# Optimality: From Neural Networks to Universal Grammar

## Alan Prince* and Paul Smolensky*

Can concepts from the theory of neural computation contribute to formal theories of the mind? Recent research has explored the implications of one principle of neural computation, optimization, for the theory of grammar. Optimization over symbolic linguistic structures provides the core of a new grammatical architecture, optimality theory. The proposition that grammaticality equals optimality sheds light on a wide range of phenomena, from the gulf between production and comprehension in child language, to language learnability, to the fundamental questions of linguistic theory: What is it that the grammars of all languages share, and how may they differ?

It is evident that the sciences of the brain and those of the mind are separated by many gulfs, not the least of which lies between the formal methods appropriate for continuous dynamical systems and those for discrete symbol structures. Yet recent research provides evidence that integration of these sciences may hold significant rewards. Research on neural computation has identified optimization as an organizing principle of some generality, and current work is showing that optimization principles can be successfully adapted to a central domain within the theory of mind: the theory of grammar. In this article, we explore how a reconceptualization of linguistic theory through optimization principles provides a variety of insights into the structure of the language faculty, and we consider the relations between optimality in grammar and optimization in neural networks.

Some of the contributions of the optimization perspective on grammar are surprising. The distinction between linguistic knowledge in the abstract and the use of this knowledge in language processing has often been challenged by researchers adopting neural network approaches to language; yet we show here how an optimization architecture in fact strengthens and rationalizes this distinction. In turn, this leads to new formal methods by which grammar learners can cope with the demands of their difficult task, and new explanations for the gap in complexity between the language children produce and the language they can comprehend. Optimization also provides a fresh perspective on the nature of linguistic constraints, on what it is that grammars of different human languages share, and on how grammars may differ. And this turns out to provide considerable analytical leverage on central aspects of the long-standing problems in language acquisition.

## Optimality Theory

Linguistic research seeks to characterize the range of structures available to human language and the relationships that may obtain between them, particularly as they figure in a competent speaker's internalized "grammar" or implicit knowledge of language. Languages appear to vary widely, but the same structural themes repeat themselves over and over again, in ways that are sometimes obvious and sometimes clear only upon detailed analysis. The challenge, then, is to discover an architecture for grammars that both allows variation and limits its range to what is actually possible in human language.

A primary observation is that grammars contain constraints on the well-formedness of linguistic structures, and these constraints are heavily in conflict, even within a single language. A few simple examples should bring out the flavor of this conflict. English operates under constraints entailing that its basic word order is subject-verb-object; yet in a sentence like *what did John see?* it is the object that stands first. This evidences the greater force of a constraint requiring question-words like *what* to appear sentence-initially. Yet even this constraint is not absolute: One must say *who saw what?* with the object question-word appearing in its canonical position; the potential alternative, *who what saw?*, with all question-words clumped at the front, which is indeed grammatical in some languages, runs afoul of another principle of clause structure that is, in English, yet stronger than the requirement of initial placement of question-words. Thus, *who saw what?* is the grammatical structure, satisfying the constraints of the grammar not perfectly, but optimally: No alternative does better, given the relative strength of the constraints in the grammar of English.

Similar conflicts abound at all levels of linguistic structure. In forming the past tense of "slip," spelled "slipped" but pronounced *slipt*, a general phonological constraint on voicing in final consonant sequences favors the pronunciation *pt* over *pd*, conflicting with the requirement that the past-tense marker be given its basic form *-d*; and the phonological constraint prevails (1). In an English sentence like *it rains*, a constraint requiring all words to contribute to meaning (unlike the element *it* in this usage) conflicts with a structural constraint requiring all sentences to have subjects; and the latter controls the outcome. Such examples indicate that a central element in the architecture of grammar is a formal means for managing the pervasive conflict between grammatical constraints.

The key observation is this: In a variety of clear cases where there is a strength asymmetry between two conflicting constraints, no amount of success on the weaker constraint can compensate for failure on the stronger one. Put another way: Any degree of failure on the weaker constraint is tolerated, so long as it contributes to success on the stronger constraint. Extending this observation leads to the hypothesis that a grammar consists entirely of constraints arranged in a strict domination hierarchy, in which each constraint is strictly more important than—takes absolute priority over—all the constraints lower-ranked in the hierarchy. With this type of constraint interaction, it is only the ranking of constraints in the hierarchy that matters for the determination of optimality; no particular numerical strengths, for example, are necessary. Strict domination thus limits drastically the range of possible strength-interactions between constraints to those representable with the algebra of total order.

Strict domination hierarchies composed of very simple well-formedness constraints can lead to surprisingly complex grammatical consequences. Furthermore, different rankings of the same set of constraints can give rise to strikingly different linguistic patterns. These properties show that strict domination, though a narrow mechanism, answers to the basic requirements on the theory of human language, which must allow grammars to be built from simple parts whose combination leads to specific kinds of complexity and diversity. Optimality theory, originally presented in 1991 (2), offers a particularly strong version of a strict-domination–based approach to grammatical op-

A. Prince is in the Department of Linguistics and Rutgers Center for Cognitive Science, Rutgers University, 18 Seminary Place, New Brunswick, NJ 08903, USA. E-mail: prince@ruccs.rutgers.edu  P. Smolensky is in the Department of Cognitive Science, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218–2685, USA. E-mail: smolensky@cogsci.jhu.edu

*Both authors contributed equally to this work.

timization. Optimality theory hypothesizes that the set of well-formedness constraints is universal: not just universally available to be chosen from, but literally present in every language. A grammar for a particular language results from imposing a strict-domination ranking on the entire universal constraint set. Also universal is the function that determines, for each input to the grammar, the set of candidate output structures that compete for optimality; every language considers exactly the same set of options for realizing an input. The observed force of a given constraint can vary from absolute (never violated) to nil (always violated), with many stops and steps along the way, depending on its position in the strict domination hierarchy for a given language, and depending on the membership in the output candidate set for a given input.

Optimality theory thus provides a direct answer to the classic questions of linguistic theory: What do the grammars of different languages have in common, and how may they differ? What they share are the universal constraints and the definition of which forms compete; they differ in how the constraints are ranked, and, therefore, in which constraints take priority when conflicts arise among them. For example, the two constraints in conflict in English *it rains* are ranked differently in Italian: The constraint against meaningless words outranks that against subjectless sentences, and the resulting grammatical sentence is simply *piove* (literally, "rains").

Optimality theory connects a number of lines of research that have occupied linguists in the last several decades: the articulation of universal formal principles of grammars; the generalization of well-formedness constraints across the outputs of formally disparate mechanisms; the descriptive use of informal notions of linguistic optimization; and output-oriented analysis (3). Such a unification is made possible by the basic notion that grammaticality means optimally satisfying the conflicting demands of violable constraints.

## Markedness and Faithfulness Constraints

Within the universal constraint set, several subclasses have been distinguished. One class of universal constraints in optimality theory formalizes the notion of structural complexity, or markedness (4). Grossly speaking, an element of linguistic structure is said to be marked if it is more complex than an alternative along some dimension; the relevant dimensions may sometimes correlate with comprehension, production, memory, or related physical and cognitive

functions. The word-final consonant cluster *pd* is more marked than *pt*; sentences lacking subjects are more marked than those with subjects. Marked elements tend to be absent altogether in certain languages, restricted in their use in other languages, later-acquired by children, and in other ways avoided. This cluster of properties diagnostic of marked elements is given a uniform explanation in optimality theory, which follows from their formal characterization: Marked structures are those that violate structural constraints. We will call the set of all such constraints STRUCTURE.

Phonological STRUCTURE constraints often induce context-dependent alteration of pronunciations. For example, the markedness of *pd* relative to *pt* is responsible for the alteration of the past-tense suffix *d* to *t* in "slipped"; this is a context in which the more marked cluster is avoided. A more dramatic alteration is common in French, driven by syllabic markedness constraints. [Our presentation simplifies somewhat for ease of exposition (5).] One such constraint, NOCODA, is violated by any syllable ending with a consonant—a closed syllable; the syllable-closing consonant is called a coda. Closed syllables are marked relative to syllables ending with a vowel. Another constraint, ONSET, is violated by syllables that begin with a vowel.

In French, the masculine form of the word for "small," written "petit," is pronounced with or without the final *t*, depending on the context. Spelling, though often merely conventional, in this case accurately represents the abstract sound-sequence that a speaker internalizes when the word is learned; we write this sequence /petit/. When the following word is vowel-initial, the final *t* is pronounced, beginning a syllable—*pe.ti.t oeuf* "little egg." Elsewhere—when the following word begins with a consonant, or when there is no following word—/petit/ is pronounced *pe.ti*, with loss of the final lexical *t*—*pe.ti. chien* "little dog." (Adjacent syllables are separated by a period in the examples.) The phonological grammar of French determines how "small" is pronounced in a given context, that is, which grammatical "output" (pronunciation) corresponds to an "input" /petit. . ./. The final *t* is not pronounced when so doing would violate NOCODA; the constraint ONSET determines that when the *t* precedes a vowel, it begins a syllable and is pronounced.

A second class of universal constraints in optimality theory, FAITHFULNESS constraints, is a direct consequence of the optimization perspective (6). An optimal (grammatical) representation is one that optimally satisfies the constraint ranking among those representations containing a

given input. The existence of many different optimal representations is due to the existence of many different inputs. The FAITHFULNESS constraints tie the success of an output candidate to the shape of the corresponding input; each FAITHFULNESS constraint asserts that an input and its output should be identical in a certain respect. For example, the constraint called PARSE asserts that every segment of the input must appear in the output; it penalizes deletion of material in the input-output mapping. [The French input-output pair (/petit/, *pe.ti*) shows a violation of PARSE.] Another constraint, known as FILL, penalizes insertion of new material that is not present in the input. Other constraints demand featural identity—one of these is violated when the English past-tense suffix *d* is pronounced *t*. As with all constraints in the universal set, these constraints are violable, and much grammar turns on resolving the tension between STRUCTURE constraints, which favor simple structures, and the FAITHFULNESS constraints, which favor exact replication of the input, even at the cost of structural complexity.

As a general illustration of this relation, consider the confrontation between PARSE and NOCODA, which must play out in every language. These constraints are in conflict, because one way to avoid a closed syllable (thereby satisfying NOCODA) is to delete any consonant that would appear in syllable-final position (thereby violating PARSE, which forbids deletion). Consider first a grammar in which NOCODA dominates PARSE, which we will write as NO-CODA ≫ PARSE. Syllabification is grammatically predictable, and therefore need not be present in the input. Suppose a hypothetical unsyllabified input word /batak/ is submitted to this grammar for syllabification and pronunciation. A large range of syllabified candidate outputs (pronunciations) is to be evaluated, among which we find the faithful *ba.tak*, and the progressively less faithful *ba.ta, bat, ba, b* and Ø [silence], as well as *a.tak, tak, ak,* and many, many others. Observe that a very wide range of candidate output options is considered; it is the universal constraint set, ranked, which handles the bulk of the selection task.

Which of these candidates is optimal, by the hierarchy NOCODA ≫ PARSE? The faithful form *ba.tak*, which ends on a closed syllable, is ruled out by top-ranked NO-CODA, because there are other competing output candidates that satisfy the constraint, lacking closed syllables. Among these, *ba.ta* is the most harmonic, because it involves the least violation of PARSE—a single deletion. It is therefore the optimal output for the given input: The grammar

certifies the input-output pair (/batak, *ba.ta*) as well-formed; the final lexical *k* is unpronounced. The optimality computation just sketched can be represented conveniently in a constraint tableau as shown in Fig. 1A. (For the sake of expositional simplicity, we are ignoring candidate outputs like *ba.ta.ki*, in which new material—the vowel *i*—appears at the end, resulting in a form that also avoids closed syllables successfully. Dealing with such forms involves ranking the anti-insertion constraint FILL with respect to PARSE; when FILL ≫ PARSE, deletion rather than insertion is optimal.) We conclude that in a language where NOCODA ≫ PARSE, all syllables must be open; for any output candidate with a closed syllable, there is always a better competitor that lacks it.

Consider now a grammar in which, contrariwise, we have PARSE ≫ NOCODA (Fig. 1B). Given the input /batak/, we have exactly the same set of output candidates to consider, because the candidate set is determined by universal principles. But now the one violation of PARSE in *ba.ta* is fatal; instead, its competitor *ba.tak*, which has no losses, will be optimal. (In the full analysis, we set FILL ≫ NOCODA as well, eliminating insertion as an option.) The dominance of the relevant FAITHFULNESS constraints ensures that the input will be faithfully reproduced, even at the cost of violating the STRUCTURE constraint NOCODA. This language is therefore one like English in which syllables will have codas, if warranted by the input.

Domination is clearly "strict" in these examples: No matter how many consonant clusters appear in an input, and no matter how many consonants appear in any cluster, the first grammar will demand that they all be simplified by deletion (violating PARSE as much as is required to eliminate the occasion for syllable codas), and the second grammar will demand that they all be syllabified (violating NOCODA as much as is necessary). No amount of failure on the violated constraints is rejected as excessive, as long as failure serves the cause of obtaining success on the dominating constraint.

Constraint interaction becomes far more intricate when crucial ranking goes to a depth of three or more; it is not unusual for optimal forms to contain violations of many constraints. Optimality-theoretic research in syllable structure expands both the set of relevant STRUCTURE constraints and the set of FAITHFULNESS constraints that ban relevant disparities between input and output. The set of all possible rankings provides a restrictive typology of syllable structure patterns that closely matches the basic empirical findings in the area, and even refines prior classifications. Many other areas of phonology and syntax have been subject to detailed investigation under optimality theory (7). Here as elsewhere in cognitive science, progress has been accompanied by disputes at various levels, some technical, others concerning fundamental matters. The results obtained to date, however, provide considerable evidence that optimization ideas in general and optimality theory in particular can lead to significant advances in resolving the central problems of linguistic theory.

## Optimality Theory and Neural Network Theory

The principal empirical questions addressed by optimality theory, as by other theories of universal grammar, concern the characterization of linguistic forms in and across languages. A quite different question is, can we explicate at least some of the properties of optimality theory itself on the basis of more fundamental cognitive principles? A significant first step toward such an explanation, we will argue, derives from the theory of computation in neural networks.

Linguistic research employing optimality theory does not, of course, involve explicit neural network modeling of language. The relation we seek to identify between optimality theory and neural computation must be of the type that holds between higher level and lower level systems of analysis in the physical sciences. For example, statistical mechanics explains significant parts of thermodynamics from the hypothesis that matter is composed of molecules, but the concepts of thermodynamic theory, like "temperature" and "entropy," involve no reference whatever to molecules. Like thermodynamics, optimality theory is a self-contained higher-level theory; like statistical mechanics, we claim, neural computation ought to explain fundamental principles of the higher level theory by deriving them as large-scale consequences of interactions at a much lower level. Just as probabilistic systems of point particles in statistical mechanics give rise to nonprobabilistic equations governing bulk continuous media in thermodynamics, so too should the numerical, continuous optimization in neural networks give rise to a qualitatively different formal system at a higher level of analysis: the nonnumerical optimization over discrete symbolic representations—the markedness calculus—of optimality theory.

To make contact with the abstract level at which mental organization like that of grammar resides, the relevant concepts of neural computation must capture rather high-level properties (8). Because of the complexity and nonlinearity of general neural network models, such concepts are in short supply; one of the few available is the method of Lyapunov functions. Such a function assigns a number to each possible global state of the dynamical system in such a way that as the system changes state over time, the value of the function continually increases. Lyapunov functions have been identified for a variety of model neural networks, and given various names, the term "energy function" being the most popular (9). We will use the term "harmony function" because the work we discuss follows most directly along the path initiated in

**Fig. 1.** A constraint tableau in optimality theory. The table in (A) displays the optimality computation in graphic form. The input is listed at the head of the first column, and (selected) output candidates occupy the cells below it. The constraints in the hierarchy are listed in domination order left-to-right across the first row. Other rows show the evaluation of a candidate with respect to the constraint hierarchy. The hand points to the optimal candidate. An asterisk indicates a constraint violation; the number of asterisks in a cell corresponds to the number of times the constraint is violated: for example, there are three asterisks in the PARSE cell of row (d) because the input-output pair (/batak/, *ba*) involves three instances on non-parsing or deletion of segments. The exclamation point marks a fatal violation—one that ensures suboptimal status. Cells after the fatal violation are shaded, indicating that success or failure on the constraint heading that column is irrelevant to the optimality status of the candidate, which has already been determined by a higher-ranked constraint. In this example, which recapitulates the discussion of the mini-grammar NOCODA ≫ PARSE in the text, the interaction of just two constraints is depicted, and only a small sampling of the candidate set is shown. Given this ranking, the word /batak/ would be pronounced *bata*, as in the optimal candidate (b). Tableau B shows the effect of reranking; in a different language, in which PARSE ≫ NOCODA, candidate (a) would be optimal; /batak/ would therefore be pronounced *batak*.

**A**

| /batak/ | | NOCODA | PARSE |
|---|---|---|---|
| **a** | | ba.tak | *! | |
| **b** ☞ | | ba.ta | | * |
| **c** | | bat | *! | ** |
| **d** | | ba | | * *! * |

**B**

| /batak/ | | PARSE | NOCODA |
|---|---|---|---|
| **a** ☞ | | ba.tak | | * |
| **b** | | ba.ta | *! | |
| **c** | | bat | *! * | * |
| **d** | | ba | *! * * | |

harmony theory (*10*).

In the particular class of model neural networks admitting a harmony function, the input to a network computation consists of an activation pattern held fixed over part of the network. Activation then flows through the net to construct a pattern of activity that maximizes—optimizes—harmony, among all those patterns of activity that include the fixed input pattern. The harmony of a pattern of activation is a measure of its degree of conformity to the constraints implicit in the network's "synapses" or connections. As illustrated in Fig. 2, A to C, an inhibitory connection between two model "neurons" or "units," modeled as a negative weight, embodies a constraint that when one of the units is active, the other should be inactive; this is the activation configuration that maximizes harmony at that connection. An excitatory connection, modeled as a positive weight, embodies the constraint that when one of the units is active, the other should be active as well. In a complex, densely interconnected network of units, such constraints typically conflict; and connections with greater numerical magnitude embody constraints of greater importance to the outcome. A complete pattern of activation that maximizes harmony is one that optimally balances the typically conflicting demands of all the constraints in the network.

An activity pattern can be understood as a representation of the information that it constitutes; the harmony of any activity pattern measures the well-formedness of that representation with respect to the constraint-system embodied in the connection weights. For a fixed input, a harmony-maximizing network produces the output it does because that is the most well-formed representation containing the input. The knowledge contained in the network is the set of constraints embodied in its synaptic connections, or equivalently, the harmony function these constraints define. This knowledge can be used in different ways during processing, by fixing input activity in different parts of the network and then letting activation flow to maximize harmony (Fig. 2D).

Because the harmony function for a neural network performs the same well-formedness–defining function as the symbol-sensitive mechanisms of grammar, it is natural to investigate harmony maximization as a means of defining linguistic grammars. In carrying out this program, two major problems arise: finding a suitable notion of optimization over linguistic structures; and finding a relation between this abstract measure and the numerical properties of neural computation. The second problem might seem sufficiently intractable to un-

dermine the enterprise, no matter how the first is executed. Linguistic explanations depend crucially on representations that are complex hierarchical structures: Sentences are built of phrases nested one inside the other; words are constructed from features of sounds, grouped to form phonetic segments, themselves grouped to form syllables and still larger units of prosodic structure. At first glance, the assumption that mental

representations have such structure does not seem compatible with neural network models in which representations are patterns of activation—vectors, mere strings of numbers. But a family of interrelated techniques developed over the past decade show that patterns of activation can possess a precise mathematical analog of the structure of linguistic representations (*11*); the basic idea is illustrated in Fig. 3.



**Fig. 2.** Harmony maximization in a neural network. The basic harmony function for a neural network is simply $H = \Sigma_{ij} a_i w_{ij} a_j$, where $a_i$ is the activation of unit (abstract neuron) $i$ and $w_{ij}$ is the strength or weight of the connection to unit $i$ from unit $j$. In (A), units $i$ and $j$ are connected with a weight of $-2$; this inhibitory connection constitutes a constraint that if one of these units is active, the other should be inactive. The microactivity pattern shown in (B) violates this constraint (marked with an asterisk): Both units have activity $+1$, and the constraint violation is registered in the negative harmony $a_i w_{ij} a_j = (+1)(-2)(+1) = -2$. The activity pattern in (C) satisfies the constraints, with harmony $+2$. Of these two micropatterns, the second maximizes harmony, as indicated by the hand. In a network containing many units, the harmony of a complete activity pattern is just the sum of all the microharmonies computed from each pair of connected units. In (D), a hypothetical network is depicted for relating English phonological inputs and outputs. The topmost units contain a pattern for the pronunciation *slipt* "slipped"; the units at the bottom host a pattern for the corresponding interpretation /slip+d/. In between, units support a pattern of activity representing the full linguistic structure, including syllables, stress feet, and so on. The connections in the network encode the constraints of English phonology. When the pattern for /slip+d/ is imposed on the lowest units, activation flows to maximize harmony, giving rise to the pattern for *slipt* on the uppermost units; this is production-directed processing. In comprehension, the pattern for *slipt* is imposed on the uppermost units, and harmony maximization fills in the rest of the total pattern, including the interpretation /slip+d/.
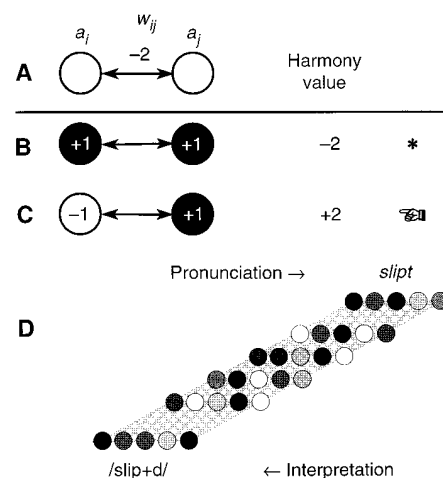
**Fig. 3.** Realizing structured representations as patterns of activity in neural networks. The top plane shows a pattern of activity **p** realizing the structure $\underset{X \quad Y}{\wedge}$ (for example, a sentence in which X is the noun-phrase subject *big dogs* and Y a verb-phrase predicate *bite*); gray levels schematically indicate the activity levels of units in a neural network (circles). This pattern is produced by superimposing a pattern **x** realizing X in the left position (middle plane) on another pattern **y** realizing Y in the right position (bottom plane). Within the middle plane, the pattern **x** is a pattern **X** for X (right edge) times a pattern **p**$^{/}$ for "left position" (bottom edge).



The product operation here is the tensor product: In **x**, the activity level of the unit in row $i$ and column $j$ is just the activation of unit $i$ in **X** times the activation of unit $j$ in **p**$^{/}$; and analogously for pattern **y**. Algebraically:

$$\mathbf{p} = \mathbf{x} + \mathbf{y}; \quad \mathbf{x} = \mathbf{p}^{/} \otimes \mathbf{X}; \quad \mathbf{y} = \mathbf{p}^{\backslash} \otimes \mathbf{Y}; \quad (\mathbf{x})_{ij} = (\mathbf{p}^{/})_i (\mathbf{X})_j; \quad (\mathbf{y})_{ij} = (\mathbf{p}^{\backslash})_i (\mathbf{Y})_j$$

Because tensor products may be nested one inside the other, patterns may realize structures embedded in other structures. Through simple neural network operations, massively parallel structure manipulation may be performed on such patterns.
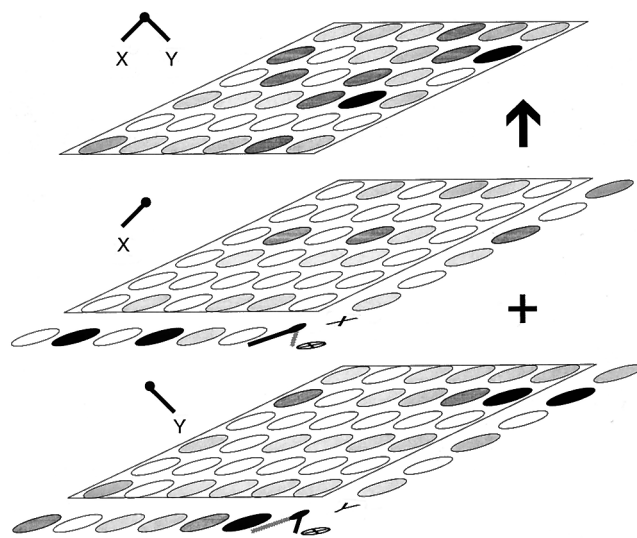
In this setting, the harmony of a linguistic structure is just the harmony of the pattern of activity realizing that structure. The connections in the network define which linguistic structures have maximal harmony—which are grammatical. This directly suggests the notion of a "harmonic grammar," a set of soft or violable constraints on combinations of linguistic elements, in which each constraint would have a numerical strength (*12*). This strength is the quantity by which the harmony of a linguistic representation is diminished when the constraint is violated; through an activation-passing computation implementing the harmony function, the strengths determine which constraints are respected, and to what degree, whenever there is conflict; a grammatical structure is then one that best satisfies the total set of constraints defining the grammar, that is, has maximal harmony.

This conception is straightforward, but obviously incomplete, for it is far from true that every weighting of the set of linguistic constraints produces a possible human language. To delimit the optimizing function narrowly enough, the strength relation between constraints must be severely regimented. And this is exactly what strict domination provides: In optimality theory, no amount of success on weaker constraints can compensate for failure on a stronger

one. This corresponds to the numerical strength of a constraint being much greater than the strengths of those constraints ranked lower than it in the hierarchy; so much so that the combined force of all the lower-ranked constraints can never exceed the force of the higher-ranked constraint. But as we have seen, strict domination means constraint interaction in grammar is highly restricted: Only the relative ranking of constraints, and not particular numerical strengths, can be grammatically relevant. The grammatical consequence is that, in many cases studied to date, the set of all rankings delimits a narrow typology of possible linguistic patterns and relations.

That strict domination governs grammatical constraint interaction is not currently explained by principles of neural computation; nor do these principles explain the universality of constraints that is central to optimality theory and related approaches. These are stimulating challenges for fully integrating optimality theory with a neural foundation. But the hypothesis that grammar is realized in a harmony-maximizing neural network rationalizes a significant set of crucial characteristics of optimality theory: Grammaticality is optimality; competition for optimality is restricted to representations containing the input; complexity arises through the interaction of simple constraints, rather than within the

constraints themselves; constraints are violable and gradiently satisfiable; constraints are highly conflicting; conflict is adjudicated via a notion of relative strength; a grammar is a set of relative strengths; learning a grammar is adjusting these strengths. OT's markedness calculus is exactly neural network optimization, specialized to the case of strict domination.

If the hypothesis that grammar is realized in a harmony-maximizing neural network is correct, we would expect that it would lead to new developments in optimality theory. We now turn to recent such work.

## Linguistic Knowledge and Its Use

Just as a numerically valued harmony function orders the activity patterns in a model neural network from highest to lowest harmony, the ranking of constraints of an optimality theoretic grammar orders linguistic structures from most to least harmonic: from those that best to those that least satisfy the constraint hierarchy. It is the constraint ranking and the ordering of structures it provides that is OT's characterization of knowledge of grammar.

Using this knowledge involves finding the structures that maximize harmony, and this can be done in several ways (*13*), directly following the lead of the corresponding neural network approach of Fig. 3. Use of grammatical knowledge for comprehending language involves taking the pronunciation of, say, a sentence, and finding the maximal-harmony linguistic structure with that pronunciation. This structure groups the given words into nested phrases, and fills in implied connections between words, such as the possible interpretive link between *John* and *him* in *John hopes George admires him* (him = John), and the necessary anti-link in *John admires him* (him ≠ John). The maximum-harmony structure projected from the pronounced sentence by the grammar plays an important role in determining its meaning.

Producing a sentence is a different use of the very same grammatical knowledge. Now the competition is among structures that differ in pronunciation, but share a given interpretation. The ordering of structures from most to least harmonic constitutes grammatical knowledge that is separate from its use, via optimization, in comprehension and production; this is depicted schematically in Fig. 4.

This view leads to a new perspective on a classic problem in child language. It is well known that, broadly speaking, young children's linguistic abilities in comprehension greatly exceed their abilities in production. Observe that this is a richer problem than many perception-action disparities—
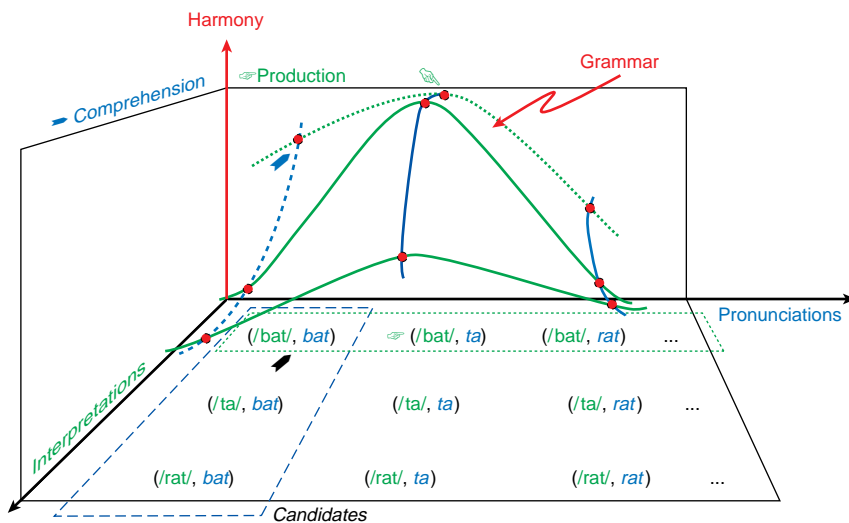


**Fig. 4.** Knowledge versus use of grammar in optimality theory. The pair (/bat/, *ta*) represents a structure in which the lexical item /bat/ is simplified and pronounced *ta*. The horizontal plane contains all such structures, and the vertical axis shows the relative harmony of each structure, an ordinal rather than a numerical scale. This harmony surface schematically depicts a young child's knowledge of grammar: STRUCTURE dominates FAITHFULNESS. This knowledge can be used by optimization in two ways. In production of "bat," the row of structures containing /bat/ compete (dotted box); the maximum-harmony structure best-satisfies top-ranked STRUCTURE with the simplified pronunciation *ta* (peak of the dotted curve): this is marked ☞. In comprehension, the pronunciation *bat* is given, and competition is between the column of structures containing *bat* (dashed box). Because these are all pronounced *bat*, they tie with respect to STRUCTURE, so lower-ranked FAITHFULNESS determines the maximum-harmony structure to be (/bat/, *bat*), marked with ➤➤ (peak of the dashed curve). Correct comprehension results from the same grammar that gives incorrect—simplified—production.

for example, we can recognize a violin without being able to play one—because real language comprehension requires sophisticated grammatical knowledge. In many cases, both the comprehension and production abilities can be captured by grammars, the "comprehension grammar" being closer to the adults' than is the "production grammar." Yet a grammar is usually seen as a characterization of linguistic competence independent of the cognitive factors involved in language use—so how can a child have two grammars, one for each type of use?

Optimality theory provides a conceptual resolution of this dilemma (14). The child has only one grammar (constraint ranking) at a given time, a grammar that is evolving toward the adult grammar by reranking of constraints. Early child grammars have an interesting property: When used for production, only the simplest linguistic structures are produced. But when used for comprehension, the same grammar allows the child to cope rather well with structures much more complex than those they can produce.

The reason is essentially this. In early child grammars, STRUCTURE constraints outrank FAITHFULNESS constraints. In production, the input is an interpretation, and what competes are different pronunciations of the given interpretation. The winner is a structure that sacrifices faithfulness to the input in order to satisfy STRUCTURE: This is a structure simpler than the corresponding adult pronunciation. (In the example of Fig. 4, the word /bat/ is simplified to *ta*.) But during comprehension, the competition is defined differently: It is between structures that all share the given adult pronunciation, which is fixed and immutable, under the comprehension regime, as the input to the grammar. These competitors are all heavy violators of STRUCTURE, but they tie in this respect; so the STRUCTURE constraints do not decide among them. The winner must then be decided by lower-ranked FAITHFULNESS constraints. (Thus in Fig. 4, the adult pronunciation *bat* is correctly comprehended as the word /bat/ even though the child's own pronunciation of /bat/ is *ta*.) Thus, production is quite "unfaithful" to adult language because FAITHFULNESS constraints are out-voted by the dominant STRUCTURE constraints. But comprehension is more "faithful" to adult language because the crucial unfaithful candidates are simply out of the competition; they do not have the given (adult) pronunciation, which is held fixed in the comprehension regime as the input to the grammar. That two such different outcomes can arise from one and the same constraint ranking is a typical effect of optimization in optimality theory: Constraints that are de-

cisive in some competitions (STRUCTURE during production) fail to decide in other competitions (comprehension), depending on the character of the candidate set being evaluated, which allows lower-ranked constraints (FAITHFULNESS) to then determine the optimal structure.

This result resolves several related difficulties of two previous conceptions of child language. In the first, a grammar is a set of rules for sequentially transforming structures, ultimately producing the correct pronunciation of a given expression. This conception fails to adequately separate knowledge and use of grammar, so that a set of rules producing correct pronunciations is incapable of operating in the reverse direction, comprehension, transforming a pronunciation into an interpretation. (Even if the rule system could be inverted, children's "unfaithful production" and relatively "faithful comprehension" are simply not inverses of one another—the challenge is to provide a principled account for this divergence with a single grammar.) Furthermore, child grammars in this conception are typically considerably more complex than adult grammars, because many more transformations must be made in order to produce the "unfaithful" distortions characteristic of child productions.

In the second nonoptimization-based conception, a grammar is a set of inviolable constraints: A structure that violates any one of the constraints is ipso facto ungrammatical. Languages differ in the values of certain "parameters" that modify the content or applicability of constraints. Thus the gap between child linguistic production and comprehension must be seen as resulting from two different sets of parameters, one for each type of use. Again, this fails to separate knowledge from use of grammar, and fails to provide any principled link between production and comprehension. By contrast, conceiving of grammar as optimization provides a natural distinction between use and knowledge of language, in such a way that a single grammar naturally provides relatively "faithful" comprehension at the same time as relatively "unfaithful" production.

The optimization perspective also offers a principled approach to a vexing fundamental problem in grammar learning. The constraints of a grammar refer to many "hidden" properties of linguistic structures, properties that are not directly observable in the data available for learning a language. For example, the way that words are grouped into nested syntactic phrases, or sounds grouped into prosodic constituents, is largely unobservable (or only ambiguously and inconsistently reflected in observables), and yet can differ from language to

language. Learning a grammar requires access to this hidden linguistic structure, so that the grammar may be adjusted to conform to the configurations of hidden structure characteristic of the language being learned. But the hidden structure itself must be inferred from prior knowledge of the grammar: It cannot be directly observed.

Within optimality theory, these coupled problems can be solved by successive approximation, as in related optimization problems outside grammar. The learner starts with an initial grammar (indeed, the early child grammar mentioned above). This grammar is used in the "comprehension direction" to impute hidden structure to the pronounced data of the target language. This hidden structure will initially be in error, because the grammar is not yet correct, but this structure can nonetheless be used to adjust the grammar so that, in the production direction, it outputs the inferred structures. With this revised grammar, the process continues with new learning data. As the grammar gets closer to the correct one, the hidden structure it assigns to learning data gets closer to the correct structure. While there are as yet no mathematical results demonstrating the success of this incremental learning method under general conditions, it has proved quite effective in related optimization problems such as speech recognition (15), and quite successful in preliminary computer simulation studies of optimality theory grammar learning (16).

The central subproblem of this incremental learning strategy is this: Given learning data including hidden structure (inferred on the basis of the current grammar), how can the grammar be improved? Here OT's optimization characterization of universal grammar provides considerable power. The grammars of human languages differ, according to the core hypothesis of optimality theory, only in the way they rank the universal constraints. Thus improving a grammar requires only reranking the constraints. Given a grammatical structure from the language to be learned, there is a straightforward way to minimally rerank constraints to make that structure optimal, hence grammatical, in the revised grammar. And this procedure can be proved to efficiently converge on a correct grammar, when one exists. "Efficient" here means that, even though there are $n!$ different constraint rankings of $n$ universal constraints, no more than $n(n-1)$ informative learning examples are needed to converge on the correct ranking (17). Corresponding results are not available within alternative general theories of how human grammars may differ; this is an indication of the learnability advantage arising from the highly

structured nature of OT's optimization characterization of universal grammar.

Will the connection between optimization in grammatical theory and optimization in neural networks lead to further progress at either level of cognitive theory? Will other theoretical connections between the sciences of the brain and of the mind prove fruitful? Of course, only time will tell. But we believe there is already in place a significant body of evidence that even a single high-level property of neural computation, properly treated, can yield a surprisingly rich set of new insights into even the most well-studied and abstract of symbol-processing cognitive sciences, the theory of grammar (18).

## REFERENCES AND NOTES

1. Basic work on this phenomenon in neural network models, and critiques, include D. E. Rumelhart and J. L. McClelland, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, *Volume 2: Psychological and Biological Models*, J. L. McClelland, D. E. Rumelhart, PDP Research Group, Eds. (MIT Press/Bradford Books, Cambridge, MA, 1986), pp. 216–271; S. Pinker and J. Mehler, Eds., *Connections and Symbols* (MIT Press, Cambridge, MA, 1988).
2. A. Prince and P. Smolensky, *Notes on Connectionism and Harmony Theory in Linguistics.* (Technical Report, Department of Computer Science, University of Colorado, Boulder, CO, 1991); *Optimality Theory: Constraint Interaction in Generative Grammar* (Technical Report, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ and Department of Computer Science, University of Colorado, Boulder, CO, 1993; also Linguistic Inquiry Monograph Series, MIT Press, Cambridge, MA, to appear); J. McCarthy and A. Prince, *Prosodic Morphology I* (Technical Report RuCCS-TR-3, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ, 1993; also Linguistic Inquiry Monograph Series, MIT Press, Cambridge, MA, to appear); J. McCarthy and A. Prince, in *Yearbook of Morphology 1993*, G. Booij and J. van Marle, Eds. (Kluwer, Boston, 1993), pp. 79–153. These are a few of the basic references for optimality theory, addressing primarily phonology. The following basic works address syntax, including the topics mentioned in the text: J. Grimshaw, *Linguist. Inq.*, in press; V. Samek-Lodovici, thesis, Rutgers University, New Brunswick, NJ (1996); G. Legendre, P. Smolensky, C. Wilson, in *Is the Best Good Enough? Proceedings of the Workshop on Optimality in Syntax*, P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis, D. Pesetsky, Eds. (MIT Press and MIT Working Papers in Linguistics, Cambridge, MA, in press). These and many other optimality theory papers may be accessed electronically [see (7)].
3. Selected basic sources: Formal universalism: N. Chomsky, *Aspects of the Theory of Syntax* (MIT Press, Cambridge, MA, 1965); _____ and M. Halle, *The Sound Pattern of English* (Harper and Row, New York, 1968). Substantive process and product universalism: D. Stampe, *A Dissertation on Natural Phonology* (Garland, New York, 1979); D. Perlmutter, Ed., *Studies in Relational Grammar 1* (Univ. of Chicago Press, Chicago, 1983). Constraints across rules: C. Kisseberth, *Linguist. Inq.* **1**, 291 (1970); N. Chomsky, *Lectures on Government and Binding* (Foris, Dordrecht, Netherlands, 1981). Markedness and informal optimization: J. Goldsmith, Ed., *The Last Phonological Rule* (Univ. of Chicago Press, Chicago, 1993); D. Archangeli and D. Pulleyblank, *Grounded Phonology* (MIT Press, Cambridge, MA, 1994); L. Burzio, *Principles of English Stress* (Cambridge Univ. Press, New York, 1994); N. Chomsky, *The Minimalist Program* (MIT Press, Cambridge, MA, 1995). Output orientation: D. Perlmutter, *Deep and Surface Structure Constraints in Syntax* (Holt, Reinhart, and Winston, New York, 1971); J. Bybee and D. Slobin, *Language* **58**, 265 (1982); J. McCarthy and A. Prince, *Prosodic Morphology 1986* (Technical Report RuCCS-TR-32, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ, 1986/1996).
4. R. Jakobson, *Selected Writings I* (Mouton, The Hague, 1962); N. Trubetzkoy, *Grundzüge der Phonologie* (1939; translation: *Principles of Phonology*, Univ. of California Press, Berkeley, CA, 1969); N. Chomsky and M. Halle, *The Sound Pattern of English* (Harper and Row, New York, 1968), chapter 9.
5. For details, see B. Tranel, *French Liaison and Elision Revisited: A Unified Account Within Optimality Theory* (ROA-15, http://ruccs.rutgers.edu/roa.html, 1994).
6. For significant extensions of FAITHFULNESS within optimality theory, see J. McCarthy and A. Prince, in *The Prosody-Morphology Interface*, R. Kager, W. Zonneveld, H. van der Hulst, Eds. (Blackwell, Oxford, UK, in press); L. Benua, in *Papers in Optimality Theory*, J. Beckman, L. Walsh Dickey, S. Urbanczyk, Eds. (Linguistics Department, Univ. of Massachusetts, Amherst, MA, 1995), pp. 77–136.
7. Examples include segmental repertories, stress patterns, vowel harmony, tonology, reduplicative and templatic morphology, syntax-phonology and morphology-phonology relations, case and voice patterns, principles of question formation, interaction of syntactic movement and clause patterns, structure of verbal complexes, order and repertory of clitic elements, the interaction between focus and the placement and retention of pronominal elements, the interpretation of anaphoric relations, the nature of constraints like the obligatory contour principle, and the compatibility of related grammatical processes. Readers interested in pursuing any of these topics may consult the Rutgers Optimality Archive at http://ruccs.rutgers.edu/roa.html, which includes many papers and an extensive bibliography.
8. P. Smolensky, *Behav. Brain Sci.* **11**, 1 (1988); S. Pinker and A. Prince, *Cognition* **28**, 73 (1988); M. McCloskey, *Psychol. Sci.* **2**, 387 (1991); A. Prince, *In Defense of the Number i: Anatomy of a Linear Dynamical Model of Linguistic Generalizations* (Technical Report RuCCS-TR-1, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ, 1993).
9. J. J. Hopfield, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2554 (1982); M. A Cohen and S. Grossberg, *IEEE Trans. Syst. Man Cybernet.* **13**, 815 (1983); P. Smolensky, *Proc. Natl. Conf. Artif. Intell. AAAI-83*, 378 (1983); G. E. Hinton and T. J. Sejnowski, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (1983), p. 448; J. J. Hopfield, *Proc. Natl. Acad. Sci. U.S.A.*, **81**, 3088 (1984); R. M. Golden, *Biol. Cybernet.* **59**, 109 (1988). For recent review articles, see M. W. Hirsch, in *Mathematical Perspectives on Neural Networks*, P. Smolensky, M. C. Mozer, D. E. Rumelhart, Eds. (LEA, Mahwah, NJ, 1996), pp. 271–323; P. Smolensky, in *ibid.*, pp. 245–270.
10. P. Smolensky, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, D. E. Rumelhart, J. L. McClelland, PDP Research Group, Eds. (MIT Press/Bradford Books, Cambridge, MA, 1986), pp. 194–281.
11. R. Pike, *Psychol. Rev.* **9**, 281 (1984); C. P. Dolan, thesis, University of California, Los Angeles (1989); _____ and P. Smolensky, *Connection Sci.* **1**, 53 (1989); P. Smolensky, *Artif. Intell.* **46**, 159 (1990); T. Plate, thesis, University of Toronto (1994).
12. G. Legendre, Y. Miyata, P. Smolensky, *Proc. Annu. Conf. Cognit. Sci. Soc.* **12**, 388, (1990); *ibid.*, p. 884.
13. B. Tesar and P. Smolensky, *Linguist. Inq.*, in press; *Learnability in Optimality Theory* (Technical Report, Cognitive Science Department, Johns Hopkins University, Baltimore, MD, 1996).
14. P. Smolensky, *Linguist. Inq.* **27**, 720 (1996).
15. L. E. Baum and T. Petrie, *Ann. Math. Stat.* **37**, 1559 (1966); L. R. Bahl, F. Jelinek, R. L. Mercer, *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-**5** (1983), pp. 179–190. For recent review articles, see A. Nádas and R. L. Mercer, in *Mathematical Perspectives on Neural Networks*, P. Smolensky, M. C. Mozer, D. E. Rumelhart, Eds. (LEA, Mahwah, NJ, 1996), pp. 603–650; P. Smolensky, in *ibid.*, pp. 453–494.
16. B. Tesar, *Lingua*, in press.
17. References in (13) and B. Tesar and P. Smolensky, *The Learnability of Optimality Theory: An Algorithm and Some Basic Complexity Results* (Technical Report, Department of Computer Science, University of Colorado, Boulder, CO, 1993).
18. As indicated in the cited publications, much of the work discussed here was carried out jointly with our collaborators G. Legendre, J. McCarthy, and B. Tesar; we are grateful to them and to our colleagues L. Burzio, R. Frank, J. Grimshaw, and C. Wilson for stimulating conversations and invaluable contributions. For support of the work presented here, we acknowledge a Guggenheim Fellowship, the Johns Hopkins Center for Language and Speech Processing, and NSF grants BS-9209265 and IRI-9213894.