algorithms, and network analysis techniques; symbolic modeling of higher cognition; elicitation and analysis of well-formedness judgements by native speakers; theoretical analysis of the structure of linguistic representations and constraints; development of novel symbolic formalizations of optimization; traditional philosophical analysis; and even the design of a new computer system to support the demands of complex argumentation.

The cognitive problems we have addressed here concern language and the foundations of cognitive science. The work on language (Section 3), while based on principles centrally involving connectionist computation, differs from most connectionist work on language in the following respects:

(43) **Comparison of HG to traditional connectionist language research**

a. The emphasis, in the initial stages, has been on competence over performance; on the development of descriptive over algorithmic grammars. The data initially in focus is that traditional in linguistics, rather than that of psycholinguistics.

b. The focus is more centered on the development of novel, general grammatical principles than on the constructions of particular accounts or models of specific data, although the former of course depends critically on the latter.

c. The contribution of formal analysis relative to that of computer simulation is greater.

d. Symbolic representations and "rules"—that is, constraints—play a much more central role.

On the other hand, compared to traditional linguistic theory, the research on language presented here differs in several important respects:

(44) **Comparison of HG to traditional linguistic theory**

a. Descriptive grammar is reconstructed on a new central organizing principle deriving from connectionist computation: Harmony maximization. While various notions of optimality are not unfamiliar in the theory of grammar (see Section 3.2), the principles presented here are formally entirely novel and have led to analyses and solutions of problems which were not previously possible.

b. The constraints comprising grammar are *soft*. In the numerical theory, this is achieved by the use of numbers which modulate the relative strengths of the constraints and explicitly assess costs to their violation; in the algebraic theory, the dominance ranking of the constraints uses non-numerical means to ensure that a constraint can be violated whenever the "cost" of doing so is compensated by avoiding the higher "cost" that would have been paid to violate a higher-ranking constraint.

c. The medium of algorithmic grammar involves purely connectionist mechanisms rather than serial symbol manipulation.

Finally, this work on grammar and the other results described here show in detail how to pursue a general solution to the Central Paradox of Cognition. On the Classical account, the mind and brain are sealed off from one another. The productivity of higher cognitive faculties is explained by positing a symbol manipulating engine which is the mind, a machine which manipulates combinatorially structured representations according to manipulation rules. How this might involve the numerical engine which is the brain is left a mystery. As are the pervasive manifestations of statistical influences on behavior. On the account proposed here, by contrast, symbolic, combinatorial representations emerge as the higher level description of vectors of distributed numerical activity of simple lower level processing elements. This is not just a vague image. It is embodied in principles that are precise enough to support mathematical analysis, computer simulations, and novel general principles and specific analyses even in some of the highest-level domains of cognition, such as the universal theory of grammar. On this account, mental symbols and the rules governing them play an essential role in the deductive explanation of the systematicity and productivity of the functions computed in higher cognition. But the algorithms that compute these functions, the causal dynamics that actually generate cognitive behavior, can be defined only at the lower level—in the form of connectionist computation.

Acknowledgements

Special thanks are due Alan Prince for many discussions of these issues and for his permission to summarize here some of his joint research with Smolensky. We would also like to thank the following people for helpful discussions, comments, and prodding: Paul Chapin, Rob Cummins, Bob Frank, Jùrgen Schmidhuber, Paul Kay, William Labov, George Lakoff, Bernard Laks, Mark Liberman, Clayton Lewis, David Perlmutter, Philip Resnick, Giorgio Satta, Peter Todd, Annie Zaenen, the members of the Boulder Connectionist Research Group, the Boulder Unaccusativity Research Group, the Institute for Research in Cognitive Science at the University of Pennsylvania, the Cognitive Science Program at UC-Berkeley, and, especially, Mike Mozer and David Rumelhart.

Appendix A: A Specific Harmonic Grammar Account

The French unaccusativity data and the HG account of it are both quite complex, and here we can examine only a small, illustrative fragment. The data we consider is summarized in Table 1, and the account of these data is given below in Tables 2 and 3. Even this fragment of the account involves the interaction of 29 numbers, in combinations of up to 11 at a time, and the sense of "understanding" of this account which we possess still falls short of our goal. This is perhaps to be expected for the first analysis in a theory like HG which involves a computational framework and hence a style of explanation which is very different from the one familiar to linguistics. We do hope, nonetheless, that by leading the reader through the accounts contained in Tables 2 and 3, we will convey some insight into Harmonic Grammar in general and our current account of French unaccusativity phenomena in particular.

Table 1: Summary of Selected Data

Argument		**Aspect**		**Constructions**					**Verb**	
VO	AN	TE	PR	OR	CR	PA	RR	ON		
−	−	−	+	+	+	+	+		bouillir	'boil'
+	+	−	+	−	−	−	−	+	rêver	'aspire'
−	−	−	−	−	−	−	+		rester	'stay'
+	+	−	−	−	+	+	+	−		
−	−	−	+	+	+	+	+		exploser	'explode'
+	+	−	+	−	−	−	−	+		
−	−	−	+	+	−	−	−		souffler	'blow'
+	+	−	+	−	−	−	−	+		

Table 1 summarizes the acceptability judgements of 36 sentences, each of which consists of an intransitive verb with its argument embedded in a syntactic construction which has been identified as a diagnostic context of unaccusativity in French. These contexts or "tests" are given in Appendix B.

The first row of Table 1 indicates that the argument of *bouillir* is non-volitional (−VO) and inanimate (−AN); that the verb is atelic (−TE) and progressivizable[30] (+PR); and that *bouillir* passes all four of the unaccusativity tests OR, CR, PA, and RR: it is acceptable in all four of the corresponding constructions. (The unergativity test, ON, is not applicable with inanimate arguments.) The behavior exhibited by *bouillir* is that of a prototypical unaccusative verb. The second row illustrates a prototypical unergative verb, *rêver*, which passes the ON test (i.e., it is acceptable when used with argument *on* under the arbitrary interpretation "someone"), but fails the unaccusativity tests—it is unacceptable in the other four constructions. The remainder of Table 1 illustrates some interesting contrasts in the behavior of several verbs which can accept either an animate volitional argument (which we'll abbreviate "+AG" = +VO, +AN) or an inanimate argument ("−AG" = −VO, −AN). (The database and full account also include animate, non-volitional arguments, but we ignore these here.)

In the data of Table 1, PA is indistinguishable from CR, and we will therefore simplify the discussion slightly by ignoring PA. And since all these data involve atelic predicates, we can also ignore TE here.

[30] That is, the verb *V* may appear in *en train de V*, 'in the process of *V*ing'

Table 2: Account of Simple Cases

Verb:		**bouillir** (+PR)			**rêver** (+PR)			
Argument:		**−AG**			**+AG**			
Construction:		OR	CR	RR	OR	CR	RR	ON
Acceptability:		+	+	+	−	−	−	+
Non-structural rules								
VO & OR	−0.4				−0.4			
AN & OR	−1.4				−1.4			
VO & CR	−0.8					−0.8		
AN & CR	0.2					0.2		
VO & RR	−0.8						−0.8	
AN & RR	−0.1						−0.1	
VO & ON	4.7							4.7
PR & OR	2.0	2.0			2.0			
PR & CR	1.0		1.0			1.0		
PR & RR	0.5			0.5			0.5	
PR & ON	2.2							2.2
VO & PR	−1.6				−1.6	−1.6	−1.6	−1.6
AN & PR	0.2				0.2	0.2	0.2	0.2
OR	−3.2	−3.2			−3.2			
CR	−3.2		−3.2			−3.2		
RR	−2.6			−2.6			−2.6	
ON	−8.1							−8.1
H_{NS}		−1.2	−2.2	−2.1	−4.4	−4.2	−4.4	−2.6
Structural rules								
OR & 2	0.4	0.4			0.4			
CR & 2	2.5		2.5			2.5		
RR & 2	2.6			2.6			2.6	
ON & 2	−2.9							−2.9
VO & 2	0.6				0.6	0.6	0.6	0.6
AN & 2	1.1				1.1	1.1	1.1	1.1
PR & 2	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
bouillir & 2	1.5	1.5	1.5	1.5				
rêver & 2	−5.9				−5.9	−5.9	−5.9	−5.9
$H_{S=2}$		2.7	4.8	4.9	−3.0	−0.9	−0.8	−6.3
$H = H_{NS} + \|H_S\|$		1.5	2.6	2.6	−1.4	−3.3	−3.6	3.7
$sign(H)$		+	+	+	−	−	−	+

A.1 Two Simple Cases

To begin the presentation of the HG account of these data, consider the simple cases of *bouillir* and *rêver*, shown in Table 2. The leftmost column of Table 2 lists all the soft rules of the HG account which are relevant for these seven sentences.[31] These soft rules and their numerical strengths were produced by the procedure given in the last paragraph of Section 3.1.1.

[31]Table 2 shows 24 grammatical soft rules and two lexical rules; in the full account, covering 885 sentences including the 36 treated here, there are 32 grammatical and 143 lexical rules. The only grammatical rules omitted from Table 2 are those pertaining to the PA construction and the TE aspectual feature; the rules involving these have exactly the same form as those shown in Table 2 involving the four other constructions and the feature PR. The other lexical rules are those specifying the argument structures of the other 141 intransitive verbs studied, and are of the same form as the two rules shown in Table 2 for *bouillir* and *rêver*. (The lexical rules for the three other verbs occurring in Table 1 are included in Table 3.)

The Relevant Soft Rules. The first rule says that if the argument of an Object Raising construction is volitional, −0.4 should be added to the Harmony H of the sentence: i.e., OR has a weak preference for non-volitional arguments. Indeed, the same is true of the other two unaccusativity tests CR and RR; the unergativity test ON, on the other hand, has a strong preference in favor of volitional arguments. These preferences are consistent with semantic accounts of unaccusativity.

The first and second blocks in the left column of Table 2 also shows the soft rules giving the tests' preferences regarding animacy and progressivizability. The rules in the third block pertain to combinations of argument and aspectual features; the first one says that if the argument is volitional and the event is +PR, 1.6 should be subtracted from H.

Correct interpretation of these rules requires a few technical details. The scale of values for H is an arbitrary one, chosen for convenience and consistency with connectionist modeling conventions. Positive total H values correspond to acceptable sentences; negative, unacceptable. Total H values between −1.4 and +1.4 are interpreted as marginal. The database includes marginal judgements, but we have not yet made systematic efforts to account for the difference between marginal and non-marginal judgements, and we ignore that distinction in the discussion here. Argument or aspectual features with value + trigger the corresponding soft rules, while − values do not; this is a choice of convenience, and it is straightforward to convert this account into an exactly equivalent notational variant with the role of + and − reversed for any set of features, or with + and − playing symmetrical roles.

The fourth block of rules in Table 2 are those triggered only by the constructions: the first says that if the construction is OR, then 3.2 should be subtracted from H. These rules essentially set a threshold which must be overcome by other positive-Harmony factors if the construction is to be acceptable: in order for an OR construction to be acceptable, the net Harmony contributed by all the other rules must exceed +3.2.

The first four blocks of rules together determine what we call the *non-structural Harmony* H_{NS} of a sentence—the component of the sentence's well-formedness that does not depend on whether the argument is assigned the deep grammatical function (GF) of subject or direct object, the "structural hidden variable" which the grammar must determine for each sentence. We will denote these two possible deep grammatical functions by DGF_1 and DGF_2, respectively (borrowing "1" and "2" from Relational Grammar). The remaining "structural" rules in Table 2 all refer to the deep GF of the argument, and together they accomplish two things: they determine whether DGF_1 or DGF_2 is assigned to the argument in a particular sentence—whichever produces greater Harmony—and they determine how well-formed the preferred choice is—the degree of "structural Harmony" H_S. The total Harmony H of a sentence is then the sum of the non-structural and structural Harmonies: $H = H_{NS} + H_S$.

The first structural rule in Table 2 says that in an OR construction, if DGF_2 (direct object) is chosen as the deep GF of the argument, then 0.4 is added to H. Not listed explicitly is a corresponding rule stating that if DGF_1 is chosen instead, −0.4 must be added to H. This block of rules distinguishes the unaccusativity tests OR, CR, and RR from the unergativity test ON: the former prefer deep direct objects, but the latter deep subjects.

The next two blocks of rules encode the preferences of correspondence between deep GF and volitionality, animacy, and progressivity. The first such rule says that if DGF_2 is chosen for a volitional argument, then 0.6 is added to H. This preference is anomalous, as is the corresponding preference of animates for DGF_2. We shall discuss these counter-intuitive preferences below.

Finally, the last block of rules encode the (soft) argument structure of the embedded intransitive verb: *bouillir* prefers its argument to be assigned DGF_2, while *rêver* prefers DGF_1—*bouillir* has unaccusative argument structure; *rêver*, unergative.

An Unaccusative Verb. The middle three columns of Table 2 show the HG account of the acceptability of *bouillir* in all three unaccusativity tests. Since the argument is non-volitional and inanimate, and the verb is progressivizable, only two non-structural rules apply in each of the three cases. Starting with the OR construction, the OR rule contributes −3.2, effectively setting a threshold of +3.2 which the other rules must meet in order to produce an acceptable sentence. The progressivizability of *bouillir* accomplishes part of this; the 'PR & OR' rule contributes +2.0. Thus the net non-structural Harmony is simply $H_{NS} = -3.2 + 2.0 = -1.2$. In order for *bouillir* to be acceptable in OR, then, the structural Harmony H_S must exceed +1.2. The cases of CR and RR with *bouillir* are similar, except that progressivizability makes somewhat less of a positive contribution, and the RR rule sets a somewhat lower threshold; the resulting non-structural Harmonies H_{NS} are respectively −2.2 and −2.1.

Moving on to the structural rules, there are three that apply for each construction. For OR, the rule 'OR & 2' says to add 0.4 to H if DGF_2 is chosen; progressivizability adds a further 0.8 for the same choice, and finally the argument structure of *bouillir* itself contributes an additional 1.5 for this choice. The net result is a structural Harmony H_S of 2.7 if DGF_2 is chosen; the alternative choice of DGF_1 would therefore produce a H_S of −2.7. So the clear Harmony-maximizing choice is DGF_2, and the corresponding structural Harmony $H_S = 2.7$ is added to the non-structural Harmony of $H_{NS} = -1.2$ to produce a total Harmony of $H = 1.5$. Since this is positive, the Harmonic Grammar declares that *bouillir* (with argument assigned DGF_2) is acceptable in OR. And the result is the same for the other unaccusativity tests CR and RR.

An Unergative Verb. The contrasting case of *rêver* is shown in the final block of columns in Table 2. Since we are again dealing with a progressivizable verb, we again have all the rules that applied with *bouillir* (except, of course, for the rule 'bouillir & 2' encoding the argument structure of *bouillir* itself); but since we now have a +AG = +VO, +AN argument, additional soft rules come into play as well. Since the unaccusativity tests OR, CR, and RR all prefer −AG arguments, the non-structural Harmony in all these cases is lower (more negative) than in the corresponding −AG contructions with *bouillir*. The unergativity test ON strongly prefers +VO (4.7), but the net H_{NS} still provides a threshold of −2.6 which must be overcome by a $H_S > +2.6$ in order to produce acceptability.

The structural rules that now apply represent conflicting preferences which must be numerically resolved. DGF_2 is preferred by all the unaccusativity tests, as well as volitionality, animacy, and progressivizability; but the argument structure of *rêver* prefers DGF_1, and with a strength greater than the combined strength of the other preferences. The net result is that the choice of DGF_2 corresponds to a negative H_S—ranging from −0.8 for RR to −6.3 for ON—so that the maximum-Harmony structure is DGF_1, corresponding to a positive H_S ranging from +0.8 to +6.3. To determine the overall Harmony H, we add H_S to H_{NS}; the result is a negative value of H for the unaccusativity tests and a positive value of H for ON. Comparing ON to the unaccusativity tests, we see that for ON both H_{NS} and H_S are greater: the former, because ON prefers +AG, and the latter, because ON's strong preference for DGF_1 agrees with *rêver*'s preference.

A.2 More Complex Examples

In the comparison of *bouillir* and *rêver*, two factors change: the argument goes from –AG to +AG, and the argument structure goes from a moderate preference (1.5) for DGF_2 to a very strong preference (5.9) for DGF_1. Both these differences contribute to explaining why the former behaves unaccusatively and the latter unergatively. More complex patterns of acceptability can result when the two factors play off against each other. A small sample of this complexity is summarized in Table 1, which illustrates the varying effects for different predicates of changing the semantic properties of the argument. As the argument changes from –AG to +AG, *exploser* 'explode' shifts its behavior from unaccusative to unergative, consistent with semantic accounts of unaccusativity. But *rester* 'stay' does the opposite, shifting from (mostly) unergative to (mostly) unaccusative behavior. And *souffler* 'blow' simply retains its unergative behavior—except for the OR test, which shows a sensitivity to AG in a direction consistent with semantic accounts. The HG account of these contrasts is shown in Table 3.

The Case of *souffler*. We begin our examination of Table 3 with *souffler* (last block of columns). The unergative behavior of *souffler* with a +AG argument is accounted for rather like that of *rêver*; the only difference in the rules is that *souffler*'s preference for a DGF_1 is weaker: 3.9, as opposed to 5.9 for *rêver*. This difference doesn't affect the sign of H, so the acceptability results are the same. As with *rêver*, with OR and ON, the grammar respects the verb's preference and assigns the argument DGF_1; however, with CR and RR, the preference of the test and of +AG for DGF_2 win out over *souffler*'s preference for DGF_1: DGF_2 is assigned. The resulting total H values for CR and RR are almost the same for *souffler* and *rêver*, however, since the H_S values are close (1.1 vs. 0.9; 1.2 vs. 0.8) even though the deep GF is reversed. For both verbs, for the unaccusative tests, the conflicting preferences for deep GF produce a small H_S, too small to overcome the negative H_{NS}; for the unergative test ON, however, there is little conflict and DGF_1 wins with a sufficiently high H_S to produce an overall positive H.

When *souffler* appears with a –AG argument (next to last block of columns in Table 3), there is less pressure for DGF_2, and DGF_1 now wins out in all constructions. Since the OR test has the weakest pressure for a DGF_2, the H_S value associated with the choice of DGF_1 is greatest for OR; large enough, in fact, to overcome the negative H_{NS}, and to produce acceptability. There are two reasons that OR breaks ranks with the other unaccusative tests here: first, it is less insistent than the others about wanting a DGF_2 (giving rise to greater H_S), and second, it is more influenced by the progressivizability of the verb (giving rise to greater H_{NS}).

The Case of *exploser*. We pass now to *exploser*, in the middle of Table 3. It exhibits the same pattern with a +AG argument as *rêver* and *souffler*, for the same reasons; the only difference is that *exploser* has an even stronger preference than *rêver* and *souffler* (6.9 vs. 5.9 and 3.9) for DGF_1. This has a somewhat surprising consequence, however, when we shift to a –AG argument: *exploser*, unlike *souffler*, passes all the unaccusativity tests, including CR and RR. Now the fairly strong preferences of CR or RR (and +AG) for DGF_2 are swamped by the even stronger preference of *exploser* for DGF_1, so that the resulting H_S is strong (3.6; 3.5), whereas for *souffler* it was weak (0.6; 0.5). So now the unaccusativity tests are passed, as this strong H_S overcomes the negative H_{NS} (H_{NS} is the same for *exploser*, *souffler*, and all other –TE, +PR predicates).

Table 3: Account of Complex Cases

Verb:		**rester** (−PR)							**exploser** (+PR)							**souffler** (+PR)						
Argument:		**−AG**			**+AG**				**−AG**			**+AG**				**−AG**			**+AG**			
Construction:		OR	CR	RR	OR	CR	RR	ON	OR	CR	RR	OR	CR	RR	ON	OR	CR	RR	OR	CR	RR	ON
Acceptability:		−	−	+	−	+	+	−	+	+	+	−	−	−	+	+	−	−	−	−	−	+
Non-structural rules																						
VO & OR	−0.4				−0.4							−0.4							−0.4			
AN & OR	−1.4				−1.4							−1.4							−1.4			
VO & CR	−0.8					−0.8							−0.8							−0.8		
AN & CR	0.2					0.2							0.2							0.2		
VO & RR	−0.8						−0.8							−0.8							−0.8	
AN & RR	−0.1						−0.1							−0.1							−0.1	
VO & ON	4.7							4.7							4.7							4.7
PR & OR	2.0								2.0			2.0				2.0			2.0			
PR & CR	1.0									1.0			1.0				1.0			1.0		
PR & RR	0.5										0.5			0.5				0.5			0.5	
PR & ON	2.2														2.2							2.2
VO & PR	−1.6											−1.6	−1.6	−1.6	−1.6				−1.6	−1.6	−1.6	−1.6
AN & PR	0.2											0.2	0.2	0.2	0.2				0.2	0.2	0.2	0.2
OR	−3.2	−3.2			−3.2				−3.2			−3.2				−3.2			−3.2			
CR	−3.2		−3.2			−3.2				−3.2			−3.2				−3.2			−3.2		
RR	−2.6			−2.6			−2.6				−2.6			−2.6				−2.6			−2.6	
ON	−8.1							−8.1							−8.1							−8.1
H_{NS}		−3.2	−3.2	−2.6	−5.0	−3.8	−3.5	−3.4	−1.2	−2.2	−2.1	−4.4	−4.2	−4.4	−2.6	−1.2	−2.2	−2.1	−4.4	−4.2	−4.4	−2.6
Structural rules																						
OR & 2	0.4	0.4			0.4				0.4			0.4				0.4			0.4			
CR & 2	2.5		2.5			2.5				2.5			2.5				2.5			2.5		
RR & 2	2.6			2.6			2.6				2.6			2.6				2.6			2.6	
ON & 2	−2.9							−2.9							−2.9							−2.9
VO & 2	0.6				0.6	0.6	0.6	0.6				0.6	0.6	0.6	0.6				0.6	0.6	0.6	0.6
AN & 2	1.1				1.1	1.1	1.1	1.1				1.1	1.1	1.1	1.1				1.1	1.1	1.1	1.1
PR & 2	0.8								0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
rester & 2	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6														
exploser & 2	−6.9								−6.9	−6.9	−6.9	−6.9	−6.9	−6.9	−6.9							
souffler & 2	−3.9															−3.9	−3.9	−3.9	−3.9	−3.9	−3.9	−3.9
$H_{S=2}$		1.0	3.1	3.2	2.7	4.8	4.9	0.6	−5.7	−3.6	−3.5	−4.0	−1.9	−1.8	−7.3	−2.7	−0.6	−0.5	−1.0	1.1	1.2	−4.3
$H = H_{NS} + \|H_S\|$		−2.2	−0.1	0.6	−2.3	1.0	1.4	−2.8	4.5	1.4	1.4	−0.4	−2.3	−2.6	4.7	1.5	−1.6	−1.6	−3.4	−3.1	−3.2	1.7
$sign(H)$		−	−	+	−	+	+	−	+	+	+	−	−	−	+	+	−	−	−	−	−	+

The Case of *rester*. Finally we come to *rester*, at the left of Table 3. Consider first the case of a +AG argument. *rester* has a slight preference for DGF_2 (0.6), and (with a +AG argument) it behaves unaccusatively on three of the four tests: it passes CR and RR, and fails ON. It passes CR and RR because all the preferences are for DGF_2, even though none are overwhelmingly strong; the resulting H_S (4.8, 4.9) is very strong and overcomes the negative H_{NS} (−3.8, −3.5). Indeed H_{NS} is also less negative than in previous cases we've considered because *rester* is not progressivizable. In contrast to the other unaccusativity tests, however, *rester* fails to pass OR, for two reasons: (a) OR's preference for a DGF_2 is considerably weaker, leading to considerably lower H_S (2.7 vs. 4.8, 4.9); and (b) H_{NS} is considerably more negative (−5.0 vs. −3.8, −3.5), because OR more strongly disprefers +AG. Finally, as expected for an unaccusative, *rester* fails ON, because the preference of ON for a DGF_1 almost entirely cancels all the other preferences for a DGF_2, and H_S is thus very weak (0.6); furthermore, the H_{NS} that must be overcome (–3.4) is more negative than for the other ON cases, because *rester* is not progressivizable.

What happens when *rester*'s argument changes to –AG? These cases are the easiest of all to compute, since for each construction, only one non-structural and two structural rules apply. In all the unaccusativity tests, the (weak) preference of *rester* for a DGF_2 agrees with the preferences of the test, and we have an uncontested choice of DGF_2. However, since OR's preference is weak as well, the total H_S for OR is not strong (1.0); for CR and RR, it is (3.1, 3.2). The question is now whether these are sufficient to overcome the thresholds for acceptability set by the three negative H_{NS}; the weakest H_S, OR, is not; the strongest, RR, is; and CR barely fails.[32] So, in going from +AG to –AG, two factors change: H_S drops by 1.7, the amount by which +AG prefers DGF_2; and H_{NS} rises, by 1.8, 0.6, and 0.9 for OR, CR, and RR, respectively—the amounts by which the tests prefer –AG. The net effect on H is that OR falls very slightly (0.1); RR falls somewhat more (0.8), and CR falls the most (1.1). So while the Harmony in all cases declines, only in the case of CR is the drop sufficient to change the sign of acceptability from + to – (and that only barely).

A.3 Discussion

The Hidden Variable. In this account, +AG arguments prefer DGF_2, which is theoretically anomalous. This is one of the problems associated with interpreting, as we have done here, the "hidden variable" assigned by the grammar[33] as DGF_2. The connectionist learning procedure really determines the meaning of this hidden variable by fixing the directions and numerical strengths of all the soft rules involving it; it is unclear in advance the degree to which the result of learning will be consistently interpretable as "deep direct object." We see here, in fact, that while such an interpretation allows us to make considerable sense of the account, it is not completely unproblematic. The main reason, we believe, is that the data presented to the connectionist network is limited entirely to embedded verbs which are intransitive: the network has no opportunity to connect the behavior of arguments of intransitive verbs with the subjects and direct objects of embedded transitive verbs, no chance to make the fundamental connection that underlies the syntactic theory of unaccusativity and motivates the distinction among deep GFs. This is the principal reason we are moving on to an account covering both embedded transitives and intransitives, as discussed in Section 3.1.5.

A second factor which may contribute to the anomalous preference of +AG arguments for DGF_2 is the kind of behavior exhibited above by *exploser*: with a –AG argument, it passes unaccusativity tests despite the assignment of DGF_1—precisely *because* this preference is so strong. Essentially, the embedded verb *exploser* achieves such a high H_S by having its argument structure satisfied through assignment of DGF_1that it no longer really matters that this GF violates the preference of the construction in which it is embedded. In the total account, behavior of this sort leads

[32]The account here actually predicts that *rester* should marginally fail CR, whereas our data indicate that it simply fails. This is an example of what we call a "minor error," in which the data and the account agree on the sign of acceptability but disagree on its marginality; a "major error" is a disagreement over the sign itself. The present account makes such minor errors in the hundreds, over the total database of 885 sentences, and, as we've said, 3 major errors. The connectionist learning procedure we designed to produce the numerical strengths of the soft rules in this account was constructed to emphasize minimization of major errors; we have not yet made serious efforts at designing procedures that will also successfully minimize minor errors. In earlier work on the four unaccusativity tests alone, we were however able to account for those 760 sentences with 0 major errors and only 30 minor errors. Clearly, the numerical nature of HG makes it well-positioned to address the issue of graded acceptability judgements, but it is too early to tell how successful it will ultimately prove in this regard.

[33]This corresponds to a "hidden unit" in the connectionist implementation.

a number of verbs to pass unaccusativity tests despite assigning DGF_1 to their arguments. In these cases, the interpretation of the hidden variable as a deep GF runs counter to the syntactic account; from the perspective of the traditional syntactic theory of unaccusativity, in these cases, the HG account is getting the right answer despite assigning the "wrong" deep GF, if we accept the deep GF interpretation of the hidden variable. There are enough of these anomalous cases that interactions between the hidden variable and AG of the argument can also be thrown out of consistency with theoretical expectations. In order to prevent the anomalous assignment of DGF_1 with some verbs that pass unaccusativity tests, it might be advisable to redo the account presented here with the following change: allow H_S to be *reduced* when a verb's preference for deep GF is violated, but do not allow H_S to be *increased* when the preference is satisfied. In the current account, the argument structure rule "exploser & 2 (−6.9)" entails an implicit rule "exploser & 1 (+6.9)"; in the revised account, the corresponding implicit rule would simply be "exploser & 1 (0)." Corresponding treatment of the implicit structural rules in the grammar, as well as those in the lexicon, might also be advisable. This is a less convenient choice from the perspective of connectionist implementation, but, understanding better the consequences of the original choice, we now see that it may well eliminate the anomalies associated with interpreting the hidden variable as a deep GF.

Testing Feature Necessity. The variables used in this account—the features VO, AN, TE, and PR, and the hidden variable we interpret as a deep GF—are all implicated in previous semantic and syntactic accounts of unaccusativity in general as well as analysis of the particular French data. It is possible, especially given the power of soft rules, that an HG account of these data might be possible using only a subset of these variables. Indeed, a purely syntactic account would invoke only the deep GF, and a purely semantic account would omit this hidden variable. These hypotheses can be tested by following the same procedure that led to the account presented here, including connectionist learning, with some subset of the variables excluded. More systematic experimentation of this sort is required before we can give definitive assessments, but we have performed some tests of this sort based on a previous account of the four unaccusativity tests alone (without the ON test) which accounted correctly for all 760 sentences. Deleting the hidden variable led to about 100 errors, deleting two of the four semantic features led to roughly 20 errors (depending on which features were excluded), and deleting one of the semantic features produced errors ranging from 2 (for AN) to 16 (for VO).

Testing Reliability. Depending on exactly how the connectionist learning is carried out, the resulting soft rule strengths vary, and with them the empirical adequacy of the account. This naturally raises the question of the degree of reliability of the approach. We have tested this in a variety of ways. For example, we have looked at the errors made by alternative accounts, and found a striking hierarchy: if account 1 made more errors than account 2, then the sentences on which account 1 erred were a strict superset of those on which account 2 erred. That is, there is a well-defined notion of one sentence being "more difficult to account for" than other sentences; more successful accounts correctly account for more difficult sentences. This is one reason we believe in the soundness of the approach—outlined in Section 3.1.5—of developing a series of more and more accurate but complex HG accounts, building out from a set of "core data" determined by HG itself. Further evidence of reliability comes from comparing across different accounts the lexical rules which specify whether a given intransitive verb prefers DGF_1 or DGF_2, and how strongly; we find a high degree of consistency.

Testing Extensibility. We have also investigated the extensibility of HG accounts as follows. We performed connectionist training using 536 sentences involving 67 verbs, getting a set of 67 lexical rules for those verbs, and a set of grammatical soft rules that do not refer to individual lexical items. We then ask: without changing these grammatical soft rules, can we accomodate 20 new verbs by an appropriate choice of one lexical rule for each of them? We found that, of the 80 new sentences involving these new verbs (in the four unaccusativity tests), 78 could in fact be correctly accounted for.

Testing a Prediction. Finally, by examining the grammatical soft rules of alternative HG accounts based on an earlier, restricted set of data, we determined that they all agreed that a certain pattern of behavior across the various tests was impossible for any intransitive verb in French—regardless of its preference for deep GF. We then examined a large, nearly exhaustive list of French intransitives, and verified this prediction.

At this point it is premature to make too much of these results concerning feature necessity, reliability, extensibility, and the ability to make novel predictions. We report these preliminary results mostly because they address some important methodological concerns about the new approach, and because they help to convey the range of questions that can already be addressed within Harmonic Grammar.

Appendix B: French Unaccusativity Tests

supply: (Legendre, 1992), Sections 2.3.1–5. (Burzio, 1986; Cinque, 1990; Legendre, 1986; Legendre, 1988; Legendre, 1989b; Legendre, 1989a; Perlmutter, 1989; Postal, 1986; Rosen, 1981; Ruwet, 1986)

References

Anderson, J. A., Silverstein, J. W., Ritz, S. A., and Jones, R. S. (1984). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, 84:413–451.

Battistella, E. L. (1990). *Markedness: The Evaluative Superstructure of Language.* State University of New York Press, Albany, NY.

Bechtel, W. and Abrahamsen, A. (1991). *Connectionism and the Mind: An Introduction to Parallel Processing in Networks.* Basil Blackwell, Cambridge, MA.

Berg, G. (1991). Learning recursive phrase structure: Combining the strengths of PDP and X-Bar Syntax. Technical Report 91–5, Department of Computer Science, State University of New York and Albany.

Bernstein, B. (1992). Euclid: Supporting collaborative argumentation with hypertext. In *Proceedings of the Macintosh Technical Conference*, Ann Arbor, MI.

Bernstein, B., Smolensky, P., and Bell, B. (1989). Design of a constraint-based hypertext system to augment human reasoning. In *Proceedings of the Rocky Mountain Conference on Artificial Intelligence*, Denver, CO.

Bromberger, S. and Halle, M. (1989). Why phonology is different. *Linguistic Inquiry*, 20:51–70.

Brousse, O. (1991). *Generativity and Systematicity in Neural Network Combinatorial Learning.* PhD thesis, Department of Computer Science, University of Colorado, Boulder, CO.

Brousse, O. and Smolensky, P. (1989). Virtual memories and massive generalization in connectionist combinatorial learning. In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, pages 380–387, Ann Arbor, MI. Lawrence Erlbaum.

Burzio, L. (1986). *Italian Syntax: A Government-Binding Approach.* Reidel, Dordrecht, Holland.

Cedergren, H. and Sankoff, D. (1974). Variable rules: Performance as a statistical reflection of competence. *Language*, 50:333–355.

Chomsky, N. (1981). *Lectures on Government and Binding.* Foris, Dordrecht, Holland.

Chomsky, N. (1991). Some notes on economy of derivation and representations. In Freidin, R., editor, *Principles and Parameters in Comparative Grammar*, pages 417–454. MIT Press, Cambridge, MA.

Chomsky, N. (1992). A minimalist program for linguistic theory. Ms.

Chomsky, N. and Halle, M. (1968/1991). *The Sound Pattern of English.* Harper and Row/MIT Press, New York/Cambridge.

Cinque, G. (1990). Ergative adjectives and the lexicalist hypothesis. *Natural Language and Linguistic Theory*, 8:1–40.

Cohen, M. A. and Grossberg, S. (1983). Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:815–825.

Dell, F. and Elmedlaoui, M. (1985). Syllabic consonants and syllabification in Imdlawn Tashlhiyt Berber. *Journal of African Languages and Linguistics*, 7:105–130.

Dolan, C. P. (1989). *Tensor Manipulation Networks: Connectionist and Symbolic Approaches to Comprehension, Learning, and Planning.* PhD thesis, Department of Computer Science, University of California, Los Angeles, CA.

Dolan, C. P. and Dyer, M. G. (1987). Symbolic schemata, role binding, and the evolution of structure in connectionist memories. In *Proceedings of the IEEE First International Conference on Neural Networks*, pages II:287–298, San Diego, CA.

Dolan, C. P. and Smolensky, P. (1989). Tensor Product Production System: A modular architecture and representation. *Connection Science*, 1:53–68.

Dowty, D. (1979). *Word Meaning and Montague Grammar.* Reidel, Dordrecht, Holland.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67:547–619.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–226.

Estes, W. K. (1988). Toward a framework for combining connectionist and symbol-processing models. *Journal of Memory and Language*, 27:196–212.

Feldman, J. A. (1981). A connectionist model of visual memory. In Hinton, G. E. and Anderson, J. A., editors, *Parallel Models of Associative Memory*, pages 49–81. Erlbaum, Hillsdale, NJ.

Feldman, J. A. (1985). Four frames suffice: A provisional model of vision and space. *The Behavioral and Brain Sciences*, 8:265–289.

Feldman, J. A. (1989). Neural representation of conceptual knowledge. In Nadel, L., Culicover, P., Cooper, L. A., and Harnish, R. M., editors, *Neural Connections, Mental Computation*, pages 68–103. MIT Press/Bradford Books, Cambridge, MA.

Fodor, J. A. (1975). *The Language of Thought.* Thomas Y. Crowell. (Paperback, Harvard University Press).

Fodor, J. A. (1987). *Psychosemantics.* MIT Press, Cambridge, MA.

Fodor, J. A. (1991). Replies. In Loewer, B. and Rey, G., editors, *Meaning in Mind: Fodor and his Critics*, pages 255–319. Basil Blackwell, Oxford.

Fodor, J. A. and McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, 35:183–204. (Reprinted in Terrence Horgan and John Tienson, editors, *Connectionism and the Philosophy of Mind*, pages 331–354. Kluwer Academic, Dordrecht, 1991.).

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28:3–71.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202.

Gazdar, G., Klein, E., Pullum, G., and Sag, I. (1985). *Generalized Phrase Structure Grammar.* Harvard University Press, Cambridge, MA.

Golden, R. M. (1986). The "Brain-State-in-a-Box" neural model is a gradient descent algorithm. *Mathematical Psychology*, 30–31:73–80.

Golden, R. M. (1988). A unified framework for connectionist systems. *Biological Cybernetics*, 59:109–120.

Goldsmith, J. A. (1990). *Autosegmental and Metrical Phonology.* Basil Blackwell, Oxford.

Goldsmith, J. A. (In press a). Local modeling in phonology. In Davis, S., editor, *Connectionism: Theory and Practice.* Oxford University Press, Oxford.

Goldsmith, J. A. (In press b). Phonology as an intelligent system. In Napoli, D. J. and Kegl, J. A., editors, *Bridges between Psychology and Linguistics: A Swarthmore Festschrift for Lila Gleitman.* Cambridge University Press, Cambridge.

Goldsmith, J. A. and Larson, G. (In press). Local modeling and syllabification. In Deaton, K., Noske, M., and Ziolkowski, M., editors, *Proceedings of the 26th Meeting of the Chicago Linguistic Society: Parasession on the Syllable in Phonetics and Phonology*, Chicago, IL.

Grossberg, S. (1982). *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control.* Reidel Press, Boston, MA.

Grossberg, S., editor (1988). *Neural Networks and Natural Intelligence.* MIT Press/Bradford Books, Cambridge, MA.

Hare, M. (1990). The role of similarity in Hungarian vowel harmony: A connectionist account. *Connection Science*, 2:123–150.

Hendler, J. A. (1989). Special issue: Hybrid systems (symbolic/connectionist). *Connection Science*, 1:227–342.

Hinton, G. E., editor (1990). *Connectionist Symbol Processing.* MIT Press/Elsevier, Cambridge, MA.

Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. (1986). Distributed representation. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, chapter 3, pages 77–109. MIT Press/Bradford Books, Cambridge, MA.

Hinton, G. E. and Sejnowski, T. J. (1983). Analyzing cooperative computation. In *Proceedings of the Fifth Annual Conference of the Cognitive Science Society*, Rochester, NY. Erlbaum Associates.

Hinton, G. E. and Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, chapter 7, pages 282–317. MIT Press/Bradford Books, Cambridge, MA.

Hopcroft, J. E. and Ullman, J. D. (1979). *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, Reading, MA.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA*, 79:2554–2558.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences, USA*, 81:3088–3092.

Hopfield, J. J. (1987). Learning algorithms and probability distributions in feed-forward and feed-back networks. *Proceedings of the National Academy of Sciences, USA*, 84:8429–8433.

Humphreys, M. S., Bain, J. D., and Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic and procedural tasks. *Psychological Review*, 96:208–233.

Jain, A. N. (1991). Parsing complex sentences with structured connectionist networks. *Neural Computation*, 3:110–120.

Jakobson, R. (1962). *Selected Writings; Vol. 1, Phonological Studies.* Mouton, The Hague.

Jakobson, R. (1971). *Selected Writings; Vol. 2, Word and Language.* Mouton, The Hague.

Kay, P. and McDaniel, C. K. (1979). On the logic of variable rules. *Language in Society*, 8:151–187.

Kayne, R. S. (1984). *Connectedness and Binary Branching.* Foris Publications, Dordrecht, Holland.

Kirkpatrick, S., Gelatt, Jr., C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220:671–680.

Labov, W. (1969). Contraction, deletion, and inherent variability of the english copula. *Language*, 45:715–762.

Lachter, J. and Bever, T. G. (1988). The relation between linguistic structure and associative theories of language learning—a constructive critique of some connectionist learning models. *Cognition*, 28:195–247.

Lakoff, G. (1988). A suggestion for a linguistics with connectionist foundations. In Touretzky, D., Hinton, G. E., and Sejnowski, T. J., editors, *Proceedings of the Connectionist Models Summer School*, pages 301–314, San Mateo, CA. Morgan Kaufmann.

Lakoff, G. (1989). Cognitive phonology. Paper presented at the UC-Berkeley Workshop on Rules and Constraints.

Larson, G. (In press). Local computational networks and the distribution of segments in the Spanish syllable. In Deaton, K., Noske, M., and Ziolkowski, M., editors, *Proceedings of the 26th Meeting of the Chicago Linguistic Society: Parasession on the Syllable in Phonetics and Phonology*, Chicago, IL.

Legendre, G. (1986). Object raising in French: A unified account. *Natural Language and Linguistic Theory*, 4:137–184.

Legendre, G. (1988). On the issue of multiple syntactic levels: Evidence from French control. In Powers, J. and de Jong, K., editors, *Proceedings of the Fifth Eastern States Conference on Linguistics*, Chicago, IL.

Legendre, G. (1989a). Inversion with certain French experiencer verbs. *Language*, 65:752–782.

Legendre, G. (1989b). Unaccusativity in French. *Lingua*, 79:95–164.

Legendre, G. (1992). Split intransitivity: A reply to Van Valin 1990. Technical Report ICS-TR-92-3, University of Colorado Institute of Cognitive Science.

Legendre, G. (In preparation). Split intransitivity: A cross-linguistic perspective.

Legendre, G., Miyata, Y., and Smolensky, P. (1990a). Harmonic Grammar—A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 388–395, Cambridge, MA. Lawrence Erlbaum.

Legendre, G., Miyata, Y., and Smolensky, P. (1990b). Harmonic Grammar—A formal multi-level connectionist theory of linguistic well-formedness: An application. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 884–891, Cambridge, MA. Lawrence Erlbaum.

Legendre, G., Miyata, Y., and Smolensky, P. (1991). Distributed recursive structure processing. In Touretzky, D. S. and Lippman, R., editors, *Advances in Neural Information Processing Systems 3*, pages 591–597, San Mateo, CA. Morgan Kaufmann. Slightly expanded version in Brian Mayoh, editor, *Scandinavian Conference on Artificial Intelligence—91*, pages 47–53. IOS Press, Amsterdam.

Legendre, G., Miyata, Y., and Smolensky, P. (In preparation). Unaccusativity, the interaction of syntax and semantics, and Harmonic Grammar.

Legendre, G., Miyata, Y., and Smolensky, P. (In press a). Can connectionism contribute to syntax? Harmonic Grammar, with an application. In Deaton, K., Noske, M., and Ziolkowski, M., editors, *Proceedings of the 26th Meeting of the Chicago Linguistic Society*, Chicago, IL.

Legendre, G., Miyata, Y., and Smolensky, P. (In press b). Unifying syntactic and semantic approaches to unaccusativity: A connectionist approach. In Sutton, L. and Johnson (with Ruth Shields), C., editors, *Proceedings of the Seventeenth Annual Meeting of the Berkeley Linguistics Society*, Berkeley, CA.

Legendre, G. and Rood, D. (1992). On the interaction of grammar components in Lakhóta: Evidence from split intransitivity. In *Proceedings of the Eighteenth Annual Meeting of the Berkeley Linguistics Society.*

Levin, B. and Rappaport, M. (1989). An approach to unaccusative mismatches. In *Proceedings of the Nineteenth Meeting of the North Eastern Linguistic Society.*

Marr, D. (1982). *Vision.* Freeman, San Francisco.

McClelland, J. L. (1991). Toward a theory of information processing in graded, random, interactive netorks. Technical Report PDP.CSN.91.1, Department of Psychology, Carnegie Mellon University.

McClelland, J. L. and Kawamoto, A. H. (1986). Mechanisms of sentence processing: Assigning roles to constituents. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*, chapter 19, pages 272–325. MIT Press/Bradford Books, Cambridge, MA.

McClelland, J. L., Rumelhart, D. E., and Hinton, G. E. (1986). The appeal of pdp. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, chapter 1, pages 3–44. MIT Press/Bradford Books, Cambridge, MA.

McCloskey, M. (1992). Networks and theories: The place of connectionism in cognitive science. *Psychological Science*.

McMillan, C., Mozer, M. C., and Smolensky, P. (1991a). The connectionist scientist game: Rule extraction and refinement in a neural network. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, Chicago, IL.

McMillan, C., Mozer, M. C., and Smolensky, P. (1991b). Learning explicit rules in a neural network. In *Proceedings of the International Joint Conference on Neural Networks*, Seattle, WA. Institute of Electrical and Electronics Engineers.

McMillan, C., Mozer, M. C., and Smolensky, P. (1992). Rule induction through integrated symbolic and subsymbolic processing. In Moody, J., Hanson, S., and Lippman, R., editors, *Advances in Neural Information Processing Systems, Volume 4*, pages 969–976, San Mateo, CA. Morgan Kaufmann.

Metcalfe-Eich, J. (1982). A composite holographic associative recall model. *Psychological Review*, 89:627–661.

Miikkulainen, R. and Dyer, M. G. (1988). Encoding input/output representations in connectionist cognitive systems. In Touretzky, D., Hinton, G. E., and Sejnowski, T. J., editors, *Proceedings of the Connectionist Models Summer School*, pages 347–356. Morgan Kaufmann, San Mateo, CA.

Miikkulainen, R. and Dyer, M. G. (1991). Natural language processing with modular pp networks and distributed lexicon. *Cognitive Science*, pages 343–399.

Minsky, M. L. and Papert, S. A. (1969). *Perceptrons*. MIT Press, Cambridge, MA.

Mozer, M. C. (1990). A focussed back-propagation algorithm for temporal sequence recognition. *Complex Systems*, 3:349–381.

Mozer, M. C. and Smolensky, P. (1989a). Skeletonization: Trimming the fat from a network via relevance assessment. In Touretzky, D. S., editor, *Advances in Neural Information Processing Systems 1*, pages 107–115, San Mateo, CA. Morgan Kauffman. Collected papers of the IEEE Conference on Neural Information Processing Systems—Natural and Synthetic, Denver, Nov. 1988.

Mozer, M. C. and Smolensky, P. (1989b). Using relevance to reduce network size automatically. *Connection Science*, 1:3–16.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89:316–338.

Newell, A. (1990). *Unified Theories of Cognition.* Harvard University Press, Cambridge, MA.

NSF Planning Workshop Report (1991). To strengthen american cognitive science for the twenty-first century.

Perlmutter, D. M. (1978). Impersonal passives and the Unaccusativity Hypothesis. In *Proceedings of the Fourth Berkeley Linguistic Society Meeting.*

Perlmutter, D. M. (1989). Multiattachment and the Unaccusative Hypothesis: The perfect auxiliary in Italian. *Probus*, 1:63–119.

Pike, R. (1984). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. *Psychological Review*, 91:281–294.

Pinker, S. and Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28:73–193.

Plate, T. (1991). Holographic reduced representations. Technical Report CRG-TR-91-1, Department of Computer Science, University of Toronto.

Pollack, J. B. (1988). Recursive auto-associative memory: Devising compositional distributed representations. In *Proceedings of the Tenth Annual Meeting of the Cognitive Science Society*, Montreal, Canada. Erlbaum Associates.

Pollack, J. B. (1990). Recursive distributed representation. *Artificial Intelligence*, 46:77–105.

Pollard, C. and Sag, I. A. (1987). *Information-Based Syntax and Semantics. Volume 1: Fundamentals.* Center for the Study of Language and Information, Stanford, CA, and University of Chicago Press, Chicago, IL.

Pollard, C. and Sag, I. A. (1991). *Information-Based Syntax and Semantics. Volume 2: Agreement, Binding, and Control.* Center for the Study of Language and Information, Stanford, CA, and University of Chicago Press, Chicago, IL.

Postal, P. (1986). *Studies of Passive Clauses.* State University of New York Press, Albany, NY.

Prince, A. (1992). Papers on the Goldsmith-Larson Dynamic Linear Model of sonority and stress structure: Convergence of the Goldsmith-Larson Dynamic Linear Model; Closed-form solution of the Dynamic Linear Model of syllable and stress structure and some properties theoreof; Remarks on the Goldsmith-Larson Dynamic Linear Model as a theory of stress with extension to the Continuous Linear Theory and additional analysis. Technical Report RuCCS TR-1, Rutgers Center for Cognitive Science.

Prince, A. and Smolensky, P. (1991). Notes on connectionism and Harmony Theory in linguistics. Technical report, Department of Computer Science, University of Colorado at Boulder. Technical Report CU-CS-533-91.

Prince, A. and Smolensky, P. (In preparation). Universal phonology through Harmony Theory: Constraint interaction and Harmony-Theoretic Grammar.

Rager, J. and Berg, G. (1990). A connectionist model of motion and government in Chomsky's government-binding theory. *Connection Science*, 2:35–52.

Resnick, P. (1992). Probabilistic tree-adjoining grammar as a framework for statistical natural langauge processing. In *Proceedings of COLING-92*. In press.

Robinson, J. A. (1965). A machine-oriented logic based on the resolution principle. *Journal of the Association for Computing Machinery*, 12:23–44.

Rosen, C. (1981). *The Relational Structure of Reflexive Clauses: Evidence from Italian.* PhD thesis, Harvard University.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, chapter 8, pages 318–362. MIT Press/Bradford Books, Cambridge, MA.

Rumelhart, D. E. and McClelland, J. L. (1986). On learning the past tenses of English verbs. In McClelland, J. L., Rumelhart, D. E., and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*, chapter 18, pages 216–271. MIT Press/Bradford Books, Cambridge, MA.

Ruwet, N. (1986). Weather-verbs and the unaccusative hypothesis. In Kirschner, C. and DeCesaris, J., editors, *Studies in Romance Linguistics*, pages 313–345. John Benjamins, Cambridge, MA.

Sag, I. A., Kaplan, R., Karttunen, L., Kay, M., Pollard, C., Shieber, S., and Zaenen, A. (1986). Unification and grammatical theory. In *Proceedings of the West Coast Conference on Formal Linguistics*, pages 238–254, Stanford, CA. Stanford Linguistics Association.

Sanger, D. (1989). Contribution analysis: A technique for assigning responsibilities to hidden units in connectionist networks. *Connection Science*, 1:115–138.

Sanger, D. (1990). *Contribution Analysis: A Technique for Assigning Responsibilities to Hidden Units in Connectionist Networks.* PhD thesis, Department of Computer Science, University of Colorado, Boulder, CO.

Servan-Schreiber, D., Cleeremans, A., and McClelland, J. L. (1991). Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning*, 7:161–194.

Shastri, L. and Feldman, J. A. (1985). Evidential reasoning in semantic networks: A formal theory. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 465–474, Los Angeles, CA.

Shieber, S. M. (1986). *An Introduction to Unification-Based Approaches to Grammar.* Center for the Study of Language and Information, Stanford, CA, and University of Chicago Press, Chicago, IL.

Smolensky, P. (1983). Schema selection and stochastic inference in modular environments. In *Proceedings of the National Conference on Artificial Intelligence*, pages 378–382, Washington, DC.

Smolensky, P. (1984a). Harmony Theory: Thermal parallel models in a computational context. Technical report, Institute for Cognitive Science, University of California at San Diego. In P. Smolensky & M. S. Riley, *Harmony theory: Problem solving, parallel cognitive models, and thermal physics*, Technical Report 8404.

Smolensky, P. (1984b). The mathematical role of self-consistency in parallel computation. In *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*, pages 319–325, Boulder, CO. Lawrence Erlbaum.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of Harmony Theory. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, chapter 6, pages 194–281. MIT Press/Bradford Books, Cambridge, MA.

Smolensky, P. (1987a). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Southern Journal of Philosophy*, 26 (Supplement):137–163. (Reprinted in Terrence Horgan and John Tienson, editors, *Connectionism and the Philosophy of Mind*, pages 281–308. Kluwer Academic, Dordrecht, 1991; and Cynthia Macdonald and Graham Macdonald, editors, *The Philosophy of Psychology: Debates on Psychological Explanation*, Basil Blackwell, Oxford, to appear.).

Smolensky, P. (1987b). On variable binding and the representation of symbolic structures in connectionist systems. Technical report, Department of Computer Science, University of Colorado at Boulder. Technical Report CU-CS-355-87.

Smolensky, P. (1988). On the proper treatment of connectionism. *The Behavioral and Brain Sciences*, 11:1–74.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist networks. *Artificial Intelligence*, 46:159–216.

Smolensky, P. (1991). Connectionism, constituency, and the Language of Thought. In Loewer, B. and Rey, G., editors, *Meaning in Mind: Fodor and his Critics*, pages 201–227. Basil Blackwell, Oxford.

Smolensky, P. (In preparation). Connectionist-grounded explanation strategies of the productivity of cognition.

Smolensky, P., Bell, B., Fox, B., King, R., and Lewis, C. (1987). Constraint-based hypertext for argumentation. In *Proceedings of Hypertext–87*, pages 215–245, Chapel Hill, NC.

Smolensky, P., Fox, B., King, R., and Lewis, C. (1988). Computer-aided reasoned discourse, or, How to argue with a computer. In Guindon, R., editor, *Cognitive Science and Its Applications For Human–Computer Interaction*, pages 109–162. Erlbaum, Hillsdale, NJ.

Smolensky, P., Mozer, M. C., and Rumelhart, D. E., editors (In preparation). *Mathematical Perspectives on Neural Networks*. Lawrence Erlbaum, Hillsdale, NJ.

St. John, M. and McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46:217–258.

Stolcke, A. (1989). Unification as constraint satisfaction in structured connectionist networks. *Neural Computation*, 1:559–567.

Touretzky, D. S. (1989). Towards a connectionist phonology: The "many maps" approach to sequence manipulation. In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, pages 188–195, Ann Arbor, MI. Lawrence Erlbaum.

Touretzky, D. S. and Wheeler, D. W. (1991). Sequence manipulation using parallel mapping networks. *Neural Computation*, 3:98–109.

Touretzky, David S., editor (1991). Special issue: Connectionist approaches to language learning. *Machine Learning*, 7:105–252.

Van Lehn, K., editor (1991). *Architectures for Intelligence.* Erlbaum, Hillsdale, NJ.

Van Valin, R. D. (1990). Semantic parameters of split intransitivity. *Language*, 66:221–260.

Vendler, Z. (1967). Verbs and time. In *Linguistics and Philosophy.* Cornell University Press, Ithaca, NY.

Weigend, A. S., Huberman, B. A., and Rumelhart, D. E. (1990). Predicting the future: A connectionist approach. *International Journal of Neural Systems*, 1:193–209.

Wheeler, D. W. and Touretzky, D. S. (In press). A connectionist implementation of cognitive phonology. In Goldsmith, J. A., editor, *The Last Phonological Rule.* University of Chicago Press, Chicago, IL.

Wiles, J., Humphreys, M. S., Bain, J. D., and Dennis, S. (1990). Control processes and cue combinations in a connectionist model of human memory. Technical Report 186, Department of Computer Science, University of Queensland.

Williamson, J. S. (1979). Patient marking in Lakhota and the Unaccusative Hypothesis. In *Proceedings of the Fifteenth Meeting of the Chicago Linguistic Society*, Chicago, IL.

Zaenen, A. (1989). Unaccusativity in Dutch: An integrated approach. Ms., Xerox PARC and CSLI–Stanford.

Appendix B: French Unaccusativity Tests [excerpt from (Legendre, 1992)]

According to Legendre (1989a), the unergative/unaccusative distinction in French manifests itself productively in several syntactic constructions, including *croire* "believe" constructions, Participial Absolutes, Reduced Relatives, Object Raising, and *on*-interpretation.

B.1. *Croire* constructions

The relevant *croire* construction (henceforth "CR") involves a participial complement. The nominal that surfaces as direct object of the main verb *croire* (henceforth the "target nominal") is typically the direct object of a complement transitive verb with a passive reading (note the optionality of the *par*-phrase, typical of French passive sentences, and the obligatory absence of the passive auxiliary *être*). The target nominal can never be the deep subject of the complement verb. This contrast is illustrated in (0):

(1) a. On avait cru l'enfant kidnappé (par son père).
We believed the child [to have been] kidnapped (by his father).
b. *On avait cru le père kidnappé l'enfant.
We believed the father [to have] kidnapped the child.

Intransitive verbs split into two classes with respect to the *croire* construction, as shown in (0) and (0):

(2) a. On croyait Marie partie/sortie avec sa mère/restée à la maison/innocente.
We believed Mary [to have] left/gone out with her mother/remained at home/innocent.
b. On croyait son père mort d'une crise cardiaque/guéri.
We believed his father [to have] died of a heart attack/(to be) cured.
c. On croyait le moment venu/ses blessures cicatrisées/la production ralentie/l'eau jaillie d'une fontaine.
We believed the moment [to have] come/his wounds [to have] healed/the production (to have) slowed down/the water [to have] sprung out of a fountain.

(3) a. *On croyait l'homme parlé/agi/téléphoné/médité.
We believed the man [to have] spoken/acted/called on the phone/meditated.
b. *On croyait le vieux roi régné toute sa vie/rêvé.
We believed the old king [to have] reigned all his life/dreamt.
c. *On croyait Pierre souffert/faibli/pelé/frémi d'horreur.
We believed Peter [to have] suffered/weakened/peeled/shudder.
d. *On croyait sa blessure enflée/la pluie continuée/la roue grincée/le froid persisté/l'épidémie sévie.
We believed his wound [to have] swollen/the rain [to have] continued/the wheel [to have] grated/the cold weather [to have] persisted/the epidemic [to have] raged.

Legendre (1989a) analyzes the CR construction as a *union* construction (the collapsing of two clauses into one) and formulates the following condition:[1]

(4) **Well-formedness Condition on *croire* Unions** (Legendre, 1989a)
Only a target nominal bearing the GF 2 at some level in the embedded clause can appear in *croire* unions.

1. The standard notation of Relational Grammar is used in the statement of the constraints: 1 = subject, 2 = direct object.

The result is a claim that the intransitive verbs in (2) are unaccusative while the ones in (3) are unergative.

Examples (2) and (3) show that animacy of the target nominal does not differentiate acceptable from unacceptable CR structures. They also show that agentivity by itself does not serve to differentiate them either: in acceptable (2) for example, "leave, go out, remain" are agentive while "(be) innocent, die, cure, heal, slow down, and spring out" are not. In unacceptable (3), we have agentive verbs such as "speak, act, call on the phone, meditate", non-agentive ones such as "suffer, get weaker, peel, shudder" and verbs like "dream" which is agentive only in some of its interpretations (*rêver* can be interpreted as unconscious dreaming during sleep, day-dreaming, or volitional aspiring to a goal).

B.2. Participial Absolutes

Participial Absolutes (PA) are preposed participial clauses followed by a main clause without a coreferential link between the two clauses, hence the name *absolute*. The participial clause itself is in several ways similar to the participial complement in the CR construction. It must have a passive reading; the understood direct object must precede the participial verb, and the understood subject can only (optionally) appear in a *par*-phrase.

(5) a. L'enfant kidnappé (par son père), la police mit la région sens dessus dessous.
The child [having been] kidnapped (by his father), the police turned the whole area upside down.
b. *Son/le père kidnappé l'enfant, la police mit la région sens dessus dessous.
His/the father [having] kidnapped the child, the police turned the whole area upside down.

Similarly, intransitives split up into two classes:

(6) a. Les Dupont partis, toute la famille se mit à table.
The Duponts gone, the whole family sat down for dinner.
b. Le père mort, ils vous retournent le champ. (La Fontaine)
Their father dead, they turned over the field.
c. Les nuages dispersés, nous decidâmes de partir en promenade.
The clouds [having] dispersed, we decided to go for a walk.
d. Une fois la plaie cicatrisée, Marie reprit ses exercices physiques réguliers.
Once the wound healed, Mary started to exercise regularly again.
e. Une fois sa femme guérie, Pierre reprit toutes ses activités.
Once his wife cured, Peter became active again.

(7) a. *Une fois the parents réagis/parlés, l'enfant cessa de rouspéter à table.
Once the parents reacted/spoke, the child stopped grumbling at the table.
b. *Pierre souffert auparavant d'une crise cardiaque, ses parents contactèrent le docteur dès les premiers malaises.
Peter [having] suffered a heart attack before, his parents contacted the doctor as soon as his first symptoms appeared.
c. *La roue grincée de plus en plus fort, Pierre acheta une burette d'huile.
The wheel [having] grated louder and louder, Peter bought an oilcan.

d. *L'enfant frémi d'horreur, le monstre ricana.
The child [having] shuddered, the monster sniggered.
e. *Le plancher vibré, on tendit l'oreille.
The wooden floor [having] vibrated, we pricked up our ears.

Legendre (1989a) argues for a well-formedness condition on PAs which requires the preposed argument in the participial clause to a) bear the GF 2 and b) to advance to 1 (by passive or unaccusative advancement).

(8) **Well-formedness Condition on Participial Absolutes** (Legendre, 1989a)
A Participial Absolute clause is well-formed only if there is 2-1 advancement in the clause.

She notes however, that this condition is a *necessary* condition only since an aspectual restriction prevents certain PAs satisfying the syntactic condition from being acceptable:

(9) a. *Ses efforts continués, tout le monde crut qu'il réussirait.
His efforts [having] continued, everybody thought he would succeed.
b. *(Une fois) l'eau coulée, tout le monde put se désaltérer.
(Once) the water [having] flowed, everybody was able to quench their thirst.
c. *(Une fois) l'état du malade empiré, le docteur ordonna son transport à l'hôpital.
(Once) his condition [having] worsened, the doctor ordered him to be taken to the hospital.

The restriction in question is that non-perfective verbs cannot occur in PAs.

As the examples above show, semantic notions like animacy or volitionality cannot by themselves differentiate unaccusatives from unergatives. We find both volitional and non-volitional participants among acceptable PAs (6) as well as unacceptable ones (7).

B.3. Reduced Relatives

The Reduced Relative (RR) construction discussed below is yet another participial construction whose meaning corresponds to a full relative clause (with relative pronoun *qui* "who/which" and auxiliary *être* "be" or *avoir* "have"): hence its name here, *reduced relatives*; Legendre (1989a) argues that such clauses are well-formed only if they involve a 2-1 advancement, the same syntactic condition as PAs.

(10) a. La petite fille kidnappée (par un inconnu) le mois dernier n'a jamais été retrouvée.
The young girl [who was] kidnapped (by a stranger) last month has never been found.
b. *Le bandit kidnappé l'enfant n'a jamais été arrêté.
The bandit [who had] kidnapped the child has never been arrested.

The contrast between (10a) and (10b) shows that initial 2s must occur in preverbal position in RRs. Initial 1s, if they appear at all, must do so in a *par*-phrase. (0) and (0) show that intransitive verbs, again, split into two classes.

(11) a. La personne morte hier soir/évanouie/évadée/assise au premier rang....
The person [who] died last night/[who] fainted/[who] escaped/[who] sat in the first row....

b. Les nuages dispersés/le moment venu/l'eau bouillie/l'eau jaillie de la fontaine...
The clouds [which] dispersed/the moment [which] came/the water [which] boiled/the water [which] sprung out of the fountain....

(12) a. *L' homme réagi/ronchonné/sévi/parlé/souffert d'une crise cardiaque/faibli/frémi d'horreur/rêvé....
The man [who] reacted/grumbled/acted ruthlessly/spoke (up)/suffered a heart attack/got weaker/shuddered/dreamt....
b. *Sa blessure enflée/la pluie continuée/la roue grincée/le froid persisté/l'épidémie sévie....
His wound [which had] swollen/the rain [which had] continued/the wheel [which had]) grated/the cold weather [which had] persisted/the epidemic [which had] raged....

The data in (11) and (12) shows that RRs, just like CR and PAs, are not sensitive to animacy or volitionality.

B.4. Object Raising

Object Raising (OR) with a limited class of predicates such as "be difficult, easy" does not share any superficial properties with the constructions discussed so far. Typically, the surface subject of the main predicate (referred to below as the "raisee") is understood to be the direct object of the embedded infinitival verb:

(13) a. Ce genre d'échec est difficile à oublier.
This type of failure is difficult to forget.
b. Un adversaire est souvent facile à éliminer.
An opponent is often easy to eliminate.

Legendre (1986, 1989a) argues at length for the following condition:

(14) **Well-formedness Condition on Object Raising** (Legendre, 1986, 1989a)
An Object Raising construction is well-formed only if the raisee bears only the GF 2 in any clause below the raising predicate.

Legendre (1986, 1989a) argues that the only intermediate clauses that can appear under this type of raising are union constructions, in particular causative *faire* constructions. Note that an embedded subject cannot raise in this construction, as shown by the causee "children" in (0b).

(15) a. La vérité n'est pas facile à faire dire aux enfants.
The truth is not easy to make children tell.
b. *Les enfants ne sont pas faciles à faire dire la vérité.
Children are not easy to make tell the truth.

Interestingly enough, intransitives split into two classes with respect to OR:

(16) a. L'eau pour le thé sera difficile à faire bouillir.
The water for tea will be difficult to make boil.
b. La bombe sera facile à faire exploser.
The bomb will be easy to make explode.

c. L' enfant sera facile à faire évanouir/à faire taire/à faire asseoir au premier rang.
The child will be easy to make faint/to make keep quiet/to make sit down in the first row.

d. Le prisonnier sera facile à faire parler.
The prisoner will be easy to make talk.

(17) a. *L'empereur sera facile à faire régner sur son territoire.
The emperor will be easy to make reign over his territory.

b. *Un homme est facile à faire méditer/réfléchir/rêver.
A man is easy to make meditate/reflect or think/dream.

c. *L' enfant sera facile à faire s'évanouir/à faire se taire/à faire s'asseoir au premier rang.
The child will be easy to make faint/to make keep quiet/to make sit down in the first row.

d. *La rivière sera impossible à faire geler/un complot sera facile à faire mijoter.
The river will be impossible to make freeze/a conspiration will be easy to make brew.

Note, in particular, the contrast between (16c) and (17c): unacceptability follows from the presence of the reflexive free morpheme *se*. Legendre (1986) argues at length that *se* registers the presence of a 1 that gets cancelled (see reference for details).

Legendre (1989a) is careful to point out that the condition on OR restricting raisees to be 2s "all the way" is only a necessary condition. Semantic factors interact, with the consequence that sometimes a raising structure satisfying the necessary syntactic condition fails to be acceptable:

(18) a. *Des ailes ne sont pas faciles à avoir.
Wings are not easy to have.

b. *Pierre est impossible à faire rester à la maison.
Peter is impossible to make stay at home.

c. *An ennemi est facile à faire décéder. (NOTE: OK with *mourir* "die")
An enemy is easy to make pass away.

Legendre (1989a) argues that the restriction is not, strictly speaking, a restriction banning stative verbs; rather, the restriction correlates highly with compatibility vs. incompatibility with the progressive aspect, expressed periphrastically in French with *en train de* "in the process of."

(19) a. *Il est en train d'avoir des ailes/de rester à la maison/de décéder.
He is in the process of having wings/staying at home/passing away.

b. Il est en train de mourir.
He is in the process of dying.

Consider next animacy and volitionality of the raisee. Take (16c) for example: *faint* is non-volitional while *keep quiet*, *sit down* are volitional. What matters here is the absence of *se*. Volitionality is the same in (17c) with *se*; yet the result is unacceptable. At first glance, one factor appears to be pragmatic in nature: the degree to which the understood agent can manipulate the raisee; this is not surprising given the meaning of these OR structures (something that is easy or not to do!). In support of this, one can consider (16d) and the reality of specific methods used to "make prisoners speak", against their will. For my consultants, when *parler* is not predicated of people that can easily be manipulated by force, the corresponding OR structure is comparatively worse:

(20) a. * ?Le président de la république sera facile à faire parler.
The president of the republic will be easy to make speak.

b. *Le conférencier sera facile à faire parler de politique.
The conference speaker will be easy to make speak about politics.

Under scrutiny, however, it is clear that this pragmatic factor cannot by itself account for all contrasts in OR structures. First of all, if one uses a stronger causative verb *forcer* "force" (instead of *faire* "make"), OR constructions are unacceptable, regardless of the semantics or pragmatics of the situation:

(21) a. *Le prisonnier sera facile à forcer à parler.
The prisoner will be easy to force to speak.
b. *Les enfants seront faciles à forcer à (se) taire.
The children will be easy to force to keep quiet.

Legendre (1986) argues for the following syntactic explanation: *forcer* governs Equi, a bi-clausal structure at all levels: one important property is that the Equi "victim" must head a final 1-arc in the complement (it is required by the Final 1 Law and supported by empirical evidence; see Legendre (1986) for details); consequently, the Equi victim which is also the raisee in (21a-b) violates the condition on OR. *Faire* governs union (i.e. the structure is underlyingly biclausal and superficially monoclausal), rendering the Final 1 Law irrelevant for the complement; the raisee thus satisfies the OR condition (which precludes the raisee from heading any arc other than a 2-arc, prior to raising).

Another reason why semantic properties of the raisee cannot be a sufficient condition for acceptability of OR structures is the unacceptability of (17b), despite the fact that one can readily conceive of contexts where a man could be manipulated into meditating, thinking, or day-dreaming.

A final, and crucial, reason to reject the semantics of the raisee as the crucial factor is that it would leave totally unexplained the contrast between OR structures acceptable without *se* (16c) and unacceptable with *se* (17c). In this respect it is important to note that in simple sentences, *se* is obligatory (satisfying the Final 1 Law) and there are reportedly dialects of French which do not allow the *se*/no *se* alternation under *faire* (why should this be the case if the presence vs. absence of *se* has an important semantic function?). I conclude from this discussion that semantic properties of the raisee may be a necessary condition for acceptability but not a sufficient one. There is no doubt that in these constructions, the restriction on *se* – however it might be formulated – overrides any other considerations.

B.5. *On*-interpretation

French has a special third person singular subject pronoun, *on* which allows two distinct interpretations: it can have a definite interpretation, i.e. first person plural "we," or an arbitrary interpretation, referring to an unspecified individual or group of individuals (translated below as "someone"). The property of *on* relevant here, defining the "ON" test, is that the arbitrary interpretation is restricted to initial 1s (of unergative and transitive structures). All data are from Legendre (1989b).

(22) a. On a téléphoné à Pierre;on a désobéi aux ordres du capitaine.
Someone/we called Peter on the phone;someone/we disobeyed [to] the captain’s orders.
b. On a chaleureusement félicité la candidate;on lui a confié une tâche délicate.
Someone/we warmly congratulated the candidate;someone/we entrusted her with a tricky task.

On the basis of the previously discussed tests that are claimed to positively identify unaccusativity, *téléphoner* is unergative (as it fails all those tests); *on* has two interpretations, as shown in the translations. In contrast, passive and unaccusative structures are not ambiguous, as shown next. *On* can only be interpreted to mean "we."

(23) a. On a été arrêté (par la police) avant même de franchir la porte d'entrée.
*We were [*someone was] arrested (by the police) before even passing the entrance door*.
b. On s'est enfin tu; on était innocent.
*We [*someone] finally shut up; we were [*someone was] innocent.*

An RG analysis provides a natural explanation for this patterning: both passive and unaccusative advancement involve 2-1 advancement. Interpretation of *on* then allows us to positively identify unergative verbs.[2]

Turning to semantic restrictions, humanness is obviously a necessary condition for *on* interpretation (*on* can never be used for non-human subjects) but volitionality is not, as the examples in (23b) demonstrate (*se taire* is volitional, *être innocent* is not).

2. To my knowledge, there is only one other unergativity test in French: Impersonal Passive with intransitive verbs that subcategorize for an indirect object or oblique argument. This somewhat peculiar additional restriction makes the number of verbs which actually occur in the impersonal passive construction very small and thus is not a productive test for identifying unergative verbs (see Postal, 1986 and Legendre, 1989b for further discussion).