2 *The Pursuit of Theory*

Alan Prince
Rutgers University

2.1 The Theory is also an Object of Analysis[*]

Common sense is often a poor guide to methodology. Any theory presents us with two fundamental and often difficult questions:

— What *is* it?
— How do you *do* it?

The first of these arises because a theory is the totality of its consequences. It must be given as the set of its defining conditions, and we may polish them, ground them, tailor them to meet various expectations, but unless we have mapped out what follows from them, the theory remains alien territory. Newton's theory of gravitation can be written on a postcard, and we might like to think of it as nothing more than what makes apples fall straight to earth and planets follow simple repetitive paths, but its actual content is strange beyond imagining and still under study hundreds of years after it was stated.[1] Once formulated, a theory has broken definitively with intuition and belief. We are stuck with its consequences whether we like them or not, anticipate them or not, and we must develop techniques to find them.

The second question arises because the internal logic of a theory determines what counts as a sound argument within its premises. General principles of rigor and validation apply, of course, but unless connected properly with the specific assumptions in question, the result can easily be oversight and gross error. Here's an example: in many linguistic theories developed since the 1960's, violating a constraint leads directly to ungrammaticality. A parochial onlooker might get the intuition that violation is somehow ineluctably synonymous with ill-formedness, in the nature of things. A grand conclusion may then be thought to follow:

(1) "… the existence of phonology in every language shows that Faithfulness [in Optimality Theory] is at best an ineffective principle that might well be done without." (Halle 1995b).

'Phonology' here means 'underlying-surface disparity'. Each faithfulness constraint forbids a certain kind of input-output disparity: *case closed*. But no version of Optimality Theory (OT) has ever been put forth that lacks a full complement of Faithfulness

[*] I'd like to thank Paul Smolensky, John McCarthy, Jane Grimshaw, Bruce Tesar, Jean-Roger Vergnaud, Vieri Samek-Lodovici, Chaim Tannenbaum, Seth Cable, Naz Merchant, and Adrian Brasoveanu for interactions which have shaped and re-shaped my views on the matters addressed here.

[1] Saari (2005) is a recent study. To get a sense of what can happen, see Ekeland (1988), esp. pp. 123-131.

constraints, because their operation – their minimal violation, which includes satisfaction as a special case – is essential to the derivation of virtually every form. The intuition behind the attempted criticism, grounded in decades of experience, is that well-formed output violates no constraints; but this precept is theory-bound and no truth of logic. It just doesn't apply to OT, or to any theory of choice where constraints function as criteria of decision between flawed alternatives.

2.1.1 Optimality Theory without Common Sense

A more telling example emerges immediately from any attempt to work within OT. At some point in the course of analyzing a given language, we have in hand a hypothesized constraint set and a set of analyses we regard as optimal. We now face the *ranking problem*: which constraint hierarchies (if any) will produce the desired optima as actual optima?

Any sophisticated problem-solvers' key tactic is to identify the simplest problem that contains the elements at play, solve it, and build up from there. Let's deploy it incautiously: since the smallest possible zone of conflict involves two constraints and two candidates (one desired optimal), gather such 2×2 cases and construct the overall ranking from the results.[2] But the alert should go up: no contact has been made with any basic notion of the theory. We actually don't know with any specificity what it is about the necessities of ranking that we can learn from such a limited scheme of comparison. A wiser procedure is to scrutinize the definition of optimality and get clear about what it is that we are trying to determine. A rather different approach to the ranking problem will emerge. What, then, does 'optimal' actually mean in OT? Let us examine this question with a certain amount of care, which will not prove excessive in the end.

Optimality is composite: the judgment of hierarchy is constructed from the judgment of individual constraints. Proceeding from local to global, definition begins with the 'better than' relation over a single constraint, proceeds to 'better than' over a constraint hierarchy, and then gets optimality out of those relations.

In the familiar way, one candidate is better than another on a constraint if it is assigned fewer violations by that constraint.

(2) **'Better than' on a constraint**

For candidates a,b and constraint C, $a \succ_C b$ iff $C(a) < C(b)$.

Here we have written $a \succ_C b$ for 'a is better than b on C', and C(x) for the (nonnegative) number of violations C assigns to candidate *x*.

To amalgamate such individual judgments, we impose a linear ordering, a 'ranking', written », on the constraint set, giving a constraint hierarchy. (We say C_1 *dominates* C_2 if $C_1 » C_2$.) Using that order, and using the definition of 'better than' on a constraint just given, we define the notion 'better than on a hierarchy'.

[2] The intuition gets a boost from previous analytical practice: in ordering rules, the analyst typically looked at two rules at time (and that worked, didn't it?).

As usual, we will say that one candidate is better than another on a hierarchy if it is better *on the highest-ranked constraint that distinguishes the two*. (This concise formulation is due to Grimshaw 1997; a constraint is said to 'distinguish' two candidates when it assigns a different number of violations to them; that is, when one is better than the other on that constraint.)

(3) **'Better than' on a constraint hierarchy.**
For candidates *a*,*b* and constraint hierarchy H,
$a \succ_H b$ iff there is a constraint C in H that distinguishes *a*, *b*, such that

(1) $a \succ_C b$
and (2) no constraint distinguishing *a* and *b* dominates C.

To be optimal is to be the best in the candidate set, and to be the best is to have none better.

(4) '**Optimal**'
For a candidate *q*, a candidate set K, with $q \in$K, and a hierarchy H, *q* is *optimal* in K according to H, iff there is no candidate $z \in$K such that $z \succ_H q$.

Now that we know what we're looking for, we can sensibly ask the key question: what do we learn about ranking from a comparison of two candidates (one of them a desired optimum)?

Since optimality is globally determined by the totality of such comparisons, and we are looking at just one of them, the best we can hope for is to arrive at conditions which will ensure that our desired optimum is *better than* its competitor on the hierarchy. This leads us right back to definition (3), and from it, we know that some constraint preferring the desired optimum must be the highest-ranked constraint that distinguishes them. The constraints that threaten this state of affairs are those that *dis*prefer the desired optimum: they must all be outranked by an optimum-preferring constraint. Let's call this the 'elementary ranking condition' (ERC) associated with the comparison.

(5) **Elementary Ranking Condition**
For *q,z*∈K, a candidate set, and S, a set of constraints, some constraint in S preferring *q* to *z* dominates all those preferring *z* to *q*.

Any constraint ranking on which candidate *q* betters *z* must satisfy the ERC. (To put it non-modally: candidate q is better than z over a ranking H of S if and only if the ranking H satisfies the ERC (5).) The ERC, then, tells us exactly what we learn from comparing two candidates.

To make use of this finding, we must first calculate each constraint's individual judgment of the comparison. A constraint measures the desired optimum against its competitor in one of just three ways: better, worse, same. We indicate these categories as follows, writing '*q~z*' for the comparison between desired optimum *q* and competitor *z*.

(6) **Constraint C assesses the comparison q *vs.* z.**

Comparative relation	Violation pattern	Gloss
$C[q\sim z]$ = **W**	$C(q) < C(z)$	‘C prefers the desired optimum’
$C[q\sim z]$ = **L**	$C(q) > C(z)$	‘C prefers its competitor’
$C[q\sim z]$ = **e**	$C(q) = C(z)$	‘C does not distinguish the pair’

Now consider a distribution of comparative values that could easily result from some such calculation. For illustrative purposes, imagine that the entire constraint set contains six constraints:

(7) **Typical two-candidate comparison**

	C_1	C_2	C_3	C_4	C_5	C_6
q ~ z	**L**	**e**	**W**	**W**	**e**	**L**

The relevant associated ERC declares this: C3 *or* C4 dominates *both* C_1 and C_6.
In any ranking of these constraints on which q is better than z, this condition must be met.

We now have the tools to examine the intuition that 2×2 comparison is the building block of ranking arguments. First, consider shrinkage of the *candidate* set. In order to narrow our focus to just 2 candidates, we exclude all the others from view. This is entirely legitimate: the hierarchical evaluation of a pair of candidates is determined entirely by the direct relation between them. Some other candidates may exist that are better than either, or worse than either, or intermediate between them, but no outsiders have any effect whatever on the head-to-head pair-internal relation. This fundamental property has been called ‘contextual independence of choice’ (Prince 2002b:iv), and is related to Arrow’s ‘irrelevance of independent alternatives’ (Arrow 1951:26). It is not a truth of logic, inherent in the notion of ‘comparison’ or ‘choice’, but the premises of OT succeed in licensing it. (It is also fragile: modify those premises and it can go away, as it does in the Targeted-Constraint OT of Wilson 2002.)

Now consider the role of the constraint set, where we find no such comfort. The form of the ERC in no way privileges 2-constraint arguments: *all* L-assessing constraints must be dominated, and *some* W-assessing constraint must do the domination. If we omit an L-assessing constraint from the calculation, the resulting ERC is incomplete, and it is no longer true that any hierarchy satisfying it will necessarily yield the superiority of the desired optimum (though the converse *is* true); further conditions may be required. Leaving out C_1 from tableau (7), for example, deprives us of the crucial information that C_1 must be dominated; if it is not, then undesired z betters q.

If we happen to omit a W-assessing constraint, the associated ERC can mistakenly exclude a successful hierarchy, leading to false assertions that cannot be remedied by merely obtaining further information. This is more dangerous than L-omission when we are arguing from optimum-suboptimum pairs to the correct ranking, as when dealing with the ‘ranking problem’ in the course of analysis. In tableau (7), for example, we have two W-assessors, C_3 and C_4. If negligence leads us to omit C_3, say, we are tempted to the conclusion that C_4 *must* dominate C_1 and C_6. This is not sound in itself, and depending on other circumstances, it could easily turn out that C_4 lies at the bottom of the correct hierarchy, dominating nothing, with C_3 doing the work of domination demanded by (7).[3]

The logic of the theory, then, allows us to discard from any particular comparison only the neutral e-assessing constraints. Tableau (7) shrinks to 2×4, and no further. In the literature, correct handling of the ERC is not ubiquitous, and omission of constraints often rests optimistically on intuitions about relevance and likely conflict. But pairwise (or intuitively restricted) examination of constraint relations has no status. This is not a matter of convenience, taste, typography, notation, presentation, or luck. We must *do* the theory as it dictates, even in the face of common sense.

2.1.2 Intuition and SPE's Evaluation Metric

Let us turn to a case where reliance on intuition leads to an interesting failure to appreciate what the theory actually claims. Consider the phonological theory put forth in *The Sound Pattern of English* (SPE: Chomsky & Halle 1968). A vocabulary is given for representing forms and for constructing rules, which are to apply in a designated order (some cyclically) to produce outputs from lexical items. Any sample of language data, even a gigantic one, is consistent with a vast, even unbounded, number of licit grammars. Which one – note the titular definite article – is correct? It is crucial to find a formal property that distinguishes the correct grammar, if linguistic theory is to claim realism and, more specifically, if it is to address the acquisition problem, even abstractly. (It is less crucial for linguistic practice, since linguists can, and indeed must, argue for grammars on grounds of evidence unavailable to the learner.) The well-known proposal is that grammars submit to evaluation in terms of their length, which is measured in terms of the number of symbols they deploy (Chomsky 1965: 37-42; SPE p.334). Shorter is better, and the shortest grammar is hypothesized to be the real one. The SPE statement runs as follows:

(8) "The 'value' of a sequence of rules is the reciprocal of the number of symbols in its minimal representation." (SPE p.334, ex. (9))

Ristad (1990) has noted a potentially regrettable consequence: the highest valued sequence of rules will have no rules in it at all. We therefore make the usual emendation, left tacit (I believe) in SPE: that we must also take account of the number of symbols expended in the lexicon. The length of the entire Lexicon+Rule System pairing determines the values we are comparing. A rule earns its keep by reducing the size of the lexicon.

The Evaluation Metric thus defined is entirely coherent (given a finite lexicon) and, as asserted by Chomsky & Halle, "provides a precise explication for the notion 'linguistically significant generalization'…" which is subject to empirical test. It seems to be the case, however, that there are literally no instances where the Evaluation Metric

[3] If an erroneously truncated ERC has excluded the correct hierarchy, there will be further information that contradicts it, yielding the impression that no correct hierarchy exists. Even if the erroneous ranking condition has not excluded the correct hierarchy, it produces a distorted account of the explanatory force of various constraint relations.

was put to use as defined. That is: no analysis in the entire literature justifies a proposed Lexicon+Rule System hypothesis by showing it to have the best evaluation of all those deemed possible by the theory. Is there even a case where the value was calculated?

The reason is not far to seek. Though defined globally, the metric was always interpreted locally. Typically, this was at the level of the rule:

(9) "… the number of symbols in a rule is inversely related to the degree of linguistically significant generalization achieved in the rule." (SPE p. 335)

But could even be extended to rule-internal contents:

(10) "…the "naturalness" of a class … can be measured in terms of the number of features needed to define it." (SPE p. 400).

Of course, nothing of the sort can legitimately be asserted without building considerable bridgework between the global metric and the behavior of the local entities out of which the grammar is composed. One has the intuition, perhaps, that it can't hurt to economize locally, and therefore that one is compelled to do so. But it can easily happen in even moderately complex optimization systems that a local splurge yields a global improvement by yielding drastic simplifications elsewhere. In a highly interactive system, the results of global optimization can be all but inscrutable locally.

We can see the local-global relation playing out variously in the other examples discussed above. The idea that Faithfulness is useless when violated represents a kind of hyperlocalism focused on one candidate and one constraint; of course, nothing follows. The local relation between 2 candidates, by contrast, is preserved intact in any set of candidates that contains them, including the entire candidate set. A relation between 2 *constraints*, though, has no such local-to-global portability to the entire constraint set. What is the situation, then, with the intuitive rule-focused evaluation of SPE phonologies?

A question not easily answered, alas: it isn't at all clear what the 'local interpretation' might be, or how it would replace the global interpretation. To evaluate, we must compare whole grammars with different lexica, different rules, and different numbers of rules. This provides no difficulty for the global metric, which doesn't see rules or lexica at all. The local interpretation wants to compare rules, though, and so must have rules in hand and some way of finding correspondences between them across grammars to render them comparable. This appears feasible for sets of adjacent rules, under the same lexicon, which perform identical mappings and collapse under the notational conventions; but beyond that … obscurity.

Stepping back from the theory, I'd suggest that the actual practice was largely based on discovering contingencies in the data, assuming that they must be reflected in rules of a specific type, and then setting out to simplify the assumed rule-types through notational collapse, ordering, and some fairly local interactional analysis; all under lexical hypotheses that sought a single underlying form for each morpheme. This is reasonable tactically, but it is a far cry from using the theory itself to compute (deterministically) which licit grammar is being evidenced by the data, and, as noted, it never involved using the theory (nondeterministically) to prove that the correct grammar had been obtained. Some such procedure of grammar discovery could even be legitimated, in principle or in part, by results clarifying the conditions under which it produces the Evaluation Metric optimum.

Overall, the effect of acting as if there were a "local interpretation" was not negative. Under its cover, attention was focused on processes, representations, their components and interactions, leading to substantive theories of great interest. Nevertheless, the divergence between theory and practice deprived the theory of the essential content that it claimed. Much effort was expended in fending off opponents who had, it seems, little knowledge of the theory they were criticizing, a faulty grasp of optimization, and little feel for how empirical consequences are derived from the actual assumptions of a theory as opposed to some general impression of them. One such defensive/offensive statement is the following:

(11) "It should be observed in this connection that although definition (9) has been referred to as the "simplicity or "economy criterion," it has never been proposed or intended that the condition defines "simplicity" or "economy" in the very general (and still very poorly understood) sense in which these terms usually appear in the philosophy of science. The only claim that is being made here is the purely empirical one ..."[4] (SPE pp.334-5)

We grant, of course, that the SPE theory is abstractly empirical in the way it characterizes linguistic knowledge; and note that the contemporary research style has profited enormously from the unprecedented daring exhibited in staking out territory where none before had imagined it possible. What's missing, though, is the sense of any *particular* empirical claim or set of claims which has been identified and tested against the facts. Worse, the failure to use the theory of evaluation means that we literally do not know what such a claim is. This is Newton's *Principia* without the equations, or with equations that have never been solved. Many rules and rule systems were put forth to describe many language phenomena; but in no case can we be sure that the system proposed is the one projected by the Evaluation Metric. But it is only the optimal system that contains the claims to test.

The Evaluation Metric imbroglio is directly due to a failure to apply the definition to the practice of the theory. The definition provided a formal front for the activities of the researcher, which proceeded on a separate, intuitive track. As with the example of erroneous but commonly applied beliefs about ranking, it is not satisfactory to point defensively to the success of some practitioners in developing interesting theories under false premises. "A long habit of not thinking a thing wrong, gives it a superficial appearance of being right, and raises at first a formidable outcry in defence of custom" (Paine 1776). We must do better.

[4] Interestingly, the actual on-the-ground interpretation of the Evaluation Metric may have been closer to the loose general sense of 'be simple' than to the formal definition of evaluation.

2.2 What is real and what is not

One need only glance at the formal literature leading up to generative grammar to grasp that we are the beneficiaries of a fundamental change in perspective. Aiming in *Methods in Structural Linguistics* (1951) for "the reduction of linguistic methods to procedures" (p.3), Zellig Harris introduces his proposals with this modest remark:

(12) "The particular way of arranging the facts about a language which is offered here will undoubtedly prove more convenient for some languages than for others." (Harris 1951:2)

He does not intend, however, to impose a "laboratory schedule" of analytical steps that must be followed sequentially, and he characterizes the value of his methodology in this way:

(13) "The chief usefulness of the procedures listed below is therefore as a reminder in the course of the original research, and as a form for checking or presenting the results, where it may be desirable to make sure that all the information called for in these procedures has been validly obtained." (Harris 1951:1-2)

These are to be "methods which will not impose a fixed system upon various languages, yet will tell more about each language than will a mere catalogue of sounds and forms."

The goal, then, is to produce useful descriptions, to be judged by such criteria as accuracy, convenience, reliability, responsiveness to variation, and independence from observer bias. No one can sensibly dispute the importance of these factors in empirical investigation of any kind. What further ends is linguistic description intended to serve? Historical linguistics and dialect geography, phonetics and semantics, the relation of language to culture and personality, and the comparison of language structure with systems of logic are cited as areas of study that will profit from "going beyond individual descriptive linguistic facts" to "the use of complete language structures" (p.3).

Largely absent from this program is a sense that the focus of study is a real object, evidenced by the arranged facts but not reducible to them, about which one makes statements that are (because it is real) *right* or *wrong* – as opposed to convenient or awkward, useful or irrelevant to one's parochial purposes. Descriptive, synchronic linguistics is a conduit for pipelining refined information to various disciplines that make use of language data. Chomsky changes all that, of course, by identifying an object that linguistics is to be *about* – competence, I-language, the internal representation of linguistic knowledge. This move is set in the context of rival conceptions of mental structure:

(14) "… empiricist speculation has characteristically assumed that only the procedures and mechanisms for the acquisition of knowledge constitute an innate property of the mind. ... On the other hand, rationalist speculation has assumed that the general form of a system of knowledge is fixed in advance as a disposition of the mind, and the function of experience is to cause this general schematic structure to be realized and more fully differentiated." (Chomsky 1965:51-52)

The ground has been shifted so fundamentally that both poles of this opposition lie outside the domain in which Harris places himself, where 'knowledge' of language is not at issue. Nevertheless, there is a clear affinity between Harris's interest in methods and the empiricist focus on 'procedures and mechanisms'. Note, too, the force of the Evaluation Metric idea in this context, since it severs the choice of grammar completely from methods and procedures of analysis: the correct grammar is defined by a formal characteristic it has, not as the result of following certain procedures.

To pursue the issue further into linguistics proper, let us distinguish heuristically between 'Theories of Data' (TOD), which produce analyses when set to work on collections of facts, and 'Free-Standing Theories' (FST), which are sufficiently endowed with structure that many predictions and properties can be determined from examination of the theory alone.

A near-canonical example of a TOD is provided by the Rumelhart-McClelland model of the English past tense (Rumelhart & McClelland 1986; examined in Pinker & Prince 1988). This is a connectionist network which can be trained to associate an input activation pattern with an output activation pattern. When trained on stem/past-tense pairs, it will produce, to the best of its ability, an output corresponding to the past tense of its input. No assumptions are made about morphology or phonology, regular or irregular, although a structured representational system (featural trigrams) is adopted which allows a word to be represented as a pattern of simultaneous activation. This is a fully explicit formal theory, which operates autonomously. And, once trained, a model will make clear predictions about what output is expected for a given input, whether that input has been seen before or not. It makes limited sense, however, to query it in advance of training, looking for guidance as to what the structure of human language might be; and a trained model is not really susceptible to fine-grained analytic dissection *post hoc* either, due to the complexity of its internal causal structure. The model only takes on predictive structure when it has been exposed to data, and that predictive structure can only be investigated by presenting it with more data.

Examples of Free-Standing Theories are not difficult to find. A theory that spells out a sufficiently narrow universal repertory of structures, constraints, or processes, and explicitly delimits their interactions, will generate an analytically investigable space of possible grammars. Clear examples range from early proposals like that of Bach (1965), Stampe (1979), Donegan & Stampe (1979) to parametrized theories in syntax and those in phonology like Archangeli & Pulleyblank (1994), Halle & Vergnaud (1987), Hayes (1995), as well as many others; Optimality Theory (Prince & Smolensky 1993/2004) falls into the Free-Standing class, both in the large and in domain-specific instantiations of constraint sets. Such theories are in no way limited to symbol-manipulation; the Dynamic Linear Model of stress and syllable structure (Goldsmith and Larson 1990, Larson 1992, Goldsmith 1994, Prince 1993), which computes with numbers, is as canonical an example of an FST as one could imagine, as we will see below in section 3.2.

The distinction is heuristic and scalar, because theories may be more and less accessible to internal analysis, and may require more or fewer assumptions about data to yield analytical results.[5] Even a dyed-in-the-wool TOD like the Rumelhart-McClelland model admits to some analysis of its representational capacities, and Pinker & Prince mount a central argument against it in terms of its apparent incapacity to generalize to variables like 'stem' which range over lexical items regardless of phonetic content (Pinker & Prince 1988, Prince & Pinker 1988; Marcus 2001). Nevertheless, it is clear that Optimality Theory, for example, or parametrized theories of linguistic form, will admit a deeper and very much more thorough explication in terms of their internal structure.

The distinction between Theories of Data and Free-Standing Theories cross-cuts the empiricist/rational distinction that Chomsky alludes to in the passage quoted above. On the empiricist side, 'procedures and methods for the acquisition of knowledge' can be so simple as to admit of detailed analysis, like that afforded to the two-layer 'perceptron' of Rosenblatt (1958) in Minsky & Papert (1969), which treats it as an FST and achieves a sharp result. But the major step forward in connectionist theory in the 1980's is generally agreed to have been the advance from linear activation functions to differentiable nonlinear activation functions, which in one step enormously enriched the class of trainable networks and rendered their analysis far more difficult.[6] On the rationalist side, SPE-type phonology has a TOD character, and investigation of its fundamental properties has shown its general finite-state character (Johnson 1972) but, to my knowledge, little of research-useful specificity.

It is perhaps not surprising that many recent versions of linguistic theory developed under the realist interpretation of its goals should fall toward the FST end of the spectrum. If the aim is to discover a 'system of knowledge' that is separate from the encounter with observables, then unless a hypothesized system has discernible properties and significant predictivity, it is unlikely to be justifiable. To the extent that it is data-dependent, and usable mostly for modeling data rather than predicting general properties, it must face off with other TOD's, particularly those offering powerful mechanisms for induction and data representation. (If compressing the lexicon is the supreme goal of phonology, expect stiff competition from the manufacturers of WinZip™ and the like.) Within the ever-expanding palette of choices available to cognitive science, it seems unlikely that rationalist theory will beat statistical empiricism on its native turf. The argument must be that the object of study is not what empiricism assumes it to be. But this must be shown; and is best shown by the quality of the theories developed from rationalist assumptions.

In the absence or failure of such theories, linguistics must recede to a Harris-like position: it might serve as a helpful guide to scientists who (for whatever reason) wish to study phenomena where language plays some role, a map of the terrain but no part of the terrain itself. What's real would be the general data-analyzing methods of empiricist cognitive science, for which language has no special identity or integrity, along with

[5] At a considerably more abstract level, there is much to be said to be said about the capacities and dynamics of connectionist networks, see Smolensky et al. eds., (1994) for a large-scale multi-perspective overview.

[6] See Rumelhart et al. (1986), McClelland et al. (1986). The general view taken there is that "the objects referred to in macrostructural [i.e. symbolic –AP] models of cognitive processing are seen as approximate descriptions of emergent properties of the microstructure" (McClelland, Rumelhart, and Hinton 1986:12). Smolensky and Legendre (2005) develop a very different view, according exact reality to both continuous (micro) and discrete (macro) processing as distinct levels.

whatever results such methods obtain when applied to the data, linguistic or other, that is fed to them.

In phonology proper, representational theory has moved from the undifferentiated featural medium of SPE to the deployment of special structures keyed to the properties of different phenomenal domains, leading naturally (though not inevitably) to contentful FST's of those domains. Increasing the structural repertory is a two-edged. Poorly handled, taken as an add-on to available resources, it can turn out to be no more than a profusion of apparatus that enriches descriptive possibilities, leading to TOD. More interestingly configured, it can yield narrow, predictive theories; but these will contain significant built-in content and hence tend toward the FST side of the spectrum.

In this context, the surprise is not the emergence of the FST but the persistence of what we might call the '**Descriptive Method**' (DM) – data description as the primary analytical methodology for determining the content of a theory. For a TOD, this is virtually inevitable; there may be no other way to get an inkling of the theory's character. As soon as an FST is given, though, its consequences are fully determined by its internal structure.

Yet by far the dominant approach to probing linguistic FST's consists of confronting them with specific data. This can be done haphazardly or with reference to a few inherited 'favorite facts', or it can be done with prodigious vigor and problem-solving prowess, as in for example Hayes (1995). Although parametric theories are plentiful, few indeed are those whose 'exponential typology' of parameter settings has been laid out in full or studied in depth.

This places linguistic theory in an odd position. The axioms or defining conditions of a theory provide a starting place, not an endpoint: a theory is the totality of its consequences. With an FST, these are available to us analytically, and claims about the theory can be decided with certainty. If we decline to pursue the consequences analytically, we impose on ourselves a limited and defective sense of what the theory actually is. This then unnecessarily distorts both further development and theory comparison. Rational arguments about two theories' comparative success, for example, depend on a broad assessment of their properties; lacking that, such discussions not infrequently descend into the cherry-picking of isolated favorable and unfavorable instances.[7] What we might call the 'Analytical Method' is essential for determining the systematic content of theory. It is particularly valuable for delimiting the negative space of prohibitions into which the Descriptive Method does not venture, but it is equally essential for finding the structure of a theory's predictions of possibility.

[7] Interestingly, competition often provokes localized analysis of rival theory, treated as an FST, even in the context where the favored theory is being laid out and investigated by the Descriptive Method. To cite merely one example: in Halle and Vergnaud (1987), an important synthetic work that brings together much prior theory under the unifying rubric of the bracketed grid (Hammond 1984), there is an argument against one of Hammond's proposals, based on an apparently false consequence derived from it (p.75). Halle & Vergnaud's system is well and even elegantly formalized, yet due to their reliance on the Descriptive Method, we have little idea of the scope of their own predictions, some of which may involve equally disturbing pathologies.

2.3 Following the Analytical Method

Analysis of Free-Standing Theories is often driven by the most basic formal questions. Perhaps the most fundamental thing we must ask of a proposed theory is — 'does it *exist*?' That is: do the proposed defining conditions actually succeed in defining a coherent entity?[8] Closely related is the question of *under what conditions* the theory exists: what conditions are required for it to give a determinate answer or an answer that makes sense formally?[9] A natural extension of such concerns, for linguistic theories, is the question of a whether the theory is *contentful* in that it excludes certain formally sensible states-of-affairs from description. It might seem to some that such questions are arid and of limited interest, since (on this view) most formal deficiencies will not show up in practice, and in the empirical hurly-burly those that do can be patched over. We have already seen how, contrary to such expectations, commanding the answers to drily fundamental questions (e.g. what *is* optimality?) is essential to the most basic acts of data-analysis. Here we examine two cases that show the very tangible value of asking the abstract questions about a theory's content and realm of existence

2.3.1 Harmonic Ascent

Let us first consider Optimality Theory in the large. Moving beyond the bare-bones definition of optimality, let us endow the constraint set with some structure: a distinction between Markedness constraints, which penalize configurations in the output, and Faithfulness constraints, which each demand identity of input and output in a certain respect by penalizing any divergence from identity in that respect. Assume that the Markedness/Faithfulness distinction partitions the constraint set, so that any licit constraint belongs to one of the categories; let's call the theory so defined 'M/F-OT'. This gives us perhaps the simplest feasible OT linguistic theory, assuming the usual generative phonological architecture in which the grammar maps a lexical form (input) to a surface form (output). We may now ask if the theory achieved at this level of generality is *contentful*, or if it requires further structure to attain predictions of interest. Exactly this question is taken up in Moreton (2004a), and the results he obtains are illuminating.[10]

[8] Nonexistence isn't the worst thing that can happen. Yang-Mills theory, for example, is said to be basic to modern particle physics, but is not known to 'exist' mathematically, i.e. to have coherent foundations. The Clay Institute offers $1,000,000 for showing its 'existence': http://www.claymath.org/millennium/Yang-Mills_Theory.

[9] For example, the theory of multiplication and division exists; but you can't divide by zero. Similarly, if you are computing probabilities, they must not be less than 0 or greater than 1. To move nearer to our concerns, note that it is crucial for OT that there be at least one *best* element in the candidate set. Suppose that a constraint was posited to offer *rewards* rather than penalties, as all do now. Let the putative constraint LONG give a reward of +1 for each syllable that a form contains. Then there is no candidate that has the maximal value on LONG, and were the constraint asked to produce the class of forms that do maximally well on it, no output would be defined. If such a constraint is admitted, the theory ceases to exist.

[10] The presentation of Moreton's results given here will be considerably more qualitative than Moreton's own, and will diverge in some points of perspective. See Moreton (2004a) for a scrupulous rendering of the details.

To begin, we note that OT has a property that we might call ‘positivity’ which it shares with certain other multiple criterion decision-making systems, though by no means all.[11] Broadly speaking, a ‘positive’ system will be one in which a candidate can do well globally only by doing well locally. If a *winning* candidate does poorly on some criteria in comparison to some particular competitor, we can infer, in a positive system, that it must be doing better than its competitor on some other criteria. OT’s positivity comes immediately from the way it defines ‘optimal’: we know that if on some *hierarchy* it happens that *q* is better than *z*, then there is some particular *constraint* on which *q* is better than *z* on (namely, the highest ranked constraint that distinguishes them). Now widen the focus: suppose we know that the inferior candidate *z* is (perversely) better than *q* on some designated subset D of the constraints, ranked as in the hierarchy as a whole. Clearly, since *q* is the overall superior candidate, it must be that *q* is better than z on some particular constraint, and that constraint must belong to *the complement set* of D.

Applying this observation to M/F-OT, we find that if *q*, the superior candidate, is worse than *z* on the Faithfulness subhierarchy, then *q* must be better than *z* on the Markedness subhierarchy (and vice versa). This observation gains particular force because it is commonly the case that there is a fully faithful candidate (FFC) in the candidate set. The FFC has a tremendous advantage, because it satisfies every F constraint and nothing can beat it over the Faithfulness constraints, no matter how they are ranked. It follows that any non-faithful mapping − any mapping introducing faithfulness-penalized input-output disparity − can be optimal only if it superior to the FFC on grounds of Markedness. Since the FFC is essentially a copy of the input, this means that in an unfaithful mapping, the output must be less marked than (the faithful copy of) the input, when it exists. We can call this property ‘harmonic ascent’, using the term ‘harmonic’ to refer to the opposite of ‘markedness’.

(15) **Harmonic Ascent**

Suppose for $y \neq x$, $x \rightarrow y$ is optimal for some hierarchy H, where $x \rightarrow x$ is also a candidate.

Then for H|M, the subhierarchy of M constraints ranked as they are in H, it must be that $y \succ x$ on H|M.

Sloganeering, we can say: if things do not stay the same, they must get better (markedness-wise).

This property severely restricts the mappings that M/F-OT can execute. A first consequence is that there can be no *circular chain shifts*. This is easiest to see in the case of the smallest possible circle: imagine a grammar that takes input /*x*/ to distinct output [*y*] and input /*y*/ to output [*x*]:

$x \rightarrow y$

$y \rightarrow x$

[11] ‘By no means all’ — this innocuous phrase hides the difficulty , in many circumstances where ordinal preference is involved, of finding a system that has the property. Common sense intuition fails dramatically here. See Saari (2001), for example, to make contact with the vast literature emerging from Arrow (1951).

(An example would be a grammar mapping /pi/ to [pe] and /pe/ to [pi].) This pair of mappings cannot be accommodated in one grammar under M/F-OT, because the 'better than' relation is a strict order. By Harmonic Ascent, the optimality of $x\rightarrow y$ requires $y\succ x$ on the Markedness subhierarchy. But $y\rightarrow x$ requires $x\succ y$. One form cannot be both *better than* and *worse than* another.

More generally, any chain shift involving a cycle cannot be expressed. For example:

(16) *Impossible chain-shift in OT*

Mapping	Markedness Relation
$x\rightarrow y$	$y\succ x$
$y\rightarrow z$	$z\succ \mathrm{y}$
$z\rightarrow x$	$x\succ z$

Here the argument is just one step more complicated. Putting all the implied Markedness relations together, we have $x \succ z \succ y \succ x$. Since 'better than' is transitive, asymmetric, and (hence) irreflexive, this set of relations is impossible: it yields $x\succ x$, as well as both $x\succ y$ and $y\succ x$.

A second consequence follows from this fact: there is an end to getting better. If OT is to exist at all, no constraint can portray the candidate set as an unbounded upward-tending sequence of better and better forms (see fn.9). This, taken with Harmonic Ascent, rules out the endless shift:

(17) *Impossible endless shifts in OT*

$x_1 \rightarrow x_2$
$x_2 \rightarrow x_3$
$x_3 \rightarrow x_4$
…
$x_k \rightarrow x_{k+1}$
…

Of these consequences, the second seems clearly right. There is, I believe, no phonological process that, for example, adds a syllable to every input. Actual augmentation processes aim to hit some target (like bimoraicity or bisyllabicity) which is clearly relatable to Markedness constraints on prosodic structure. There is no sense in which longer is better regardless of the outcome (McCarthy & Prince 1993, Prince & Smolensky 1993/2004).

The first is perhaps more interesting, perhaps because it characterizes rather than merely excludes. Chain shifts are well-attested, and almost always noncircular. Moreton & Smolensky (2002) review some 35 segmental cases, of which 3 are doubtful, 4 inferred from distribution, and 28 robustly evidenced by alternations; none are circular. The famous counterexample is the 'Min tone circle' of Taiwanese (Xiamen, Amoy) tone sandhi, examined in Moreton (1996/99; 2004a) and much discussed in the literature (see e.g. Chen 1987, 2000, Yip 2002 and references therein). The details of the case, Moreton argues, are such that it does not invite analysis in terms of "simple, logical, plausibly innate constraints," and, as a phenomenon that is "synchronically speaking, completely arbitrary and idiosyncratic," it must be understood as a nonphonological "paradigm replacement" (Moreton 2004a:159), an intriguing possibility in need of further specification (but see Mortensen 2004 for more cases and a different view). In the end, if the circular cases prove to fall under special generalizations outside the reach of core phonology, then the prediction is vindicated. At this point, the matter must be regarded as somewhat unsettled, absent a compelling analysis of the tone circle.

Whatever the fate of circularity, it remains remarkable that a theory as simple as M/F-OT, at a level of analysis that lacks any characterization of constraints other than the formal, should show a property like Harmonic Ascent, which governs and severely restricts what it can do. We need theories that have such properties if we are to establish the rationalist perspective that Chomsky enunciated in his foundational work. The Descriptive Method of theory investigation, and its typically particularized results, can give no hint that such a property is obtainable without stipulation. Equally remarkable is the abstractness of the question that led to its discovery: 'what limitations does the theory place on the mappings a grammar can accommodate?' One might expect the answer to be so negative ('no limit') or so abstract (for example, registering them with respect to automata theory) that no obvious practical consequences ensue. Theoretically, we learn that expanding the repertory of constraint types to include *anti*-Faithfulness constraints (Alderete 1999b, 2001b) is more than an aesthetic complication; if unrestricted, it imperils the core emergent property of M/F-OT. And empirically, we find ourselves steered directly toward an entirely central phenomenon and informed that it is not merely of descriptive interest, but that its character actually determines the kind of theory we can have.

A further consequence of major analytical significance follows immediately from Moreton's work. Suppose we have a chain shift, [1] $x \to y$, [2] $y \to z$; this can only be obtained by preventing x from going all the way to z. We know from [2] that z is better than y on the Markedness subhierarchy. Thus, only Faithfulness can prevent x from leaping all the way to z; it is futile to seek a Markedness explanation for the fact that x halts at y.

More exactly, the ungrammatical candidate *$x \to z$, which we wish to avoid, must be *better* on Markedness than licit $x \to y$ and, if it is to lose, must be *worse* on Faithfulness. This means that we need a Faithfulness constraint forbidding *$x \to z$ *which does not forbid* x→y. The analysis of M/F-OT not only tells us in general terms that circular shifts are disallowed; it specifically characterizes the kind of Faithfulness constraints that must exist if *non*circular chain shifts are to be admitted. It is far from trivial to develop a respectable theory of Faithfulness that contains such constraints; see, for example, Kirchner (1996), Gnanadesikan (1997), Moreton & Smolensky (2002), Mortensen (2004); and for other approaches, Alderete (1999b), (2001b) for antifaithfulness, and Łubowicz (2002), who aims to put the issue entirely outside the M/F distinction.

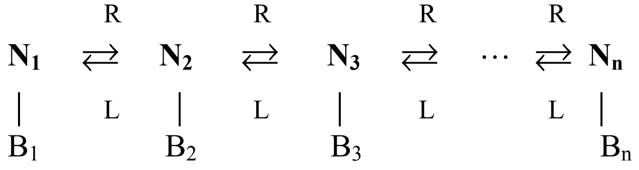
2.3.2 The Barrier Models

Goldsmith and Larson have proposed a spreading-activation account of linguistic prominence, which they have vigorously pursued through encounters with many attested patterns of stress and syllable structure — the Descriptive Method (Goldsmith & Larson 1990, Larson 1992, Goldsmith 1994). The model is, however, entirely self-contained as a formal object and susceptible to treatment as a Free-Standing Theory whose key properties can be determined analytically (Prince 1993).[12] The aim of this section is to illustrate once again, in a very different context, how pursuing the basic formal questions leads not to an exercise in logical purification, but quite directly to properties of notable empirical significance.

The model works like this: the basic structure is a sequence of N 'nodes', each of which carries an 'activation' level, represented numerically. This gives it the power to represent ordinal properties of segments and syllables like sonority and prominence. Each node also has an unvarying bias, which may be interpreted as the intrinsic sonority or prominence of the linguistic unit that it represents. Rather than make a single calculation over these values to determine the output activation, the model calculates repeated interactions between adjacent nodes — the same mode of interaction repeated over and over. When the process converges on stable values, the model has calculated an activation profile that corresponds to a prominence structure such as a stress pattern or assignment of syllable peaks and margins. Nodes which bear greater activation than their closest neighbors − local maxima − are interpreted as having peaks of prominence.[13] Since the updating scheme is linear and iterative, we will call it the Dynamic Linear Model (DLM).

The neighborly interaction is mediated by two numerical parameters, which we designate L and R, each of which governs the character of the interaction in one of the two directions. The parameter L governs leftward spreading of activation; R, rightward spreading. Diagrammatically, we can portray the situation like this:

(18) **DLM Network**



The model starts out with each node bearing zero activation. In the first step, each node gains the activation donated by its own bias; and then the serious trading begins. At each stage, the new activation of a node is determined from the current activation of its neighbors taken together with its own intrinsic bias level. The update scheme, in which we write $\mathbf{a_k}$ for the activation of N_k, can be represented like this:

[12] Discussion is based on "In defense of the number *i*" (Prince 1993, henceforth IDN), improved notationally and formally in a few respects.

[13] Although the model operates internally on numbers, it does not strive to compute an empirically-determined numerical value; its interpreted output is fully discrete and indeed binary, discriminating only peaks from nonpeaks.

(19) $\mathbf{a_k} \leftarrow \frac{1}{2}\, L\cdot\mathbf{a_{k+1}} + \frac{1}{2}\, R\cdot\mathbf{a_{k-1}} + B_k$

A node's own current activation plays no role in determining its next state: only its bias, which never changes. Since L, R, and B_k are all constants, this is a linear scheme: each node's new activation is a weighted sum of its neighbor's activations, with its own bias added in.

Here are some examples to give a sense of how it works. Suppose we start out with a bias sequence (1,1,1,1,1,1), representing a string of 6 undifferentiated syllables. Let L=R= −1. The result is approximately (1.1, −0.3, 1.4, −0.6, 1.7, −0.9). This may look like nothing more than a mess of numbers, but the significant fact is the location of the local maxima – those nodes greater than their neighbors (or neighbor, if at an edge). Marking those, we see that the DLM has calculated this mapping, which we write using x for 'unstressed' and X for 'stressed': x x x x x x → X x X x X x
A familiar kind of alternating pattern has been imposed.

Now suppose we start out with a bias sequence (0,0,1,0,0,0) and set L=1.333 and R=.75. The result comes out approximately like this: (2.0, 3.0, **3.4**, 1.9, 1.0, 0.4). Identifying the one maximum (bolded), we see that this is the IO relation:

x x X x x x → x x X x x x

which is naturally interpreted to express a case in which an accent marked in the lexical input has been preserved on the surface.

If we alter the L,R parameters, we get a different result: for L=1.6, R=.635, we get approximately (2.9, **3.7**, 3.4, 1.6, 0.7, 0.2). The significant configuration now centers on the second entry, and we have portrayed the map

x x X x x x→ x X x x x x

in which an underlying accent has been over-ridden.

A variety of linguistic and nonlinguistic patterns may be produced from such experimentation, suggesting the value of further systematic research.[14] What, then, are the general properties of the theory? At this point, two paths diverge. We may follow the Descriptive Method, with Goldsmith and Larson, aiming to deal with a wide range of known prominence phenomena in specific languages by finding L, R values and biases that will accommodate them. Or we may attempt to see what we can learn by interrogating the formal structure of theory, trying to classify its parameter space and look for characterizing properties.[15]

Let's start with one of the most fundamental questions we can ask: under what conditions does the theory *exist*? In the context of an iterative scheme like the DLM, this question takes a clear and exact form: when does the model converge, producing stable finite values as output? Specifically, what values of the parameters L and R lead to convergence? The fine-grained convergence limit is tied to a specific model's length in nodes; but generalizing over all models, we have this pleasing result, which will prove quite useful: if the absolute (unsigned) value of the product L·R is less than or equal to 1, any model of any length will converge.

[14] Such experimentation with the parameters of a theory is a part of what we are calling the Analytic Method, though here we are emphasizing the aspects of analysis that yield provable results.

[15] In noting this methodological divergence, we are of course not asserting that only one path should be pursued.

(20) **Convergence of the DLM**
Any Dynamic Linear Model M_n with $|LR| \leq 1$ converges, for all n, n the number of nodes in the model.[16] (IDN:53)

From the descriptive point of view, this result has its uses – it tells us where not to look for parameter values – though, in practical terms, if we start our search near zero for both L and R, an astute prospector armed with a spreadsheet program ought to be able to find suitable values experimentally, when they exist. Analytically, its interest emerges when we ask a further question, targeted at finding the content of the theory in its realm of existence: given L, R, and a sequence of biases, is there a *formula* that describes the output of the iterative scheme? The goal is not merely to shorten the process of calculation (pointless in the Excel™ era), but to have a characterization of the model's output that may be scrutinized for general properties.

For the vast majority of networks, 'solving the model' in this way is not an option, and the Descriptive Method is essential to finding out what's going on; this is why we classified the Rumelhart & McClelland model as a TOD, and why people tend to think of network models as TOD on arrival. But the simple structure of the DLM renders it amenable to analysis.

Because the function computed by the DLM is linear in the biases, it is natural formally to inquire about the fate of bias sequences that consist entirely of 0's except for a single 1. Any other sequence can be built up from a weighted sum of such basic sequences. Here linguistics lines up happily with algebra – it is also linguistically natural to regard such sequences as representing a form with a single lexical accent.

We want to describe the value assumed by each node, given that the 'underlying accent' occurs in a certain place. The local maximum in the output, which is fully determined by these values, is where the surface accent lies. Calculation produces a formula which is a bit messy though not intractable (involving hyperbolic sines and cosines and the occasional complex number; see IDN:62). But a remarkable simplification occurs when we restrict the parameters to the curves LR=1, on which convergence is universally guaranteed.[17] Because of their simplicity, we may call these the 'Canonical Models'. The Canonical Models come in two kinds. Either L and R are both negative, in which case we have alternation of prominence, as we always do when both parameters are negative; or both parameters are positive.

[16] For a specific length, we have convergence iff $|LR| < 1/\cos^2(\pi/(N+1))$, which is always greater than 1. If L and R have the same sign, a model diverges to infinity at and beyond the limiting value; if they have different signs, the model enters an oscillatory regime of period 4 at the limiting value, and diverges beyond it.

[17] The resulting formula turns out to involve the product of two linear terms, each reflecting distance to the edge, and an exponential term based on either of the L or R parameters, whose exponent reflects the distance between the underlying accent and the node whose value is being computed, Schematically, we can write it like this:
$\mathbf{a_j} = C \cdot \text{dist-k\#(j)} \cdot \text{dist-j\#(k)} \cdot R^{\text{dist(j,k)}}$
where C is a length-based constant 2/(n+1), the 'tilt' $\sqrt{(R/L)} = R$, dist-k#(j) gives the distance of j from the edge where k is not in the j-to-edge path, dist-j#(k) *mutatis mutandis*; dist(j,k) is the signed distance j–k between j and k.

The behavior of the general DLM when both L and R are positive is straightforward: accent is culminative, with a single maximum occurring in the activation function.[18] The same will be true in the Canonical Models. But when we seek the location of that maximum in the Canonical Models, a striking property emerges: there is a *window* at one edge or the other into which the surface accent must fall.

Given any value of R greater than 1, the surface accent can fall no further than a certain distance from the right edge, regardless of where the underlying accent is placed. The same is true for L (corresponding to values of R less than 1), with respect to the beginning of the word. Within the window, underlying accent is preserved. Outside the window, it is lost and in its place, as it were, the accent shows up at inner edge of the window – the closest unit to the underlying accent that can be surface-accented.

We can name each model by the farthest internal location at which an accent can fall, (given single accented input), indicating by subscript the edge it measures from: thus, 3-Model$_R$ is the model in which the accent can fall no further into the string than the 3rd node from the end. Let us call these Canonical Models the 'barrier models', since in a k-Model, the kth node provides a kind of barrier beyond which surface accent may not venture. The parameter space divides up as in table (21). NB: the cited ranges *exclude* the end points.

(21) **Right Barrier Models**

Model #	"range" of R	Length of Range	Accent no further from end than
1-Model$_R$	∞ to 2	∞	final syllable
2-Model$_R$	2 to 3/2	1/2	penult
3-Model$_R$	3/2 to 4/3	1/6	antepenult
4-Model$_R$	4/3 to 5/4	1/12	preantepenult
5-Model$_R$	5/4 to 6/5	1/20	prepreantepenult
…	…	…	
j-Model$_R$	$j/(j-1)$ to $(j+1)/j$	$1/j(j-1)$	(pre)$^{j-3}$ antepenult

Symmetrically, the Left Barrier Models determine a window at the *beginning* of the string. The Right Barrier Models charted above occupy the parameter span where $R\in (1, \infty)$. The Left Barrier Models lie within the positive line segment $L\in(1, \infty)$, or equivalently $R\in (0,1)$, since $R=1/L$.[19]

This result is multiply remarkable. First, the barrier/windowing behavior is fully emergent from assumptions which make no mention of anything like that property. The alternating pattern that comes about when L and R are both negative has a kind of resonance with structural formulations like *CLASH [Kager §9.2.1]. Both, in their

[18] Caveat: what we are calling a 'maximum' can be spread across two adjacent nodes that have identical activation values.

[19] For R=L=1, we simply reproduce the input accent, no matter where it is located, on any string of any length.; this is the ∞-Model. The behavior at the other end points of the ranges is not entirely welcome: we get adjacent pairs of nodes with equal activation at the window boundary when the input accent lies at or beyond the barrier. In the R Models, for example, when R=2, we get equal activation on the final and penult when the input accent is penult or earlier. When R=3/2, we get equal activation on penult and antepenult when the input accent is antepenult or earlier.

different ways, seek to suppress prominence on adjacent units. And when L and R are both positive, it is perhaps not naively expected that the result should be a single maximum in the activation function, but it doesn't seem like an unusual outcome. It is the particularity of the windowing effect, and its lack of reducibility to some obvious local characteristic of the network, that makes it surprising.

Second, it is remarkable that the parameter ranges are valid for any length of string.[20] The number of nodes plays a role in the formula describing the output, and in other situations it figures in empirically anomalous dependencies (IDN:17). In this case, though, we have conditions that are valid across all forms, fully independent of form size.

Third, although nontrivial barrier/windowing behavior, with non-peripheral accents allowed, goes on outside the Canonical Models, it is restricted to a relatively small portion, a little less than 1/6, of the parameter space in the first quadrant. This means that random prospecting could easily miss it. Crucial to finding it is investigation along the hyperbola LR=1; but this curve presents itself as particularly interesting only because of its role in delimiting convergence.[21] The abstract, airless-seeming question with which we began – under what conditions does the model exist? – has led us right to one of its central properties.

Finally, it is striking that this fundamental result connects directly with a major phenomenon in stress and accent systems. The DLM overshoots the mark in a couple of respects – it is totally left-right symmetric, and allows windows of any size, while known windowing systems typical range up to no more than 3 syllables in length at the end of words, and 2 syllables at the beginning.[22] Whatever the remaining questions, the model opens the way to an entirely novel account of the windowing effect, unlike anything seen before. This renders the DLM worth studying alongside the other contentful accounts of prosodic structure that occupy linguistic attention, while vindicating the analytic method that reveals its structure.

2.4 Description and Descriptivism

In a recent essay, Larry Hyman asks and answers the question "Why Describe African Languages?" (Hyman 2004). He argues that there is irreducible value in describing "complex phenomena using the ordinary tools of general linguistics," and that this goal stands in opposition to, and is at least as worthy as, developing grammars within current "theories [that] are not description-friendly," such as Minimalism and OT.

With the main thrust of his argument there can be little dissent: deep empirical work discovering the facts and generalizations of human languages is the very basis of linguistics, and it is essential that that there be sound descriptions to convey them to the community of researchers. Why then the question? In part, Hyman's concern is driven by disciplinary attitudes toward 'theory' and 'description' – where, it seems, a certain class of person expects one to make a 'theoretical contribution' in every outing and will

[20] Hence the celebratory appellation *Theorema Egregium* applied to its announcement (IDN:85).

[21] In the original expression of the model, the Canonical Models were defined by LR=1/4, which is even less obvious as a condition to pursue.

[22] One could imagine that the drastic shrinking of the parameter range with increase in window size might support a more detailed account of the empirical restrictions, at least in part (IDN:91).

disdain or suppress work that lacks that key ingredient.[23] As for what a 'theoretical contribution' might be, Hyman cites an unidentified commentator:

(22) "The shared belief of many in the field appears to be that a paper making a theoretical contribution must (a) propose some new mechanism, which adds to or replaces part of some current theory, or (b) contradicts some current theory. Papers that do neither, or those that do neither but in a relatively minor way, are not looked at as making a theoretical contribution." Quoted in Hyman (2004:25).

This is very much a matter of 'mind your labels' – and we shouldn't be led to abandon the idea of 'theoretical contribution' because an obtunded version is instrumental in the intercollegial jostling and jousting of the field. In the present context, where a theory is taken to be an object in grave need of explication and analysis, it should be clear that an authentic 'theoretical contribution' can involve deepening the understanding of a theory's consequences or of the proper methods of using it, without a hint of replacement or contradiction.[24] We reject the 'shared belief' identified in the quote, and deny the privileged status it accords to certain types of work, to advocate a broader though not boundaryless account of what a contribution, including a 'theoretical contribution', may be. Hyman's move, by contrast, is to argue toward a unification of theory with description, neutralizing the distinction: "description and theory are very hard to disentangle – and when done right, they have the same concerns" (p.25). He goes on to clarify:

(23) "Description is *analysis* and should ideally be
a. rigorous …
b. comprehensive …
c. rich …
d. insightful …
e. interesting …" (Hyman 2004:25)

No one would dispute either the importance of the cited criteria or the claim that they apply to theory as well as description. A closer look, though, is profitable, and suggests some important divergences. Criteria (c), (d), and (e) are contentful but difficult to assess intersubjectively, and perhaps connect more closely with Harris's 'convenience' than with questions of truth and falsity. We therefore focus on (a) *rigor* and (b) *comprehensiveness*.

Of rigor, the key remark is the one made in section 1 above: there is no general sense of rigor that can be directly applied without regard for the specific assumptions at play in a given case. Work is therefore required. To design a successful ranking argument, as in our example, you must build from the actual definition of 'optimality'. It is necessary to ask 'what can be learned from the comparison of two candidates, one

[23] Stepping through the looking glass, we can easily discern the antitype who demands an 'empirical contribution' as the prerequisite for admissibility.

[24] Just as in certain regions of physics, to risk an extravagant comparison, finding a solution to a known equation, or a method for solving a type of equation, can net a Nobel Prize or an office at the Institute for Advanced Study.

assumed optimal?' If the Evaluation Metric is to be employed seriously, you must inquire about the relation between local reduction of symbol consumption and the eventual global symbol count of the entire grammar. To achieve 'rigor', there is a range of questions that must be asked about the theory itself, and these questions differ in character from those asked of data (e.g. what is the distribution of downstepped high tone in Bangangte Bamileke?) or of the data-analysis relation (e.g. how are floating tones interpreted? how are they manipulated in Bangangte Bamileke?).[25] And different methods are required to answer them.[26]

Comprehensiveness – the inclusion of all relevant material – is a systematic notion and therefore presupposes a notion of 'system' which delimits relevance. Just like rigor, then, it takes on different colorations in different contexts. Contrast the questions to be asked and the techniques required to attain and evaluate, say, a full account of a language's verbal paradigm[27] with those used to derive and characterize the consequences of a formal theory. It makes sense to classify these as different 'contributions', if we are classifying things, though the inevitable ensuing scuffle to hierarchize them socially is better explicated by primatology than by the philosophy of science.

In the present context, the interpretation of *comprehensiveness* also marks an important divide between appropriate strategies for descriptive work and for theory development. Much can be gained theoretically by explicitly failing to be comprehensive over the data in ways that would be absurd descriptively. The study of idealized, delimited problems is a familiar and essential tool for exploring theories. At the grand level: the de Sitter cosmology imagines a universe that lacks matter entirely (it expands); Schwarzschild solves the field equations of General Relativity under the assumption of strict spherical symmetry of matter distribution (local collapse can result).[28] To cite a case considerably humbler and closer to home: much can be learned by working with a simplified Jakobsonian typology of syllable structure (Clements & Keyser 1983, Prince & Smolensky 1993/2003), although it would be grossly inappropriate to claim comprehensiveness for a *description* of natural language syllable patterns that overlooks long vowels, diphthongs, and intrasyllabic consonant clusters.

Investigation of theories, even via the Descriptive Method, is tied to the availability of research strategies that idealize and delimit, deferring comprehensiveness. In the case of FST, this is particularly crucial because it opens up possibilities for obtaining analytical results when the general situation is complex and its structure obscure. Attitudes toward comprehensiveness therefore play a subtle but central role in

[25] The questions are drawn from Hyman's discussion of Vorhoeve (1991).

[26] Those methods require analysis and development in themselves, since they call on statistics, formal language theory, ordinal preference theory, recursive function theory, logic, and so on.

[27] This casual and overly certain-sounding allusion to 'verbal paradigm' should remind us that the categories of the presupposed 'system' are almost always under contention, and can be wrong, leading to failure of comprehensiveness and the missing of generalizations. Is a phonological description comprehensive without reference to aspects of speech perception and speech production? Is a syntactic analysis comprehensive that overlooks pragmatics? In some such cases, the answer must be *yes*, or we are done for; but which?

[28] Interestingly, Einstein neither expected nor was happy with these results. Pais (1983) is the authoritative account of the life and works, though its perspective has been somewhat outdated by the intense subsequent growth (unexpected, perhaps, by Pais) of black hole studies and String Theory with its higher-dimensional space-times.

estimating the relative promise of different research directions. One line of thinking finds expression in "Why Phonology is Different" (Bromberger and Halle 1989). The authors are concerned to justify their belief that phonology is intrinsically not amenable to being understood as the interaction of universal principles, distinguishing it in their view from syntax; the key, they argue, is the availability of stipulated language-specific rule-ordering in phonology alone:

(24) "Rule ordering is one of the most powerful tools of phonological description, and there are numerous instances in the literature where the ordering of rules is used to account for phonetic effects of great complexity." (Bromberger & Halle 1989: 59).

The perspective here is determinedly descriptive; the theory is to be justified by its ability to portray "complex" cases, for which much "power" is thought to be needed. There is no hint of an ambition to find and derive general properties of the language faculty, and consequently no willingness to tolerate the local costs of such ambition — idealization; plurality of theoretical lines; openness to ideas that limit rather than expand descriptive options; empirical lacunae and anomalies; admission of uncertainty. Their argument continues:

(25) "Until and unless these accounts are refuted and are replaced by better-confirmed ones, we must presume that Principle (7) [extrinsic ordering – AP] is correct." (Bromberger & Halle 1989:59).

One can only admire the authors' willingness to take on the entire literature in an area before rejecting its premises, but there are sound reasons why this strategy has never had much purchase on the field, which has been more notable for innovation than uniformity. At bottom, providing unsteady foundations, is an unexamined notion of 'confirmation', without which such qualifiers as 'better-confirmed' and 'correct' risk vacuity. More concretely, there are so many active, promising lines of investigation into every aspect of the enterprise, from the nature of the data to the identity of the targets of explanation, that it seems premature to shut them down on the basis of a presumption.

Whatever the ultimate status of their imperative, its interest in the present context is its orthogonality to the kind of theoretical concerns we have been probing. There is no sense in their work that a theory is an opaque object, whose content and proper handling must be discovered before we can declare success and failure, even descriptively, or compare it properly with other theories. Supreme is the goal of 'accounting for', and given a disposition to regard the facts as a fixed body, the approach merges with classic descriptivism. The real threat to their favored theory, then, is not provided by those versions of generative phonology which pursue very different explanatory goals, but rather by statistical empiricism, which also avails itself of 'powerful tools' to gain even more comprehensive models of their data.

2.5 Conclusion

The encounter with fact is essential to the validation, falsification, and discovery of theories. But as soon as a theory comes into existence, it must also be encountered on its own terms. A theory cannot even be faced with fact – we cannot *do* it properly – if we don't know how to construct valid arguments from its premises. And since a theory's content is the set of its consequences, which are typically far from legible in its defining conditions, we are obliged to interrogate its structure to find out what it *is*. Asking the fundamental formal questions, and finding or developing techniques to answer them, is an irreplaceable aspect of linguistic research that identifies the major predictions and particularly meaningful empirical challenges associated with a theory.

Linguistic theory has shown a notable tendency to develop what we have called Free-Standing Theories, those which have an internal structure susceptible to detailed analysis independent of the factual encounter. The reasons for doing so may be, as suggested above, intrinsic to the realist project, since rationalist theories requires an abstract object of study whose existence is likely to be justifiable only in terms of deep, nonobvious properties. In the absence of such properties, empiricist inductivism exerts a strong claim to the territory.

It is reasonable to ask, then, why the 'Analytic Method' of confronting theories on their own terms does not play a more conspicuous role in the current ecology of the field, which could be argued to conserve, largely, an intuitive methodology more properly rooted in the descriptive ambitions of pre-generative work. An important factor may be the sense that formal analysis can be successfully replaced by approaches more closely allied to facts and to techniques for dealing with facts – 'the ordinary tools of general linguistics'. Invaluable in empirical assessment of claims, the Descriptive Method has often been taken as the primary mode of exploring a theory's structure and content, where it has severe limitations. Adhered to strictly, it cannot distinguish between a superset theory ("too powerful") and a proper subset theory; it has no particular relation to a theory's systematic properties; and it is unable to provide certainty in the assessment of claims about predictions and exclusions.

A more recent development which is sometimes taken to provide a feasible substitute for analysis is 'grounding' – in the case of phonology, pointing to phonetics as supporting the correctness of theoretical assertions. In much work, the term has a specific well-defined sense which gives it theoretical status (Archangeli & Pulleyblank 1994, Hayes 2004a:299), but it also leads a second, more fluid life as a motivator and recipient of intuitive appeals. Some of this may be discerned in the following statement from Hayes (2004a:291), who is asking "what qualifies a constraint as an authentic markedness principle:"

(26) "The currently most popular answer, I think, relies on typological evidence: a valid constraint "does work" in many languages, and does it in different ways.

However, a constraint could also be justified on functional grounds. In the case of phonetic functionalism, a well-motivated phonological constraint would be one that either renders speech easier to articulate or renders contrasting forms easier to distinguish perceptually. From the functionalist point of view, such constraints are a priori plausible, under the reasonable hypothesis that language is

> a biological system that is designed to perform its job well and efficiently." (Hayes 2004a:291).

But the symmetry is illusory. A constraint, in the intended sense, is a principle within a theory and, like any other principle in any other theory, is justified by its contribution to the consequences of that theory. Since OT is a theory of grammar, the consequences are displayed in the grammars predicted and disallowed – 'typological evidence'. A constraint which cannot be justified on those grounds cannot be justified. Further, 'justifying' a constraint functionally (or in any other extrinsic way) can have no effect whatever on its role within the theory. A constraint, viewed locally, can appear wonderfully concordant with some function, but this cannot supplant the theory's logic or compel the global outcome ('efficiency') that is imagined to follow from the constraint's presence, or even make it more likely.

A ranking argument based on two candidates, one desired optimal, remains valid whether the constraints are grounded or not; and in Targeted Constraint OT, where grounding is invoked to support the notion of targeting (Wilson 2002:156-160), such two-candidate arguments lose their validity because of the formal structure of the theory, and phonetic function cannot restore it. The property of Harmonic Ascent cannot be abrogated, amended, or influenced by grounding or its lack. The choice of *Markedness* constraints, no matter how grounded, cannot by itself predict grammatical behavior, because mappings are determined by the interaction of Markedness with Faithfulness constraints, whose properties are crucial to the range of possible outcomes.

When stated explicitly (p.299), Hayes's 'inductive grounding' is not an exercise in the plausible,[29] but a concrete proposal for the generation of certain kinds of constraints from specific data, which relies on finding the local maxima in a certain space of possibilities. Its fate is in the hands of geometry and logic. As an actual theory, it has left behind any hopes that attended its conception and birth, and now lives in the realm of the issues explored here.

Such considerations suggest a bright future for linguistic research as it grows beyond its origins. Analysis is deaf to our desires, but it can tell us what we want to know, if we know how to ask.

[29] Terms like 'plausible' or 'reasonable' seem to diagnose what we might call 'conceptual orientation' in the discourse participants. The implicit contrast is with *possible* — if something is said to be X-ologically possible, the implication is that we know enough about the theory of X-ology to calculate with it; the comforts of the X-ologically *plausible* are those of intuition and common sense.